

文章编号 1004-924X(2025)02-0324-13

针对 RGBT 跟踪的特殊属性的跨模态交互融合网络

邵小强, 李浩*, 吕植越, 马博, 刘明乾, 韩泽辉
(西安科技大学 电气与控制工程学院, 陕西 西安 710054)

摘要: 因为不易受光照、遮挡影响以及极强的鲁棒性, RGBT 目标跟踪在视频监控、自动驾驶等领域得到了广泛的应用。利用在红外图像和可见光图像中常见的挑战属性, 将两种模态信息充分交互, 构建了一个有效的跟踪网络, 能够克服跟踪过程中一些常见的不利场景带来的影响。该网络由特殊属性融合模块、共同属性融合模块和跨模态交互模块三部分组成, 特殊属性融合模块让网络能够提取两种模态所特有的挑战属性信息, 充分利用不同模态信息的优势; 共同属性融合模块对目标跟踪中两种模态都匹配的属性提取特征并自适应聚合信息, 给每个共同挑战属性分配对应的权重, 增强跟踪器的自适应性; 模态交互模块将红外图像和可见光图像进行模态交互, 在两个特定模态信息中融合共同模态信息, 让网络具有更强的鲁棒性; 针对不同模态的信息损失异源问题, 优化了损失函数, 增强了对每个模态的关注, 加快了网络的收敛速度。该网络分别在 GTOT, RGBT234, LasHer 数据集进行了测试, 在 RGBT234 数据集上达到了 84.1% 的精确率和 57.3% 的成功率, 在 Lasher 数据集上达到了 52.3% 的精确率和 39.1% 的成功率。结果表明跟踪器达到了较好的性能, 证明了该方法的有效性。

关键词: 目标跟踪; 热红外跟踪; 挑战属性; 模态交互

中图分类号: TP394.1; TH691.9 **文献标识码:** A

10. 37188/OPE. 20253302. 0324 CSTR: 32169. 14. OPE. 20253302. 0324

Special attribute-based cross-modal interactive fusion network for RGBT tracking

SHAO Xiaoqiang, LI Hao*, LÜ Zhiyue, MA Bo, LIU Mingqian, HAN Zehui

(College of Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710054, China)

* Corresponding author, E-mail: 2670815399@qq.com

Abstract: RGBT target tracking has gained widespread application in fields such as video surveillance and autonomous driving due to its robustness and resistance to illumination and occlusion. By leveraging the common challenging attributes in infrared and visible light images and fully interacting between the two modalities, an effective tracking network was constructed, capable of overcoming the impacts of various adverse scenarios encountered during the tracking process. This network was composed of three modules: the specific attribute fusion module, the common attribute fusion module, and the cross-modality interaction module. The specific attribute fusion module enabled the network to extract modality-specific chal-

收稿日期: 2024-07-09; 修订日期: 2024-09-09.

基金项目: 国家自然科学基金 (No. 52174198)

lenging attributes, effectively utilizing the advantages of different modalities. The common attribute fusion module extracted features that were matched in both modalities during target tracking and adaptively aggregated this information. It assigned corresponding weights to each common challenging attribute, thereby enhancing the tracker's adaptability. The cross-modality interaction module incorporated common modality information into the specific modality information of infrared and visible light images, thus improving the network's robustness. To address the issue of information loss across different modalities, the traditional cross-entropy loss was optimized to enhance focus on each modality and accelerate network convergence. The proposed network is tested on the GTOT, RGBT234, and LasHeR datasets, achieving an accuracy of 84.1% and a precision of 57.3% on the RGBT234 dataset, 52.3% and a precision of 39.1% on the Lasher dataset. The results demonstrate that the tracker has achieved commendable performance, which validates the effectiveness of the proposed method.

Key words: object tracking; thermal infrared tracking; challenge attributes; modal interaction

1 引言

目标跟踪是根据目标第一帧的位置进行初始定位来预测后续帧目标的位置,在智能机器人、自动驾驶^[1]等领域中发挥了极大的作用。随着深度学习的发展,人们根据神经网络设计了很多的跟踪算法,但现在大多数目标追踪算法都是基于可见光数据集进行跟踪,当在夜晚、浓雾、遮挡等恶劣环境下,针对可见光的跟踪器很难取得较好的效果;而红外目标跟踪器对纹理、细节信息的挖掘不明显,容易丢失关键的细节信息,由此看来,单模态跟踪具有很多局限性。而可见光结合红外光(RGB and Thermal, RGBT)跟踪同时具有细节纹理信息明显、强穿透性等两种模态的优点,具有极强的鲁棒性,因此,将可见光与红外光进行融合跟踪具有很好的应用前景。

一般将多模态融合分为像素级融合、特征级融合和决策级融合。传统的图像融合方法主要采用的是像素级融合^[2],但是当融合低分辨率的图像时,会遇到巨大的困难,因为需要严格配准图像;决策级融合^[3]具有较快的速度,但是由于舍弃了大量的细节特征信息,导致最后跟踪的精度一般较差;而特征级融合^[4]将从图像中提取出的特征信息进行融合,不需要将两种模态信息进行严格配准,同时也能保证一定的效率,特征级融合还可以将红外光与可见光两种模态的信息进行特征互补,极大地提高了跟踪结果的精度,因此,特征级的融合算法在国内外已经得到了广泛的应用。

在最近几年,陆续出现了一些RGBT目标追踪算法,但还是有很多需要解决的问题。针对目标跟踪领域中一些常见的难点和场景,比如尺度变化(Size Change-SC)、目标遮挡(Occlusion-OCC)、快速移动(Fast Move-FM)等,Zhang^[5]提出了利用4个特殊的挑战属性提高跟踪的鲁棒性,Xiao^[6]在此基础上引入了聚合模块,增强了各个模态信息的表征能力;针对两种模态的交互问题,Li^[7]提出了残差连接让模型同时考虑局部和全局信息,Zhang^[8]分开处理搜索图像和模板图像。但是,这些方法都没有尽可能地利用两种模态的信息进行交互,所以文中提出了一种新的模态信息交互框架,并且同时考虑了特征的位置信息和空间信息,这样可以减少网络对某种特定模态信息的依赖性,让跟踪器在不同的场景下都能保证较高的精度。

网络以MDNet^[9]作为基本框架,将跟踪过程中的困难场景称作挑战属性,针对这些挑战属性的数据进行单独训练,跟踪过程中常见的挑战属性分别是遮挡(OCC)、光照变化(Illumination variation-ILL)、尺度变化(SC)、热红外交叉(Thermal Cross-TC)、快速移动(FM)。

该文还提出了一种新的损失函数,在减小预测结果与真实框误差的同时,尽可能地让网络考虑两种模态的交互信息,以此来补足各单模态信息的不足,这样相对单模态可以取得更好的跟踪效果。

