

中图法分类号: TP389.1 文献标识码: A 文章编号: 1006-8961(2023)01-0037-20

论文引用格式: Li C L, Lu A D, Liu L and Tang J. 2023. Multi-modal visual tracking: a survey. Journal of Image and Graphics, 28(01): 0037-0056 (李成龙, 鹿安东, 刘磊, 汤进. 2023. 多模态视觉跟踪方法综述. 中国图象图形学报, 28(01): 0037-0056) [DOI: 10.11834/jig.220578]

多模态视觉跟踪方法综述

李成龙¹, 鹿安东², 刘磊², 汤进^{2*}

1. 安徽大学人工智能学院, 合肥 230601; 2. 安徽大学计算机科学与技术学院, 合肥 230601

摘要: 目标跟踪是计算机视觉研究中的前沿和热点问题, 在安全监控、无人驾驶等领域中有着重要的应用价值。然而, 目前基于可见光数据的视觉跟踪方法, 在光照变化、恶劣天气下因数据质量受限难以实现鲁棒跟踪。因此, 一些研究者提出了多模态视觉跟踪任务, 通过引入其他模态数据, 包括红外模态、深度模态、事件模态以及文本模态, 在一定程度上弥补了可见光模态在恶劣天气、遮挡、快速运动和外观歧义等条件下的不足。多模态视觉跟踪旨在挖掘可见光和其他模态数据的互补优势, 在视频中实现鲁棒的目标定位, 对全天候感知有着重要的价值和意义, 受到越来越多的研究和关注。由于主流的多模态视觉跟踪方法针对可见光—红外跟踪展开, 因此, 本文以阐述可见光—红外跟踪方法为主, 从信息融合的角度将现有方法划分为结合式融合和判别式融合, 分别进行了详细介绍和分析, 并对不同类方法的优缺点进行了分析和比较。然后, 本文对其他多模态视觉跟踪任务的研究工作进行了介绍, 并对不同多模态视觉跟踪任务的优缺点进行了分析和比较。最后, 本文对多模态视觉跟踪方法进行了总结并对未来发展进行展望。

关键词: 信息融合; 视觉跟踪; 多模态; 结合式融合; 判别式融合

Multi-modal visual tracking: a survey

Li Chenglong¹, Lu Andong², Liu Lei², Tang Jin^{2*}

1. School of Artificial Intelligence, Anhui University, Hefei 230601, China;

2. School of Computer Science and Technology, Anhui University, Hefei 230601, China

Abstract: Visual tracking can be as one of the key tasks in computer vision applications like surveillance, robotics and automatic driving in the past decades. The performance issue for visual tracking is still challenged of the quality of visible light data in adverse scenes, such as low illumination, background clutter, haze and smog. To deal with the imaging constraints of visible light data, current researches are focused on multiple modal data-introduced in common. The visible and modal data integration can be effective in tracking performance in terms of the manner of thermal infrared, depth, event and language. Benefiting from the integrated capability of visible and multi-modal data, multi-modal trackers have been developing intensively in such complicated scenarios of those are low illumination, occlusion, fast motion and semantic ambiguity. Nowadays, our executive summary is focused on reviewing the RGB and thermal infrared (RGBT) tracking algorithms, which is oriented for the popular visible-infrared visual tracking towards multi-modal visual tracking. Existing multi-modal visual tracking-based summaries are concerned of the segmentation of tracking algorithms in terms of multi-framework tracking or multi-level based fusions derived of pixel, feature, and decision. With respect of the information fusion plays a key

收稿日期: 2022-06-02; 修回日期: 2022-07-08; 预印本日期: 2022-07-15

* 通信作者: 汤进 tangjin@ahu.edu.cn

基金项目: 国家自然科学基金项目(61976003); 模式识别国家重点实验室开放课题

Supported by: National Natural Science Foundation of China (61976003)

role in multi-modal visual tracking, we divide and analyze existing RGBT tracking methods from the perspective of information fusion, including synthesized and specific-based fusions. Specifically, the fusion-integrated can be used to combine all multimodal information together via different fusion methods, including: 1) sparse representation fusion, 2) collaborative graph representation fusion, 3) modality-synthesized and modality-specific information fusion, and 4) attribute-based feature decoupling fusion. First, sparse representation fusion has a good ability to suppress feature noise, but most of these algorithms are restricted by the time-consuming online optimization of the sparse representation models. In addition, these methods can be used as target representation via pixel values, and thus have low robustness in complex scenes. Second, collaborative graph representation fusion can be used to suppress the effect of background clutter in terms of modality weights and local image patch weights. However, these methods are required for multi-variables optimization iteratively, and the tracking efficiency is quite lower. Furthermore, these models are required to use color and gradient features, which are better than pixel values but also hard to deal with challenging scenarios. Third, modality-synthesized and modality-specific information fusion can be used to model modality-synthesized and modality-specific representations based on different sub-networks and provide an effective fusion strategy for tracking. However, these methods are lack of the information interaction in the learning of modality-specific representations, and thus introduce noises and redundancy easily. Fourth, attribute-based feature decoupling fusion can be applied to model the target representations under different attributes, and it alleviates the dependence on large-scale training data more. However, it is difficult to cover all challenging problems in practical applications. Although these fusion-synthesized methods have achieved good tracking performance, all multiple modalities information-synthesized have to introduce the information redundancy and feature noises inevitably. To resolve these problems, some researches have been concerned of fusion-specific methods in RGBT tracking. This sort of fusion is aimed to mine the specific features of multiple modalities for effective and efficient information fusion, including: 1) feature-selected fusion, 2) attention-mechanism-based adaptive fusion, 3) mutual-enhanced fusion, and 4) other fusion-specific methods. Feature selection fusion is employed to select specific features in regular. It can not only avoid the interference of data noises and is beneficial to improving tracking performance, but also eliminate data redundancy. However, the selection criteria are hard to be designed, and unsuitable criterion often removes useful information under low-quality data and thus limits tracking performance. Adaptive fusion is aimed to estimate the reliability of multi-modal data in term of attention mechanism, including the modality, spatial and channel reliabilities, and thus achieves the adaptive fusion of multi-modal information. However, there is no clear supervision to generate these weights-reliable, which is possible to mislead the estimated results in complex scenarios. Mutual enhancement fusion is focused on data noises-related suppression for low-quality modality and its features can be enhanced between the specific information and other modality. These methods can be implemented to mine the specific information of multiple modalities and improve target representations of low-quality modalities. However, these methods are complicated and have low tracking efficiency. The task of multi-modal vision tracking has three sub-tasks besides RGBT tracking, including: 1) visible-depth tracking (called RGB and depth (RGBD) tracking), 2) visible-event tracking (called RGBE (RGB and event) tracking), 3) visible-language tracking (called RGB and language (RGBL) tracking). We review these three multi-modal visual tracking issues in brief as well. Furthermore, we predict some academic challenges and future directions for multi-modal visual tracking.

Key words: information fusion; visual tracking; multiple modalities; combinative fusion; discriminative fusion

0 引言

视觉跟踪是计算机视觉研究中的前沿和热点问题,在视频监控、智能交通和无人驾驶等领域中有着重要的应用价值。然而,基于可见光成像的视觉跟踪方法受光照变化、雨雪雾等恶劣天气的影响较大,很难满足复杂场景和环境下的应用需求。为了克服

可见光成像的局限,学者们引入其他模态数据,包括红外、深度、事件和语言数据,通过协同利用不同模态的互补优势,实现复杂场景和环境下的鲁棒跟踪,对全天时全天候感知有着重要的价值和意义。例如,可见光波段对直径小于 $2.5\ \mu\text{m}$ 的微粒穿透能力差,在雾霾环境下成像质量急剧下降,甚至失效,极大限制了视觉跟踪性能。而红外热像是通过感知场景物体热辐射而形成的,反映了场景物体表面的

温度分布,对雾霾穿透能力强,对可见光成像具有较好的补充作用。因此,多模态视觉跟踪受到了越来越多的关注和研究。

多模态视觉跟踪主要包含4个子任务:可见光—红外(RGB and thermal infrared, RGBT)跟踪、可见光—深度跟踪(RGB and depth, RGBD)、可见光—事件跟踪(RGB and event, RGBE)和可见光—语言跟踪(RGB and language, RGBL),如图1所示。与可见光成像相比,其他模态成像具有特定的成像优势和劣势。红外传感器是一种被动成像系统,反映了场景物体表面温度分布,基本不受光线变化、雾霾和雨雪等恶劣天气因素影响,但红外成像分辨率低、易出现热交叉。深度传感器已广泛应用于计算机视觉领域中,与可见光传感器获得图像的颜色信息不同,它可以获得图像的深度信息,反馈的是图像中各像素点和传感器之间的距离信息,能够较好地感知遮挡状态,但深度成像距离近、噪声多且受室外光线影响大。与可见光相机一帧一帧地记录图像内容不同,受生物启发的事件相机以更低的延迟产生异步和稀疏的事件流,可以避免运动模糊,具有更大的动

态范围,在快速运动和低照度下能够很好地工作,但事件成像缺乏物体纹理和细节信息。基于语言描述的目标表示方法能够提供更明确的目标语义信息,可以有效增加目标表示的判别能力以及消除矩形框表示的歧义现象,但语言描述无法提供准确的外观细节和定位信息。因此,本文在表1中总结了4种模态的优劣势。可以看出,可见光模态和其他模态具有良好的互补优势。而如何充分挖掘和利用不同模态信息是多模态视觉跟踪领域的核心问题。

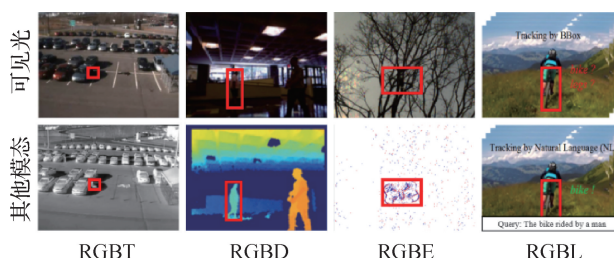


图1 可见光和其他模态成像及其互补性示意图

Fig. 1 Illustration of the complementarity of visible data against other modality data

表1 不同模态成像的优劣势

Table 1 Advantages and disadvantages of different modalities

模态类型	优势	劣势
可见光	分辨率高,纹理清晰,色彩信息丰富	光照敏感,对雾霾等穿透能力差
红外	基本不受光照和雾霾等因素影响	纹理细节弱,对玻璃材质穿透能力差
深度	易对遮挡状态进行推理	成像距离有限,受环境光照影响大
事件	低延迟,对运动感知强	低速物体感知弱,纹理细节弱
语言	提供确定的语义信息,消除一定的外观歧义	对高精度定位不敏感,有一定的描述歧义性

目前,也有一些工作对多模态视觉跟踪方法进行了整理和回顾,如Zhang等人(2020d)对RGBT视觉跟踪方法的发展历程进行了全面梳理,并根据跟踪器框架的不同以及融合层次(如像素级、特征级和决策级)的不同对RGBT视觉跟踪算法进行了划分。Zhang等人(2020b)同时回顾了RGBT和RGBD视觉跟踪方法,并根据辅助模态的作用和跟踪框架的不同对多模态视觉跟踪算法进行了划分。尽管这些工作梳理了一些多模态视觉跟踪方法,但是多模态视觉跟踪不仅包含有RGBT和RGBD跟踪,还包含RGBE和RGBL跟踪,而现有的多模态跟踪综述缺少对这些多模态跟踪方法的介绍。此外,考虑到多模态

信息融合是多模态视觉跟踪研究的核心,本文从信息融合的角度对现有算法进行归类和分析,以期帮助读者更容易、更深刻地掌握多模态视觉跟踪领域发展的基本脉络,能够为读者带来更多的启发。

