



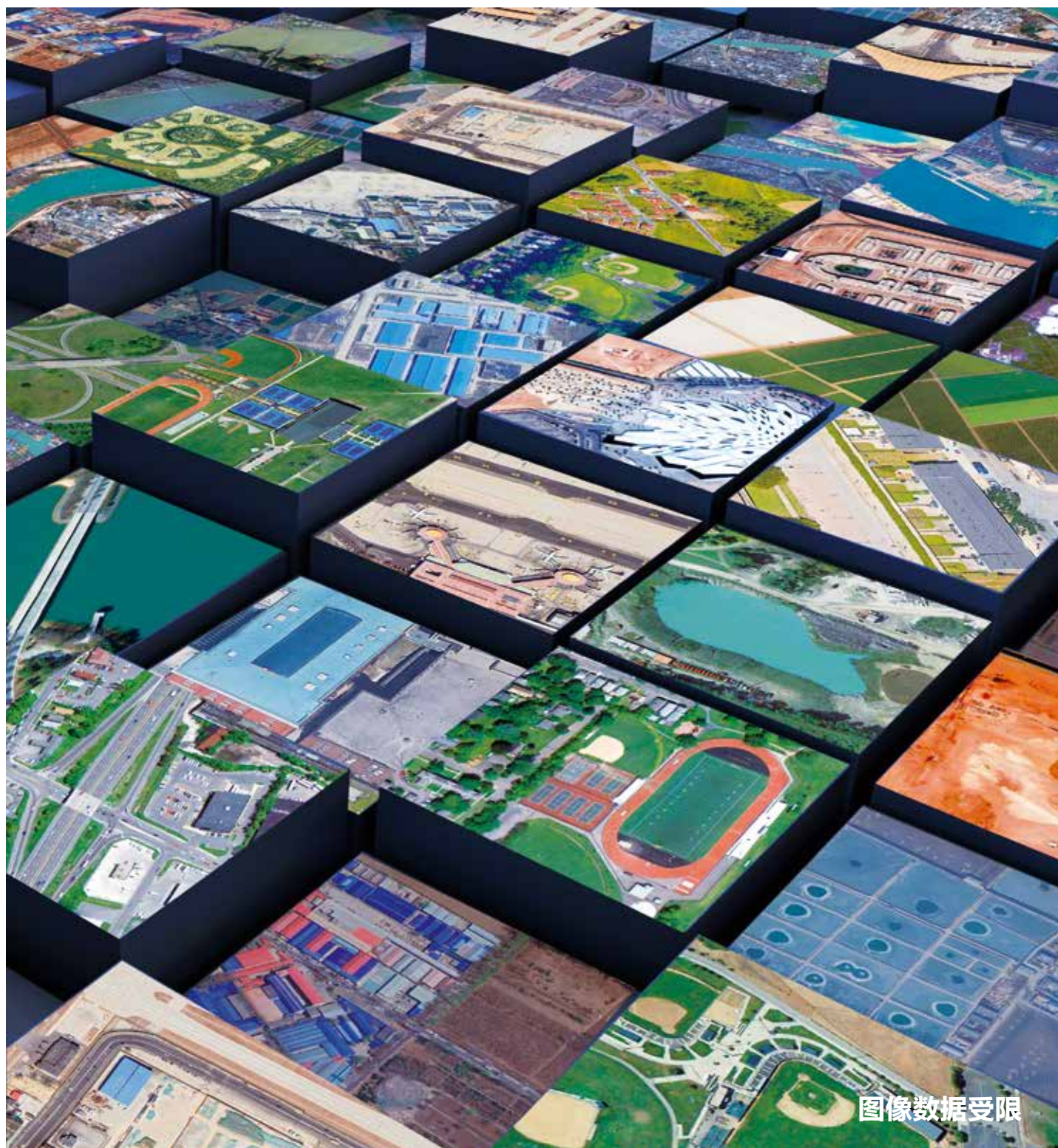
JOURNAL OF IMAGE AND GRAPHICS

主办: 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
10
VOL.27

ISSN1006-8961
CN11-3758/TB



图像数据受限



第27卷第10期（总第318期）
2022年10月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

版权声明

凡向《中国图象图形学报》投稿，均视为同意在本刊网站及CNKI等全文数据库出版，所刊载论文已获得著作权人的授权。本刊所有图片均为非商业目的使用，所有内容，未经许可，不得转载或以其他方式使用。

Copyright

All rights reserved by Journal of Image and Graphics, Institute of Remote Sensing and Digital Earth, CAS. The content (including but not limited text, photo, etc) published in this journal is for non-commercial use.

