面向不同挑战及同异质信息分离的RGBT跟踪

方 鑫,陈 柘*,刘占文,李小鹏,宿雨心
(长安大学信息工程学院,陕西西安 710064)

摘 要: 可见光热红外(RGB and Thermal infrared,RGBT)跟踪是一种结合了可见光和热红外光两种不同传感器信息的多模态目标跟踪方法.这种方法旨在克服单一传感器在特定环境下的局限性,通过融合多种传感器的数据来提高目标跟踪的鲁棒性和准确性.然而,在现有的RGBT跟踪算法中,大多将可见光与热红外图像提取的特征直接进行融合,忽略了两种模态间的同质性与异质性.此外,RGBT跟踪还经常受到目标快速运动、尺度变化、光照变化、热交叉和遮挡等多种挑战因素的影响,现有工作往往是通过研究单一结构来同时解决所有问题,但这需要足够复杂的模型和足够多的训练数据.本文提出了一种新的面向不同挑战并结合多模态同异质信息分离与融合的网络,用于RGBT跟踪.在该网络的每层主干中都设计了一个挑战感知模块用于融合每种挑战下来自可见光与热红外两种不同模态的特征,并自适应地聚合所有挑战下的融合特征.此外,还加入了注意力增强模块及多尺度辅助模块对主干网络所提取的特征进行增强.最后根据可见光与热红外的同质性与异质性,分别提取它们的特有特征与共有特征并进行自适应融合.在GTOT、RGBT234和LasHeR数据集上的大量实验表明,与现有RGBT跟踪方法相比,本文提出的跟踪器显示出非常强的竞争力.

关键词: RGBT跟踪;挑战感知;同异质信息分离;自适应聚合;注意力机制;多尺度特征
基金项目: 国家自然科学基金(No.52172302);陕西省重点研发计划项目(No.2022GY-063)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2025)03-0910-16
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240713

Facing Different Challenges and Separating Homogeneous and Heterogeneous Information for RGBT Tracking

FANG Xin, CHEN Zhe*, LIU Zhan-wen, LI Xiao-peng, SU Yu-xin
(*School of Information Engineering*, *Chang'an University*, *Xi'an*, *Shaanxi* 710064, *China*)

Abstract: RGB and Thermal infrared (RGBT) tracking is a multi-modal object tracking method that integrates different information from visible light and thermal infrared sensors. This method aims to overcome the limitations of single sensor in a specific condition and increase the robustness and accuracy of object tracking by fusing data from multiple sensors. However, the majority of RGBT tracking methods in use today directly fuse features extracted from thermal infrared and visible light images, ignoring the homogeneity and heterogeneity of the two modalities. In addition, RGBT tracking is often affected by multiple challenging factors such as objects fast motion, scale variation, illumination variation, thermal crossover, and occlusion. Existing work often focuses on a single model to solve all challenges simultaneously, which requires highly complex model and extensive training data. This paper proposes a novel network called CMHHNet (facing different Challenges and combining Multi-modal Homogeneous and Heterogeneous information separation and integration Network) for RGBT tracking. In this network, a challenge-aware module is deployed in each layer of the backbone to fuse the visible light and thermal infrared features from two different modalities under each challenge separately, and adaptively aggregate the fused features under all challenges. In addition, an attention enhancement module and a multi-scale auxiliary module are added to strengthen the features that the backbone network has extracted. Finally, according to the homogeneity and heterogeneity of thermal infrared and visible light, their unique and common features are extracted separately and adaptively fused. Extensive experiments on GTOT, RGBT234 and LasHeR datasets demonstrate that the tracker proposed in this paper shows quite strong competitiveness compared with existing RGBT tracking methods.

收稿日期:2024-07-31;修回日期:2025-02-26;责任编辑:朱梅玉
*通讯作者:陈柘

Key words: RGBT tracking; challenge-aware; separation of homogeneous and heterogeneous information; adaptive aggregation; attention mechanism; multiscale features

Foundation Item(s): National Natural Science Foundation of China (No.52172302); Key Research and Development Program of Shaanxi Province (No.2022GY-063)

1 引言

目标跟踪是计算机视觉研究领域中一项重要且基础性的研究课题,要求已知目标的第一帧位置后,在后续视频序列中确定该目标的位置[1]. 目前,目标跟踪技术在许多领域都有广泛的应用,包括监控、自动驾驶、无人机、虚拟现实、人机交互等[2].

近几年,借助Transformer[3]等理论框架开发的跟踪算法[4,5]使基于可见光(RGB)图像的目标跟踪取得了显著的进步. 这些算法相比于早期算法在精确度和速度方面都有明显的提高. 然而,当前的RGB目标跟踪技术在恶劣天气、弱光条件、遮挡和迷雾等复杂环境中仍然面临困难. 因此,这种仅依赖RGB图像的跟踪算法在处理复杂场景时性能有所降低,无法有效满足实际需求.

热红外图像因其对光照变化不敏感、穿透雾霾能力强等优势而备受关注,并逐渐被应用于许多计算机视觉任务中,如显著性检测[6]、行人检测[7]、语义分割[8]等. 然而,热红外图像中通常缺失物体的边缘、纹理和几何等细节信息. 考虑到可见光和热红外图像的互补性,基于可见光-热红外(RGB and Thermal infrared, RGBT)图像的目标跟踪技术逐渐引起人们的关注. 多项研究[9~15]表明,将可见光和热红外两种模态的数据结合使用可以有效提升目标跟踪性能. 这两种模态的数据能够相互增强,提供互补信息,从而增强跟踪器在应对不同跟踪挑战时的鲁棒性和准确性.

现有的RGBT目标跟踪算法大多采用深度学习方法,如Zhang等人[16]首次将MDNet[17]扩展为双流结构,并将其应用于RGBT目标跟踪领域,该网络作为RGBT目标跟踪的基线网络被广泛使用. Li等人[18]提出了一种多适配器架构MANet,该架构考虑了模态共享信息和实例感知信息的潜在价值,但该网络不包含任何融合方案,且设计的两阶段学习算法容易带来过拟合的风险. 为此,Lu等人[19]在MANet的基础上提出了一种利用分层散度损失的多适配器网络的升级版本MANet++,在实例适配器中设计了动态融合模块来实现不同模态的质量感知融合. 此外,为了解决大多数RGBT跟踪算法由于优化过程复杂而存在较长延迟的问题,Gao等人[20]提出了DAFNet,通过一种轻量级深度学习融合框架自动计算不同层和不同模态的权重来聚合特征,但其容易受到低质量模态的影响. 为了解决这个问题,Zhu等人[21]提出了FANet,通过设计一个分层的自动聚合策略来提高权重的可靠性,从而学习到更鲁棒的特征表示.

然而,这些方法只使用一些简单的操作,如拼接或注意力加权进行特征融合,无法充分挖掘RGBT跟踪的潜力. 此外,RGBT跟踪常常受到各种挑战因素的影响. 现有的工作往往通过研究一种模型来同时解决所有的挑战,如Zhu等人[22]提出的DAPNet使用递归策略以密集的方式递归聚合所有层的特征,充分利用了从浅到深的空间和语义特征,同时提出了一种协同特征修剪方法来去除噪声和冗余的特征映射,以获得更强的鲁棒性跟踪. Zhang等人[14]则在DiMP[23]的基础上,提出了探寻不同可见光与热红外融合方法的mfDiMP,但这需要一个大规模的数据集进行训练,费时费力.

为了解决这些问题,本文提出了一种新的基于不同挑战及多模态同异质信息分离与融合的网络CMHHNet(facing different Challenges and combining Multimodal Homogeneous and Heterogeneous information separation and integration Network),用于RGBT跟踪. 首先,采用VGG-M[24]网络的前三层作为骨干网络,并将其拓展为双流结构,分别用于提取可见光和热红外图像的目标特征. 其次,由于可见光图像具有较高的空间分辨率,能够更好地表示目标的纹理、轮廓和颜色等细节信息,而热红外图像不易受光照变化的影响,且对遮挡的鲁棒性较高,为了充分挖掘RGBT数据的互补性,本文设计了一个特征分离与融合模块,对可见光与热红外两个模态进行同异质信息分离,分别提取它们的特有特征与共有特征并进行自适应融合. 此外,RGBT跟踪还经常会受到目标快速运动(Fast Motion, FM)、尺度变化(Scale Variation, SV)、光照变化(Illumination Variation, IV)、热交叉(Thermal Crossover, TC)和遮挡(OCClusion, OCC)等各种挑战因素的影响,受到文献[25]和文献[26]的启发,本文在每层网络中都嵌入了一个挑战感知模块,设计了5个挑战处理分支,并将这5种挑战划分为模态共有挑战与模态特有挑战,针对每种挑战的特性对不同的挑战设计了不同的网络结构来进行可见光与热红外特征的融合处理,最后将所有挑战下处理的融合特征进行聚合,以获得鲁棒的特征表示. 为了抑制特征提取时的噪声传播,捕获不同尺度的目标特征,在每层网络中还加入了注意力增强模块和多尺度辅助模块来进行特征增强. 最后,使用3个全连接层进行二值分类和边界框回归,以确定目标的位置.

本文的主要贡献可以总结如下:首先,为有效利用

可见光与热红外图像的互补性,提出一种多模态同异质信息分离和融合网络,提取并融合可见光和热红外图像的共有特征和特有特征,提升RGBT跟踪的鲁棒性;其次,为了应对RGBT跟踪中所面临的不同挑战,减少对大规模训练数据的依赖,设计了一个挑战感知模块来分别处理每种挑战,并自适应地聚合所有挑战下处理的结果特征.与现有方法不同的是,考虑到不同挑战下的目标外观具有不同的特性,对网络结构做了针对性设计;最后,加入了注意力及多尺度模块来提升网络的性能与泛化能力.注意力增强模块用于降低网络对噪声或无用信息的敏感度,多尺度辅助模块则有利于处理尺度变化和识别小尺寸目标.在GTOT[27]、RGBT234[12]和LasHeR[28]数据集上的大量实验表明,所提方法可以有效应对多种挑战,实现目标的有效跟踪,性能优异.