该文主要贡献有:设计了一个新的多模态属性融合网络,可以有效地利用红外光和可见光的

模态信息,增强网络跟踪的鲁棒性。提出了一种模态独有特征和共有特征分开处理的模块,利用 ILL 和 TC 分别增强可见光和红外图像的独有特征表达能力,再与两个模态的共有特征属性融合,让每个模态的独有特征能够共享。提出了一个模态交互模块,让两种模态能充分利用对方的信息;提出了一种新的损失函数,增强目标的特殊模态辨别能力。所提出的网络分别在 GTOT, RGBT234, Lasher 数据集上进行了测试,验证了网络的有效性。

2 相关工作

随着深度学习的发展,视觉跟踪已经是计算机视觉中一项基本且重要的任务之一,在最近几年也取得了巨大的突破。该节将简要介绍目标跟踪、RGBT 目标跟踪和属性驱动跟踪。

2.1 目标跟踪

目标跟踪因为其巨大的应用前景,吸引了众多学者的研究。Nam^[9]提出了一种基于判别式的卷积神经网络的视觉跟踪方法,通过判断生成的候选框评分来预测目标下一时刻的位置。Liu^[10]提出了一种基于相关滤波器的多层卷积特征集成跟踪器用于热红外跟踪,取得了良好的效果。Li^[11]采用孪生网络框架,利用结构和语义两个方面的相似度来提高网络的判别能力。Li^[12]提出了一种包含红外特殊判别特征和细粒度相关特征的双特征模型用于红外目标跟踪,具有较好的鲁棒性。Transformer 最开始在自然语言处理领域中提出,因其可以通过独特的注意力机制捕获到全局信息,很好地解决了全局与局部信息交互的问题,一些研究人员将其应用到计算机视觉领域^[13]并证明了其有效性。相关滤波算法^[14]在目标追踪中也得到了广泛使用,后续也有很多研究人员致力于由粗到细预测目标下一帧的位置以此来实现追踪的目的。

该文中,将训练过程采用了三阶段的卷积网络结构进行特征的提取,在每个阶段进行对应模块参数的学习,最大程度地利用了训练数据。在测试阶段,只对最后的全连接层进行在线更新,以此增加跟踪的准确性。

2.2 RGBT 目标跟踪

由于其利用两种模态的信息增强了特征的

表达能力,并且对环境有极强的适应性,RGBT 目标跟踪得到了越来越多的关注。Li^[7]通过编码器提取特征和解码器重建图像设计了一种 RGBT 图像融合方法;Zhang^[8]利用判别模型预测分别考虑了像素级、特征级和决策级融合;Zhou^[15]针对可见光图像单独提出了一个光照感知特征对齐模块,增强了可见光的特殊感知能力;Peng^[16]用双流网络做基准,单独考虑了独有特征和共同特征的不同动态贡献,对不同模态的图像自适应加权求和;Zhu^[17]提出了一种三叉戟结构,即分别对两种模态的特征和融合模态的特征进行特征信息的挖掘,在最后将所有信息进行聚合;Zhang^[18]提出了一种特征重选模块从单模态和融合模态特征中自适应选择判别特征用于回归和分类;Zhu^[19]通过学习相关模态的提示调整冻结的预训练基础模型,以此适应多模态任务;Li^[20]提出了一个动态记忆 transformer 追踪网络,结合了全局信息增强网络性能。

该文将网络扩展成双流网络,一部分负责处理可见光信息,另一部分负责处理红外光信息,在最后阶段将两种模态信息进行融合。

2.3 属性驱动跟踪

与其他的计算机视觉任务相比,跟踪更容易遇到快速移动、目标遮挡等问题,会对最后的跟踪结果产生影响,有一些属性场景是红外图像和可见光图像共有的,有些则是模态的独有特征,比如 ILL 和 TC,因此,单独考虑这些特殊属性可以增强跟踪器在这些极端环境下的性能,提高网络的自适应性。

Li^[21]利用属性的模态特定场景信息指导另一种模态学习挑战属性,增强在另一模态的困难场景的判别能力;Zhang^[5]给每个挑战属性设计了一个残差分支来构建更好的特征表示,最后自适应的聚合这些特征表示。因此该文设计了一个属性聚合网络,对模态的共同属性和特有属性单独进行处理,同时考虑空间与通道维度特征,在训练阶段产生对应挑战属性权重,以此来增加网络的表征能力。

3 网络方法

3.1 网络整体结构

该网络的整体框架如图 1 所示,“+”代表矩

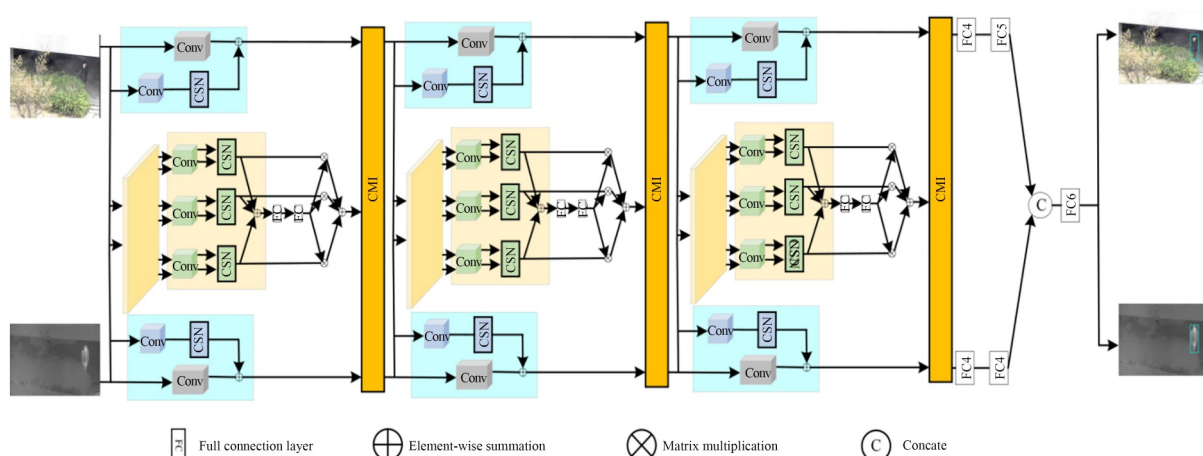


图1 整体网络框架图

Fig. 1 Proposed network framework

阵相加,“ $\times$ ”代表矩阵相乘,蓝色的CSN代表ILT和TC两种特别挑战属性融合模块,绿色代表OCC,SC,FM三种共同挑战属性的融合,CMI表示跨模态交互模块。该网络整体框架是由CSN(Channel Space Fusion),CMI(Cross-Modal Interaction)和输出头分类模块组成(彩图见期刊电子版)。具体来说,前向特征提取模块是由VGG-M^[22]的前三层网络作为基础网络,将其扩展为双流网络以处理两种模态的信息,整个网络分三个阶段训练,每一个阶段分开训练,共同特征融合模块负责融合FM,SC,OCC这三个两种模态都共有的信息,并自适应地分配两个模态在其中的权重占比。在之后通过扩展的SKNet^[23]将三个挑战属性的特征信息自适应的聚合成为共同模态信息,作为两种模态的共同特征。将ILT和TC这两种特定模态才具有的特征分别加入可见光和热红外的主干网络中,以此来增加两种模态的独有模态信息。经过主干网络后,使用两个全连接层分别提取两种模态的信息,最后通过FC6层将两种模态信息加权得出最后的分数。