1 RGB视觉跟踪方法

本文仅简单阐述与多模态视觉跟踪方法具有较高相关性的RGB视觉跟踪方法,更全面和深入地介绍可参考相关综述论文。主要将现有RGB视觉跟踪方法分为两大类,即基于传统模型的跟踪方法和基于深度模型的跟踪方法。

1.1 基于传统模型的跟踪方法

1.1.1 稀疏表示跟踪方法

稀疏表示可以有效降低数据维度且减小特征向量各维度之间的依赖性,且稀疏约束可以使得计算后得到的各个基对于解释数据具有同等的重要性,因此,稀疏表示模型成功应用于 RGB 视觉跟踪中。Mei 和 Ling(2009)较早地提出了基于稀疏表示的目标跟踪算法。他们引入了一个模板集,将稀疏表示优化转换成一个 L1 范数最小二乘问题,并采用 LASSO (least absolute shrinkage and selectionator operator) 算法进行求解,较好地解决了遮挡和噪声带来的问题。受该工作启发,研究者们提出了一系列基于稀疏表示的跟踪算法。例如, Liu 等人(2010)提出了一种两步稀疏优化算法,极大提高了跟踪效率。为了进一步提升鲁棒性, Li 等人(2012)将 Gabor 特征引入到基于稀疏表示的目标跟踪算法中,有效地解决了光照变化带来的影响。为了平衡实时性和鲁棒性,侯跃恩等人(2013)提出一种将目标与背景联合稀疏表示的跟踪方法,把稀疏表示模板字典与背景信息相结合,有效地增强了跟踪算法对背景的判别能力。

1.1.2 结构化支持向量机跟踪方法

该类方法基于检测算法,其核心是通过结构化支持向量机将目标样本和背景样本进行区分,从而实现目标的定位。鉴于支持向量机的良好泛化能力、标签噪声鲁棒性和核函数表示的灵活性, Hare 等人(2011)引入结构化支持向量机到跟踪任务中,提出了一种有效的判别式跟踪模型,因其较好的跟踪性能受到了广泛关注。Kim 等人(2015)提出了基于空间有序加权特征的结构化支持向量机跟踪方法,通过随机游走模型计算局部图像块的重要性权重,结合加权特征和结构化支持向量机实现稳健跟踪。Ning 等人(2016)提出了双线性的结构化支持向量机优化方法和显式特征映射,显著降低了模型计算复杂度。此外, Li 等人(2017a)在 Kim 等人(2015)工作基础上引入动态图模型更为鲁棒的加权特征表示,有效融合了图像块之间的局部和全局关系,并结合结构化支持向量机显著提升了跟踪性能。

1.2 基于深度模型的跟踪方法

1.2.1 深度相关滤波跟踪方法

相关滤波跟踪方法是通过循环矩阵的特性快速学习一个高判别力的相关滤波器,从而快速、稳健跟

踪。由于其计算效率高、性能好的特点,引起了研究者的广泛研究,但前期的一些工作仅使用了手工设计特征,处理复杂场景和环境因素的能力有限。一些研究者结合深度特征和相关滤波模型,取得了显著的性能提升。Ma 等人(2015)使用多层深度特征训练相关滤波器,有效地利用了多尺度信息,实现了由粗到细的融合和定位。后续有一系列研究工作如何更好地融合不同尺度特征下的滤波器及其预测的响应图(Qi 等, 2016; Wang 等, 2018b; Danelljan 等, 2016, 2017; Valmadre 等, 2017; 张艳琳 等, 2020)。

1.2.2 分类网络跟踪方法

这类方法一般通过多域卷积神经网络提取出通用的深度特征,然后利用第 1 帧的正负样本训练域分类器,并通过后续的正负样本进行微调域分类器,从而实现目标实例的有效跟踪。Nam 和 Han (2016)首次提出基于多域卷积神经网络的 RGB 视觉跟踪方法,取得了优越的跟踪性能,后续许多研究工作基于该模型进行了拓展。Song 等人(2018)引入生成对抗网络生成具有遮挡属性的样本,使得跟踪模型可以更好地应对遮挡挑战。Jung 等人(2018)引入感兴趣池化对齐(region of interest (ROI) align) 操作,避免了重复的特征提取过程,在保持跟踪性能的同时极大提高了跟踪效率。

1.2.3 孪生网络跟踪方法

该类方法将视频目标跟踪视为匹配任务,通过使用目标模板特征在后续帧中匹配出正确的目标位置。其基本架构是由共享参数的双分支卷积神经网络和互相关层组成,其中双分支网络用以分别提取模板特征和搜索区域特征,互相关层执行相关操作获得目标响应图,响应图上的峰值点就是目标位置所在。Bertinetto 等人(2016)较早提出基于模板匹配的孪生网络方法,促进了 RGB 视觉跟踪领域的快速发展,涌现了一大批基于孪生网络的跟踪算法(Li 等, 2018a; Zhang 和 Peng, 2019; Xu 等, 2020; 陈志良和石繁槐, 2022)等。

1.2.4 IoU 优化跟踪方法

为了充分发挥分类网络跟踪方法和孪生网络跟踪方法的优势, Bhat 等人(2019)提出了一种多任务跟踪方法,同时执行分类和回归任务。他们首先通过离线训练一个 IoU(intersection over union) 预测器实现尺度估计,然后通过一个在线训练的分类器来用于目标定位,这种分类和回归两个任务的协同工

作有效地提升了跟踪器的鲁棒性。后续也有一些工作对其进行了改进,有效地提升了对目标和背景的判别性(Bhat 等,2020)以及对相似干扰物的处理能力(Mayer 等,2021)。

1.2.5 Transformer 跟踪方法

Transformer 的注意力机制可以更好地建模全局依赖关系,在许多计算机视觉任务中都有着很好的应用,在 RGB 视觉跟踪中也取得了较大的成功。Wang 等人(2021a)利用 Transformer 结构建模了不同时刻的时序信息,丰富了跟踪模型中的时序关系。Chen 等人(2021)使用 Transformer 结构替代了孪生网络跟踪方法中的互相关操作,将模板信息融合到搜索区域中,可以更好地进行目标定位和尺度回归。Yan 等人(2021a)使用 Transformer 结构来预测目标的角点进行跟踪,此外还引入了时空信息,可以更好地适应目标的外观变化,进一步提升了基于 Transformer 跟踪方法的鲁棒性。然而,上述这些跟踪器仍然依赖于卷积神经网络提取特征,不能充分发挥 Transformer 全局建模的能力。为了解决这个问题,Cui 等人(2022)使用 Transformer 结构同时进行特征提取和信息交互,进一步简化了跟踪器的结构,并极大地提升了跟踪性能。

2 RGBT 视觉跟踪方法

从信息融合角度,本文把现有 RGBT 视觉跟踪算法划分为两大类:结合式融合和判别式融合。结合式融合是通过不同的融合方法结合不同模态信息,而判别式融合是挖掘不同模态的判别信息进行融合。图2展示了两类融合方式的过程。

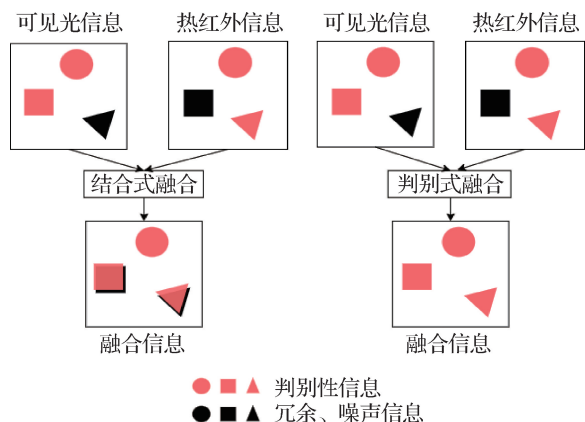


图2 两类融合策略示意图

Fig.2 Illustration of two kinds of fusion methods

2.1 结合式融合

现有结合式融合的 RGBT 跟踪算法可以分为5类:稀疏表示融合、协同图表示融合、模态共享和特定信息融合、基于属性的特征解耦融合以及其他结合式融合方法,下面分别予以介绍。

2.1.1 稀疏表示融合

由于稀疏表示模型具有非常好的抑制特征噪声的能力,因此一些研究者将稀疏表示引入到 RGBT 跟踪中,旨在联合建模多模态特征的稀疏表示并抑制特征噪声。这些方法通常将传统贝叶斯滤波框架拓展成多模态贝叶斯滤波框架实现跟踪,其中运动模型和传统方法一致,而似然模型一般是基于融合多模态信息的稀疏表示模型。

Wu 等人(2011)首次将稀疏表示模型应用到 RGBT 跟踪中,他们将可见光和红外粒子(图像块)的特征拼接成1维的特征向量,然后采用稀疏表示方法实现多模态跟踪。Liu 和 Sun(2012)设计一种联合稀疏表示模型,分别计算每种模态的重构系数,并通过最小化操作融合两种模态的重构系数,然后计算粒子的置信度作为似然分数。Li 等人(2016b)提出一种拉普拉斯稀疏表示模型,利用了粒子之间的相似性提高稀疏表示的鲁棒性。Lan 等人(2020)设计了一种基于模态相关性感知稀疏表示模型,通过低秩正则化和稀疏正则化分别利用不同模态之间的相关性和自适应选择具有代表性的粒子处理外观变化。

然而,这些方法并没有考虑到模态可靠性对跟踪性能的影响,同时也忽略了前景和背景的判别信息,限制了跟踪算法的鲁棒性。为了解决这些问题, Li 等人(2016a)提出了一种协同稀疏表示模型,对不同的模态数据通过传统的联合稀疏表示模型进行建模,同时为每个模态引入一个可靠性权重,实现多个模态的自适应融合。此外,该工作构建了一个 RGBT 跟踪数据集 GTOT (grayscale-thermal object tracking),为该领域的研究提供了一个统一的评价平台。为了增加模态权重计算的可靠性, Li 等人(2017b)设计了一种多任务拉普拉斯稀疏表示模型,利用图像块之间的相似性优化稀疏系数。为了克服遮挡、嘈杂背景等因素的影响, Lan 等人(2019a)研究了如何有效地从潜在污染样本中学习可靠的多模态稀疏表示,并将非负约束引入优化框架,增强了模型鲁棒性。

尽管这些稀疏表示融合方法在抑制特征噪声方

面表现良好,但是由于稀疏表示模型的在线优化比较耗时,导致这些算法大都不能满足实时性要求,且该类模型一般基于像素的特征表示,对复杂场景和环境具有较差的鲁棒性。

2.1.2 协同图表示融合

基于图模型的视觉跟踪方法能够较好地抑制嘈杂背景对特征表示的影响,受到了一定的关注。这类方法一般是将样本区域划分成多个不重叠的图像块,

把图像块的颜色和梯度特征连接在一起作为该样本的特征表示,然后在结构化支持向量机跟踪框架进行跟踪。但该特征易受背景成分的影响,为此,学者们通过为每个图像块分配一个权重抑制背景块的影响,并以图像块为图结点构建图模型,优化图像块权重的计算。基本流程如图3(Li等,2018c)所示。在RGBT跟踪中,学者们主要考虑如何设计鲁棒的协同图模型,有效利用多模态信息提高权重计算精度和鲁棒性。

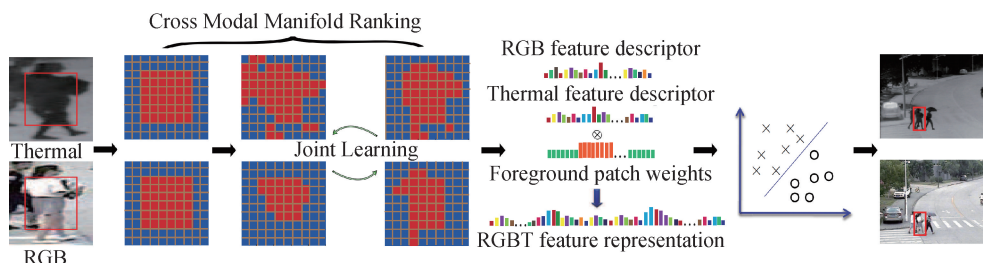


图3 基于跨模态排序算法的RGBT跟踪流程图(Li等,2018c)