2 相关工作

在本节中,主要介绍现有RGBT目标跟踪算法,以及注意力机制和多尺度特征的相关工作.

2.1 RGBT跟踪

RGBT目标跟踪是一种结合了可见光图像和热红外图像的信息来实现对目标跟踪的技术.早期的RGBT目标跟踪算法[13,29,30]是基于相关滤波、稀疏表示等传统模型的,这些方法仅使用手工特征,无法处理复杂场景中目标外观的显著变化.近年来,基于深度学习技术的RGBT跟踪器开始普遍得到应用,按照基线方法的不同,大致可以分为以下3类.

第一类是基于多域网络(Multi-Domain Network,MDNet)[17]的目标跟踪算法.其中,Zhang等人[31]利用模态感知网络MaCNet自动学习模态贡献的差异.然而,这种方法并没有充分挖掘低质量模态的潜力,如大雾环境的可见光模态提取的特征中虽然会有大量噪声,但仍然包含对跟踪有利的信息.针对这一缺陷,Lu等人[32]提出DMCNet,以一个模态的目标特征作为参考,指导另一个模态目标特征的学习,并采用基于光流算法的重采样策略来处理跟踪失败.但是复杂环境容易导致跟踪漂移,该方法不能在产生跟踪漂移后提供有效的候选框,导致跟踪器性能有限.为此,Xia等人[33]提出了CIRNet,执行多级共有模态、特有模态和最后的目标概率预测学习,通过设计对象感知分支来预测跟踪漂移后物体的中心状态以优化跟踪.

第二类是基于孪生网络(Siamese Network)的目标跟踪算法.如Zhang等人[34]提出的SiamFT使用一种基于全卷积孪生网络的融合跟踪方法,并提出了一种基于孪生网络响应图的模态权重计算方法,以便更好地利用多模态信息.之后,Zhang等人[35]在SiamFT的基础上又提出了DSiamMFT,将孪生网络升级为动态孪生网络来分别处理可见光和热红外图像.然而,当时缺乏大规模成对的RGBT跟踪数据集,导致该网络跟踪性能较差.为此,Zhang等人[36]提出了SiamCDA,使用一种语义感知图像到图像的转换方法从真实的可见光图像中生成热红外图像,此外使用视频着色方法从真实的热红外视频中生成可见光视频,从而解决了数据集规模较小的问题.

第三类则是基于判别式相关滤波器(Discriminative Correlation Filter,DCF)的目标跟踪算法.如Zhao等人[37]提出的CEDiMP首次使用通道交换的方法来融合可见光与热红外数据,融合效率得到了极大的提高,但同时却丢失了很多有效的多模态信息.为了解决这个问题,Zhang等人[38]提出MFNet,设计了一个模态差异补偿模块,通过选择性地促进特征之间的模态交互来减少模态差异并融合多模态特征.

本文针对可见光与热红外两种模态的同质性与异质性,设计了特征分离与融合模块,通过提取它们的共有特征与特有特征来使融合过程变得更加合理有效.

2.2 注意力机制

注意力机制是深度学习中的关键技术,它模拟了人类处理信息时大脑注意力的分配过程,如今已被广泛应用于机器学习领域,旨在提升模型对输入数据的关注度,使其更聚焦于输入中的重要部分,从而增强模型的性能.其中,空间注意力机制和通道注意力机制便是典型代表,它们分别关注特征图中的空间和通道信息,以提高对图像特征的建模能力.如Woo等人[39]提出卷积块注意模块CBAM,重新校准了特征的通道和空间维度,以增强卷积神经网络(Convolutional Neural Network,CNN)的特征表达能力.而在RGBT跟踪中,文献[40]通过引入通道注意机制来明确建模特征通道之间的相互依赖性,显著提升了跟踪性能.此外,文献[3]提出的Transformer中的自注意力机制和交叉注意力机制也被广泛应用.如Mei等人[41]提出AGMINet,使用自注意力与交叉注意力机制设计了一个全局-局部交互模块,旨在探索模态之间的全局相关性和单个模态的上下文信息,同时还对挖掘的局部特征进行补充,以增强特征表示.类似地,Mei等人[42]还提出了DRGCNet,设计了全局挖掘协作模块GMCM分别提取模态内和跨模态的全局信息,然后进行自适应加权运算来平衡二者.在本文中,引入空间和通道注意机制来细化所提取的特征,减少冗余信息;此外,采用自注意力及交叉注意力来实现两个模态共有特征与特有特征的提取.

2.3 多尺度特征

在深度学习中,多尺度特征指的是在不同尺度下捕捉图像或其他数据中的信息.这种方法可以帮助模型更全面地理解和解释输入数据,从而使网络获得丰

富的感受野和多层次信息. 在RGBT目标跟踪中,文献[43]就是通过使用多尺度特征来更好地挖掘两种模态内的多层级特征. 该网络设计了一个多尺度适配器来获取输入图像的多尺度信息,采用不同大小的卷积核提取图像的多尺度特征,此外还提出了一个多分支融合模块来集成不同的尺度信息和前一层的特征,从而抑制了特征噪声和冗余. 受此启发,本文将多尺度特征模块放置在前两层特征提取网络中,以此来增加感受野,识别小尺寸目标.

3　本文方法

在本节中,首先介绍所提出的CMHHNet的整体结构,然后分别解释其中的分离与融合模块、挑战感知模块以及注意力增强和多尺度辅助模块的细节,最后给出了网络训练与跟踪的实施细节.

3.1　总体网络结构

本文所提出的CMHHNet主要由特征提取器、分离与融合模块(Separation and Fusion Module,SFM)、挑战感知模块(Challenge Aware Module,CAM)、注意力增强模块(Attention Enhancement Module,AEM)、多尺度辅助模块(Multi-scale Auxiliary Module,MAM)以及3个全连接层组成,整体网络结构如图1所示,其中FM、SV、OCC、TC和IV分别代表快速运动、尺度变化、遮挡、热交叉和光照变化5个特定挑战.

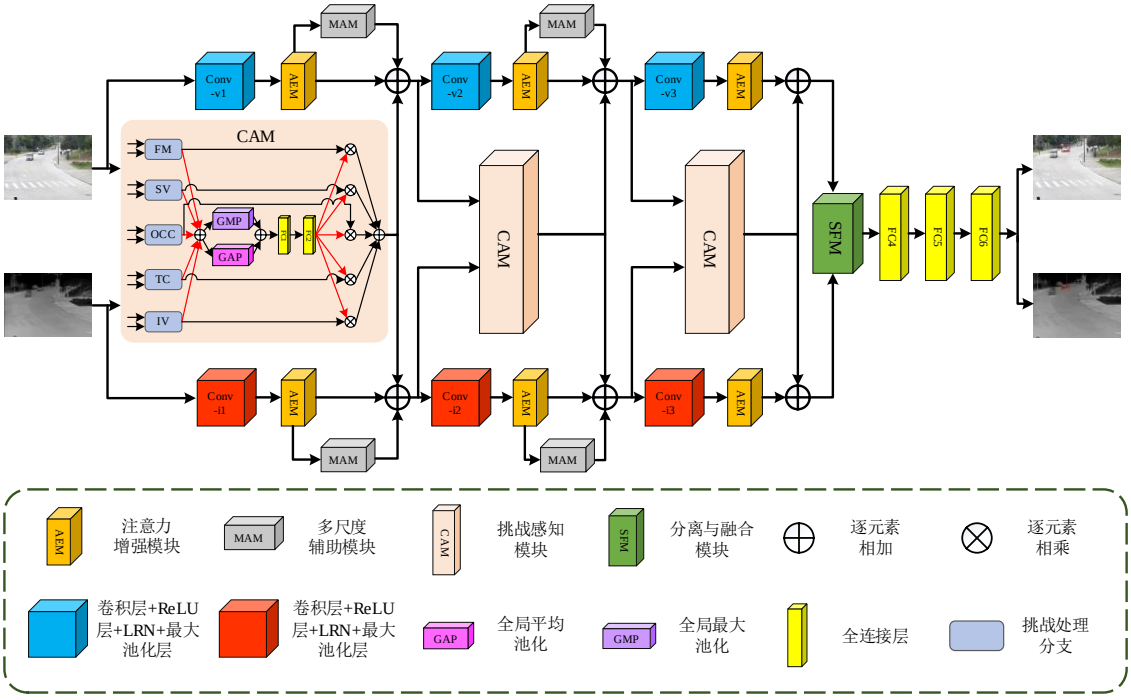


图1　面向不同挑战及多模态同异质信息分离网络(CMHHNet)的整体结构

具体来说,本文采用MDNet[17]的双流结构作为基线网络,同时选用一种轻量级的VGG-M网络[24],将其前三层作为特征提取网络,分别提取两种模态的特征. 3个卷积核的大小分别为7×7、5×5和3×3. 在每一层里,都嵌入了一个挑战感知模块,用于处理跟踪过程中所面临的不同挑战. 此外,在骨干网络中的每层都加入了注意力增强模块用于细化特征,降低冗余,而在前两层加入多尺度辅助模块来增加感受野. 最后,设计了一个特征分离与融合模块,对两种模态进行同异质信息分离后再融合所提取的可见光与热红外特征.

整个网络的运行流程如下. 首先,将可见光与热红外图像均裁剪为107×107的尺寸大小,之后输入处理后的图像,利用主干网络对其进行特征提取,然后将提取到的特征通过注意力增强模块与多尺度辅助模块进行特征增强. 与此同时,挑战感知模块将输入的图像特征在每个挑战下进行融合处理,并将所有处理后的特征发送到挑战聚合模块进行自适应聚合,之后将其加入特征提取网络中,以实现更鲁棒的特征表示. 接下来,将提取并得到增强的可见光与热红外特征输入到分离与融合模块中进行融合. 最后,将融合的特征输入3个全连接层来适应不同视频和帧实例外观的变化,通过二值分类与边界框回归来预测目标的位置.