与APFNet网络不同的是,分开对两种模态信息所针对的特殊场景单独进行了特征提取,当面对特殊场景时,跟踪器可以更多依赖更有优势的模态信息,同时,采用的跨模态交互模块相比encoder和decoder极大地减少了参数量,使跟踪器在效率上也有一定的提升。

3.2 属性融合模块

该文的通道空间融合模块如图2和图3所示。

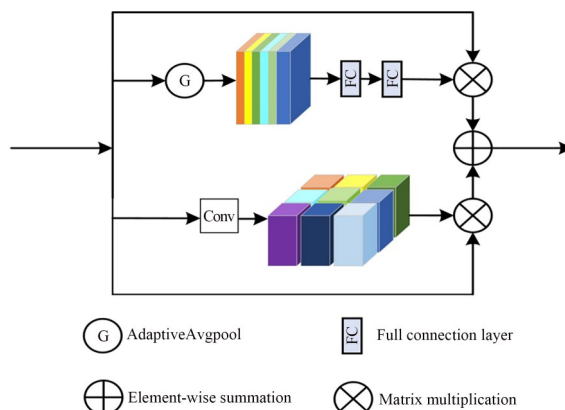


图2 特殊挑战属性融合模块

Fig. 2 Special Challenge Attribute Fusion Module

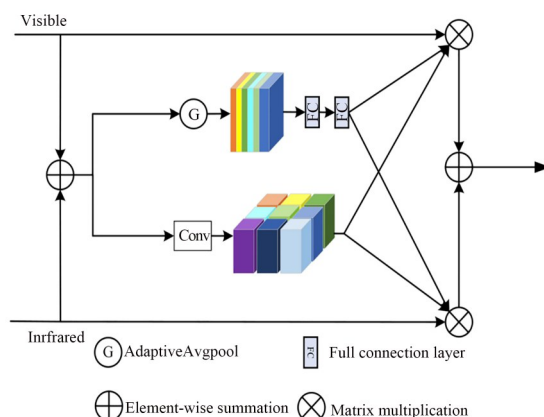


图3 共同挑战属性融合模块

Fig. 3 Common Challenge Attribute Fusion Module

对于ILT和TC两个挑战属性,只对其输入相应特定模态的信息。这里以ILT为例,为了同

时关注空间与通道上的信息^[23-24],先经过全局平均池化,在减少计算量的同时尽可能地保留特征,将输入特征分为两部分,一部分提取通道上的注意力,即将特征图信息进行压缩,通过线性转换得到通道维度的权重,另一部分利用 1×1 的卷积压缩通道维度提取空间上的注意力权重,再分别将权值与原始可见光信息相乘得到对应的特征信息,最后再将两部分特征相加以此增加跟踪器对小的正样本的关注度。具体的计算公式如式(1)所示:

$$Output = w_c * V_{in} + w_s * V_{in}, \quad (1)$$

其中: w_c 表示计算出的通道注意力权重, w_s 表示空间注意力权重, V_{in} 是原始可见光特征,如果是TC为挑战属性,则输入为红外光信息。

对于共同的挑战属性,同时输入可见光和红外光信息,将两者信息直接相加分别提取两个模态对应属性的通道和空间注意力,权重相乘后输出融合的特征信息,分配注意力的同时给予两个模态不同的权重,以应对不同的实际场景,通用挑战属性的计算公式为:

$$Output = w_{c1} * V_{FM} + w_{s1} * V_{FM} + w_{c2} * I_{FM} + w_{s2} * I_{FM}, \quad (2)$$

其中: V_{FM} 表示输入的可见光, I_{FM} 表示输入的红外光, w_{c1} 和 w_{s1} 分别表示计算出的可见光的通道和空间权重, w_{c2} 和 w_{s2} 则分别对应热红外的权重。

通过CSN模块,对跟踪过程中常见的五种困难场景数据着重进行了关注,同时进行了通道与空间维度的权重计算,极大地提高了网络在困难场景下的跟踪精度。

3.3 跨模态交互模块

考虑到两个模态的信息具有一定的差异性,直接将提取出的特征信息进行相加会忽略一些细节信息,所以设计了一个跨模态交互模块(CMI)。跨模态交互模块如图4所示,输入分别为可见光特征,融合特征,红外光特征,输出分别为增强后的可见光和红外光特征。

跨模态交互模块的输入为红外特征、可见光特征和融合特征,利用注意力机制^[13],借用融合特征的信息指导两种模态的特征生成,也将红外模态信息和可见光模态信息进行了模态交互,让两种模态得到更多的互补信息。以可见光为例,用可见光信息产生查询值 Q ,将其与融合信息进

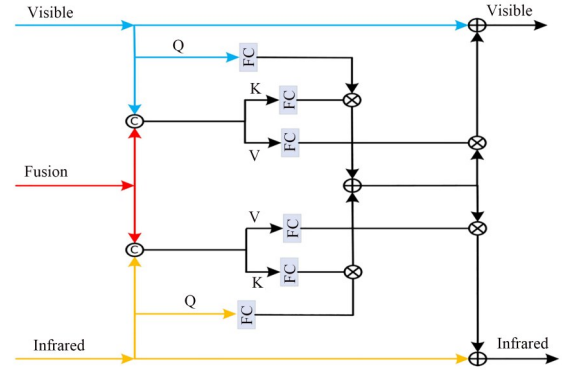


图4 跨模态交互模块的整体结构

Fig.4 Overall structure of the cross-modal interaction module

行拼接得到更全面的模态信息,经过全连接产生 K 和 V ,接着将查询值与键值相乘得到模态信息的权重值,通过权重计算模态信息期望值,最后与原始信息相加得到互补特征。这样,在保留原始模态信息的同时,充分利用了融合特征的信息。输出权重计算公式为:

$$Output_v = V + (V_q * V_k + I_q * I_k) * V_v, \quad (3)$$

$$\begin{cases} I_v = FC1(Cat(V, F)) \\ I_k = FC2(Cat(I, F)) \\ V_v = FC3(Cat(V, F)) \end{cases}, \quad (4)$$