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主 编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮 编 100190
电子信箱 jig@aircas.ac.cn
电 话 010-58887035
网 址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订 购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
(邮政信箱: 北京399信箱 邮编: 100048)
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

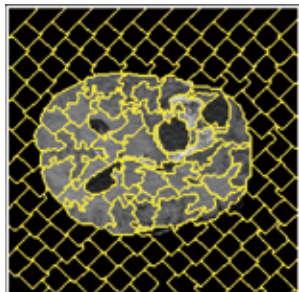
Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048, P.R.China)
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

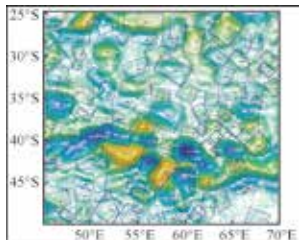
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元



MRI脑肿瘤图像的超像素/体素分割及发展现状(第2897页)



融合多尺度特征与全局上下文信息的X光违禁物品检测(第3043页)



融合多尺度旋转锚机制的海洋中尺度涡自动检测(第3092页)

图像数据受限

《中国图象图形学报》图像数据受限专栏简介

刘怡光, 孙显, 赵启军, 魏秀参, 王琦, 陈秀妍 2801

数据受限条件下的多模态处理技术综述

王佩瑾, 闫志远, 容雪娥, 李俊希, 路晓男, 胡会扬, 严启炜, 孙显 2803

图像数据受限下的处理与分析

刘怡光 2835

面向跨模态行人重识别的单模态自监督信息挖掘

吴岸聪, 林城桔, 郑伟诗 2843

小样本条件下的RGB-D显著性物体检测

何静, 傅可人 2860

综述

面向目标检测的对抗样本综述

袁珑, 李秀梅, 潘振雄, 孙军梅, 肖蕾 2873

MRI脑肿瘤图像的超像素/体素分割及发展现状

方玲玲, 王欣 2897

图网络层级信息挖掘分类算法综述

魏文超, 蔺广逢, 廖开阳, 康晓兵, 赵凡 2916

个性化图像美学评价的研究进展与趋势

祝汉城, 周勇, 李雷达, 赵佳琦, 杜文亮 2937

视盘和视杯分割在计算机辅助青光眼诊断中的应用综述

方玲玲, 张丽榕 2952

图像处理和编码

多监督损失函数光滑化图像超分辨率重建

孟志青, 张晶, 邱健数 2972

轻量级注意力约束对齐网络的视频超分重建

靳雨桐, 宋慧慧, 刘青山 2984

面向图像修复的增强语义双解码器生成模型

王倩娜, 陈焱 2994

图像分析和识别

动态模态交互和特征自适应融合的RGBT跟踪

王福田, 张淑云, 李成龙, 罗斌 3010

采用Transformer网络的视频序列表情识别

陈港, 张石清, 赵小明 3022

对抗型半监督光伏面板故障检测

卢芳芳, 牛然, 杜海舟, 杨振辰, 陈菁菁 3031

融合多尺度特征与全局上下文信息的X光违禁物品检测

李晨, 张辉, 张邹铨, 车爱博, 王耀南 3043

高速公路场景的车路视觉协同行车安全预警算法

汪长春, 高尚兵, 蔡创新, 陈浩霖 3058

结合卷积神经网络与曲线拟合的人体尺寸测量

马燕, 殷志昂, 黄慧, 张玉萍 3068

医学图像处理

U-Net支气管超声弹性图像纵膈淋巴结分割

刘羽, 吴蓉蓉, 唐璐, 宋宁宁 3082

遥感图像处理

融合多尺度旋转锚机制的海洋中尺度涡自动检测

杜艳玲, 刘倩倩, 王丽丽, 徐鑫, 魏泉苗, 宋巍 3092

集成注意力机制和扩张卷积的道路提取模型

王勇, 曾祥强 3102

空域协同自编码器的高光谱异常检测

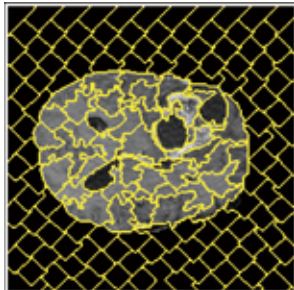
樊港辉, 马泳, 梅晓光, 黄珺, 樊凡, 李峰 3116

自适应权重金字塔和分支强相关的SAR图像舰船检测

郭伟, 申磊, 曲海成, 王雅萱, 林畅 3127

CONTENTS

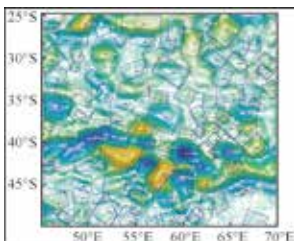
JOURNAL OF IMAGE AND GRAPHICS



The review of superpixel/voxel segmentation of MRI brain tumor images(P2897)



Integrated multi-scale features and global context in x-ray detection for prohibited items(P3043)



Multi-scale rotating anchor mechanism based automatic detection of ocean mesoscale eddy(P3092)

Limited Image Data

Review of multimodal data processing techniques with limited data

Wang Peijin, Yan Zhiyuan, Rong Xuee, Li Junxi, Lu Xiaonan, Hu Huiyang, Yan Qiwei, Sun Xian ... 2803