Fig. 3 Flow chart of RGBT tracking based on the cross-modal ranking algorithm(Li et al., 2018c)

Li等人(2017c)首次将图模型应用到RGBT跟踪中,提出了加权稀疏表示正则化协同图模型,联合优化模态权重、稀疏表示系数、图结点相似性和图结点权重。进一步,Li等人(2019a)提出了一种基于种子点稀疏优化的协同图模型,通过L1范数约束权重变量和种子点权重的关系,以此降低噪声种子点的影响。为了同时考虑模态间的协同性和异质性,Li等人(2018c)设计了跨模态排序图模型,引入了模态之间的跨模态软一致性有效融合多模态信息。Li等人(2022a)改进了跨模态排序图模型,通过分离不同特征(颜色和梯度特征)的排序过程,在多任务学习框架下实现更为精准的权重计算。为了联合利用局部信息和全局信息,Li等人(2019c)提出了一种局部—全局多图模型,通过利用全局和局部线索挖掘多模态图像块之间的内在关系,实现多模态信息的协同融合。为了提高对初始化权重的鲁棒性,Li等人(2018d)提出了一种两阶段排序模型,通过第1阶段的排序结果重新初始化结点权重,以此降低初始化权重的噪声影响。Shen等人(2022)提出了一种协同低秩图模型,将输入特征分解成低秩特征部分和稀疏噪声部分,利用协同图学习算法对其进行动态更新。Lan等人(2019c)提出了一种稀疏正则化锚点协同图模型,通过图结点亲和性的学习构建锚点协同图,进而设计了标签预测模型,通过少量标注样本准确预测样本标签。

协同图表示融合模型使用局部特征表示目标,能够对目标变形、部分遮挡等挑战有着较好的鲁棒性,且通过权重抑制背景成分影响,能够较好地应对嘈杂背景等因素。但这类算法一般需要迭代优化多个变量,因此具有较差的跟踪效率。此外,虽然使用了颜色和梯度特征,相对于稀疏表示融合方法有了较高的表示能力,但与深度特征相比,这类方法的表示能力仍旧较弱。

2.1.3 模态共享和特定信息融合

一些学者设计专门的网络协同提取模态共享和模态特定信息,以此提高多模态特征的表征能力,然后结合模态共享和特定特征实现鲁棒的视觉跟踪(Lu等,2021)。通过在特征提取阶段构建不同的子网络分别提取模态共享特征和模态特定特征。

Li等人(2019b)提出一种多适配网络,包括通用适配模块、模态特定适配模块和实例适配模块,分别提取模态共享特征、模态特定特征和实例感知特征。具体地,通用适配模块使用多模态数据学习模态共享表示,模态特定适配模块则分别利用单模态数据学习模态特定表示。通过渐进式训练策略实现了对多模态特征的解耦建模,有效提升了多模态特征的表征能力。此外,实例适配模块通过在线更新机制能够有效应对目标及其周围环境的动态变化。为了更好地解耦模态共享和模态特定特征,Lu等人(2021)进一步设计了基于多核均方差函数的层级

差异损失,以监督模态共享特征和模态特定特征的解耦学习,提升模型判别能力。具体地,该方法通过在多个层级上施以最大均值方差损失约束,最大化模态共享特征和模态特定特征之间的分布差异,同时最小化模态共享特征分布距离。此外,该方法在实例适配模块中引入在线的模态权重预测网络实现了更鲁棒的自适应融合。类似地,Peng 等人(2021)提出了一种具有耦合滤波器的双流卷积互补特征融合网络,有效提取红外和可见光图像之间的共享特征和特定特征。

尽管模态共享和模态特定信息融合方法提供了一种有效和通用的多模态结合式融合框架,但在模态特定特征学习中缺少模态间信息交互,一定程度上限制了融合性能。

2.1.4 基于属性的特征解耦融合

深度跟踪模型往往依赖大规模标注数据,且很难克服多种多样的挑战因素。Qi 等人(2019)提出了一种基于属性表示学习的跟踪算法,通过基于属性的多分支网络学习目标相应属性的特定表示,可以有效解耦目标外观的建模过程,达到仅需少量参数就可以有效学习复杂目标外观变化的目的,并克服了对大规模训练数据的依赖。借鉴该解耦的方式,一些学者提出了基于属性的多模态特征解耦融合算法,一般采用多阶段训练方法建模不同属性的参数,如图4所示。

Li 等人(2020)提出了一种挑战感知网络,实现适用于 RGBT 跟踪任务的解耦框架。具体地,针对可见光和红外数据既面临着共同的挑战(如快速运动、遮挡和尺度变化等挑战),又面临着特定的挑战(如光照变化和热交叉等挑战),Li 等人(2020)设计了一些参数共享的挑战感知分支解决模态共享挑战和一些参数独立的挑战感知分支解决模态特定挑战。由于模态特定挑战通常包含模态互补信息,因此还设计了指导模块显式地将较好模态的判别性信息传播到较差的模态中,有效增强了较差模态的特征判别能力并抑制噪声。此外,针对跟踪过程中挑战不可知的问题,他们还设计了一种自适应聚合模块融合所有挑战感知分支的输出结果,学习过程如图4所示。由于这些挑战感知分支能够在特定挑战下对目标外观进行建模,因此即使在训练数据不足的情况下,也可以有效地学习目标表示。

然而上述工作只考虑了5种特殊的挑战,而真

实场景会包含更多的挑战因素。为了解决这个问题,Zhang 等人(2021b)提出了一种自适应属性驱动表示的 RGBT 跟踪算法,设计了一个通用分支学习其他挑战因素下的特征表示。为了训练这些不同的属性分支,他们设计了一种数据生成策略,为每种属性生成对应的训练数据,并设计了一种属性集成网络,在通道级和空间级进行自适应聚合,以适应属性不可知的跟踪过程。考虑到融合过程的复杂性,Xiao 等人(2022)通过属性信息解耦融合过程,提出了一种基于属性的渐进式融合网络,设计了参数独立的属性融合分支来捕捉不同属性下的多模态融合方式,设计了一种基于注意力机制的挑战聚合模块来自适应地聚合不同挑战下的特征。考虑到 Transformer 可以捕捉全局信息,他们设计了自注意力增强模块,用于增强聚合特征和主干网络特征,并使用交叉注意力将聚合特征以更有效的方式与主干网络特征进行自适应融合。

基于属性的特征解耦方法提供了一种有效的目标表示方式,并克服了对大规模训练数据的依赖。但在实际应用中,挑战属性是多种多样的,如何充分挖掘属性信息实现更为精准的目标表示,以及设计合适的网络分支建模不同属性下的目标表示还有待进一步探索和研究。

2.1.5 其他结合式融合方法

多模态信息的像素级融合可以直接使用单模态跟踪方法实现跟踪过程。Schnelle 和 Chan(2011)利用13种空间域和基于金字塔的像素级融合算法融合可见光和红外图像,提高了现有视频监控系统的跟踪性能,并采用基于背景减法的算法进行目标定位。Chan 和 Schnelle(2012)在此基础上进一步进行了研究,使用13个基于像素的图像融合算法进行图像融合,确定了5种可以显著提升跟踪性能的基于金字塔的方法,并在后续的工作中分析了这些融合方法的计算效率。然而,像素级融合要求严格配对多模态数据,且很难充分挖掘多模态有效信息。为了验证像素级、特征级和决策级融合的有效性,Zhang 等人(2019a)提出了一种端到端的多模态融合跟踪算法,采用数据生成策略解决训练数据不足问题,通过对比不同融合策略的跟踪结果,最终采取了在特征级融合多模态特征,并在国际权威的目标跟踪比赛 VOT2019 中 RGBT 跟踪赛道获得冠军。

由于相关滤波模型具有速度快、性能高的特点,

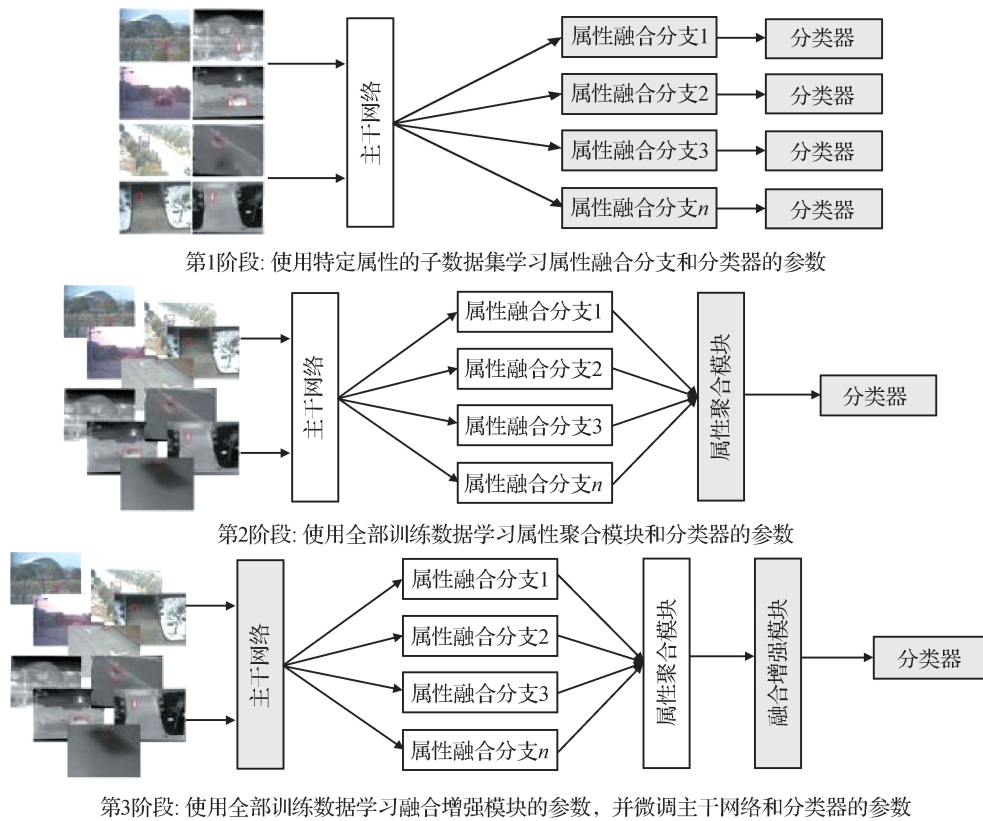


图4 基于属性的特征解耦融合网络模型及渐进式学习策略

Fig. 4 Network of feature decoupling fusion based on attributes and progressive learning strategy

因此一些学者在该框架下研究 RGBT 跟踪算法。Wang 等人 (2018b) 首次将相关滤波方法引入到 RGBT 跟踪任务, 提出了一种软一致性相关滤波算法, 充分考虑了多模态数据之间的协同性和异质性, 并设计了一种加权融合机制来预测最终的响应图。Zhai 等人 (2019) 利用低秩约束联合学习不同模态的相关滤波模型, 实现模型的协同学习, 对光照变化和恶劣天气等挑战具有很好的鲁棒性。Feng 等人 (2020) 提出了一种自适应时空正则化相关滤波模型, 采用自适应加权融合方法和多尺度搜索机制, 实现了决策级的多模态信息融合。Yun 等人 (2019) 提出了一种判别融合相关滤波模型, 通过有效地融合可见光和热红外模态的特征及响应图提高跟踪性能。基于相关滤波的 RGBT 跟踪算法虽然具有较快的跟踪速度, 但由于一般使用传统特征, 因此性能相对深度模型还有很大的差距。