3.2　挑战感知模块

在RGBT跟踪中,常常会受到各种挑战因素的干扰. 为了有效应对这些挑战,引入了挑战感知模块. 该模块选择了RGBT跟踪中最常见的尺度变化(SV)、光照变化(IV)、遮挡(OCC)、热交叉(TC)和快速运动(FM)5种挑战进行处理. 通过多任务训练解决训练数

据集规模不足,以及只用一种模型来应对所有挑战而导致的模型复杂度过高的问题. 每个挑战分支专注于一种挑战下的特征融合处理,因此,可以使用少量的参数和训练数据集来设计模型.

该模块主要分为挑战处理模块与挑战聚合模块两部分. 挑战处理模块包含5个分支,每个分支又包含基于挑战的通用特征提取部分与特征融合部分. 通用特征提取部分用于提取输入到挑战感知模块的可见光与热红外图像特征,学习一种初级、基本的目标表示,以保留原始模态图像中的基本信息,为后续模块对其进行深入的特征处理做准备. 每个分支均采用相同的特征提取结构,但不共享参数,从而能更好地提取每个挑战下的特征. 为了更清晰地展示该操作的合理性,本文从RGBT234数据集中选取了BlueCar视频序列的其中一帧可见光与热红外图像,分别输入网络第一层的3个挑战分支中,经过通用特征提取模块进行特征提取并求和后,将输出的特征选取第一个通道进行可视化,结果如图2所示. 可视化的结果表明,即便每个挑战分支采用相同的通用特征提取结构,但由于每个分支独立训练且不共享参数,输出的特征也会各不相同,能够针对不同的挑战进行差异化处理.

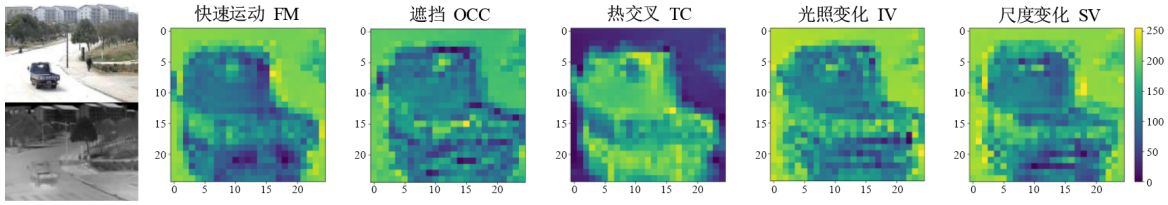


图2 选取的图像通过不同挑战分支通用特征提取模块后的可视化特征图

此外,基于挑战的通用特征提取结构在主干网络的每一层略有不同,具体结构如图3所示,其中ReLU为激活函数.

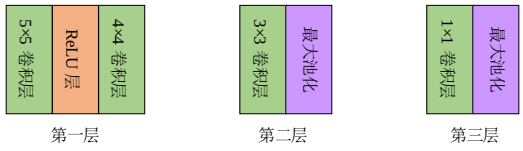
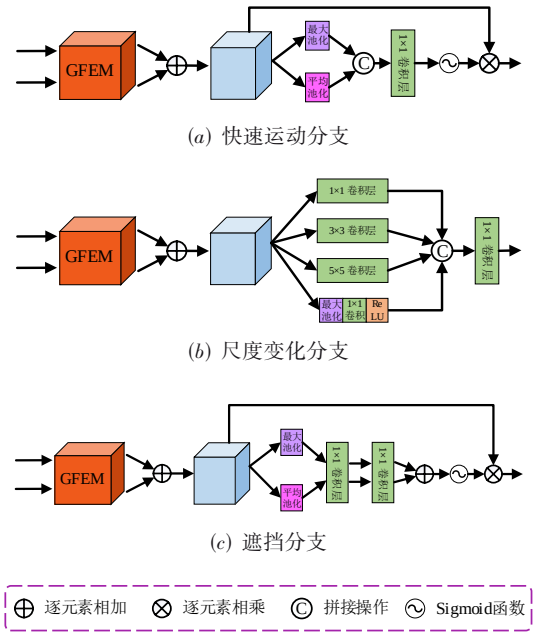


图3 通用特征提取结构的各层结构

在对每个挑战分支的输入进行特征提取建模后,还需要对提取的可见光与热红外特征进行融合. 本文根据不同挑战属性下的目标外观特性,开发了不同的融合结构来建模目标外观,以学习鲁棒的目标表示. 具体来说,将这5种挑战根据其特性划分为模态共有挑战与模态特有挑战. 其中,快速运动(FM)、尺度变化(SV)和遮挡(OCC)挑战在可见光与热红外两种模态下均存在,故划分为模态共有挑战;而光照变化(IV)挑战主要存在于可见光模态中,对热红外模态影响较小,同理,热交叉(TC)挑战则主要存在于热红外模态中,因此将这两种挑战划分为模态特有挑战.

在模态共有挑战中,首先采用求和的方式来融合两种模态的特征,该操作能够保留原始的可见光与热红外特征信息,为下游模块提供有效的目标特征. 从图2中可以看出,每个分支模态求和后的特征图各不相同,但都保留了图像中原始的特征信息. 此外,本文在后续的实验部分还进行了模态共有挑战分支不同融合方式的对比实验,能够更好地说明求和操作的有效性与高效性. 之后针对不同的挑战设计不同的结构来建模目标外观. 具体来说,对于快速运动挑战,由于目标在短时间内速度较快,易产生运动模糊、重影等问题. 而空间注意力机制能够有效捕捉图像中不同区域的重要性,可以根据目标的位置变化自适应调整权重,且能够过滤掉背景噪声,只关注目标所在的区域,这对于跟踪快速运动的目标非常关键,因此采用空间注意力结构来处理该挑战,从而能够确保目标在快速移动时仍能被准确定位,减少误检和漏检的发生,提高跟踪的稳定性. 在尺度变化挑战中,目标在视野中的尺寸会因距离、视角和运动而变化,本文则采用多尺度特征结构来解决该问题,首先设计了多个不同感受野的分支来模拟不同尺度下的目标表征,然后通过拼接操作进行集成,最后使用一个1×1的卷积来调整通道维度. 该结构可以捕捉目标在不同尺度下的信息,使得模型能够更好地适应目标的尺度变化. 最后针对遮挡挑战,目标会被其他物体部分或完全挡住,导致目标的特征信息丢失并产生遮挡噪声. 本文使用一个通道注意力结构来重新加权特征通道,使得目标在被遮挡时可以增强未被遮挡部分的关键特征,模型仍然能够从未被遮挡的部分中提取到有用信息,从而提升跟踪的鲁棒性. 图4展示了3个模态共有挑战分支的详细结构,包括快速运动分支、尺度变化分支和遮挡分支,其中GFEM为通用特征提取模块.

对于光照变化和热交叉两种模态特有挑战,本文采用同一种结构但不共享参数来建模目标外观. 由于可见光与热红外两种模态在模态特有挑战中有一方为低质量模态,如光照变化挑战中的可见光模态与热交叉挑战中的热红外模态,受到文献[44]的启发,采用



(a) 快速运动分支

(b) 尺度变化分支

(c) 遮挡分支

图4　3个模态共有挑战分支的详细结构

SKNet进行注意力加权操作自适应地从两种模态中选择对跟踪有利的通道相关特征,根据输入模态的质量动态调整不同模态的权重来进行特征融合. 与文献[26]不同的是,本文在加权融合之前同时进行了全局平均池化(Global Average Pooling,GAP)和全局最大池化(Global Max Pooling,GMP)操作. GAP注重特征图的整体信息,而GMP则更注重特征图的突出信息. 同时使用GAP与GMP不仅可以增强特征的代表性,丰富提取的卷积特征,而且可以减少对局部噪声的敏感,从而实现更准确的目标跟踪. 通过这种结构,可以有效处理RGBT目标跟踪中模态质量差异的问题,动态调整和优化特征融合过程,从而提升跟踪性能和鲁棒性. 图5展示了模态特有挑战分支的详细结构,其中包括光照变化分支与热交叉分支.

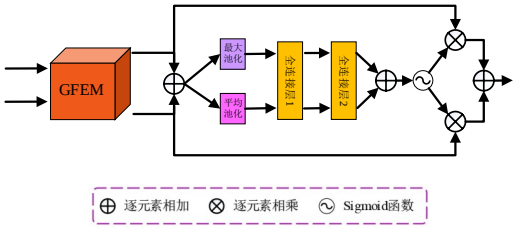


图5　模态特有挑战分支的详细结构

在数据集里对每个视频序列都手工标注了相应的挑战,这些标注在训练时可以使用,但在测试阶段却不能使用,因此在跟踪过程中无法确定应该激活哪些挑战分支. 为了解决这一问题,设计了挑战聚合模块,对所有的挑战分支进行自适应聚合. 仍然使用SKNet[44]进行聚合,能够动态调整每个分支的聚合权重,以获得更鲁棒的聚合特征. 详细结构如图1所示.

3.3　分离与融合模块

为了融合主干网络所提取的可见光与热红外特征,设计了一个特征分离与融合模块. 先对两个模态进行同异质信息分离,将可见光与热红外的共有特征与特有特征分离开,之后再通过注意力机制进行自适应融合,以达到使融合过程更加合理、有效的目的. 该模块的网络结构如图6所示,其中的K、Q、V分别表示自注意力与交叉注意力中的Key、Query和Value.