The processing and analyzing derived of limited image data

Liu Yiguang 2835

Single-modality self-supervised information mining for cross-modality person re-identification

Wu Ancong, Lin Chengzhi, Zheng Weishi 2843

RGB-D salient object detection of using few-shot learning

He Jing, Fu Keren 2860

Review

Review of adversarial examples for object detection

Yuan Long, Li Xiumei, Pan Zhenxiong, Sun Junmei, Xiao Lei 2873

The review of superpixel/voxel segmentation on MRI brain tumor images

Fang Lingling, Wang Xin 2897

Survey of graph network hierarchical information mining for classification

Wei Wenchao, Lin Guangfeng, Liao Kaiyang, Kang Xiaobing, Zhao Fan 2916

The review of personalized image aesthetics assessment

Zhu Hancheng, Zhou Yong, Li Leida, Zhao Jiaqi, Du Wenliang 2937

The review of optic disc and optic cup segmentation applications in computer-aided glaucoma diagnosis

Fang Lingling, Zhang Lirong 2952

Image Processing and Coding

Multi-supervision loss function based smoothed super-resolution image reconstruction

Meng Zhiqing, Zhang Jing, Qiu Jianshu 2972

Super-resolution Video frame reconstruction through lightweight attention constraint alignment network

Jin Yutong, Song Huihui, Liu Qingshan 2984

Enhanced semantic dual decoder generation model for image inpainting

Wang Qianna, Chen Yi 2994

Image Analysis and Recognition

RGBT tracking based on dynamic modal interaction and adaptive feature fusion

Wang Futian, Zhang Shuyun, Li Chenglong, Luo Bin 3010

Video sequence-based human facial expression recognition using Transformer networks

Chen Gang, Zhang Shiqing, Zhao Xiaoming 3022

Generative adversarial networks based semi-supervised fault detection for photovoltaic panel

Lu Fangfang, Niu Ran, Du Haizhou, Yang Zhenchen, Chen Jingjing 3031

Integrated multi-scale features and global context in x-ray detection for prohibited items

Li Chen, Zhang Hui, Zhang Zouquan, Che Aibo, Wang Yaonan 3043

Vehicle-road visual cooperative driving safety early warning algorithm for expressway scenes

Wang Changchun, Gao Shangbing, Cai Chuangxin, Chen Haolin 3058

The convolution neural network and curve fitting based human body size measurement

Ma Yan, Yin Zhiang, Huang Hui, Zhang Yuping 3068

Medical Image Processing

U-Net-based mediastinal lymph node segmentation method in bronchial ultrasound elastic images

Liu Yu, Wu Rongrong, Tang Lu, Song Ningning 3082

Remote Sensing Image Processing

Multi-scale rotating anchor mechanism based automatic detection of ocean mesoscale eddy

Du Yanling, Liu Qianqian, Wang Lili, Xu Xin, Wei Quanmiao, Song Wei 3092

Road extraction model derived from integrated attention mechanism and dilated convolution

Wang Yong, Zeng Xiangqiang 3102

Spatial-coordinated autoencoder for hyperspectral anomaly detection

Fan Ganghui, Ma Yong, Mei Xiaoguang, Huang Jun, Fan Fan, Li Hao 3116

Ship detection in SAR images based on adaptive weight pyramid and branch strong correlation

Guo Wei, Shen Lei, Qu Haicheng, Wang Yaxuan, Lin Chang 3127

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2022)10-3010-12

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

动态模态交互和特征自适应融合的 RGBT 跟踪

王福田^{1,2}, 张淑云¹, 李成龙^{1*}, 罗斌¹

1. 安徽大学计算机科学与技术学院多模态认知计算实验室, 合肥 230000;

2. 合肥综合性国家科学中心人工智能研究院, 合肥 230000

摘要: **目的** 可见光和热红外模态数据具有很强的互补性, RGBT (RGB-thermal) 跟踪受到越来越多的关注。传统 RGBT 目标跟踪方法只是将两个模态的特征进行简单融合, 跟踪的性能受到一定程度的限制。本文提出了一种基于动态交互和融合的方法, 协作学习面向 RGBT 跟踪的模态特定和互补表示。**方法** 首先, 不同模态的特征进行交互生成多模态特征, 在每个模态的特定特征学习中使用注意力机制来提升判别性。其次, 通过融合不同层次的多模态特征来获得丰富的空间和语义信息, 并通过设计一个互补特征学习模块来进行不同模态互补特征的学习。最后, 提出一个动态权重损失函数, 根据对两个模态特定分支预测结果的一致性和不确定性进行约束以自适应优化整个网络中的参数。**结果** 在两个基准 RGBT 目标跟踪数据集上进行实验, 数据表明, 在 RGBT234 数据集上, 本文方法的精确率 (precision rate, PR) 为 79.2%, 成功率 (success rate, SR) 为 55.8%; 在 GTOT (grayscale-thermal object tracking) 数据集上, 本文方法的精确率为 86.1%, 成功率为 70.9%。同时也在 RGBT234 和 GTOT 数据集上进行了对比实验以验证算法的有效性, 实验结果表明本文方法改善了 RGBT 目标跟踪的结果。**结论** 本文提出的 RGBT 目标跟踪算法, 有效挖掘了两个模态之间的互补性, 取得了较好的跟踪精度。