2.2 判别式融合

结合式融合方法通过结合不同模态的所有信息融合多模态数据, 但也往往会带来特征的冗余和噪声。而判别式融合方法旨在挖掘每个模态的判别性

信息, 通过仅聚合判别性信息降低了特征冗余和噪声, 实现多模态数据的有效融合。现有判别式融合的 RGBT 跟踪算法可以分为 4 类: 特征选择融合、基于注意力机制的自适应融合、双向增强融合以及其他判别式融合方法。

2.2.1 特征选择融合

特征选择融合方法通过一些准则选择多模态判别性特征用于视觉跟踪。一方面避免了数据噪声干扰, 有利于提高跟踪性能; 另一方面消除了数据冗余, 有利于提高跟踪效率。

Li 等人 (2018b) 提出了一种基于双流卷积神经网络的 RGBT 跟踪算法, 通过双流卷积网络用于提取不同模态的深度特征, 通过选择融合网络选择判别性特征、去除冗余和噪声特征, 用于后续跟踪。其中, 特征图的重要性是根据训练过程的损失变化大小定义的。不同于上述方法, Zhu 等人 (2019b) 提出了基于自适应剪枝策略的融合方法, 通过一种多模态协作的方式来修剪所有模态的联合特征, 并采用全局平均池化和加权随机选择操作对每个通道进行打分, 最后通过一种硬选择机制保留高分数的特征通

道,以此去除特征冗余和噪声。进一步,Zhu 等人(2022)将该剪枝策略分别应用于单模态分支和多模态融合分支,在更充分利用所有模态特征的同时消除了特征冗余和噪声。

判别性特征选择的标准是这类方法的关键,现有基于损失函数变化大小和损失函数驱动的选择算法由于选择标准的精准度和鲁棒性问题,往往会去除有用信息,尽管显著降低了计算开销但也限制了跟踪性能。

2.2.2 基于注意力机制的自适应融合

不同模态数据包含的信息量和判别性都是不同的,如何挖掘和利用有用信息、抑制噪声信息对于 RGBT 跟踪是非常重要的。最有代表性的一类方法是通过注意力机制计算数据的可靠性,包括模态可靠性、空间可靠性和通道可靠性,从而实现多模态信息的自适应融合,有效抑制数据噪声。基于注意力机制的自适应融合基本流程如图 5 所示。

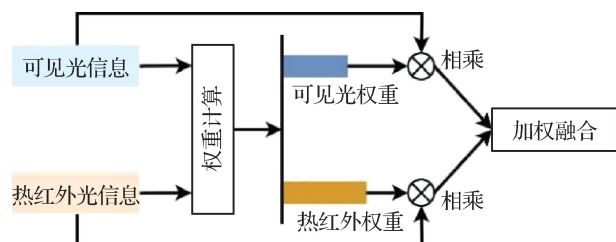


图5 基于注意力机制的自适应融合示意图

Fig.5 Illustration of attention based adaptive fusion

早期工作一般通过单类型注意力机制实现自适应融合,Zhu 等人(2019a)提出了一种质量感知特征聚合网络,同时考虑了不同模态的可靠性和不同层特征的可靠性,对多层深度特征进行自适应聚合。

为了建模简单样本与困难样本之间的关系,Tu 等人(2022)提出了一种多模态多边界度量学习框架,通过探索和利用简单正样本、简单负样本、困难正样本和困难负样本之间的关系提升特征的鲁棒性,并设计了一个基于注意力机制的质量感知聚合模块。为了提高算法实时性,Zhang 等人(2019b)在多模态孪生跟踪模型中引入了基于注意力机制的特征融合网络,用不同模态的响应图计算不同模态的可靠性权重。进一步,Zhang 等人(2020e)提出了一种基于动态孪生网络的多层融合跟踪算法,利用注意力机制计算多层特征的可靠性权重,实现多层次多模态特征的自适应融合。

然而,上述方法只考虑了模态可靠性,虽然能够抑制噪声模态的影响,但也会抑制模态中的判别信息,因此,一些学者通过引入通道注意力和空间注意力凸显更为细粒度的判别性特征。Gao 等人(2019)提出了一种渐进式的深度自适应融合网络,通过基于通道注意力机制的自适应融合模块估计不同模态特征不同通道的可靠性权重,有效地降低了特征噪声和冗余信息。Zhang 等人(2021c)进一步考虑了空间位置注意力对多模态信息融合的影响,提出了一个基于空间注意力机制的多模态融合网络,估计不同模态响应图不同空间位置的可靠性权重。

为了进一步挖掘注意力机制在多模态信息融合上的潜力,一些学者使用混合注意力机制,实现鲁棒的多模态信息融合,并广泛应用于 RGBT 跟踪网络中。Xu 等人(2022)提出了一种基于通道注意力和空间注意力的多模态跨层双线性池化的 RGBT 跟踪算法,使用通道注意力实现多层特征通道的统一,然后通过基于空间注意力机制的质量感知融合模块预测两个模态特征中每个空间位置的可靠性权重,自适应地聚合不同模态不同层交互的双线性池化特征。Zhang 等人(2022a)提出了一种多层级多模态融合跟踪算法,同时考虑了基于分布约束的互补图像融合、基于通道注意力的判别特征融合和基于空间注意力的自适应决策融合,进一步提升了跟踪性能。

2.2.3 双向增强融合

特征选择融合和基于注意力机制的融合方法可以分别看成硬选择和软选择策略,而选择策略有可能会抑制一些有用信息。例如,为每个模态进行可靠性加权,那么该模态的所有特征都将被削弱。为此,一些学者通过双向增强融合策略避免有用信息的抑制,主要是利用一个模态的判别性信息增强另外一个模态的特征,同时会抑制该模态的数据噪声。

Zhang 等人(2022b)设计了一种互补感知的多模态特征融合网络,通过权重生成模型选择一个模态的判别性特征,利用元素级相加的方式将判别性特征传播到另一个模态中。Mei 等人(2021)提出了一种层级双传感器交互网络,通过设计数据编码模块和特征交互模块实现模态互增强。特征交互模块结合两个模态的特征计算可靠权重,以选择每个模态的判别性特征,然后采用元素级相加方式实现模态互增强。Zhang 等人(2020a)提出了一种基于模

态感知的注意网络,通过跨模态融合子网络引导多模态判别性特征的充分交互,然后通过元素级相加将交互特征传播到原始特征上,增强各个模态的特征表示。Wang 等人(2020a)基于自注意机制设计了一个跨模态模式传播模块,该模块首先对两个模态特征执行自相关注意力操作,获得每个模态的自相关关系矩阵,而后在两个关系矩阵中执行融合操作,以实现模态间的互增强。此外该方法还引入时间域信息,通过长时上下文传播模块进一步增强当前的模态特征。为了充分利用低质量模态判别性信息,Lu 等人(2022)提出了一种对偶门控互条件网络,通过对偶门控模块提取模态判别性特征,对另一个模态特征学习进行指导,同时通过双门控机制过滤冗余和噪声信息。这类方法虽然取得了较好的跟踪性能,但模型稍显复杂,跟踪效率偏低。

2.2.4 其他判别式融合方法

以上判别式融合方法均是基于深度学习框架的,而一些学者在传统跟踪框架中也探索了判别式融合方法。Lan 等人(2018)提出了一种判别性学习框架,在统一的计算模型中联合优化特征表示、可靠性权重和分类器,从而使这 3 个学习任务相互受益,获得更好的性能。为了缓解模态差异问题,Lan 等人(2019a)提出了一种模态一致性的稀疏表示模型,并通过引入特征判别性一致性约束,实现多模态信息的协同建模和判别。

基于传统模型的判别式融合方法尽管一定程度上解决了 RGBT 跟踪中的部分挑战,但其跟踪性能难以与基于深度学习的多模态视觉跟踪方法相比,而且这类方法往往会涉及较为复杂的迭代优化过程,使得跟踪效率偏低。

2.3 融合策略总结和比较

得益于可见光和红外的强互补优势,当前的 RGBT 跟踪领域引起了众多学者的关注,并得到了快速发展。尽管各种多模态融合策略可以解决多模态跟踪的一些问题,但是这些方法都有着各自的优缺点。因此,本文在表 2 中总结了当前主流融合方法以及它们优缺点的比较。

2.4 RGBT 跟踪数据集

大规模数据集对 RGBT 视觉跟踪领域有着至关重要的作用,一方面可以帮助训练深度学习算法,另一方面可以综合评价不同跟踪器的性能,有效促进该领域的研究和发展。Li 等人(2016a)提出了第 1 个

大规模 RGBT 视觉跟踪数据集 GTOT,包含有 50 对不同场景和条件下的 RGBT 视频序列,并提供了 7 种挑战属性标注,用来评价算法在不同挑战属性下的性能。为了进一步丰富 RGBT 视觉目标跟踪数据集的多样性,Li 等人(2017a)提出了一个更大规模的 RGBT 视觉跟踪数据集 RGBT210,包含有 210 对 RGBT 视频序列。然而,RGBT210 数据集的标注不够精细。为了解决这一问题,Li 等人(2019a)对其进行了改善,提出了更大规模的 RGBT 跟踪数据集 RGBT234,并提供了 12 种挑战属性标注。虽然上述 3 个数据集对于评估不同算法的性能已经足够大,但是还不能满足深度跟踪器对大规模训练数据的需求。为了解决这一问题,Li 等人(2022a)提出当前最大规模的 RGBT 跟踪数据集 LasHeR,提供了 1 224 对 RGBT 视频序列,并标注了 19 种挑战属性,这将有利于推动 RGBT 跟踪领域的方法研究。此外,Zhang 等人(2022a)提出了面向无人机平台的 RGBT 跟踪数据集 VTUAV(visible-thermal unmanned aerial vehicle),并提出了 RGBT 长时跟踪问题,这为 RGBT 领域拓展了新的研究空间。为了更为清晰地理解不同数据集的特点,本文在表 3 中总结了主流 RGBT 跟踪数据集的细节。

3 其他多模态视觉跟踪方法

除了 RGBT 跟踪任务,多模态视觉跟踪还包括 RGBD 跟踪、RGBE 跟踪和 RGBL 跟踪。由于这些领域的研究相对较少,因此本文将简要阐述这 3 种多模态视觉跟踪任务的研究进展。

3.1 RGBD 视觉跟踪

RGBD 跟踪是引入与可见光图像对应的场景深度图,有效克服可见光模态的局限性,从而提高跟踪鲁棒性。特别是在低光照和遮挡情形下,深度图能够更好地跟踪目标,并推理目标的遮挡状态,如图 6(Liu 等,2019)所示,图中,中间一列表示特征响应,最后一列表示通过深度信息感知遮挡状态。2012 年,García 等人(2012)提出了第 1 个 RGBD 跟踪方法,通过浓缩算法融合了灰度特征、颜色特征和深度图特征,用以区分前背景目标。2013 年,Song 和 Xiao(2013)提出了第 1 个大规模 RGBD 数据集,包含了 100 条配对的 RGB 和深度视频序列,引起了计算机视觉领域研究者的广泛关注。在早期 RGBD

表 2 主要 RGBT 跟踪方法总结与优缺点比较

Table 2 Overview of main RGBT tracking methods and their advantages and disadvantages