其中: $Output_v$ 表示交互后的可见光信息, V 表示输入可见光, V_q 表示可见光的查询, V_k 表示可见光的键, I_q 表示红外光的查询, I_k 表示红外光的键, V_v 表示可见光的值, Cat 表示拼接操作, FC 表示线性转换。

与一般的注意力机制不同的是,该文的跨模态交互模块极大地利用了共同特征,让单个模态在困难场景下利用另一个模态的信息,两种模态的信息充分交互,以此解决单个模态的弊端。

3.4 输出头分类模块

在最后一层主干网络中,将模态交互后的红外模态、可见模态的特征信息分别进行全链接特征提取,最后再拼接到一起进行处理。这样可以提取到更多的细节信息,增加了两个模态的特征转换,增强了网络的鲁棒性。同时,对传统的交叉熵损失进行了优化,MDNet^[9]的损失函数是一个简单的二分类损失,只考虑了最后融合分数的损失,该网络加入了另外两个模态特征的信息,用特定信息约束可见光与红外光的损失,使函数

考虑更多的模态独有信息。

$$L = \frac{L_{pos} + L_{neg}}{\text{num}(pos_{\text{example}} + neg_{\text{example}})}, \quad (5)$$

$$L_{pos} = k11 * L_{pos_v} + k12 * L_{pos_i} + k13 * L_{pos_j}, \quad (6)$$

$$\begin{cases} k11 = 1 - |0.5 - \text{softmax}(pos_{score_i})| \\ k12 = 1 - |0.5 - \text{softmax}(pos_{score_v})| \\ k13 = 1 + |\text{softmax}(pos_{score_v}) - \text{softmax}(pos_{score_i})| \end{cases}, \quad (7)$$

其中: L_{pos} 表示正样本的损失, L_{neg} 表示负样本损失, pos_{score_v} 和 pos_{score_i} 分别表示网络输出的可见光与红外光的正样本得分。

这里添加的 $k11$ 和 $k12$ 权重减少了网络预测的不确定性,让网络更趋向预测为正样本或者负样本, $k13$ 则是让两种模态最后的输出更加接近,更好地利用两个模态的信息,使跟踪器具有更强的鲁棒性。从图5可以看出,与传统的交叉熵损失相比,该文所提出的损失函数收敛速度更快,且最后的收敛效果更好,可以使网络预测达到更精准的结果。

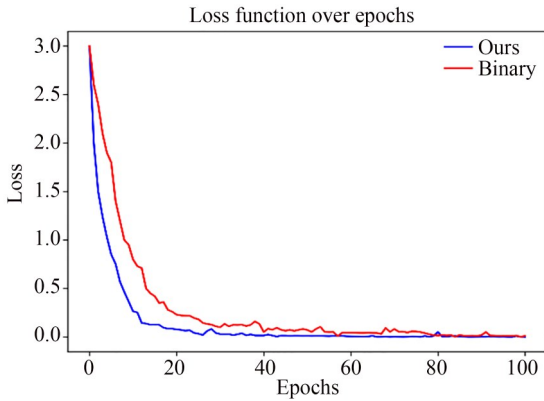


图5 优化的损失与二分类损失的比较结果

Fig. 5 Result of comparing our loss with binary loss

3.5 训练跟踪细节

由于在实际情况中无法预料会遇到哪些挑战属性,所以该文沿用APFNet^[6]的思想,将部分网络参数在训练后直接冻结,整个网络分三个阶段训练,在预训练阶段,利用imagenet^[25]上的权重只训练主干网络,去掉其他所有分支,其他挑战属性模块以及跨模态交互模块的权重采用从均值为0方差为1的正态分布中随机采样。在第一训练阶段,分别输入对应挑战属性的数据,一共5

个挑战属性;在第二训练阶段,冻结属性融合参数,学习属性聚合参数;在第三阶段,学习跨模态交互参数的同时微调其他参数;在测试阶段,冻结前面所有参数,每一个视频输出一个FC6层预测不同的视频序列。

该文跟踪器在3个阶段的特征图输出大小分别为 25×25 , 5×5 , 3×3 ,图6展示了第一阶段的红外光和可见光经过特殊挑战属性模块和共同挑战属性模块得到的特征图,ILL是可见光对应的特殊挑战属性模块,TC是红外光对应的挑战属性融合模块,Fusion是经过共同挑战属性融合模块得到的特征图。

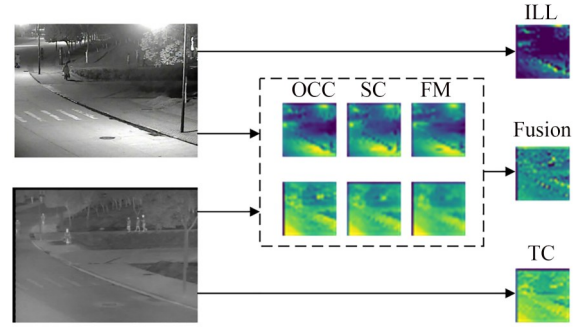


图6 第一阶段的特征提取结果

Fig. 6 Feature extraction results from the first stage

从图6可以看出,经过共同挑战属性提取后的特征图同时具有两种模态信息的共同点,这也是跨模态交互模块输入的共同模态信息。

在该文的跟踪网络中,训练过程只对相应挑战属性的数据进行单独训练,高效地利用了有限的数,在线跟踪时,第一帧产生500个正样本,5000个负样本,在之后的跟踪过程中,每一帧产生50个正样本,200个负样本,同时进行困难负样本挖掘操作,取得分最高的5个正样本进行回归操作,以此微调跟踪框。

4 实验评估

4.1 数据集和评估指标

4.1.1 数据集和评估指标

该文将设计的跟踪器与现在主流的一些跟踪器分别在GTOT^[26],RGBT234^[27],Lasher^[28]三个数据集上进行比较,都取得了较好的性能。

GTOT, RGBT234, Lasher是由李成龙课题组提出的专门针对RGBT问题的数据集,分别包含50对、234对和1 224对视频序列及其真实框,且每一帧图片都对其属性做了标注,但是Lasher没有挑战属性方面的标注,所以该文采用GTOT和RGBT234两个数据集作为训练集,三个数据集作为测试集分别对比。

4.1.2 评估指标

RGBT领域一般使用两个主要的评估指标,精准率(PR)表示预测框和真实框之间的中心位置距离在一定阈值下的比例,成功率(SR)表示预测框与真实框之间的交并比在一定阈值内的比例。PR和SR的计算公式如式(8)和式(9)所示:

$$PR = \frac{Num(\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} < \tau)}{N}, \quad (8)$$

$$SR = \frac{Num\left(\frac{A \cap B}{A \cup B} > \tau\right)}{N}, \quad (9)$$

其中:Num表示总帧数, (x_1, y_1) 表示真实框中心位置, (x_2, y_2) 表示预测框中心位置, τ 表示给定的阈值, A和B分别为预测框和真实框的面积。