关键词: 模态交互; 模态融合; 互补特征学习; 模态特定信息; RGBT 目标跟踪

RGBT tracking based on dynamic modal interaction and adaptive feature fusion

Wang Futian^{1,2}, Zhang Shuyun¹, Li Chenglong^{1*}, Luo Bin¹

1. Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230000, China; 2. Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230000, China

Abstract: **Objective** Visual target tracking can be applied to the computer vision analysis, such as video surveillance, unmanned autopilot systems, and human-computer interaction. Thermal infrared cameras have the advantages of long-range of action, strong penetrating ability, hidden objects. As a branch of visual tracking, RGBT (RGB-thermal) tracking aims to estimate the status of the target in a video sequence by aggregating complementary data from two different modalities given the groundtruth bounding box of the first frame of the video sequence. Previous RGBT tracking algorithms are constrained of traditional handcraft features or insufficient to explore and utilize complementary information from different modalities. In

收稿日期: 2021-04-21; 修回日期: 2021-09-18; 预印本日期: 2021-09-25

* 通信作者: 李成龙 lcl1314@foxmail.com

基金项目: 国家自然科学基金项目 (62076003); 安徽高校协同创新项目 (GXXT-2019-007); 安徽省自然科学基金项目 (1908085MF206)

Supported by: National Natural Science Foundation of China (62076003); University Synergy Innovation Program of Anhui Province (GXXT-2019-007); Natural Science Foundation of Anhui Province, China (1908085MF206)

order to explore the complementary information between the two modalities, we propose a dynamic interaction and fusion method for RGBT tracking. **Method** Generally, RGB images capture visual appearance information (e.g., colors and textures) of target, and thermal images acquire temperature information which is robust to the conditions of lighting and background clutter. To obtain more powerful representations, we can introduce the useful information of another modality. However, the fusion of different modalities is opted from addition or concatenation in common due to some noisy information of the obtained modality features. First, a modality interaction module is demonstrated to suppress clutter noise based on the multiplication operation. Second, a fusion module is designed to gather cross-modality features of all layers. It captures different abstractions of target representations for more accurate localization. Third, a complementary gate mechanism guided learning structure calculates the complementary features of different modalities. As the input of the gate, we use the modality-specific features and the cross-modality features obtained from the fusion module. The output of the gate is a numerical value. To obtain the complementary features, we carry out a dot product operation on this value and the cross-modality features. Finally, a dynamic weighting loss is presented to optimize the parameters of the network adaptively in terms of the constraints of the consistency and uncertainty of the prediction results of two modality-specific branches. Our method is evaluated on two standard RGBT tracking datasets, like GTOT (grayscale thermal object tracking) and RGBT234. The two evaluation indicators (precision rate and success rate) is illustrated to measure the performance of tracking. Our model is built through the open source toolbox Pytorch, and the stochastic gradient descent method is optimized. Our implementation runs on the platform of PyTorch with 4.2 GHz Intel Core I7-7700K and NVIDIA GeForce GTX 1080Ti GPU. **Result** We conduct many comparative experiments on the RGBT234 and GTOT datasets. Based on the GTOT dataset analysis, our method (86.1%, 70.9%) exceeds baseline tracker (80.6%, 65.6%) by 5.5% in precision rate (PR) and 5.3% in success rate (SR). our method (79.2%, 55.8%) is 7.0% higher in PR and 6.3% higher in SR than the baseline tracker in terms of the RGBT234 dataset analysis. Compared to the second-performing tracking method, on the RGBT234 dataset, our method is 2.6% higher than DAPNet (76.6%) in PR, and 2.1% higher than DAPNet (53.7%) in SR. At the same time, we conduct component analysis experiments on two datasets. Our experimental results illustrate that each module can improve the performance of tracking. **Conclusion** Our RGBT target tracking algorithm obtains rich semantic and spatial information through modal interaction and fusion modules, and uses the gate mechanism to explore the complementarity between different modalities. Dynamic weighting loss is illustrated to adaptively optimize the parameters in the model in accordance with the constraints of the prediction results of two modality-specific branches.