	融合策略	优点	缺点	相关文献	与单模态跟踪的关系
结合式融合	稀疏表示融合	对特征噪声有良好的抑制能力	在线优化较为耗时,使用传统特征,跟踪精度和鲁棒性较低	Lan 等, 2019a; Lan 等, 2020; Li 等, 2016a, b, 2018e; Liu 和 Sun, 2012; Liu 等, 2019	稀疏表示模型(侯跃恩 等, 2013)的多模态拓展,一般通过多模态数据协同优化稀疏表示系数,达到多模态融合的目的
	协同图表示融合	对目标变形、部分遮挡等挑战有着较好的鲁棒性	需要迭代优化多个变量,跟踪效率较差	Lan 等, 2019; Li 等, 2022a, 2017c, 2018c, 2019c; Shen 等, 2022	动态图表示模型(Li 等, 2019a)的多模态拓展,通过多模态数据构建协同图表示模型,联合优化图结构、边权和点权
	模态共享和特定信息融合	有效建模模态协同和抑制信息,多模特征的表征能力强	模态间信息缺乏交互,性能受到一定限制	Li 等, 2019b; Lu 等, 2021; Peng 等, 2021	基于多域卷积神经网络(Nam 和 Han, 2016),设计了共享卷积模型和特定卷积模块,分别提取模态共享特征和特定特征
	基于属性的特征解耦融合	仅需少量参数即可处理复杂外观变化,不依赖大规模多模态训练数据	属性划分依赖额外标注,且训练策略较为烦琐	Li 等, 2020; Xiao 等, 2022; Zhang 等, 2021b	基于多域卷积神经网络(Nam 和 Han, 2016),嵌入了基于属性的外观建模分支,极大缓解了对大规模多模态训练数据的依赖
判别式融合	特征选择融合	噪声和冗余信息低,计算开销小	部分有用信息丢失	Li 等, 2018b; Zhu 等, 2019b, 2022	基于全卷积跟踪网络(Wang 等, 2015),设计了双分支网络模型,并根据通道重要性选择标准,选择高判别性特征
	基于注意力的自适应融合	突出判别性信息并抑制噪声和冗余信息	部分有用信息有可能被抑制,特别对低质量模态信息的利用不充分	Gao 等, 2019; Tu 等, 2022; Xu 等, 2022; Zhang 等, 2021c, 2022d, 2020d, 2012; Zhu 等, 2019a	基于多域卷积神经网络(Nam 和 Han, 2016)和 DiMP(Bhat 等, 2019)等跟踪模型,设计了双分支模型,并嵌入注意力机制实现多模态特征的自适应融合
	双向增强融合	判别性信息的利用程度高,噪声抑制效果显著	模型结构较为复杂,效率偏低	Lu 等, 2022; Mei 等, 2021; Wang 等, 2020a; Zhang 等, 2020a, 2022b	基于多域卷积神经网络(Nam 和 Han, 2016),设计了特征交互模块和互增强传播策略,以此实现对所有模态判别性信息的充分利用

表 3 RGBT 跟踪数据集比较

Table 3 Comparison of RGBT tracking datasets

数据集	序列数	分辨率/像素	最小帧数	最大帧数	总帧数	目标类别数	挑战属性数	出版年份
GTOT	50	384 × 384	40	376	7.8 K	7	7	2016
RGBT210	210	630 × 460	40	4 140	104.7 K	12	12	2017
RGBT234	234	630 × 460	40	4 140	116.7 K	12	12	2019
LasHeR	1 224	630 × 460	57	12 862	734.8 K	32	19	2021
VTUAV	500	1 920 × 1 080	196	27 213	1.7 M	27	13	2021

跟踪研究中,通常是针对单模态跟踪器的拓展展开的,通过从深度图中获取手工设计特征或跟踪结果,解决特定挑战下的跟踪问题。例如, Hannuna 等人(2019) 基于现有的该相关滤波算法(kernel correlation filter, KCF),联合颜色和深度线索实现鲁棒跟

踪,同时利用深度数据调整目标尺寸。尽管最近的工作引入深度学习技术,但大都基于现有单模态跟踪器的拓展。本文大致将当前 RGBD 跟踪方法分为两类:早期融合和后期融合,下面予以分别介绍。

早期融合是指像素级和特征级实现两种模态

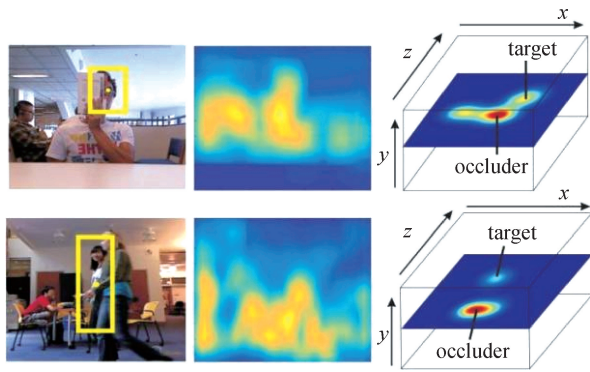


图6 RGBD 目标定位示意图(Liu 等,2019)

Fig. 6 Illustration of RGBD target localization
(Liu et al., 2019)

的信息融合。基于像素级融合的 RGBD 跟踪算法(Liu 等,2019;Kart 等,2019;Xie 等,2019;Gutiev 和 Debono,2019)通常是将深度图和可见光图像形成一种四通道输入,然后使用单模态跟踪算法实现目标跟踪;基于特征级融合的 RGBD 跟踪算法(Wang 等,2020b;Bibi 等,2016;Meshgi 等,2016;Ma 和 Xiang,2017;Xiao 等,2018;Liu 等,2020;Zhao 等,2020;Yan 等,2021b)通常是对深度图和可见光图像分别提取手工特征或者深度特征,然后进行特征融合,并利用融合特征进行跟踪。例如,Wang 等人(2014)同时结合光流、颜色和深度图特征预测目标位置。此外,基于深度跟踪框架 ATOM(arate tracking by orlap maximization)和 DiMP(discriminative model prediction),Yan 等人(2021b)提出了一种双流网络,通过构建一个额外的深度网络分支提取两种模态的深度特征。

在 RGBD 跟踪领域中,大部分研究者更侧重于在跟踪结果中进行修正式融合,也称为后期融合。一些研究算法通过使用深度模态的几何或深度先验信息来修正当前 RGB 跟踪器的结果,这在早期的工作中得到广泛使用(Zhong 等,2015;Chen 等,2015;Shi 等,2015;Kart 等,2018;Zhai 等,2018;Ding 和 Song,2015)。Shi 等人(2015)提出使用来自可见光和深度图像 HoG(histogram of oriented gradient)的检测结果,联合可见光和深度图像各自的跟踪结果,并以加权融合的方式聚合上述结果。此外,根据深度图像的直方图判断当前帧遮挡情况,从而对聚合后的跟踪结果进行修正。为了进一步解决跟踪中的遮挡问题,Zhai 等人(2018)提出了一种基于深度图像

和定向梯度直方图特征的目标遮挡判断机制,并利用相关最大似然估计粒子滤波器算法,建立了目标预测—跟踪—优化—再检测的跟踪机制,极大地提升了模型在遮挡场景中的跟踪性能。类似地,Ding 和 Song(2015)同样使用深度直方图来判断是否发生遮挡,但是额外引入了基于深度图像的分割图重新定位目标。

在 RGBD 多模态视觉跟踪中,由于深度模态的成像质量的限制,现有的大多数 RGBD 跟踪工作仍然以深度模态作为一种辅助的信息使用。无论是早期融合或者后期融合,通常聚焦于挖掘深度模态图像的几何或空间性质,获得跟踪目标的先验信息。然而缺乏对深度模态特征信息的充分挖掘,这限制了该领域的基于多模态信息融合的跟踪方法的发展。

3.2 RGBE 视觉跟踪

事件相机由于其低延迟、高动态范围的特点,使得异步跟踪成为可能。为了解决可见光相机存在的运动模糊问题,一些学者联合可见光和事件相机的信息流,实现更可靠的跟踪结果,图 7(Wang 等,2022)展示了 4 个典型场景下可见光和事件信息的互补优势。为了提供统一评测基准,Liu 等人(2016)构建了一个小规模 RGBE 跟踪数据集 Ulster,并提出了一个基准方法。他们将跟踪过程分为 3 个步骤:1)使用聚类的跟踪方法和事件相机数据生成感兴趣区域;2)使用卷积神经网络和可见光数据将这些感兴趣区域划分为前景和背景;3)使用粒子滤波从这些感兴趣区域推断出目标的位置。为了获取更丰富的事件数据,Hu 等人(2016)在现有跟踪数据集的基础上生成对应的事件流,构建了一个大规模 RGBE 跟踪数据集 VOT-DVS(visual object tracking dynamic vision sensors)和 TD-DVS(tracking dataset dynamic vision sensors),可以应用于目标跟踪、行为识别和目标识别等领域。

早期的 RGBE 跟踪方法通常是提取 RGB 的特征,然后利用事件流辅助检测和跟踪,如 Tedaldi 等人(2016)提出的基于迭代的几何配准方法、Kuang 等人(2016)提出的基于事件的视觉测程算法和 Gehrig 等人(2018,2020)提出的最大似然生成事件模型等。

为了更充分融合多模态信息,Huang 等人(2018)提出了一种基于支持向量机的事件引导跟踪模型,利用在线自适应区域搜索实现更准确的目

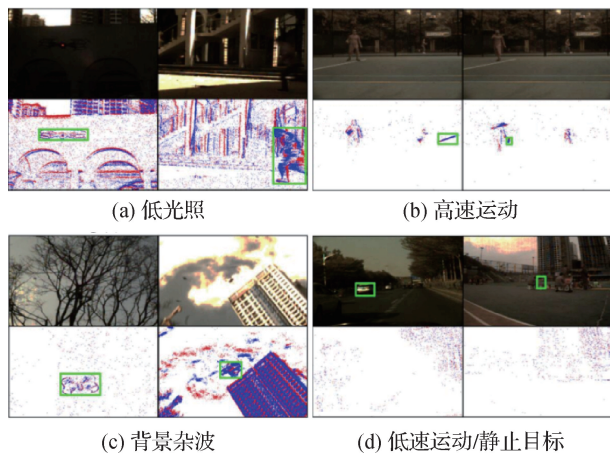


图7 不同挑战因素下的可见光成像和事件成像优势 (Wang 等, 2022)

Fig. 7 Advantages of visible and event cameras under different challenges (Wang et al., 2022) ((a) low illumination; (b) high speed motion; (c) clutter background; (d) slow motion/static target)

标定位,利用事件数据和可见光数据重建样本以实现多模态信息融合。Yang 等人(2019)设计了一种时间互补滤波器和注意力机制联合处理可见光数据和事件流,并构建了一个更大规模的 RGBE 跟踪数据集 NFS-DAVIS (need for speed-dynamic and active pixel vision sensor)。

随着深度学习技术的发展,一些学者也探索了基于深度学习的 RGBE 跟踪框架,并取得了较好的跟踪效果。Zhang 等人(2021a)提出了一种跨域注意力机制实现特征增强,并设计了一种加权方案自适应地平衡两种模态的贡献,能够有效地、自适应地融合两种模态的有效信息。此外,他们也提出了一个大规模 RGBE 跟踪数据集 FE108。为了进一步推进 RGBE 跟踪领域的发展,Wang 等人(2022)构建了一个更为完善的 RGBE 跟踪数据集 VisEvent,并提出了一种基于跨模态 Transformer 的多模态信息融合方法,以实现可见光数据和事件数据之间更有效的特征融合。

在 RGBE 视觉跟踪中,由于可见光和事件数据格式的不同,为了实现多模态信息融合,现有做法通常将事件流转换为事件图像,这不可避免地带来了信息损失,对事件流数据进行有效建模,进而挖掘可见光和事件流数据的互补优势,对提升 RGBE 跟踪器的性能具有重要意义。此外,由于可见光和事件相机捕获信息的频率不同,在跟踪中关联可见光和