由于GTOT数据集中的目标一般比较小,所以将阈值设为5 pixels,其他数据集设为20 pixels。

4.2 实验细节

文中的训练设备CPU是在15 vCPU AMD EPYC 7542 32-Core Processor, GPU是RTX

3090(24 GB),整个训练过程分三个阶段,首先使用imagenet的权重进行初始化,在第一阶段,只保留挑战属性模块和主干网络,在第二阶段,冻结挑战属性参数,对聚合模块进行训练,在第三阶段,训练网络融合模块,同时将输出分类头进行微调。

4.3 与其他跟踪器的比较

通过表1可以看出该网络在各个数据集的精确率和成功率已经达到了较好的效果。

表1 跟踪器在GTOT, RGBT234, Lasher数据集的性能评估

Tab. 1 Performance evaluation of our tracker in GTOT, RGBT234, Lasher datasets

数据集	PR	SR
GTOT	91.7%	73.9%
RGBT234	84.1%	57.3%
Lasher	52.3%	39.1%

4.3.1 数据集和评估指标

以RGBT234数据集为训练集,将该文的跟踪器与13个主流的跟踪器在GTOT数据集上进行了比较测试:MDNet^[29]+RGBT, MANet^[30], ECO^[31], ADNet^[32], MEEM^[33], SiamDW^[34]+RGBT, DAPNet^[35], RT-MDNet^[29], DAT^[36], STRUCK^[37], SGT^[38], APFNet^[6], ADRNet^[5]。评估结果如图7所示。

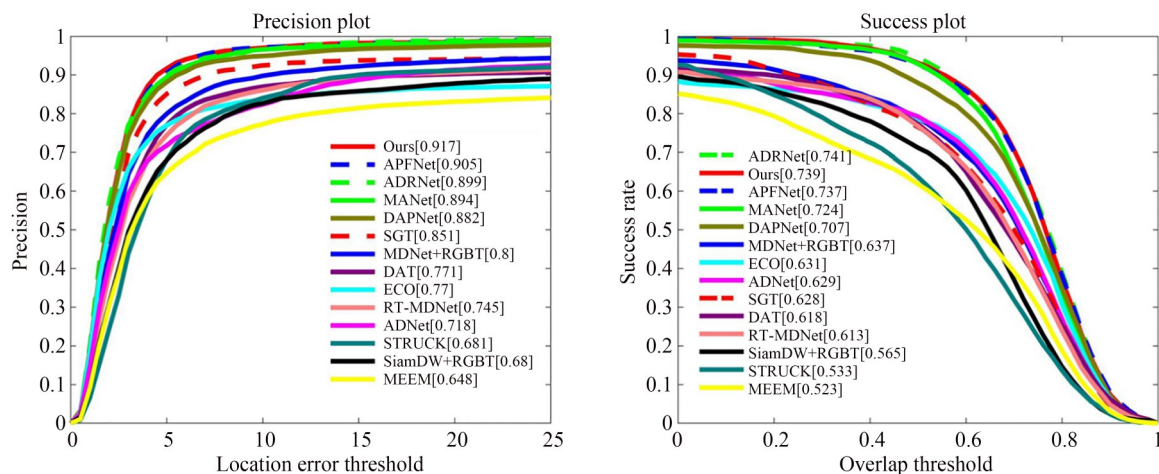


图7 在GTOT数据集上的评估结果

Fig. 7 Results of the evaluation on the GTOT dataset

从图7中可以看出,该跟踪器表现出优越的性能,精确率和成功率分别达到了91.7%和73.9%,相对APFNet网络有1.2%的提升,优于现在主流的跟踪器。

4.3.2 RGBT234数据集

图8是该文的跟踪器分别与RGB跟踪器(包括DAT^[36], RT-MDNet^[29], SOWP^[28], ECO^[31], CSR-DCF^[39], DSST^[40], CFnet^[41])和RGBT跟踪器(包括APFNet^[6], ADRNet^[5], DAPNet^[35], DAT^[36], MDNet^[29]+RGBT, SOWP^[38]+RGBT, SiamDW^[34]+RGBT, CFnet^[41]+RGBT, KCF+

RGBT^[42]) 在RGBT234数据集上的比较结果,通过与两种类型跟踪器的比较,可以更好地展示RGBT跟踪器相对单模态跟踪器的效果提升,更能说明该跟踪器的有效性。这里以GTOT数据集作为训练集。

从图8(a)可以看出,该文的跟踪器效果极大地领先现在主流的单模态RGB跟踪器,精准率和成功率达到了84.1%和57.3%,证明了RGBT跟踪器的优越性。从图8(b)可以看出,与RGBT跟踪器比较,也在精确率上领先了APFNet^[6]1.4%,达到了很好的跟踪效果。

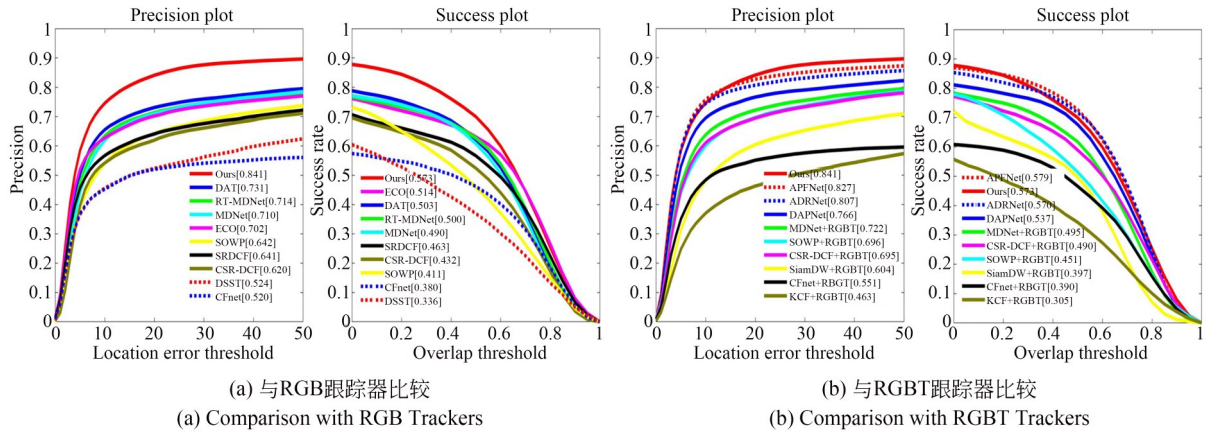


图8 在RGBT234数据集上的评估结果

Fig. 8 Evaluation results on the RGBT234 dataset

4.3.3 RGBT234数据集

为了验证该文基于挑战属性模块的有效性,与不同跟踪器比较了在RGBT234数据集中不同挑战属性下的性能:尺度变化(SC),复杂背景(Background Complex),部分遮挡(Partial Occulsion),低分辨率(Low Resolution),无遮挡(No Occulsion),变形(Deflection),低照度(Low Illumination),运动模糊(Move Blur),热红外交叉(TC),严重遮挡(Heavy Occlusion),相机运动(Camera Motion)和快速移动(FM)。