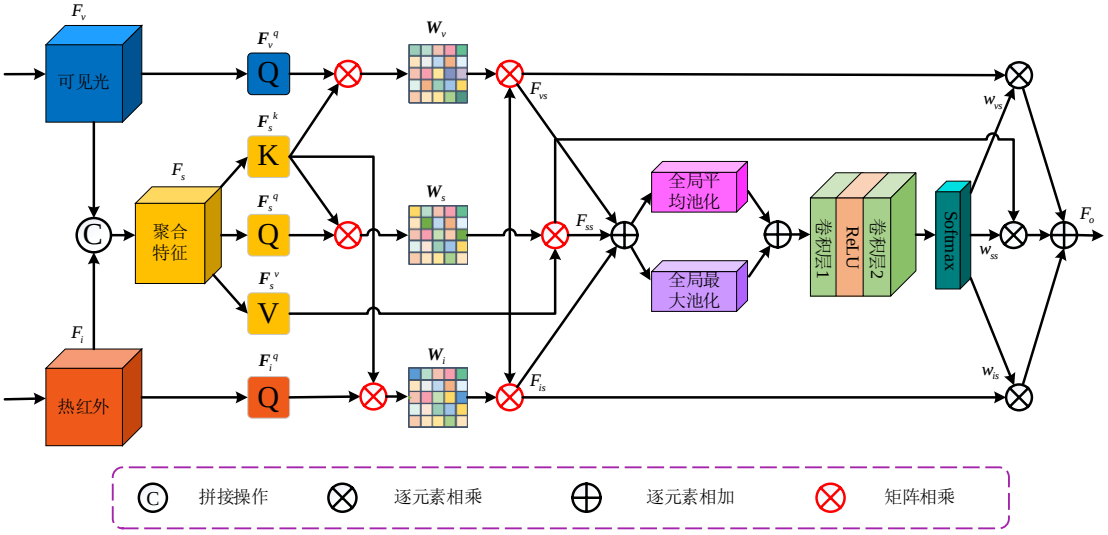


图6　分离与融合模块的网络结构图

受到文献[3]的启发,采用自注意力及交叉注意力来完成两个模态共有特征与特有特征的提取.具体来说,首先通过拼接操作将从可见光和热红外图像中提取的特征进行聚合,聚合特征包含了可见光和热红外的特征信息.之后通过1×1卷积生成可见光及热红外特征的Query以及聚合特征的Key、Query、Value.分别将可见光及热红外特征的Query与聚合特征的Key相乘来计算全局相关性,然后再与聚合特征的Value相乘来计算交叉注意力,从而得到可见光与热红外的特有特征.同时对聚合特征进行自注意力计算得到可见光与热红外的共有特征.最后分别将它们的特有特征与共有特征进行自适应加权融合,从而得到最终的融合特征.上述过程用公式表示如下:

$$F_s=\varepsilon(F_v,F_i) \tag{1}$$

$$\begin{cases}\boldsymbol{W}_v=\boldsymbol{F}_v^q\otimes\boldsymbol{F}_s^k=f_{vq}^{1\times1}(F_v)\otimes f_{sk}^{1\times1}(F_s)\\ \boldsymbol{W}_i=\boldsymbol{F}_i^q\otimes\boldsymbol{F}_s^k=f_{iq}^{1\times1}(F_i)\otimes f_{sk}^{1\times1}(F_s)\\ \boldsymbol{W}_s=\boldsymbol{F}_s^q\otimes\boldsymbol{F}_s^k=f_{sq}^{1\times1}(F_s)\otimes f_{sk}^{1\times1}(F_s)\end{cases} \tag{2}$$

$$\begin{cases}F_{vs}=\delta(\boldsymbol{W}_v)\otimes\boldsymbol{F}_s^v\\ F_{is}=\delta(\boldsymbol{W}_i)\otimes\boldsymbol{F}_s^v\\ F_{ss}=\delta(\boldsymbol{W}_s)\otimes\boldsymbol{F}_s^v\end{cases} \tag{3}$$

$$\begin{aligned}w_{vs},w_{is},w_{ss}=\delta\Big(\xi\big(&\Phi_1(F_{vs}+F_{is}+F_{ss})\\&+\Phi_2(F_{vs}+F_{is}+F_{ss})\big)\Big)\end{aligned} \tag{4}$$

$$F_o=w_{vs}\times F_{vs}+w_{is}\times F_{is}+w_{ss}\times F_{ss} \tag{5}$$

其中,$F_v,F_i\in\mathbf{R}^{C\times H\times W}$表示经过第三层特征提取网络输出的可见光与热红外特征;$F_s\in\mathbf{R}^{2C\times H\times W}$表示聚合特征;$F_o\in\mathbf{R}^{2C\times H\times W}$则表示最终的融合特征;+和×分别代表逐元素相加和相乘操作;⊗则表示矩阵相乘;此外,ε为按通道维度进行拼接操作;$f^{1\times1}$为1×1卷积层;δ为Softmax操作;Φ_1与Φ_2分别表示全局最大池化与全局平均池化;$\xi=f^2\big(\gamma(f^1)\big)$,$f^1$和$f^2$均表示1×1卷积层,$\gamma$则表示ReLU激活函数.

3.4 注意力增强模块与多尺度辅助模块

为了增强主干网络所提取的特征,使其对后续的跟踪更加有利,受到文献[39]和文献[41]的启发,在主干网络中加入了注意力增强模块与多尺度辅助模块.下面将详细描述这两个组件.

3.4.1 注意力增强模块

加入该模块的主要目的是使网络更加关注对跟踪有利的特征,从而抑制噪声的传播,减少冗余特征.在双流主干网络的每一层都嵌入了该模块,其主要包含通道注意力与空间注意力两部分.该模块的详细结构如图7所示,其中的浅绿色区域表示通道注意力,浅蓝色区域表示空间注意力.

通道注意力的目标是增强每个通道的特征表达.

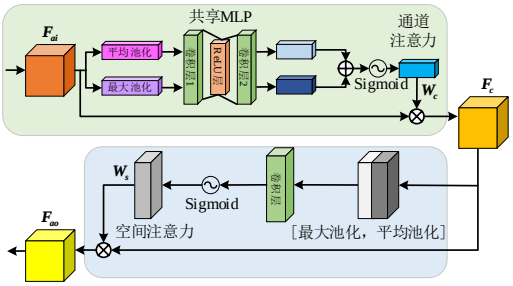


图7 注意力增强模块的网络结构图(图中的所有符号含义与图6相同)

具体步骤:对于输入的特征图,首先对每个通道执行全局最大池化和全局平均池化操作,之后将经过池化后的特征向量输入一个共享全连接层中得到注意力权重向量.再通过Sigmoid函数进行权重归一化,最后将得到的注意力权重与原始特征图的每个通道相乘,得到注意力加权后的通道特征图.空间注意力的目的则是强调图像中不同位置的重要性.详细操作:首先将输入的特征图沿通道维度进行最大池化和平均池化操作,生成不同上下文尺度的特征;之后将其沿通道维度进行拼接并通过一个卷积层和一个Sigmoid函数来生成空间注意力权重;最后将得到的空间注意力权重应用于原始特征图,对每个空间位置的特征进行加权.经过通道与空间注意力操作后,特征的有效性得到了显著的提高.上述操作用公式表示如下:

$$W_c=\sigma\Big(f_m\big(\varphi_{c1}(F_{ai})\big)+f_m\big(\varphi_{c2}(F_{ai})\big)\Big) \tag{6}$$

$$F_c=W_c\times F_{ai} \tag{7}$$

$$W_s=\sigma\Big(f_s\big(\varepsilon\big(\varphi_{s1}(F_c),\varphi_{s2}(F_c)\big)\big)\Big) \tag{8}$$

$$F_{ao}=W_s\times F_c \tag{9}$$

其中,$F_{ai}\in\mathbf{R}^{C\times H\times W}$,$F_{ao}\in\mathbf{R}^{C\times H\times W}$分别表示该模块的输入与输出特征;$F_c\in\mathbf{R}^{C\times H\times W}$表示经过通道注意力强化后的特征;$W_c\in\mathbf{R}^{C\times1\times1}$与$W_s\in\mathbf{R}^{1\times H\times W}$则分别表示通道注意力与空间注意力的权重;$\varphi_{c1}$、$\varphi_{c2}$、$\varphi_{s1}$、$\varphi_{s2}$分别表示沿空间维度和通道维度的最大池化与平均池化操作;$f_m=f_m^2\big(\gamma(f_m^1)\big)$,$f_m^1$和$f_m^2$均为1×1卷积层,$\gamma$为ReLU函数;$f_s$为7×7的卷积层,$\varepsilon$为通道维度的拼接操作;$\sigma$则表示Sigmoid函数.

3.4.2 多尺度辅助模块

在跟踪过程中,经常会遇到小尺寸的目标,直接提取它们的特征往往效果不佳,而多尺度特征有利于处理尺度变化和识别小尺寸目标.因此,为了更好地提取对捕获不同尺度目标信息至关重要的多尺度特征,本文在特征提取网络的前两层加入了多尺度辅助模块,以促进多尺度信息的传播.该模块结构如图8所示,图中d表示膨胀率.

为了减少网络的参数,本文采用小尺寸的卷积核来代替大尺寸卷积核.这样不仅可以保持感受野不变,

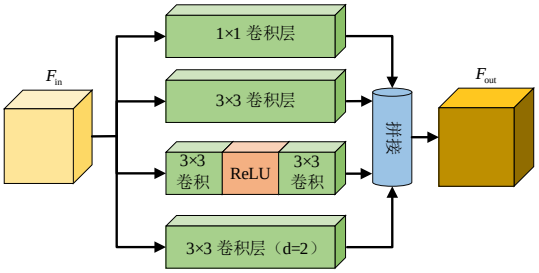


图8　多尺度辅助模块的网络结构图

而且由于堆叠的卷积层中包含的激活函数的非线性更大,使得决策函数更具判别性[45]. 因此,在多尺度辅助模块的第三层采用两个3×3的卷积和一个ReLU函数,而在第四层则采用一个膨胀率为2的3×3空洞卷积. 该操作用公式表示如下:

$$\begin{aligned} F_{\text{out}} &= \varepsilon(F_1, F_2, F_3, F_4) \\ &= \varepsilon\left(f^1(F_{\text{in}}), f^2(F_{\text{in}}), f^3(F_{\text{in}}), f^4(F_{\text{in}})\right) \end{aligned} \tag{10}$$

其中,$F_{\text{in}} \in \mathbf{R}^{C\times H\times W}$和$F_{\text{out}} \in \mathbf{R}^{C\times H\times W}$分别表示输入特征与输出特征;$F_1 \in \mathbf{R}^{C/4\times H\times W}$、$F_2 \in \mathbf{R}^{C/4\times H\times W}$、$F_3 \in \mathbf{R}^{C/4\times H\times W}$、$F_4 \in \mathbf{R}^{C/4\times H\times W}$是提取的不同尺度的特征;此外,$f^1$和$f^2$分别表示1×1和3×3卷积操作,$f^3 = f^{32}\left(\gamma(f^{31})\right)$,$f^{31}$和$f^{32}$均为3×3卷积层,$\gamma$表示ReLU操作,$f^4$为膨胀率为2的3×3空洞卷积;$\varepsilon$则代表沿通道维度进行拼接操作.