Key words: modality interaction; modality fusion; complementary features learning; modality-specific information; RGBT object tracking

0 引言

视觉目标跟踪是计算机视觉领域中的热门课题之一,在多个领域有着广泛的应用,如视频监控、无人驾驶和人机交互(Li等,2019a;Ruan等,2019;Yao等,2017;Yuan等,2015;Zhang等,2015)。红外成像具有成像距离长、对烟雾的穿透力强和对光照变化不敏感等优点。随着热红外技术的不断成熟,RGBT(RGB-thermal)目标跟踪受到越来越多学者的关注(Wu等,2011;Liu和Sun,2012;Li等,2016,2017)。RGBT目标跟踪是视觉目标跟踪的一个分支,其任务是在给定某视频序列第1帧目标状态的情况下,通过聚合不同模态数据来估计后续帧中目标的状

态。虽然RGBT目标跟踪近年来取得了很大的进展,但是如何充分探索和利用不同模态的信息仍然是一个挑战。

早期的RGBT跟踪方法(Li等,2016;Liu和Sun,2012)主要依赖于传统的手工特征。例如,Liu和Sun(2012)通过在基于粒子滤波器的框架中联合稀疏表示和对稀疏系数的最小化操作融合可见光和热红外模态的跟踪结果。Li等人(2016)引入可靠的权重来自适应地融合可见光和热红外模态的信息。但是,这些方法很难应对复杂场景和环境。近年来,深度学习方法成功地应用于RGBT目标跟踪领域,并逐渐在性能上超越传统方法。Gao等人(2019)提出了一种新的机制来自适应地融合不同模态的有效信息。Zhu等人(2019)提出了一种新颖

的特征聚合网络以此聚合不同层的特征,同时利用修剪技术消除冗余信息。Tu 等人(2020)提出了一种多边际度量学习框架 M^5L (multi-modal multi-margin metric learning) 来解决困难样本问题。Xu 等人(2022)提出了一种跨层双线性池化网络来自适应地聚合两种模态的不同信息。但是,这些基于深度学习的跟踪方法要么仅利用高层的语义信息,要么将不同层的特征进行逐层聚合,很少去有效挖掘不同模态的互补特征,因此无法有效探索 RGBT 数据的优越性。

本文提出了一种基于动态交互和融合的跟踪框架来有效挖掘两个模态之间的互补性。一般来说,RGB 图像主要用来捕获目标的视觉外观信息,例如颜色和纹理等,而热红外图像在光照变化剧烈和背景杂乱等条件下能提供有效信息。两个模态特征的融合方式通常是元素级相加或通道上级联。如果某种模态的信息中包含一些噪声或者是错误的信息,那么经过上述方法融合的特征里面难免会引入噪声,这样会影响跟踪的性能。因此,本文利用了基于乘法操作的模态交互模块来抑制杂波噪声,同时设计了一个融合模块来聚合不同层的多模态特征以捕获丰富的语义信息和空间信息。最后,设计了一个动态权重损失函数来优化网络参数。

图 1 展示了 RGB 模态下的特定特征和互补特征的可视化结果。本文的主要贡献有:

1) 提出了一种新颖的动态交互和融合模块,有效挖掘 RGB 和热红外数据的互补性;

2) 提出了一种动态权重损失函数,通过对两个模态特定分支预测结果进行一致性和不确定性约束来优化整个网络中的参数;

3) 在 RGBT234 (Li 等, 2019) 和 GTOT (grayscale-thermal object tracking) (Li 等, 2016) 数据集上进行了实验,结果表明本文方法在跟踪性能上

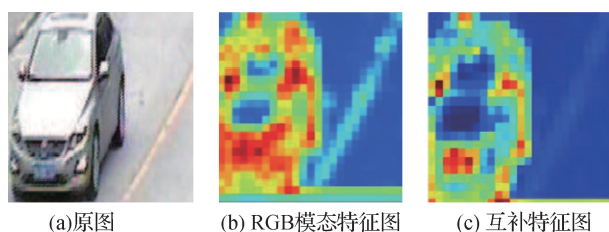


图 1 特征图的可视化结果

Fig. 1 Visualization of feature map((a) original map; (b) RGB-specific feature map; (c) complementary feature map)

优于其他主流的跟踪方法,验证了有效性。

1 相关工作

1.1 RGBT 目标跟踪算法

可见光信息和热红外信息的融合能够有效改善跟踪性能,且随着热红外相关产品的价格下降,越来越多的学者开始研究 RGBT 目标跟踪。

一些早期的工作(Li 等, 2016; Lan 等, 2018)通过为每种模态引入可靠的权重实现自适应融合。Li 等人(2016)在贝叶斯框架的基础上提出了一种基于稀疏表示的自适应跟踪框架。Lan 等人(2018)提出了一种新颖的判别性学习框架,以自适应和协作地学习分类器和不同模态的可靠性权重。但是,当权重不能可靠地反映模态的重要性时,跟踪器将会面临跟踪失败的状况。同时,一些工作(Li 等, 2017; Zhu 等, 2020)探索如何利用多模态数据来生成更加鲁棒的特征表示。Li 等(2018c)提出了一种跨模态排序模型,该模型考虑了两种模态的异质性以及初始化排序种子点时所引起的噪声,同时使用结构化的 SVM (support vector machines) 来进行跟踪。这些 RGBT 跟踪方法依赖于传统的手工特征,而传统的手工特征无法鲁棒地表示目标对象,所以这些方法复杂场景下的跟踪性能会受到限制。