事件相机的信息存在着很大的挑战。

3.3 RGBL 视觉跟踪

RGBL 跟踪是通过在第 1 帧中引入额外的对跟踪目标的语言描述,结合目标的 RGB 图像来执行后续的跟踪任务。Li 等人(2017d)提出这一研究问题的一个基准算法 LSNNet,并且该工作基于现有的跟踪数据集 OTB (object tracking benchmark) 构建了第 1 个 RGBL 数据集 Lingual OTB99,其中跟踪示例如图 8 (Li 等, 2017d) 所示。



图8 RGBL 跟踪过程示例 (Li 等, 2017d)

Fig. 8 Examples of RGBL tracking (Li et al., 2017d)

图像和语言模态之间存在较大的差异,现有 RGBL 跟踪方法 (Yang 等, 2021; Wang 等, 2021b) 主要是将目标的两种模态特征进行融合,通过充分利用模态间的互补信息丰富目标表示。Yang 等人(2021)提出了一个多任务学习框架,将 RGBL 跟踪解耦成 3 种子任务:视觉定位、跟踪和融合,从而实现在第 1 帧跟踪框缺失条件下的鲁棒跟踪。此外也有一些工作 (Feng 等, 2021a, b) 关注于后期融合。Feng 等人(2021a)通过共享搜索分支,将语言和图像分布作为模板分支分别进行跟踪,然后聚合两者的响应图获得最后的跟踪结果。

在 RGBL 多模态视觉跟踪中,由于其图像和文本间的巨大差异,现有的工作通常都集中于特征空间中的多模态融合方法。此外在对目标描述中,静态的文本模态和动态的图像模态之间的不一致性对该领域是一个重要的挑战,但是鲜有探索。

4 结 语

本文对 RGBT, RGBD, RGBE, RGBL 不同的多模态视觉跟踪方法进行了总结和分析。由于目前主流的多模态跟踪方法是基于 RGB 和热红外数据,所以

本文从数据融合的角度详细阐述了 RGBT 视觉跟踪方法,并简述了其他 3 种多模态跟踪算法的研究进展。

4.1 工作总结

基于对现有研究工作的梳理,RGBT 视觉跟踪方法可以归纳为结合式融合和判别式融合两大类。其中,结合式融合的研究可以分为两个阶段,早期研究者聚焦于稀疏表示融合和协同图表示融合,这些工作对 RGBT 跟踪的研究有着重要的推动意义。但是,这些方法受限于较弱的特征表示能力导致跟踪性能不佳。随着深度学习技术的发展,当前结合式融合方法得到了更深入的研究。例如,基于模态共享和模态特定信息融合策略以及基于属性的特征解耦融合策略,通过充分利用数据驱动的解耦表征学习模式极大提升了 RGBT 跟踪性能。然而这些方法往往依赖于对训练数据的人工划分和标注。基于判别式融合方法的 RGBT 跟踪研究工作在深度学习时代兴起并取得极大的关注。特征选择融合策略和基于注意力机制的自适应融合策略是多模态融合领域中常用的融合方法并在许多领域都取得了一定的成功,然而在 RGBT 跟踪中由于跟踪场景和模态质量都是不断变化的,现有的研究工作往往侧重于抑制低质量或者低贡献的模态信息,限制了跟踪性能的进一步提升。为此,一些研究者探索了在模态质量分布差异场景下的多模态融合算法,即双向互增强融合策略。该类方法通过引导不同质量下模态的判别性信息进行双向增强模态特征表示能力。尽管判别式融合的跟踪算法取得了较好的性能表现,但是在融合方法的设计上往往会引入较为复杂的融合模块,对跟踪效率有着较大的影响。尽管其他多模态视觉跟踪任务也取得了较快的发展,但研究工作相对较少,对相应模态信息(如事件模态和语言模态)的挖掘和利用还有较大的空间。

4.2 未来展望

尽管多模态跟踪领域发展蓬勃,跟踪性能得到了极大的提升,但还存在着一些问题亟待解决,主要包括以下几点。

1)复杂多模态数据。真实场景和环境中,多模态数据复杂多样,具有不确定性、不一致性和数据噪声,如何充分挖掘和利用多模态信息建立有效融合模型仍是多模态视觉跟踪中的非常关键而具有挑战的问题。

2)非配准数据。现有多模态跟踪方法大都需要高精度配准的多模态数据。然而,人工配准的代价十分昂贵,且人工或设备配准仍无法保证完全配准。因此,研究非配准数据下的多模态视觉跟踪模型有着重要意义和应用价值。此外,如何利用非配准数据或者非配对数据实现多模态视觉跟踪模型的训练也是值得探索的研究方向。

3)模态缺失问题。在真实应用场景中,由于成像设备等问题,某些模态往往会存在缺失现象。而现有多模态视觉跟踪方法是在数据完整的前提下进行的,无法处理模态缺失的问题。因此,研究适应各种缺失情形的多模态视觉跟踪模型有着重要的价值。

4)无监督多模态融合。现有多模态跟踪方法通常依赖大规模标注的训练数据,具有较大的人工标注和配准成本,且现有的多模态跟踪数据集难以满足深度模型的需要,导致跟踪模型对新场景的适应能力弱。因此,如何设计无监督多模态融合方式,实现对大规模无标注多模态数据的充分利用,对提升跟踪模型的适应能力和泛化能力具有重要的研究价值。

5)融合模型轻量化。现有多模态视觉跟踪方法为了充分挖掘多模态信息,设计了较为复杂的多模态融合模型,但也带来了较大的计算量和存储负担,无法在实时应用场景和嵌入式设备上部署。因此,如何设计轻量化、高精度多模态融合模型是未来研究的一个方向。

6)辅助模态学习。现有视觉系统大多基于单一模态成像,且嵌入多模态成像设备往往带来较大的成本和代价。因此,研究多模态数据训练、单模态数据测试的辅助模态学习算法,在不增加额外代价的基础上有效提高视觉跟踪算法鲁棒性有着重要的价值和意义。

7)更多模态集成。目前的多模态视觉跟踪方法均是针对两种模态(RGB 和其他模态)进行研究的,而不同模态对不同场景和环境的作用会有较大差异。因此,集成更多模态信息对多模态视觉跟踪系统的鲁棒性和适应性会有较大的帮助,是未来值得研究的一个方向。

8)统一平台构建。目前,多模态视觉跟踪的研究相对散乱,缺乏统一的数据集和评测平台,为相关研究带来极大不便,不利于该领域的研究和发展。

因此,构建多模态视觉跟踪的统一平台,是该领域研究和发展的迫切需求。

参考文献 (References)

- Bertinetto L, Valmadre J, Henriques J F, Vedaldi A and Torr P H S. 2016. Fully-convolutional siamese networks for object tracking//Proceedings of the European Conference on Computer Vision. Amsterdam, the Netherlands; Springer: 850-865 [DOI: 10.1007/978-3-319-48881-3_56]
- Bhat G, Danelljan M, Gool L V and Timofte R. 2019. Learning discriminative model prediction for tracking//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South); IEEE: 6181-6190 [DOI: 10.1109/ICCV.2019.00628]
- Bhat G, Lawin F J, Danelljan M, Robinson A, Felsberg M, Gool L V and Timofte R. 2020. Learning What to Learn for Video Object Segmentation//Proceedings of 2020 European Conference on Computer Vision. Springer: 777-794 [DOI: 10.1007/978-3-030-58536-5_46]
- Bibi A, Zhang T Z and Ghanem B. 2016. 3D part-based sparse tracker with automatic synchronization and registration//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA; IEEE: 1439-1448 [DOI: 10.1109/CVPR.2016.160]
- Chan A L and Schnelle S R. 2012. Target tracking using concurrent visible and infrared imageries//Proceedings Volume 8392, Signal Processing, Sensor Fusion, and Target Recognition XXI. Baltimore, United States; SPIE: #918373 [DOI: 10.1117/12.918373]
- Chan A L and Schnelle S R. 2013. Fusing concurrent visible and infrared videos for improved tracking performance. Optical Engineering, 52(1): #017004 [DOI: 10.1117/1.OE.52.1.017004]
- Chen X, Yan B, Zhu J W, Wang D, Yang X Y and Lu H C. 2021. Transformer tracking//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 8122-8131 [DOI: 10.1109/CVPR46437.2021.00803]
- Chen Y, Shen Y J, Liu X and Zhong B N. 2015. 3D object tracking via image sets and depth-based occlusion detection. Signal Processing, 112: 146-153 [DOI: 10.1016/j.sigpro.2014.08.046]
- Chen Z L and Shi F H. 2022. Double template fusion based siamese network for robust visual object tracking. Journal of Image and Graphics, 27(4): 1191-1203 (陈志良, 石繁槐. 2022. 结合双模板融合与孪生网络的鲁棒视觉目标跟踪. 中国图象图形学报, 27(4): 1191-1203) [DOI: 10.11834/jig.200660]
- Cui Y T, Jiang C, Wang L M and Wu G S. 2022. MixFormer: end-to-end tracking with iterative mixed attention//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA; IEEE: 13608-13618 [DOI: 10.1109/CVPR52688.2022.01324]
- Danelljan M, Bhat G, Khan F S and Felsberg M. 2017. ECO: efficient convolution operators for tracking//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA; IEEE: 6931-6939 [DOI: 10.1109/CVPR.2017.733]
- Danelljan M, Robinson A, Khan F S and Felsberg M. 2016. Beyond correlation filters: learning continuous convolution operators for visual tracking//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands; Springer: 472-488 [DOI: 10.1007/978-3-319-46454-1_29]
- Ding P and Song Y. 2015. Robust object tracking using color and depth images with a depth based occlusion handling and recovery//Proceedings of the 12th International Conference on Fuzzy Systems and Knowledge Discovery. Zhangjiajie, China; IEEE: 930-935 [DOI: 10.1109/FSKD.2015.7382068]
- Feng M Z, Song K C, Wang Y Y, Liu J and Yan Y H. 2020. Learning discriminative update adaptive spatial-temporal regularized correlation filter for RGB-T tracking. Journal of Visual Communication and Image Representation, 72: #102881 [DOI: 10.1016/j.jvcir.2020.102881]
- Feng Q, Ablavsky V, Bai Q Y and Sclaroff S. 2021a. Robust visual object tracking with natural language region proposal network[EB/OL]. [2022-06-02]. <https://arxiv.org/pdf/1912.02048v1.pdf>
- Feng Q, Ablavsky V, Bai Q Y and Sclaroff S. 2021b. Siamese natural language tracker: tracking by natural language descriptions with siamese trackers//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 5847-5856 [DOI: 10.1109/CVPR46437.2021.00579]
- 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 the 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea (South); IEEE: 91-99 [DOI: 10.1109/ICCVW.2019.00017]
- García G M, Klein D A, Stückler J, Frintrop S and Cremers A B. 2012. Adaptive multi-cue 3D tracking of arbitrary objects//The Joint 34th DAGM and 36th OAGM Symposium. Graz, Austria; Springer: 357-366 [DOI: 10.1007/978-3-642-32717-9_36]
- Gehrig D, Rebecq H, Gallego G and Scaramuzza D. 2018. Asynchronous, photometric feature tracking using events and frames//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany; Springer: 766-781 [DOI: 10.1007/978-3-030-01258-8_46]
- Gehrig D, Rebecq H, Gallego G and Scaramuzza D. 2020. EKLt: Asynchronous photometric feature tracking using events and frames. International Journal of Computer Vision, 128(3): 601-618 [DOI: 10.1007/s11263-019-01209-w]
- Gutierrez A and Debono C J. 2019. Exploiting depth information to increase object tracking robustness//Proceedings of the 18th International Conference on Smart Technologies. Novi Sad, Serbia; IEEE: 1-5 [DOI: 10.1109/EUROCON.2019.8861628]