3.5　网络训练

由于本文的网络包含挑战感知模块,如果直接使用所有的训练数据来训练网络,那么训练数据的损失都会向后传播到所有挑战的聚合分支中[26]. 因此,本文采用了三阶段的训练方法,具体操作如下.

首先,在基线网络MDNet+RGBT[16]中分别添加每个挑战的挑战处理分支,采用具有不同挑战标签的数据集来逐个训练对应挑战的处理分支. 具体而言,首先利用imageNetvid[17]上得到的预训练模型参数对双流CNN网络进行初始化,该模型包括3个卷积层和2个全连接层FC4与FC5,然后再初始化所有挑战处理分支的参数并添加新的分类分支FC6. 对于基线网络模型中的卷积层和FC4与FC5,将其参数学习率设置为0.000 1,而将FC6与新添加的挑战处理分支中的参数学习率分别设置为0.000 5和0.001. 之后采用随机梯度下降法(Stochastic Gradient Descent,SGD)作为优化策略,动量为0.9,权值衰减设为0.000 5,损失函数则采用与MDNet相同的二分类交叉熵[17]. 每个挑战的训练轮数均为200,最后只保存每个挑战处理分支的参数.

在第二阶段中,将所有的挑战处理分支及挑战聚合模块添加到基线网络中,使用所有训练数据来训练挑战聚合模块. 之后随机初始化挑战聚合模块和全连接层FC6的参数,并将学习率分别设置为0.001和0.000 5. 该阶段训练轮数为300,其他设置与第一阶段相同,最后保存挑战聚合模块与FC4和FC5的参数.

在最后一个阶段,将所有模块均添加到网络中,然后使用所有的数据来训练新添加的分离与融合模块、注意力增强模块及多尺度辅助模块. 随机初始化新增模块和全连接层FC6的参数,并将它们的学习率分别设置为0.001和0.000 5,其他已有的模块学习率则设置为0.000 1进行参数微调. 该阶段的训练轮数为500,其他设置与第一阶段相同,最后保存整个模型的参数.

3.6　在线跟踪

对于每个新的视频序列,在跟踪过程中会随机初始化一个新的全连接层(FC6)分支,然后固定之前训练的所有模型参数. 接着,对FC4、FC5和FC6进行微调,以适应跟踪过程中出现的各种外观变化. 在第一帧中,首先针对给定的初始目标收集样本,包括500个正样本和5 000个负样本,用于微调全连接层. 定义交并比(Intersection over Union,IoU)大于0.7的样本为正样本,小于0.5的样本为负样本. 为了提高跟踪准确性,在第一帧中额外收集1 000个样本,用于训练回归器以适应新的目标域. 在后续帧中,根据前一帧的跟踪结果从256个候选样本中进行采样,并将其发送给网络用于当前帧的跟踪. 然后,选择得分最高的5个候选样本,对它们边界框的宽和高及中心位置坐标求平均值作为当前帧的跟踪结果. 如果跟踪得分高于0,则视为跟踪成功,并使用回归器对结果进行微调,以进一步提高定位精度. 最后,收集20个正样本和100个负样本,动态更新网络,以进一步适应目标的变化.

4　实验结果及分析

实验中,使用PyTorch深度学习框架,操作系统为Ubuntu 20.04,GPU采用NVIDIA GeForce RTX3090 24 GB,对所提出的网络CMHHNet与近年来提出的性能先进的RGBT跟踪器进行了比较,分别考察了模型在GTOT[27]、RGBT234[12]和LasHeR[28]这3个常用数据集上的表现. 此外,对网络结构做了消融实验,以验证网络中每个模块的有效性. 为了直观说明本文所提方法的有效性,除了给出评测指标结果,还展示了与其他网络的可视化对比结果.

4.1　评估设置

4.1.1　数据集

实验所用数据集均为RGBT跟踪常用数据集. GTOT是由Li等人[27]提出的首个RGBT跟踪基准. 它由50对在不同场景和条件下采集的具有真值注释的可见光和热红外视频对组成. 此外,该数据集还对视频进行了属性标注,并根据这些属性将其划分为7个子集,旨在评估算法对这些属性的感知能力. RGBT234是Li等人[12]提出的另一个大型RGBT跟踪数据集. 它是

RGBT210[46]数据集的扩展，包括234个RGBT视频序列对及其对应的真值注释. 视频序列标注中有22个目标类别和12个挑战属性，总的帧数为234K，最长的视频序列帧数为8K. LasHeR[28]则是一个更大更全面的RGBT跟踪数据集，它分为测试集和训练集，共包含1 224个对齐的RGBT视频序列，共有730K个图片对，其中包括了32个目标类别和19个挑战属性.

4.1.2 评估指标

GTOT和RGBT234数据集上的跟踪性能通常由两个指标来评估：精确率(Precision Rate，PR)和成功率(Success Rate，SR). PR表示算法预测的目标框中心与标注的目标框中心之间的欧氏距离小于规定阈值的帧的百分比，GTOT数据集上的目标较小，因此将该数据集上的阈值设定为5个像素，而将RGBT234、LasHeR数据集上的阈值设定为20个像素. SR则表示输出边界框与真值边界框之间的重叠率大于阈值的帧的百分比，改变阈值可以获得SR曲线图，本文采用SR曲线下的面积作为代表性的SR. 在LasHeR数据集上，由于精确率会受到视频分辨率的影响，本文又添加了标准化精确率(Normalized Precision Rate，NPR)这个指标来进行评估.

4.2 GTOT与RGBT234数据集上的评估

本文在GTOT与RGBT234数据集上进行交叉训练与测试. 具体来说，使用RGBT234数据集上训练的网络模型在GTOT上进行评估，同理，采用在GTOT上训练的模型来测试RGBT234数据集. 将本文的方法与目前现有的最先进的RGBT跟踪器在精确率及成功率指标上进行了比较，包括MDNet+RGBT[16]、CAT[25]、JMMAC[47]、ADRNet[48]、DFNet[49]、MFGNet[50]、APFNet[26]、HMFT[51]、CIRNet[33]、MMMPT[52]、MSIFNet[53]、DECFNet[54]、DRGCNet[42]、SiamCAF[55]、JTPMA[56]共15个跟踪器，对比结果如表1所示，加粗结果表示对比的最佳性能.

由表1可以看出，CMHHNet在GTOT数据集上的精确率和成功率分别是91.1%和73.8%，在RGBT234数据集上则达到了82.3%和58.8%. 相比于基线的MDNet+RGBT算法分别提高了11.1个百分点/10.1个百分点和10.1个百分点/9.3个百分点，遥遥领先于基线网络. 与同样基于属性挑战的网络CAT、ADRNet与APFNet相比在不同数据集上的各个指标均有不同程度的提高. 相较于APFNet采用统一的网络结构来应对不同的挑战，本文网络则将跟踪所面临的挑战划分为模态共有与模态特有，根据每个挑战的特点，量身设计了独特的网络结构，从而能够更好地建模每个挑战下的目标外观. 此外，相较于CAT网络，CMHHNet在最后设计了分离与融合模块代替传统的拼接操作来融合两种模态的特征，显著提高了跟踪的精确率.

表1 CMHHNet与其他15个跟踪器在GTOT与RGBT234数据集上的精确率及成功率对比结果 单位：%

跟踪器	年份	GTOT		RGBT234	
		PR	SR	PR	SR
MDNet+RGBT[16]	2018	80.0	63.7	72.2	49.5
CAT[25]	2020	88.9	71.7	80.4	56.1
JMMAC[47]	2021	90.1	73.2	79.0	57.3
ADRNet[48]	2021	90.4	73.9	80.9	57.1
DFNet[49]	2022	88.1	71.9	77.2	51.3
MFGNet[50]	2022	88.9	70.7	78.3	53.5
APFNet[26]	2022	90.5	73.7	**82.7**	57.9
HMFT[51]	2022	90.8	73.8	78.8	56.8
CIRNet[33]	2022	90.1	72.8	81.0	54.4
MMMPT[52]	2023	90.3	73.6	79.8	54.6
MSIFNet[53]	2023	90.4	74.1	81.7	57.0
DECFNet[54]	2023	90.4	71.8	82.3	57.6
DRGCNet[42]	2023	90.5	73.5	82.5	58.1
SiamCAF[55]	2023	90.6	73.0	77.1	53.7
JTPMA[56]	2023	90.7	**75.1**	80.3	56.2
CMHHNet	—	**91.1**	73.8	82.3	**58.8**

4.3 基于不同挑战的评估

本节中，将本文的网络与其他9种RGBT跟踪算法在RGBT234数据集上针对运动模糊(Motion Blur，MB)、局部遮挡(Partial Occlusion，PO)、低分辨率(Low Resolution，LR)、快速移动(Fast Motion，FM)、背景杂乱(Background Clutter，BC)、严重遮挡(Heavy Occlusion，HO)、镜头移动(Camera Moving，CM)、无遮挡(No Occlusion，NO)、变形(DEFormation，DEF)、热交叉(Thermal Crossover，TC)、弱光照(Low Illumination，LI)和尺度改变(Scale Variation，SV)这12种不同挑战下的表现进行了对比. 表2和表3分别给出了精确率PR和成功率SR的对比结果.