由于深度学习方法成功应用跟踪领域,因此 RGBT 跟踪取得了很大的进步(Xu 等, 2018; Zhang 等, 2018; Zhu 等, 2019; Li 等, 2020a)。Xu 等人(2018)提出了一种像素级融合的 RGBT 跟踪方法,该方法直接将热红外图像作为可见光图像的额外通道。该融合方式容易引入大量噪声,从而影响跟踪的性能。Zhang 等人(2018)使用 VGG-M (Visual Geometry Group-middle network) 网络提取两个模态的特征,然后对两个模态的特征进行级联,最后对目标进行前景和背景的分类。Zhu 等人(2019)提出了一种新的跟踪框架递归的聚合不同层的特征和抑制其中的冗余信息。但上述方法没有考虑到不同模态的特征在不同场景下的重要程度。Li 等人(2020a)考虑到不同模态中存在不同的和共同的挑战因素,提出了一个基于挑战感知的 RGBT 跟踪框架,取得了很好的效果。虽然上述 RGBT 跟踪方法有效地利用了两个模态的信息,但是没有对两个模态之前的互补性进行充分探索。

1.2 注意力机制

注意力机制一开始应用于自然语言领域。2014年,谷歌大脑团队利用注意力机制进行图像分类,取得了很好的效果。由于注意力机制的有效性,其广泛用于各种任务中,如目标跟踪、目标检测和语义分割。近年来,注意力机制在提高网络性能方面取得了很大的进步。Hu 等人(2018)提出了一种有效的通道注意力机制,该机制通过自适应地调节各通道的特征响应值来选择性地加强有用信息和抑制无用信息。Chen 等人(2018)提出了一种双重注意力机制,从输入图像或视频的时空空间中聚合和传递全局特征,从而使得后续的卷积层有效地提取特征。Fu 等人(2019)引入空间注意力和通道注意力机制来聚合更多有用的信息。上述这些方法为了提升性能,都致力于研究复杂的注意力模块,但同时不可避免地增加了计算开销。为了平衡性能和资源消耗,Wang 等人(2020)提出了一个

轻量级的注意力模块,其通过1维卷积运算生成通道关注图。

2 网络框架及其实现

2.1 网络框架

在 RGBT 跟踪任务中,输入可见光和相应的热红外图像,本文网络框架如图2所示。从图2中可以看到,除了可见光和热红外分支之外,还包含跨模态交互模块和融合模块、互补特征学习模块,跨模态交互模块用来获取不同级别的多模态特征,模态融合模块用来融合不同级别的多模态特征,互补特征学习模块用来获取不同模态的互补特征。本文使用 VGG-M (Simonyan 和 Zisserman 等,2015) 的前3个卷积层作为骨干网络来提取可见光和热红外模态的特定特征,其中卷积核大小分别为 $7 \times 7 \times 96$ 、 $5 \times 5 \times 256$ 、 $3 \times 3 \times 512$ 。