分别与MANet^[30], MDNet^[9], DAPNet^[35], CAT^[21], ADRNet^[5], APFNet^[6]这几个主流的跟踪器进行了比较,从表2(彩图见期刊电子版)中可以看出,该文的跟踪器在大部分属性上表现出较好的性能,且在总体的数据下也表现出最

好的效果,证明了该文的跟踪器在面对困难环境下也有较好的鲁棒性,说明了针对不同属性设计的挑战属性模块的有效性。

同时可以看出,基于MDNet网络设计的跟踪器FPS较低,这是因为MDNet采取的在线跟踪的方法,相对而言,此网络采用了较轻量化的模块,在此系列中取得了较快的速度。

为了检验文中选择的五个挑战属性的有效性,分别比较了各个挑战属性在GTOT和RGBT234数据集上对应属性的结果,只用对应的特定挑战属性数据进行结果的验证。

在表3和表4中展示了不同数据集上的各个挑战属性结果,COMMON表示FM,OCC,SC三种属性数据,设计的特定模块在跟踪相应数据时都能取得较好的结果,表明了设计的挑战属性模块是有效的。

表 2 所提出的跟踪器在 RGBT234 数据集不同属性下的 PR/SR

Tab. 2 PR/SR of the proposed tracker for different attributes of the RGBT234 dataset

	Num	MANet	MDNet	DAPNet	CAT	ADNet	APFNet	Our
NO	41	91.4/64.9	81.2/59.0	90.4/64.4	93.2/66.8	89.8/64.8	93.4/66.4	93.7/66.3
PO	96	83.1/58.4	74.7/50.9	82.1/57.4	85.1/59.3	82.7/58.2	85.0/58.7	88.8/63.4
HO	96	66.5/45.9	63.3/43.2	66.0/45.7	70.0/48.0	70.9/49.4	72.9/49.0	72.5/48.2
LI	63	81.7/56.0	58.9/39.6	77.5/53.0	81.0/54.7	79.8/54.5	82.3/54.4	87.3/60.6
LR	50	78.5/51.3	66.0/44.5	75.0/51.0	82.0/53.9	80.2/54.2	82.9/54.8	86.7/60.8
TC	28	72.9/52.9	74.8/53.0	76.8/54.3	80.3/57.7	79.7/57.8	82.1/57.3	82.9/57.5
DEF	76	72.3/52.4	66.4/46.8	71.7/51.8	76.2/54.1	72.1/51.2	77.1/54.6	76.5/53.6
FM	32	71.1/45.5	63.2/39.3	67.0/44.3	73.1/47.0	73.9/48.8	78.2/49.2	78.6/49.3
SC	120	77.8/54.8	73.9/51.9	78.0/54.2	79.7/56.6	78.5/56.1	82.1/56.5	82.4/56.7
MB	55	66.8/48.2	62.4/44.2	65.3/46.7	68.3/49.0	70.6/50.6	72.8/53.0	71.5/52.6
CM	89	69.9/50.5	61.3/43.3	66.8/47.4	75.2/52.7	73.9/52.3	76.3/54.5	83.3/59.0
BC	54	75.6/49.8	62.5/41.8	71.7/48.4	81.1/51.9	74.3/50.2	80.6/52.4	80.4/51.3
FPS	—	1.1	1	2	20	25	1.9	3.5
ALL	234	77.8/54.4	71.0/49.0	76.6/53.7	80.4/56.1	79.2/55.8	82.7/57.9	84.1/57.3

表 3 GTOT 数据集挑战属性数据结果

Tab. 3 Results of challenge attribute in GTOT dataset

数据集	ILL-Branch	TC-Branch	CONNON-Branch
ILL	43.4%	37.3%	39.7%
TC	88.5%	89.2%	87.2%
COMMON	72.8%	72.4%	74.9%

表 4 RGBT234 数据集挑战属性数据结果

Tab. 4 Results of challenge attribute in RGBT234 dataset

数据集	ILL-Branch	TC-Branch	CONNON-Branch
ILL	65.3%	57.3%	59.4%
TC	68.9%	75.3%	67.6%
COMMON	62.3%	68.4%	70.3%

4.3.4 Lasher 数据集

Lasher 数据集因其庞大的种类,并且很多视频序列都是由监控拍摄,相对而言跟踪难度更大,能很好地检验跟踪器的性能,因此现在成为

主流的 RGBT 方法验证集。

图 9 是该跟踪器与主流的一些算法的精确率与成功率比较结果,从图中可以明显看出该跟踪器在 Lasher 数据集中的跟踪效果优于其他跟踪器,这是因为该跟踪器针对每个模态比较难跟踪的特殊场景数据进行了训练,因此表现出较好的性能。

图 10 中,用四个视频序列不同帧可视化了不同跟踪器的跟踪结果,同时为了更加直观,保留了前 10 帧的轨迹在跟踪框的正下方展示,每个序列里面,左边是可见光图片,右边是红外光图片,其中,蓝色表示真实框,红色表示 MANet++^[40],绿色表示 CAT,深蓝色表示 APFNet,紫色表示该文的跟器(彩图见期刊电子版)。从图 10 可以看出,因为在第一帧的目标位置是已知的,也进行了初始预训练,所以一般跟踪器在前面几帧都有比较好的追踪结果,但是随着环境的变化,会出现遮挡、光照变化等恶劣环境,这时大部分的跟踪器就不能很好地跟踪到目标了,但是该文的跟踪器由于针对这些挑战属性进行了特殊训练,依然可以较好地跟踪到指定目标。