从表2和表3中可以看出，对于精确率，本文的网络在9项挑战中都排名第一，11项挑战位于前三名，仅有FM一项挑战排名第四，仅比前一名低了0.8个百分点；在成功率指标上则是达到了7个挑战居于第一，所有挑战全部处于前三名的优异成绩，该对比结果证明了本文方法在处理各种挑战场景时的鲁棒性，尤其是HO、LI、CM、MB等更复杂的挑战，其中的挑战感知模块能够很好地模拟这些挑战属性下目标的外观表示. 相比于ADRNet仅采用求和方式来融合不同的挑战分支，CMHHNet则对每个分支采用自适应加权的方法进行聚合，因此其性能在大部分挑战上都有所提升. 此外，CMHHNet相比于基线网络MDNet+RGBT在所有挑战上都全方面超越，说明了网络中各个模块的有效性.

表2　RGBT234数据集上CMHHNet与其他9个跟踪器在不同挑战下的精确率对比　　单位:%

跟踪器	NO	PO	HO	LI	LR	TC	DEF	FM	SV	MB	CM	BC	ALL
MDNet+RGBT[16]	86.2	76.1	61.9	67.0	75.9	75.6	66.8	58.6	73.5	65.4	64.0	64.4	72.2
DAPNet[22]	90.0	82.1	66.0	77.5	75.0	76.8	71.7	67.0	78.0	65.3	66.8	71.7	76.6
SiamCAF[55]	85.3	80.5	69.8	78.3	73.7	70.7	71.2	64.4	75.8	68.3	71.7	73.1	77.1
HDINet[57]	88.4	84.9	67.1	77.7	80.1	77.2	76.2	71.7	77.5	70.8	69.7	71.1	78.3
FANet[21]	88.2	86.6	66.5	80.3	79.5	76.6	72.2	68.1	78.5	70.0	72.4	75.7	78.7
MaCNet[31]	92.7	81.1	70.9	77.7	78.3	77.0	73.1	72.8	78.7	71.6	71.7	77.8	79.0
ADRNet[48]	91.7	86.3	70.8	80.2	83.1	78.9	74.3	77.6	79.0	72.7	75.7	78.9	80.9
MBAFNet[43]	91.3	86.4	68.8	81.5	82.3	77.0	76.2	75.2	79.7	71.6	75.2	78.2	80.1
JTPMA[56]	92.0	84.8	70.5	82.3	83.5	79.8	72.0	75.9	79.8	70.3	70.8	81.8	80.3
CMHHNet	92.8	86.6	73.2	85.3	82.4	80.8	77.8	74.4	83.2	73.4	77.5	79.8	82.3

注:排名前三的数据分别用红色、绿色和蓝色突出显示.

表3　RGBT234数据集上CMHHNet与其他9个跟踪器在不同挑战下的成功率对比　　单位:%

跟踪器	NO	PO	HO	LI	LR	TC	DEF	FM	SV	MB	CM	BC	ALL
MDNet+RGBT[16]	61.1	51.8	42.1	45.5	51.5	51.7	47.3	36.3	50.5	46.3	45.4	43.2	49.5
DAPNet[22]	64.4	57.4	45.7	53.0	51.0	54.3	51.8	44.3	54.2	46.7	47.4	48.4	53.7
SiamCAF[55]	61.2	56.1	47.9	52.7	50.1	50.9	51.7	41.7	53.8	49.2	51.1	48.1	53.7
HDINet[57]	65.1	60.4	47.3	53.2	54.5	57.5	56.5	47.5	55.8	52.6	51.4	47.8	55.9
FANet[21]	65.7	60.3	45.8	54.8	53.2	55.1	52.7	43.8	56.3	50.3	52.3	50.3	55.3
MaCNet[31]	66.5	57.2	48.8	52.7	52.3	56.3	51.4	47.1	56.1	52.5	51.7	50.1	55.4
ADRNet[48]	65.8	61.2	49.1	55.1	55.6	58.9	52.9	50.3	56.2	53.0	53.5	52.7	57.1
MBAFNet[43]	68.7	63.9	48.5	57.2	56.0	57.4	57.0	50.2	59.7	53.2	55.6	51.3	58.5
JTPMA[56]	65.8	60.8	47.4	56.9	55.8	58.1	50.5	48.8	56.5	51.3	50.6	53.5	56.2
CMHHNet	68.0	62.4	51.1	59.2	56.4	59.0	56.9	48.9	59.4	54.7	56.9	54.2	58.8

注:排名前三的数据分别用红色、绿色和蓝色突出显示.

4.4　可视化评估

为了更加直观地将CMHHNet与其他跟踪器进行比较,选用MDNet+RGBT[16]、FANet[21]、TFNet[58]和SGT[46]这4种方法与本文的方法进行了可视化对比.通过对RGBT234数据集中elecbike2、boundaryandfast、man24和woman89这4对视频序列做跟踪测试,CMHHNet表现出了非常优秀的性能.具体细节如图9所示,其中,每个视频序列中上面一行代表可见光图像,下面一行为热红外图像.

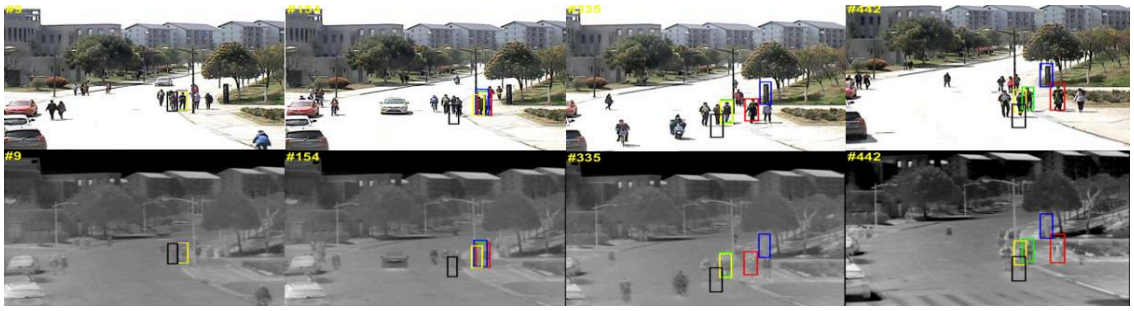
在elecbike2视频序列中,跟踪目标为一位黑衣男孩所骑的电动自行车.在视频第38帧时,目标开始与行人发生重叠并受到了行人的遮挡,从而导致在第66帧时,目标与行人分离,但SGT与FANet均发生了跟踪漂移.之后在第253帧中,目标开始与大量行人发生交汇,TFNet开始逐渐偏离目标,最后在第581帧中,MDNet+RGBT的跟踪结果也从目标电动车偏移到了骑车的男孩,但CMHHNet仍能够精确定位到目标电动车.整个视频序列中包含遮挡、热交叉、小尺寸目标跟踪等多种挑战,跟踪难度高,但本文网络仍能够有效应对这些挑战,实现精准跟踪.在boundaryandfast序列中,目标人物快速运动,画面产生模糊,在第35帧中,FANet首先偏离目标,之后在49帧与164帧之间,其他跟踪器也陆续跟丢目标,到了最后一帧时,只有CMHHNet还在保持着很高的精准度.对于man24序列,第9帧时MDNet+RGBT已经偏离目标,154帧时目标人物开始受到遮挡,在后面的335帧与442帧中,目标人物几乎被完全遮挡,其他跟踪器都已偏离目标,只有本文的跟踪器仍能够准确跟踪.最后,在woman89视频中,第308帧时镜头出现了快速移动,导致画面模糊,其他跟踪器都有不同程度的偏离,最后在388帧中,目标人物受到车辆的部分遮挡,也只有CMHHNet能够紧跟目标,没有受到这些挑战的影响,表明了挑战感知模块的有效性.

4.5　LasHeR数据集上的评估

为了进一步说明CMHHNet对不同数据集的鲁棒性,本文在LasHeR的测试子集上也进行了测试,LasHeR是比RGBT234规模更大的数据集,同时也具有更加复杂的挑战.将本文提出的跟踪器与FANet[21]、CAT[25]、MaCNet[31]、MANet++[19]、AGMINet[41]、DMCNet[32]、APFNet[26]、MSIFNet[53]、DRGCNet[42]、DECF-

(*a*) elecbike2

(*b*) boundaryandfast



(*c*) man24

(*d*) woman89

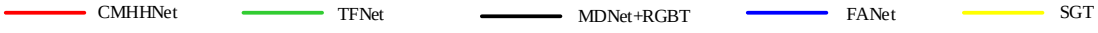


图 9 CMHHNet 与其他 4 个 RGBT 跟踪器在 4 对视频序列上的可视化对比图

Net[54]、EANet[59]、LMINet[60]、CAT++[61]共 13 个最先进的 RGBT 跟踪器在精确率 PR、成功率 SR 及标准化精确率 NPR 这 3 个指标上进行了比较，对比结果如表 4 所示.