- Hannuna S, Camplani M, Hall J, Mirmehdi M, Damen D, Burghardt T, Paiement A and Tao L L. 2019. DS-KCF: a real-time tracker for RGB-D data. *Journal of Real-Time Image Processing*, 16 (5): 1439-1458 [DOI: 10.1007/s11554-016-0654-3]
- Hare S, Saffari A and Torr P H S. 2011. Struck: structured output tracking with kernels//Proceedings of the 2011 IEEE/CVF International Conference on Computer Vision. Barcelona, Spain; IEEE: 263-270 [DOI: 10.1109/ICCV.2011.6126251]
- Hou Y E, Li W G, Rong A Q and Ye G Q. 2013. Tracking algorithm of block sparse representation with background information. *Journal of South China University of Technology (Natural Science Edition)*, 41(8): 21-27 (侯跃恩, 李伟光, 容爱琼, 叶国强. 2013. 融合背景信息的分块稀疏表示跟踪算法. 华南理工大学学报(自然科学版), 41(8): 21-27) [DOI: 10.3969/j.issn.1000-565X.2013.08.004]
- Hu Y H, Liu H J, Pfeiffer M and Delbruck T. 2016. DVS benchmark datasets for object tracking, action recognition, and object recognition. *Frontiers in Neuroscience*, 10: 405 [DOI: 10.3389/fnins.2016.00405]
- Huang J, Wang S Z, Guo M H and Chen S S. 2018. Event-guided structured output tracking of fast-moving objects using a CeleX sensor. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9): 2413-2417 [DOI: 10.1109/TCSVT.2018.2841516]
- Jung I, Son J, Baek M and Han B. 2018. Real-time MDNeT//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany; Springer: 89-104 [DOI: 10.1007/978-3-030-01225-0_6]
- Kart U, Kämäräinen J K, Matas J, Fan L X and Cieri F. 2018. Depth masked discriminative correlation filter//Proceedings of the 24th International Conference on Pattern Recognition. Beijing, China; IEEE: 2112-2117 [DOI: 10.1109/ICPR.2018.8546179]
- Kart U, Lukežič A, Kristan M, Kämäräinen J K and Matas J. 2019. Object tracking by reconstruction with view-specific discriminative correlation filters//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA; IEEE: 1339-1348 [DOI: 10.1109/CVPR.2019.00143]
- Kim H U, Lee D Y, Sim J Y and Kim C S. 2015. SOWP: spatially ordered and weighted patch descriptor for visual tracking//Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile; IEEE: 3011-3019 [DOI: 10.1109/ICCV.2015.345]
- Kueng B, Mueggler E, Gallego G and Scaramuzza D. 2016. Low-latency visual odometry using event-based feature tracks//Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems. Daejeon, Korea (South); IEEE: 16-23 [DOI: 10.1109/IROS.2016.7758089]
- Lan X Y, Ye M, Shao R, Zhong B N, Jain D K and Zhou H Y. 2019a. Online non-negative multi-modality feature template learning for RGB-assisted infrared tracking. *IEEE Access*, 7: 67761-67771 [DOI: 10.1109/ACCESS.2019.2916895]
- Lan X Y, Ye M, Zhang S P and Yuen P C. 2018. Robust collaborative discriminative learning for RGB-infrared tracking//Proceedings of the 32nd AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans, USA; AAAI: 7008-7015
- Lan X Y, Ye M, Zhang S P, Zhou H Y and Yuen P C. 2020. Modality-correlation-aware sparse representation for RGB-infrared object tracking. *Pattern Recognition Letters*, 130: 12-20 [DOI: 10.1016/j.patrec.2018.10.002]
- Lan X Y, Zhang W, Zhang S P, Jain D K and Zhou H Y. 2019c. Robust multi-modality anchor graph-based label prediction for RGB-infrared tracking. *IEEE Transactions on Industrial Informatics*: #2947293 [DOI: 10.1109/TII.2019.2947293]
- Li B, Yan J J, Wu W, Zhu Z and Hu X L. 2018a. High performance visual tracking with siamese region proposal network//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 8971-8980 [DOI: 10.1109/CVPR.2018.00935]
- Li C L, Cheng H, Hu S Y, Liu X B, Tang J and Lin L. 2016a. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing*, 25(12): 5743-5756 [DOI: 10.1109/TIP.2016.2614135]
- Li C L, Hu S Y, Gao S H and Tang J. 2016b. Real-time grayscale-thermal tracking via Laplacian sparse representation//Proceedings of the 22nd International Conference on Multimedia Modeling. Miami, USA; Springer: 54-65 [DOI: 10.1007/978-3-319-27674-8_6]
- 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, Lin L, Zuo W M and Tang J. 2017a. Learning patch-based dynamic graph for visual tracking//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA; AAAI: 4126-4132
- 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-adaptor RGBT tracking//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea (South); IEEE: 2262-2270 [DOI: 10.1109/ICCVW.2019.00279]
- Li C L, Sun X, Wang X, Zhang L and Tang J. 2017b. Grayscale-thermal object tracking via multitask Laplacian sparse representation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(4): 673-681 [DOI: 10.1109/TSMC.2016.2627052]
- Li C L, Wu X H, Zhao N, Cao X C and Tang J. 2018b. 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, Xiang Z Q, Tang J, Luo B and Wang F T. 2022. RGBT tracking via noise-robust cross-modal ranking. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9): 5019-5031 [DOI: 10.1109/TNNLS.2021.3067107]
- Li C L, Zhao N, Lu Y J, Zhu C L and Tang J. 2017c. 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 C L, Zhu C L, Huang Y, Tang J and Wang L. 2018c. Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking//*Proceedings of the 15th European Conference on Computer Vision*. Munich, Germany: Springer: 831-847 [DOI: 10.1007/978-3-030-01261-8_49]
- Li C L, Zhu C L, Zhang J, Luo B, Wu X H and Tang J. 2019c. Learning local-global multi-graph descriptors for RGB-T object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(10): 2913-2926 [DOI: 10.1109/TCSVT.2018.2874312]
- Li C L, Zhu C L, Zheng S F, Luo B and Tang J. 2018d. Two-stage modality-graphs regularized manifold ranking for RGB-T tracking. *Signal Processing: Image Communication*, 68: 207-217 [DOI: 10.1016/j.image.2018.08.004]
- Li L, Li C L, Tu Z Z and Tang J. 2018e. A fusion approach to grayscale-thermal tracking with cross-modal sparse representation//*Proceedings of the 13th Conference on Image and Graphics Technologies and Applications*. Beijing, China: Springer: 494-505 [DOI: 10.1007/978-981-13-1702-6_49]
- Li W G, Hou Y E, Lou H D and Ye G Q. 2012. Robust visual tracking based on Gabor feature and sparse representation//*Proceedings of 2012 IEEE International Conference on Robotics and Biomimetics*. Guangzhou, China: IEEE: 1829-1835 [DOI: 10.1109/ROBIO.2012.6491234]
- Li Z Y, Tao R, Gavves E, Snoek C G M and Smeulders A W M. 2017 d. Tracking by natural language specification//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, USA: IEEE: 7350-7358 [DOI: 10.1109/CVPR.2017.777]
- Liu B Y, Yang L, Huang J Z, Meer P, Gong L G and Kulikowski C. 2010. Robust and fast collaborative tracking with two stage sparse optimization//*Proceedings of the 11th European Conference on Computer Vision*. Heraklion, Greece: Springer: 624-637 [DOI: 10.1007/978-3-642-15561-1_45]
- Liu H J, Moeys D P, Das G, Neil D, Liu S C and Delbrück T. 2016. Combined frame- and event-based detection and tracking//*2016 IEEE International Symposium on Circuits and Systems*. Montreal, Canada: IEEE: 2511-2514 [DOI: 10.1109/ISCAS.2016.7539103]
- Liu H P and Sun F C. 2012. Fusion tracking in color and infrared images using joint sparse representation. *Science China Information Sciences*, 55(3): 590-599 [DOI: 10.1007/s11432-011-4536-9]
- Liu W C, Tang X A and Zhao C L. 2020. Robust RGBD tracking via weighted convolution operators. *IEEE Sensors Journal*, 20(8): 4496-4503 [DOI: 10.1109/JSEN.2020.2964019]
- Liu Y, Jing X Y, Nie J H, Gao H, Liu J and Jiang G P. 2019. Context-aware three-dimensional mean-shift with occlusion handling for robust object tracking in RGB-D videos. *IEEE Transactions on Multimedia*, 21(3): 664-677 [DOI: 10.1109/TMM.2018.2863604]
- Lu A D, Li C L, Yan Y Q, Tang J and Luo B. 2021. RGBT tracking via multi-adaptor network with hierarchical divergence loss. *IEEE Transactions on Image Processing*, 30: 5613-5625 [DOI: 10.1109/TIP.2021.3087341]
- Lu A D, Qian C, Li C L, Tang J and Wang L. 2022. Duality-gated mutual condition network for RGBT tracking. *IEEE Transactions on Neural Networks and Learning Systems*:#3157594 [DOI: 10.1109/TNNLS.2022.3157594]
- Ma C, Huang J B, Yang X K and Yang M H. 2015. Hierarchical convolutional features for visual tracking//*Proceedings of 2015 IEEE International Conference on Computer Vision*. Santiago, Chile: IEEE: 3074-3082 [DOI: 10.1109/ICCV.2015.352]
- Ma Z A and Xiang Z Y. 2017. Robust object tracking with RGBD-based sparse learning. *Frontiers of Information Technology and Electronic Engineering*, 18(7): 989-1001 [DOI: 10.1631/FITEE.1601338]
- Mayer C, Danelljan M, Paudel D P and van Gool L. 2021. Learning target candidate association to keep track of what not to track//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*. Montreal, Canada: IEEE: 13424-13434 [DOI: 10.1109/ICCV48922.2021.01319]
- Mei J T, Zhou D M, Cao J D, Nie R C and Guo Y B. 2021. HDINet: hierarchical dual-sensor interaction network for RGBT tracking. *IEEE Sensors Journal*, 21(15): 16915-16926 [DOI: 10.1109/JSEN.2021.3078455]
- Mei X and Ling H B. 2009. Robust visual tracking using l_1 minimization//*Proceedings of the 12th IEEE International Conference on Computer Vision*. Kyoto, Japan: IEEE: 1436-1443 [DOI: 10.1109/ICCV.2009.5459292]
- Meshgi K, Maeda S I, Oba S, Skibbe H, Li Y Z and Ishii S. 2016. An occlusion-aware particle filter tracker to handle complex and persistent occlusions. *Computer Vision and Image Understanding*, 150: 81-94 [DOI: 10.1016/j.cviu.2016.05.011]
- 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]
- Ning J F, Yang J M, Jiang S J, Zhang L and Yang M H. 2016. Object tracking via dual linear structured SVM and explicit feature map//*Proceedings of 2016 IEEE Conference on Computer Vision and Pat-*