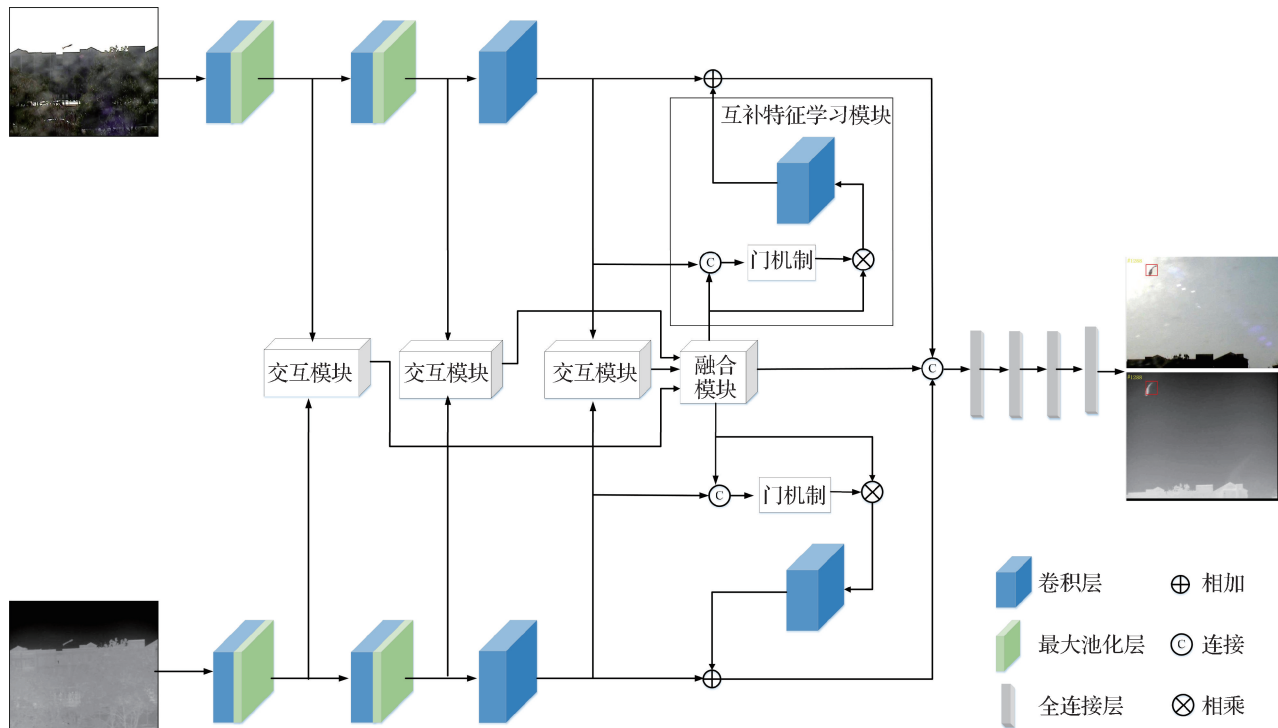


图2 网络框架图

Fig.2 Pipeline of our method

2.2 交互和融合模块

2.2.1 交互模块

为了减少噪声影响,本文设计了一种模态交互模块。首先,使用通道注意模块从通道方面增强不同模态的特定特征,然后将可见光和热红外特征进

行元素级相乘。在某些情况下,热红外数据中有一些噪声,可见光数据也存在噪声。如果直接将RGB特征和热红外特征进行相加或级联,可能会引入许多噪声。本文将不同模态的特征进行相乘操作,可以在一定程度上抑制噪声。整个交互过程可以表

示为

$$x_{RT}^1 = M(\varepsilon(x_R^1), \varepsilon(x_T^1)) \quad (1)$$

$$x_{RT}^2 = M(\varepsilon(x_R^2), \varepsilon(x_T^2)) \quad (2)$$

$$x_{RT}^3 = M(\varepsilon(x_R^3), \varepsilon(x_T^3)) \quad (3)$$

式中, M 代表相乘操作, ε 代表通道注意力机制, x_R^i 和 x_T^i 分别表示第 i 层的可见光模态的特征和热红外模态的特征, x_{RT} 表示融合之后的特征。

考虑上下文信息有助于模型捕获目标的位置, 而推理跟踪中的上下文信息需要知道哪些位置是与目标相关的关键点。本文网络模型利用了通道注意力模块 (Wang 等, 2020) 从通道方面增强不同模态的特定特征。Wang 等人 (2020) 的实验结果表明捕获所有通道之间的依赖关系是低效的, 而捕获局部通道之间的信息是高效的。式 (4) 能够为每个通道捕获其周围的邻居信息以预测该通道的注意力预测。分组卷积在固定组数量的情况下, 高维 (低维) 通道与长距离 (短距离) 卷积核大小成正比; 因此, 局部通道信息交互作用的范围 (即 1 维卷积的核大小 k) 与通道维数 C 也存在一定的映射关系。

具体来说, 首先将输入的特征图经过全局平均池化 (global average pooling, GAP), 再执行卷积核大小为 k 的快速 1 维卷积来生成通道权值, 最后将输入的特征和通道权值做相乘操作

$$k = \varphi(C) = \left\lfloor \frac{\log_2 C}{\gamma} + \frac{b}{\gamma} \right\rfloor_{\text{odd}} \quad (4)$$

式中, k 表示 1 维卷积的内核大小, C 表示通道数, $\lfloor \cdot \rfloor_{\text{odd}}$ 表示奇数, b 和 γ 的值分别设为 2 和 1。

2.2.2 融合模块

首先将前两层获得的多模态特征图调整为与最后一层的多模态特征图相同尺寸, 然后进行元素级相加。最后将融合后的特征输入到卷积核大小为 3×3 的深度可分离卷积中, 从而减少参数量。具体为

$$x_{RT}^{\text{new}} = S(x_{RT}^1 + x_{RT}^2 + x_{RT}^3) \quad (5)$$

式中, S 表示深度可分离的卷积操作, x_{RT}^i 表示第 i 层的多模态特征。

2.3 互补特征学习模块

现有方法 (Cheng 等, 2017; Li 等, 2020b) 大多使用门机制来控制信息流的输出。Cheng 等人 (2017) 使用加权门机制来自适应地融合不同模态的特征。Li 等人 (2020b) 通过使用门机制自适应融合多级特