从表 4 可以看出，CMHHNet 在 LasHeR 数据集上的表现要优于在 GTOT 和 RGBT234 上的表现，在 3 个指标上均超过了其他的跟踪器，PR、NPR 及 SR 分别达到了

表4　CMHHNet与其他13个跟踪器在LasHeR数据集上的性能对比结果　　单位:%

跟踪器	年份	LasHeR		
		PR	NPR	SR
FANet[21]	2020	44.2	38.4	30.9
CAT[25]	2020	45.1	39.8	31.7
MaCNet[31]	2020	48.3	42.3	35.2
MANet++[19]	2021	46.7	40.8	31.7
AGMINet[41]	2022	48.8	42.9	34.3
DMCNet[32]	2022	49.1	43.1	35.7
APFNet[26]	2022	50.0	43.9	36.2
MSIFNet[53]	2023	—	42.8	35.2
DRGCNet[42]	2023	48.3	42.3	33.8
DECFNet[54]	2023	49.6	—	35.0
EANet[59]	2023	50.6	—	36.7
LMINet[60]	2024	49.0	43.3	34.8
CAT++[61]	2024	50.9	44.4	35.6
CMHHNet	—	**51.1**	**45.7**	**37.6**

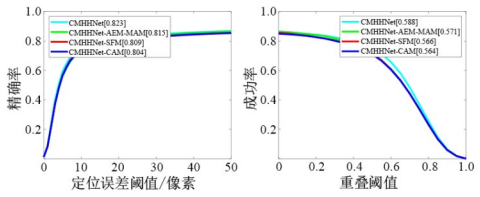
51.1%、45.7%和37.6%. 与第二名的CAT++相比,改进幅度分别为0.2、1.3和2.0个百分点. 与同样基于属性挑战的网络CAT、APFNet相比则是在各个指标上均得到了全面提升. 与最新的网络LMINet和CAT++相比,CMHHNet的整体参数规模更小,训练速度更快,除此之外,在训练时对大规模数据集的依赖也更小,可以仅使用少量的数据集就训练出一个泛化性能更强的模型,能够有效降低网络过拟合的风险. 这也证明了该网络在应对复杂挑战场景下的可靠性,能够使用不同的网络结构有效处理不同的挑战,对每个挑战下的目标外观特性进行合理建模.

从以上3个典型的RGBT跟踪数据集上的表现来看,CMHHNet对不同的数据集都具有很强的鲁棒性,在RGBT跟踪算法中展示出很强的竞争力.

4.6　消融实验

为了验证CMHHNet中各个模块的有效性,在RGBT234数据集上进行了消融实验,在CMHHNet中依次去除各个模块得到变体模型. 具体来说,CMHHNet-AEM-MAM表示去除网络中的注意力增强及多尺度辅助模块,CMHHNet-CAM表示去除挑战感知模块,CMHHNet-SFM则表示去除分离与融合模块. 实验结果如图10所示.

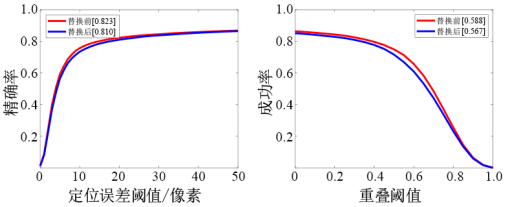
从图10可以看出,CMHHNet相比于它的3个变体在精确率和成功率上均有较大提升,说明了该网络的各个组件是有效的. 此外,CMHHNet相比于CMHHNet-CAM的提升最大,达到了1.9%和2.4%,则表明本文设计的挑战感知模块相比于其他模块更有优势,对网络的贡献更大,也间接说明了本文针对不同挑战设计不同的网络结构进行处理并自适应聚合所有挑战分支的思想是有效的.



(a) 精确率曲线图　　(b) 成功率曲线图

图10　CMHHNet与3个变体在RGBT234数据集上的精确率及成功率曲线图

为了进一步说明挑战感知模块CAM的有效性,本节利用文献[25]中的挑战分支设计替换本文设计的CAM,在RGBT234数据集上进行了对比实验,实验结果如图11所示. 图11中的替换前与替换后分别代表采用CAM与采用文献[25]中的挑战分支设计.



(a) 精确率曲线图　　(b) 成功率曲线图

图11　将CAM替换为文献[25]中的挑战分支设计在RGBT234数据集上的精确率及成功率曲线图

由结果可以看出,本文设计的CAM相较于文献[25]中的挑战分支设计在RGBT234数据集上的精确率及成功率分别提高了1.3%和2.1%,进一步证明了挑战感知模块的有效性,且对于网络的性能提升是巨大的.

4.7　挑战分支评估实验

本节对挑战感知模块中每个挑战处理分支的有效性进行验证,以证明每个挑战分支都是特定的,能够很好地解决对应的挑战. 从GTOT数据集中选择了FastMotor、Jogging、occBike、BlueCar、Crossing共5个典型的视频序列,每个序列都对应一个典型挑战. 在网络基本框架上,每次只添加一个挑战处理分支,依次对所选择的每个序列进行测试. 测试结果在表5中展示,每个视频序列的名称之后标记了该视频具有的典型挑战.

实验结果表明,与其他挑战处理分支相比,每个挑战处理分支在与其挑战对应的视频序列上都取得了最佳性能,验证了挑战处理分支的有效性.

4.8　网络参数规模与跟踪实时性对比

在目标跟踪领域,网络参数规模与跟踪实时性是影响实用性的两个重要因素. 本节针对这两个方面与其他网络进行了对比,表6与表7分别展示了CMHHNet与5种优秀RGBT算法的参数规模与跟踪实时性的比

表5 每个挑战处理分支在5个GTOT视频序列上的跟踪成功率

单位:%

	快速运动分支	光照变化分支	遮挡分支	尺度变化分支	热交叉分支
FastMotor:快速运动	**64.3**	61.5	62.6	64.1	61.8
Jogging:光照变化	80.4	**83.9**	82.1	79.2	80.8
occBike:遮挡	55.7	60.0	**62.4**	58.8	56.9
BlueCar:尺度变化	56.2	56.8	55.7	**56.9**	53.7
Crossing:热交叉	79.8	78.9	79.7	79.4	**80.3**

较结果,其中FPS为每秒跟踪帧数.

从表6和表7可以看出,本文网络的整体参数规模相较于其他网络要少,模型更加轻量化,因此所需要的计算资源及占用的内存资源更少. 此外,更小的模型有助于快速训练与部署,还能够减少过拟合的风险,有助于提高模型在未知数据上的泛化能力. 对于跟踪速度,CMHHNet与使用相同基线网络的5种跟踪器相比最快,效率是MaCNet的3倍. 总的来说,本文网络在跟踪性能和效率方面均取得了良好的效果.

表6 CMHHNet与5种RGBT跟踪器的参数规模对比

单位:百万

跟踪器	参数规模
mfDiMP[14]	175
HMFT[51]	127
ViPT[62]	93
TBSI[63]	202
BAT[64]	92
CMHHNet	31

表7 CMHHNet与5种RGBT跟踪器的跟踪实时性对比

单位:帧/秒

跟踪器	FPS
CMPP[65]	1.3
DAPNet[22]	2.0
HDINet[57]	0.9
APFNet[26]	1.9
MaCNet[31]	0.8
CMHHNet	2.3

4.9 共有挑战分支融合方式对比实验

在挑战感知模块的模态共有挑战分支中,本文采用了求和操作来融合可见光与热红外两种模态的特征. 为了说明这一操作的合理性,本文将求和操作替换为拼接操作,并在GTOT数据集上分别测试了网络的精确率、成功率及跟踪速度,得到的对比实验结果如表8所示.

实验结果表明,拼接操作相较于求和在精确率和成功率上分别降低了1.7个百分点和1.1个百分点,跟踪速度也有所降低. 该结果说明采用求和进行特征融合在网络性能与跟踪速度上具有优势.

表8 模态共有挑战分支中不同融合方式在GTOT上的精确率、成功率及跟踪速度对比

融合方式	求和	拼接
精确率/%	91.1	89.4
成功率/%	73.8	72.7
FPS/(帧/秒)	2.3	2.2

4.10 计算量对比实验

在深度学习中,网络的计算量是指神经网络在进行前向传播时所需的算术运算的次数. 它主要用于衡量网络运行时的计算成本,是评估模型复杂度的一个重要指标. 通常用浮点运算次数(Floating Point Operations,FLOPs)来衡量,它表示整个网络需要执行的浮点运算总次数. 在本节中,为了探寻分离与融合模块SFM的计算量,将SFM替换为直接拼接操作,分别进行计算量对比实验,得到的实验结果如表9所示.

表9 添加SFM与直接拼接模型的计算量对比 单位:GFLOPs

融合方式	添加SFM	直接拼接
计算量	1.499	1.477

由结果可以看出,添加SFM后的模型,即整个CMHHNet模型的计算量仅有1.5 G(1 G=10^9)FLOPs左右,整个模型较为轻量化,这也与前面的网络参数规模实验结果相符. 此外,分离与融合模块SFM的计算量仅有0.022 GFLOPs,仅占整个模型的1.5%,体现了该模块的简洁高效性.

5 结论

本文提出一种新的RGBT跟踪网络CMHHNet,设计了多模态同异质信息分离与融合模块,用于提取可见光与热红外两个模态的共有特征与特有特征并进行融合,使融合结果变得更加合理有效. 此外,为应对不同的挑战,设计了不同的结构对不同挑战做适应性处理,减轻了用大量数据集进行模型训练的压力;增添注意力增强模块和多尺度辅助模块,分别用于降低特征提取时的噪声传播和拓宽感受野,强化小尺寸目标识别能力. 在3个常用数据集GTOT、RGBT234及LasHeR上进行了大量实验,验证了本文方法的有效性及鲁棒性. 然而,尽管相比基于相同基线模型的工作,本文模型具备一定优势,但受到基线网络MDNet+RGBT的制约,本文的方法目前尚未实现实时跟踪. 在未来,会继续探索更完善的目标跟踪网络,着力提升复杂网络RGBT目标跟踪的实时性,充分发挥多模态目标跟踪的优势.

参考文献

[1] 张天路, 张强. 基于深度学习的RGB-T目标跟踪技术综述[J]. 模式识别与人工智能, 2023, 36(4): 327-353.

ZHANG T L, ZHANG Q. A survey of RGB-T object tracking technologies based on deep learning[J]. Pattern Recognition and Artificial Intelligence, 2023, 36(4): 327-353. (in Chinese)

[2] WANG Q R, YUAN C, LIN Z H. Learning attentional recurrent neural network for visual tracking[C]//Proceedings of 2017 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE, 2017: 1237-1242.

[3] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of Advances in Neural Information Processing Systems (NeurIPS). Long Beach, 2017: 5998-6008.

[4] CUI Y T, JIANG C, WANG L M, et al. MixFormer: End-to-end tracking with iterative mixed attention[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 13598-13608.