- tern Recognition. Las Vegas, USA; IEEE: 4266-4274 [DOI: 10.1109/CVPR.2016.462]
- Peng J C, Zhao H T, Hu Z W, Yi Z and Wang B F. 2021. Siamese infrared and visible light fusion network for RGB-T tracking[EB/OL]. [2022-06-02]. <https://arxiv.org/pdf/2103.07302.pdf>
- Qi Y K, Zhang S P, Qin L, Yao H X, Huang Q M, Lim J and Yang M H. 2016. Hedged deep tracking//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA; IEEE: 4303-4311 [DOI: 10.1109/CVPR.2016.466]
- Qi Y K, Zhang S P, Zhang W G, Su L, Huang Q M and Yang M X. 2019. Learning attribute-specific representations for visual tracking//Proceedings of the 33rd AAAI Conference on Artificial Intelligence and the 31st Innovative Applications of Artificial Intelligence Conference and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence. Honolulu, USA; AAAI: 8835-8842 [DOI: 10.1609/aaai.v33i01.33018835]
- Schnelle S R and Chan A L. 2011. Enhanced target tracking through infrared-visible image fusion//Proceedings of the 14th International Conference on Information Fusion. Chicago, USA; IEEE: 1-8
- Shen L F, Wang X X, Liu L, Hou B, Jian Y L, Tang J and Luo B. 2022. RGBT tracking based on cooperative low-rank graph model. Neurocomputing, 492: 370-381 [DOI: 10.1016/j.neucom.2022.04.032]
- Shi H Z, Gao C X and Sang N. 2015. Using consistency of depth gradient to improve visual tracking in RGB-D sequences//Proceedings of 2015 Chinese Automation Congress. Wuhan, China; IEEE: 518-522 [DOI: 10.1109/CAC.2015.7382555]
- Song S R and Xiao J X. 2013. Tracking revisited using RGBD camera: unified benchmark and baselines//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Australia; IEEE: 233-240 [DOI: 10.1109/ICCV.2013.36]
- Song Y B, Ma C, Wu X H, Gong L J, Bao L C, Zuo W M, Shen C H, Lau R W H and Yang M H. 2018. VITAL: visual tracking via adversarial learning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 8990-8999 [DOI: 10.1109/CVPR.2018.00937]
- Tedaldi D, Gallego G, Mueggler E and Scaramuzza D. 2016. Feature detection and tracking with the dynamic and active-pixel vision sensor (DAVIS)//Proceedings of the 2nd International Conference on Event-based Control, Communication, and Signal Processing. Krakow, Poland; IEEE: 1-7 [DOI: 10.1109/EBCCSP.2016.7605086]
- Tu Z Z, Lin C, Zhao W, Li C L and Tang J. 2022. M5L: multi-modal multi-margin metric learning for RGBT tracking. IEEE Transactions on Image Processing, 31: 85-98 [DOI: 10.1109/TIP.2021.3125504]
- Valmadre J, Bertinetto L, Henriques J, Vedaldi A and Torr P H S. 2017. End-to-end representation learning for correlation filter based tracking//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA; IEEE: 5000-5008 [DOI: 10.1109/CVPR.2017.531]
- Wang C Q, Xu C Y, Cui Z, Zhou L, Zhang T, Zhang X Y and Yang J. 2020a. Cross-modal pattern-propagation for RGB-T tracking//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 7062-7071 [DOI: 10.1109/CVPR42600.2020.00709]
- Wang L J, Ouyang W L, Wang X G and Lu H C. 2015. Visual tracking with fully convolutional networks//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile; IEEE: 3119-3127 [DOI: 10.1109/ICCV.2015.357]
- Wang N, Zhou W G, Wang J and Li H Q. 2021a. Transformer meets tracker: exploiting temporal context for robust visual tracking//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 1571-1580 [DOI: 10.1109/CVPR46437.2021.00162]
- Wang Q, Fang J W and Yuan Y. 2014. Multi-cue based tracking. Neurocomputing, 131: 227-236 [DOI: 10.1016/j.neucom.2013.10.021]
- Wang X, Li J N, Zhu L, Zhang Z P, Chen Z, Li X, Wang Y W, Tian Y H and Wu F. 2022. VisEvent: reliable object tracking via collaboration of frame and event flows [EB/OL]. [2022-06-02]. <https://arxiv.org/pdf/2108.05015.pdf>
- Wang X, Shu X J, Zhang Z P, Jiang B, Wang Y W, Tian Y H and Wu F. 2021b. Towards more flexible and accurate object tracking with natural language: algorithms and benchmark//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 13758-13768 [DOI: 10.1109/CVPR46437.2021.01355]
- Wang Y, Wei X, Shen H, Ding L and Wan J Q. 2020b. Robust fusion for RGB-D tracking using CNN features. Applied Soft Computing, 92: #106302 [DOI: 10.1016/j.asoc.2020.106302]
- Wang Y L, Li C L and Tang J. 2018b. Learning soft-consistent correlation filters for RGB-T object tracking//Proceedings of the 1st Chinese Conference on Pattern Recognition and Computer Vision. Guangzhou, China; Springer: 295-306 [DOI: 10.1007/978-3-030-03341-5_25]
- Wu Y, Blasch E, Chen G S, Bai L and Ling H B. 2011. Multiple source data fusion via sparse representation for robust visual tracking//Proceedings of the 14th International Conference on Information Fusion. Chicago, USA; IEEE: 1-8
- Xiao J J, Stolkin R, Gao Y Q and Leonardis A. 2018. Robust fusion of color and depth data for RGB-D target tracking using adaptive range-invariant depth models and spatio-temporal consistency constraints. IEEE Transactions on Cybernetics, 48(8): 2485-2499 [DOI: 10.1109/TCYB.2017.2740952]
- 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 AAAI Conference on Artificial Intelligence, 36(3): 2831-2838

- [DOI: 10.1609/aaai.v36i3.20187]
- Xie Y J, Lu Y and Gu S. 2019. RGB-D object tracking with occlusion detection//Proceedings of the 15th International Conference on Computational Intelligence and Security. Macao, China: IEEE: 11-15 [DOI: 10.1109/CIS.2019.00011]
- Xu Q, Mei Y M, Liu J P and Li C L. 2022. Multimodal cross-layer bilinear pooling for RGBT tracking. *IEEE Transactions on Multimedia*, 24: 567-580 [DOI: 10.1109/TMM.2021.3055362]
- Xu Y D, Wang Z Y, Li Z X, Yuan Y and Yu G. 2020. SiamFC++: towards robust and accurate visual tracking with target estimation guidelines. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7): 12549-12556 [DOI: 10.1609/aaai.v34i07.6944]
- Yan B, Peng H W, Fu J L, Wang D and Lu H C. 2021a. 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]
- Yan S, Yang J Y, Käpylä J, Zheng F, Leonardis A and Kämäräinen J K. 2021b. DepthTrack: unveiling the power of RGBD tracking//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 10705-10713 [DOI: 10.1109/ICCV48922.2021.01055]
- Yang Z Y, Kumar T, Chen T L, Su J S and Luo J B. 2021. Grounding-tracking-integration. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(9): 3433-3443 [DOI: 10.1109/TCSVT.2020.3038720]
- Yang Z Y, Wu Y J, Wang G R, Yang Y K, Li G Q, Deng L, Zhu J and Shi L P. 2019. DashNet: a hybrid artificial and spiking neural network for high-speed object tracking [EB/OL]. [2022-06-02]. <http://arxiv.org/pdf/1909.12942.pdf>
- Yun X, Sun Y J, Yang X X and Lu N N. 2019. Discriminative fusion correlation learning for visible and infrared tracking. *Mathematical Problems in Engineering*, 2019: #2437521 [DOI: 10.1155/2019/2437521]
- Zhai S L, Shao P P, Liang X Y and Wang X. 2019. Fast RGB-T tracking via cross-modal correlation filters. *Neurocomputing*, 334: 172-181 [DOI: 10.1016/j.neucom.2019.01.022]
- Zhai Y Y, Song P, Mou Z L, Chen X X and Liu X J. 2018. Occlusion-aware correlation particle filter target tracking based on RGBD data. *IEEE Access*, 6: 50752-50764 [DOI: 10.1109/ACCESS.2018.2869766]
- Zhang H, Zhang L, Zhuo L and Zhang J. 2020a. Object tracking in RGB-T videos using modal-aware attention network and competitive learning. *Sensors*, 20(2): #393 [DOI: 10.3390/s20020393]
- Zhang J Q, Yang X, Fu Y K, Wei X P, Yin B C and Dong B. 2021a. Object tracking by jointly exploiting frame and event domain//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 13023-13032 [DOI: 10.1109/ICCV48922.2021.01280]
- Zhang K H, Zhang L and Yang M H. 2012. Real-time compressive tracking//Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer: 864-877 [DOI: 10.1007/978-3-642-33712-3_62]
- Zhang L C, Danelljan M, Gonzalez-Garcia A, van de Weijer J and Khan F S. 2019a. Multi-modal fusion for end-to-end RGB-T tracking//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea (South): IEEE: 2252-2261 [DOI: 10.1109/ICCVW.2019.00278]
- Zhang P Y, Wang D and Lu H C. 2020b. Multi-modal visual tracking: review and experimental comparison [EB/OL]. [2022-06-02]. <https://arxiv.org/pdf/2012.04176.pdf>
- Zhang P Y, Wang D, Lu H C and Yang X Y. 2021b. Learning adaptive attribute-driven representation for real-time RGB-T tracking. *International Journal of Computer Vision*, 129(9): 2714-2729 [DOI: 10.1007/s11263-021-01495-3]
- Zhang P Y, Zhao J, Bo C J, Wang D, Lu H C and Yang X Y. 2021c. 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. 2022a. 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, Liu X R, Zhang Q and Han J G. 2022b. SiamCDA: complementarity- and distractor-aware RGB-T tracking based on Siamese network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3): 1403-1417 [DOI: 10.1109/TCSVT.2021.3072207]
- Zhang X C, Ye P, Leung H, Gong K and Xiao G. 2020d. Object fusion tracking based on visible and infrared images: a comprehensive review. *Information Fusion*, 63: 166-187 [DOI: 10.1016/j.inffus.2020.05.002]
- Zhang X C, Ye P, Peng S Y, Liu J and Xiao G. 2020e. DSiamMFT: an RGB-T fusion tracking method via dynamic Siamese networks using multi-layer feature fusion. *Signal Processing: Image Communication*, 84: #115756 [DOI: 10.1016/j.image.2019.115756]
- Zhang X C, Ye P, Peng S Y, Liu J, Gong K and Xiao G. 2019b. Siam-FT: an RGB-infrared fusion tracking method via fully convolutional siamese networks. *IEEE Access*, 7: 122122-122133 [DOI: 10.1109/ACCESS.2019.2936914]
- Zhang Y L, Qian X Y, Zhang M and Ge H J. 2020. Correlation filter target tracking algorithm based on adaptive multifeature fusion. *Journal of Image and Graphics*, 25(6): 1160-1170 (张艳琳, 钱小燕, 张森, 葛红娟. 2020. 自适应多特征融合相关滤波目标跟踪. *中国图象图形学报*, 25(6): 1160-1170) [DOI: 10.1016/j.inf-fus.2020.05.002]
- Zhang Z P and Peng H W. 2019. Deeper and wider siamese networks for

- real-time visual tracking//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE; 4586-4595 [DOI: 10.1109/CVPR.2019.00472]
- Zhao P Y, Liu Q L, Wang W and Guo Q. 2020. TSDM: tracking by SiamRPN++ with a depth-refiner and a mask-generator//Proceedings of the 25th International Conference on Pattern Recognition. Milan, Italy: IEEE; 670-676 [DOI: 10.1109/ICPR48806.2021.9413315]
- Zhu Y B, Li C L, Lu Y J, Lin L, Luo B and Tang J. 2019a. FANet: quality-aware feature aggregation network for RGB-T tracking [EB/OL]. [2022-06-02]. <https://arxiv.org/pdf/1811.09855v1.pdf>
- Zhu Y B, Li C L, Luo B, Tang J and Wang X. 2019b. 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, 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

李成龙,男,教授,主要研究方向为计算机视觉和深度学习。

E-mail:lcl1314@foxmail.com

汤进,通信作者,男,教授,主要研究方向为计算机视觉、模式识别和视频分析。E-mail:tangjin@ahu.edu.cn

鹿安东,男,博士研究生,主要研究方向为多模态跟踪和行人重识别。E-mail:adlu_ah@foxmail.com

刘磊,男,博士研究生,主要研究方向为多模态跟踪。

E-mail:liulei970507@163.com