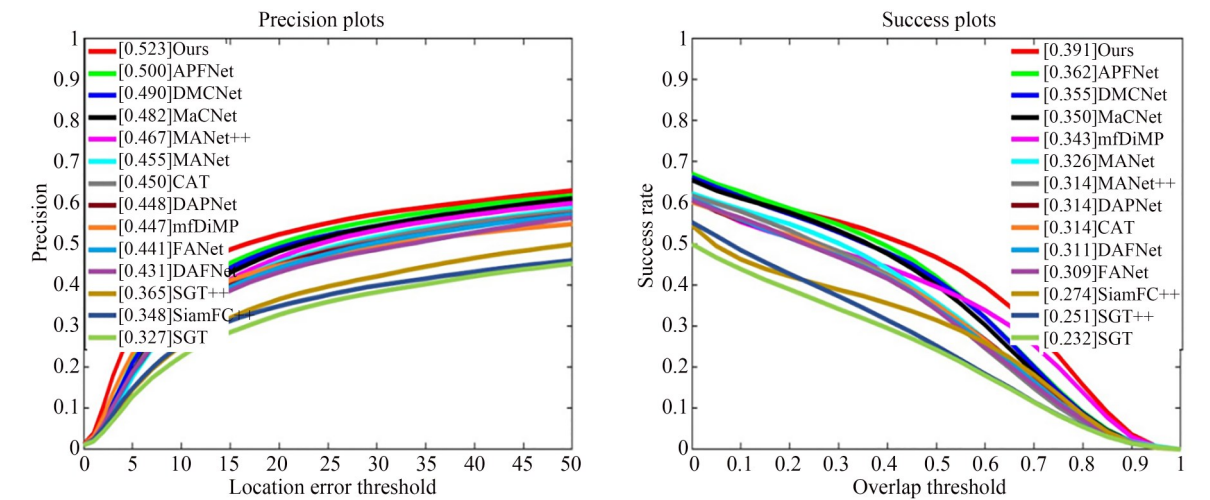


图 9 在 Lasher 数据集上的评估结果

Fig. 9 Results of the evaluation on the Lasher dataset

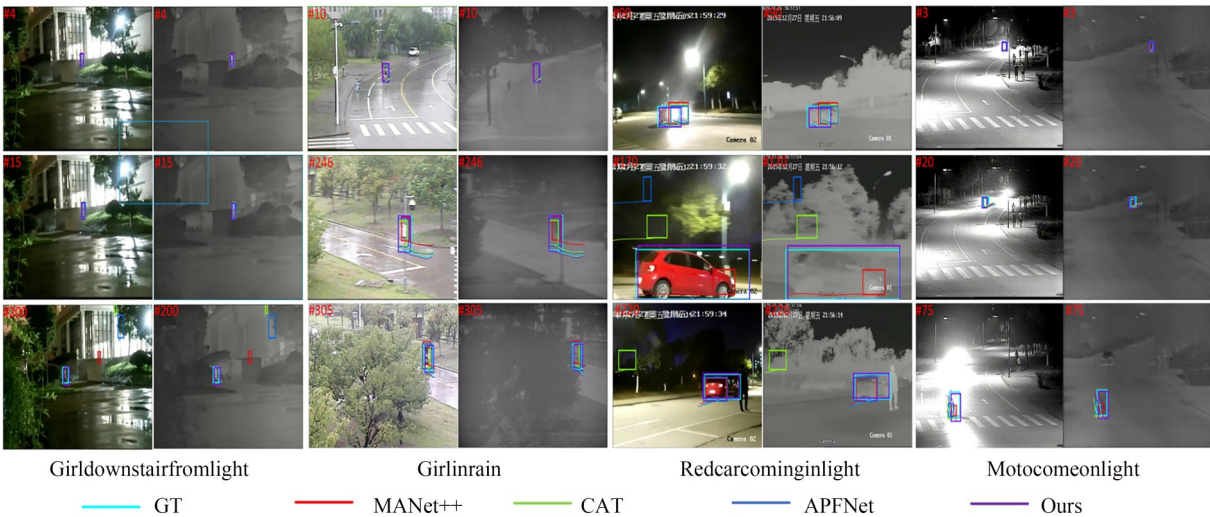


图 10 在 Lasher 数据集不同视频序列下不同帧的可视化跟踪结果

Fig. 10 Visualization of tracking results for different frames under different video sequences in the Lasher dataset

4.4 消融实验

为了验证每个模块的有效性,对网络进行了消融实验,分别验证每个模块的性能。

从表 5 中可以看出,在增加了融合模块和模态交互模块后,跟踪器的精准率和成功率均有提升,可以证明提出模块的有效性。

表 5 消融实验表					
Tab. 5 Ablation experiments					
数据集	CSN	CMI	FPS	PR	SR
GTOT		✓	4.2	90.5%	72.5%
	✓		3.8	91.1%	73.4%
	✓	✓	3.5	91.7%	73.9%
RGBT234		✓	4.2	81.3%	56.4%
	✓		3.8	82.6%	56.8%
	✓	✓	3.5	84.1%	57.3%

5 结 论

在该文中,提出了一种新的 CSN 模块,同时考虑了空间和通道上面的信息,并针对特殊属性分开处理,设计了一个有效的跨模态交互模块,增强了各模态的交互性,优化了损失函数,增加了对模态独有特征的关注,通过设计的网络,对每个挑战属性分别进行训练,减少计算量的同时

增加了跟踪器的准确率。通过消融实验证明,在 RGBT234 数据集中,增加了 CSN 模块后,准确率和成功率分别提升了 2.8% 和 0.9%,增加了 CMI 模块后,精确率和成功率分别提升了 1.5% 和 0.5%,在将两个模块结合后,也完成了一定指标的提升。在三个公开的数据集上也取得了较好的指标,特别是在 RGBT234 数据集上达到了 84.1% 和 57.3% 的准确率和成功率,证明了该文跟踪器的优越性。在未来,将会探索更有效的多模态融合方法来提高 RGBT 跟踪器的性能,并

且会进一步研究实时处理能力,提高网络的跟踪速度。

作者贡献声明:

邵小强:测量方法的提出,论文的审核;

李浩:论文构思和撰写,实验的验证;

吕植越:论文审核与数据分析;

马博:相关工作的调研;

韩泽辉:论文校对与审核;

刘明乾:测量实验数据分析。

参考文献:

- [1] LAURENSE V A, GOH J Y, GERDES J C. Path-tracking for autonomous vehicles at the limit of friction [C]. 2017 *American Control Conference (ACC)*. May 24-26, 2017, Seattle, WA, USA. IEEE, 2017: 5586-5591.
- [2] PAJARES G, MANUEL DE LA CRUZ J. A wavelet-based image fusion tutorial[J]. *Pattern Recognition*, 2004, 37(9): 1855-1872.
- [3] AL-JARRAH M A, YASEEN M A, AL-DWEIK A, *et al.* Decision fusion for IoT-based wireless sensor networks[J]. *IEEE Internet of Things Journal*, 2020, 7(2): 1313-1326.
- [4] ILHAN H O, SERBES G, AYDIN N. Decision and feature level fusion of deep features extracted from public COVID-19 data-sets[J]. *Applied Intelligence*, 2022, 52(8): 8551-8571.
- [5] 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.
- [6] 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.
- [7] LI H, WU X J. DenseFuse: a fusion approach to infrared and visible images[J]. *IEEE Transactions on Image Processing*, 2019, 28(5): 2614-2623.
- [8] ZHANG L C, DANELLJAN M, GONZALEZ-GARCIA A, *et al.* Multi-modal fusion for end-to-end Rgb-T tracking[C]. 2019 *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. October 27-28, 2019, Seoul, Korea (South). IEEE, 2019: 2252-2261.
- [9] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking[C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 27-30, 2016, Las Vegas, NV, USA. IEEE, 2016: 4293-4302.
- [10] LIU Q, LU X H, HE Z Y, *et al.* Deep convolutional neural networks for thermal infrared object tracking [J]. *Knowledge-Based Systems*, 2017, 134: 189-198.
- [11] LIU Q, LI X, HE Z Y, *et al.* Learning deep multi-level similarity for thermal infrared object tracking [J]. *IEEE Transactions on Multimedia*, 2020, 23: 2114-2126.
- [12] LIU Q, YUAN D, FAN N N, *et al.* Learning dual-level deep representation for thermal infrared tracking [J]. *IEEE Transactions on Multimedia*, 2022, 25: 1269-1281.
- [13] LIU Z, LIN Y T, CAO Y, *et al.* Swin transformer: hierarchical vision transformer using shifted windows[C]. 2021 *IEEE/CVF International Conference on Computer Vision (ICCV)*. October 10-17, 2021, Montreal, QC, Canada. IEEE, 2021: 9992-10002.
- [14] KELLY K. *The inevitable: Understanding the 12 Technological Forces That Will Shape Our Future* [M]. New York: Penguin Publishing Group, 2016.
- [15] ZHOU K L, CHEN L S, CAO X. *Improving Multispectral Pedestrian Detection by Addressing Modality Imbalance Problems*[M]. Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 787-803.
- [16] PENG J C, ZHAO H T, HU Z W. Dynamic fusion network for RGBT tracking[J]. *IEEE Trans-*

- actions on Intelligent Transportation Systems, 2023, 24(4): 3822-3832.
- [17] 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*, 2021, 32(2): 579-592.
- [18] 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.
- [19] ZHU J W, LAI S M, CHEN X, *et al.* Visual prompt multi-modal tracking [C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 17-24, 2023, Vancouver, BC, Canada. IEEE, 2023: 9516-9526.
- [20] LI M Y, ZHANG P, YAN M, *et al.* Dynamic feature-memory transformer network for RGBT tracking [J]. *IEEE Sensors Journal*, 2023, 23(17): 19692-19703.
- [21] 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.
- [22] CHATFIELD K, SIMONYAN K, VEDALDI A, *et al.* Return of the devil in the details: delving deep into convolutional nets[EB/OL]. 2014: arXiv: 1405.3531. <https://arxiv.org/abs/1405.3531v4>.
- [23] LI X, WANG W, HU X, *et al.* Selective kernel networks[C]. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019: 510-519.
- [24] WOO S, PARK J, LEE J Y, *et al.* CBAM: Convolutional block attention module [C]. *Computer Vision-ECCV 2018*. Cham: Springer International Publishing, 2018: 3-19.
- [25] DENG J, DONG W, SOCHER R, *et al.* ImageNet: a large-scale hierarchical image database [C]. 2009 *IEEE Conference on Computer Vision and Pattern Recognition*. June 20-25, 2009, Miami, FL, USA. IEEE, 2009: 248-255.
- [26] 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.
- [27] LI C L, LIANG X Y, LU Y J, *et al.* RGB-T object tracking: Benchmark and baseline[J]. *Pattern Recognition*, 2019, 96: 106977.
- [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*, 2021, 31: 392-404.
- [29] JUNG I, SON J, BAEK M, *et al.* Real-time md-net [C]. *Proceedings of the European conference on computer vision (ECCV)*. 2018: 83-98.
- [30] LI C L, LU A D, ZHENG A H, *et al.* Multi-adapter RGBT tracking[C]. 2019 *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. October 27-28, 2019. Seoul, Korea (South). IEEE, 2019: 0-0.
- [31] DANELLJAN M, BHAT G, KHAN F S, *et al.* ECO: efficient convolution operators for tracking[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 21-26, 2017, Honolulu, HI, USA. IEEE, 2017: 6931-6939.
- [32] YUN S, CHOI J, YOO Y, *et al.* Action-decision networks for visual tracking with deep reinforcement learning[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 21-26, 2017, Honolulu, HI, USA. IEEE, 2017: 1349-1358.
- [33] ZHANG J M, MA S G, SCLAROFF S. MEEM: robust tracking via multiple experts using entropy minimization [M]. *Computer Vision - ECCV 2014*. Cham: Springer International Publishing, 2014: 188-203.
- [34] ZHANG Z P, PENG H W. Deeper and wider siamese networks for real-time visual tracking [C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 15-20, 2019, Long Beach, CA, USA. IEEE, 2019: 4586-4595.
- [35] 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*. October 21 - 25, 2019, Nice, France. ACM, 2019: 465-472.
- [36] PU S, SONG Y B, MA C, *et al.* Deep attentive tracking via reciprocative learning [C]. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. December 3 - 8, 2018, Montréal, Canada. ACM, 2018: 1935-1945.
- [37] HARE S, GOLODETZ S, SAFFARI A, *et al.*

- Struck: structured output tracking with kernels[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, 38(10): 2096-2109.
- [38] 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*. October 23-27, 2017, Mountain View, California, USA. ACM, 2017: 1856-1864.
- [39] LUKEŽIC A, VOJÍR T, ZAJC L C, *et al.* Discriminative correlation filter with channel and spatial reliability[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 21-26, 2017, Honolulu, HI, USA. IEEE, 2017: 4847-4856.
- [40] DANELLJAN M, HÄGER G, KHAN F S, *et al.* Discriminative scale space tracking[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, 39(8): 1561-1575.
- [41] ZHANG G, LI Z, LI J, *et al.* Cfnet: cascade fusion network for dense prediction [EB/OL]. 2023: *arXiv*: 2302.06052. <https://doi.org/10.48550/arXiv.2302.06052>
- [42] HENRIQUES J F, CASEIRO R, MARTINS P, *et al.* High-Speed tracking with kernelized correlation filters [J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2015, 37(3): 583-596.

作者简介:



邵小强(1976—),男,陕西商洛人,博士,副教授,2005年于西安科技大学获得硕士学位,2018年于西安科技大学获得博士学位,主要从事图像处理方面的研究。E-mail: shaoxq@163.com

通讯作者:



李浩(2001—),男,四川南充人,硕士研究生,2022年于西南财经大学天府学院获得学士学位,主要从事机器视觉及多模态融合方向的研究。E-mail: 2670815399@qq.com