[5] LIN L T, FAN H, ZHANG Z P, et al. Swintrack: A simple and strong baseline for transformer tracking[C]//Proceedings of Advances in Neural Information Processing Systems (NeurIPS). New Orleans, 2022: 16743-16754.

[6] TU Z Z, XIA T, LI C L, et al. RGB-T image saliency detection via collaborative graph learning[J]. IEEE Transactions on Multimedia, 2020, 22(1): 160-173.

[7] XU D, OUYANG W L, RICCI E, et al. Learning cross-modal deep representations for robust pedestrian detection[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 4236-4244.

[8] SUN Y X, ZUO W X, LIU M. RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes[J]. IEEE Robotics and Automation Letters, 2019, 4(3): 2576-2583.

[9] LI C L, ZHU C L, ZHENG S F, et al. Two-stage modality-graphs regularized manifold ranking for RGB-T tracking[J]. Signal Processing: Image Communication, 2018, 68: 207-217.

[10] LAN X Y, YE M, ZHANG S P, et al. Robust collaborative discriminative learning for RGB-infrared tracking[C]// Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. New York: ACM, 2018: 7008-7015.

[11] LAN X Y, YE M, SHAO R, et al. Learning modality-consistency feature templates: A robust RGB-infrared tracking system[J]. IEEE Transactions on Industrial Electronics, 2019, 66(12): 9887-9897.

[12] LI C L, LIANG X Y, LU Y J, et al. RGB-T object tracking: Benchmark and baseline[J]. Pattern Recognition, 2019, 96: 106977.

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

[14] ZHANG L C, DANELLJAN M, GONZALEZ-GARCIA A, et al. Multi-modal fusion for end-to-end RGB-T tracking[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Piscataway: IEEE, 2019: 2252-2261.

[15] KANG B, LIANG D, DING W, et al. Grayscale-thermal tracking via inverse sparse representation based collaborative encoding[J]. IEEE Transactions on Image Processing, 2019, 29: 3401-3415.

[16] ZHANG X M, ZHANG X H, DU X D, et al. Learning multi-domain convolutional network for RGB-T visual tracking[C]//Proceedings of 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). Piscataway: IEEE, 2018: 1-6.

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

[18] LI C L, LU A D, ZHENG A H, et al. Multi-adapter RGBT tracking[C]//Proceedings of IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Piscataway: IEEE, 2019: 1-9.

[19] LU A D, LI C L, YAN Y Q, et al. RGBT tracking *via* multi-adapter network with hierarchical divergence loss[J]. IEEE Transactions on Image Processing, 2021, 30: 5613-5625.

[20] GAO Y, LI C L, ZHU Y B, et al. Deep adaptive fusion network for high performance RGBT tracking[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Piscataway: IEEE, 2019: 1-9.

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

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

[23] BHAT G, DANELLJAN M, VAN GOOL L, et al. Learning discriminative model prediction for tracking[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 6181-6190.

[24] CHATFIELD K, SIMONYAN K, VEDALDI A, et al. Re-

turn of the devil in the details: Delving deep into convolutional nets[EB/OL]. (2014-05-14) [2025-03-11]. https://arxiv.org/abs/1405.3531v4.

[25] LI C L, LIU L, LU A D, et al. Challenge-Aware RGBT Tracking[M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 222-237.

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

[27] LI C L, CHENG H, HU S Y, et al. Learning collaborative sparse representation for grayscale-thermal tracking[J]. IEEE Transactions on Image Processing, 2016, 25(12): 5743-5756.

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

[29] LI C L, WU X H, ZHAO N, et al. Fusing two-stream convolutional neural networks for RGB-T object tracking[J]. Neurocomputing, 2018, 281: 78-85.

[30] LI C L, ZHU C L, HUANG Y, et al. Cross-Modal Ranking with Soft Consistency and Noisy Labels for Robust RGB-T Tracking[M]//Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 831-847.

[31] 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.

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

[33] XIA W D, ZHOU D M, CAO J D, et al. CIRNet: An improved RGBT tracking via cross-modality interaction and re-identification[J]. Neurocomputing, 2022, 493: 327-339.

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

[35] ZHANG X C, YE P, PENG S Y, 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.

[36] ZHANG T L, LIU X R, 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, 2022, 32(3): 1403-1417.

[37] ZHAO L, ZHU M, REN H E, et al. Channel exchanging for RGB-T tracking[J]. Sensors, 2021, 21(17): 5800.

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

[39] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional Block Attention Module[M]//Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.

[40] XU Q, MEI Y M, LIU J P, et al. Multimodal cross-layer bilinear pooling for RGBT tracking[J]. IEEE Transactions on Multimedia, 2021, 24: 567-580.

[41] MEI J T, LIU Y Y, WANG C C, et al. Asymmetric global-local mutual integration network for RGBT tracking[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 5017417.

[42] MEI J T, ZHOU D M, CAO J D, et al. Differential reinforcement and global collaboration network for RGBT tracking[J]. IEEE Sensors Journal, 2023, 23(7): 7301-7311.

[43] LI Y D, LAI H C, WANG L J, et al. Multibranch adaptive fusion network for RGBT tracking[J]. IEEE Sensors Journal, 2022, 22(7): 7084-7093.

[44] LI X, WANG W H, HU X L, et al. Selective kernel networks[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 510-519.

[45] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 2818-2826.

[46] LI C L, ZHAO N, LU Y J, et al. Weighted sparse representation regularized graph learning for RGB-T object tracking[C]// Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM, 2017: 1856-1864.

[47] ZHANG P Y, ZHAO J, BO C J, et al. Jointly modeling motion and appearance cues for robust RGB-T tracking[J]. IEEE Transactions on Image Processing, 2021, 30: 3335-3347.

[48] ZHANG P Y, WANG D, LU H C, et al. Learning adaptive attribute-driven representation for real-time RGB-T tracking[J]. International Journal of Computer Vision, 2021, 129(9): 2714-2729.

[49] PENG J C, ZHAO H T, HU Z W. Dynamic fusion network for RGBT tracking[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(4): 3822-3832.

[50] WANG X, SHU X J, ZHANG S L, et al. MFGNet: Dynamic modality-aware filter generation for RGB-T tracking[J]. IEEE Transactions on Multimedia, 2022, 25: 4335-4348.

[51] ZHANG P Y, ZHAO J, WANG D, et al. Visible-thermal UAV tracking: A large-scale benchmark and new baseline[C]// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 8876-8885.

[52] CAI Y J, SUI X B, GU G H, et al. Learning modality fea-

ture fusion via transformer for RGBT-tracking[J]. Infrared Physics & Technology, 2023, 133: 104819.
[53] XIAO X B, XIONG X Z, MENG F Q, et al. Multi-scale feature interactive fusion network for RGBT tracking[J]. Sensors, 2023, 23(7): 3410.
[54] YANG J R, DONG E Z, TONG J G, et al. Differential enhancement and commonality fusion for RGBT tracking[C]//Proceedings of 2023 IEEE International Conference on Mechatronics and Automation (ICMA). Piscataway: IEEE, 2023: 351-356.
[55] XUE Y J, ZHANG J W, LIN Z J, et al. SiamCAF: Complementary attention fusion-based Siamese network for RGBT tracking[J]. Remote Sensing, 2023, 15(13): 3252.
[56] CAI Y J, SUI X B, GU G H. Multi-modal multi-task feature fusion for RGBT tracking[J]. Information Fusion, 2023, 97: 101816.
[57] MEI J T, ZHOU D M, CAO J D, et al. HDINet: Hierarchical dual-sensor interaction network for RGBT tracking[J]. IEEE Sensors Journal, 2021, 21(15): 16915-16926.
[58] ZHU Y B, LI C L, TANG J, et al. RGBT tracking by trident fusion network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(2): 579-592.
[59] TÜRKOĞLU A, AKAGUNDUZ E. EANet: Enhanced attribute-based RGBT tracker network[C]//Proceedings of Sixteenth International Conference on Machine Vision (ICMV 2023). Armenia: SPIE, 2024: 363-370.
[60] MEI J T, ZHOU J X, WANG J, et al. Learning multifrequency integration network for RGBT tracking[J]. IEEE Sensors Journal, 2024, 24(9): 15517-15530.
[61] LIU L, LI C L, XIAO Y, et al. RGBT tracking *via* challenge-based appearance disentanglement and interaction[J]. IEEE Transactions on Image Processing, 2024, 33: 1753-1767.
[62] ZHU J W, LAI S M, CHEN X, et al. Visual prompt multi-modal tracking[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 9516-9526.
[63] HUI T R, XUN Z Z, PENG F G, et al. Bridging search region interaction with template for RGB-T tracking[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 13630-13639.
[64] CAO B, GUO J L, ZHU P F, et al. Bi-directional adapter for multimodal tracking[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(2): 927-935.
[65] WANG C Q, XU C Y, CUI Z, et al. Cross-modal pattern-propagation for RGB-T tracking[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 7064-7073.

作者简介

方　鑫　男,2000年出生于陕西省商洛市.现为长安大学信息工程学院硕士研究生.主要研究方向为视觉目标跟踪.
E-mail: fangxin@chd.edu.cn

陈　柘　男,1969年出生于陕西省西安市.博士,现为长安大学信息工程学院副教授、硕士生导师.主要研究方向为计算机视觉与机器学习.
E-mail: zchen@chd.edu.cn

刘占文　女,1983年出生于河南省安阳市.博士,现为长安大学信息工程学院教授、博士生导师.主要研究方向为车路云一体化交通环境感知与在环测试.
E-mail: zwliu@chd.edu.cn

李小鹏　男,2000年出生于福建省福州市.现为长安大学信息工程学院硕士研究生.主要研究方向为图像超分辨率.
E-mail: xiaopengli@chd.edu.cn

宿雨心　女,1999年出生于内蒙古自治区呼和浩特市.现为长安大学信息工程学院硕士研究生.主要研究方向为图像复原.
E-mail: yuxinsu@chd.edu.cn