征。这种门机制不会像掩码一样导致 0 或 1 的输出, 而是充当限制器并控制数值。本文互补特征学习模块采用门机制来控制信息的输出。如图 2 所示, 模态特定特征和多模态特征作为门机制的输入, 然后门机制对输入的信息进行选择增强或抑制, 输出一个数值, 数值范围 0 ~ 1, 该值和多模态特征执行点乘运算获得模态互补特征, 然后执行 1×1 卷积运算对特征进行降维。Gate 模块是由两个卷积层和一个 Sigmoid 神经网络层组成, 卷积层的大小分别为 1×1 和 3×3 , Sigmoid 层的输出值的范围 0 ~ 1, 它代表输入信息通过的程度。

2.4 动态权重损失函数

RGBT 目标跟踪方法 (Zhu 等, 2019; Li 等, 2017) 一般使用传统的二分类损失函数 (L) 来训练模型, 即

$$L(y', y) = \begin{cases} -\log(y') & y = 1 \\ -\log(1 - y') & \text{其他} \end{cases} \quad (6)$$

式中, y 表示样本的标签, 取值范围为 $\{+1, -1\}$, $+1$ 表示正样本, -1 表示负样本; y' 表示预测样本为正样本的概率 (范围 0 ~ 1)。

本文提出了一个新的损失函数来动态优化模型参数。不同模态的样本输入到相应的模态特定分支中, 样本的预测结果可能有所不同。通过对两个模态特定分支的预测结果进行一致性和不确定性的约束来动态优化整个网络的参数, 从而提高最终预测结果的置信度。提出的动态损失函数为

$$L = L_1 + \lambda \times L_2 \quad (7)$$

$$L_1 = L_{\text{rgb}}(y_1', y) + L_{\text{thermal}}(y_2', y) \quad (8)$$

$$L_2 = L_{\text{fusion}}(y', y) \quad (9)$$

式中, y_1' 和 y_2' 分别表示可见光分支和热红外分支的预测结果, y' 表示最终的预测结果, L_{rgb} 、 L_{thermal} 和 L_{fusion} 分别代表 RGB、thermal 和融合分支的损失函数。 λ 表示自适应权重, 取值范围 0 ~ 1, 计算为

$$\lambda = \frac{|y_1' - y_2'| + 1 - |y_1' - 0.5| - |y_2' - 0.5|}{2} \quad (10)$$

式中, 第 1 项 $|y_1' - y_2'|$ 的 L1 损失是为了约束两个分支的预测结果, 使其更加一致, 而第 2 项 $1 - |y_1' - 0.5| - |y_2' - 0.5|$, 用于限制两个特定分支的预测结果的不确定性。如果预测结果得分接近 0.5, 则表示网络不确定检测到的候选框是否为前景。从图 3 可以看出, 在训练参数设置相同的情况

下,本文提出的动态加权损失下降得更快,收敛得更好。这也进一步表明本文提出的动态加权损失函数的有效性。

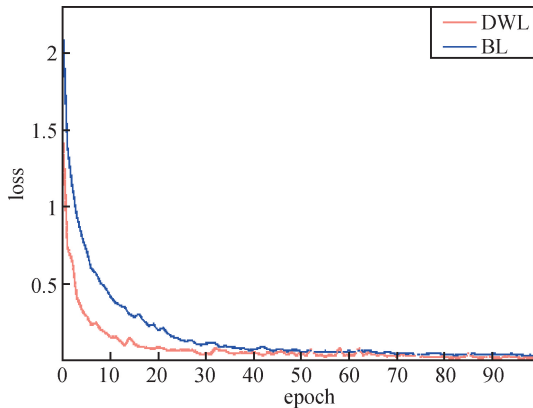


图3 动态权重损失与二分类损失在训练期间的比较结果

Fig. 3 Comparison results of dynamic weight loss(DWL) and the binary loss(BL) during training

2.5 训练过程

本文网络可以进行端到端的训练。在初始化时,网络主干的前3层相应权重从VGG-M网络(Simonyan和Zisserman,2015)中加载得到,全连接层和其他卷积层的权重随机初始化。采用SGD(stochastic gradient descent)算法来对整个网络进行优化。同时参考MDNet(multi-domain convolutional neural networks)(Nam等,2016),在使用 K 个视频的多领域学习中,一共执行100个循环,其中每次循环包括 K 个迭代,每个迭代都对应一个视频。在每次训练迭代中,最小批处理由视频序列中随机选取的8帧图像组成。在每帧图像上随机采样32个正样本(与帧值的IoU(intersection over union)重叠率不小于0.7)和96个负样本(与帧值的IoU重叠率不大于0.5)。因此,每个最小批处理都包含256个正样本和768个负样本。在训练过程中,全连接层和卷积层的学习率分别设置为0.005和0.0005,权重衰减、动量和梯度剪切阈值分别固定为0.0005、0.9和10。在GTOT(Li等,2016)数据集上进行测试时,从RGBT234(Li和Liang,2019)数据集中随机选择了78个视频作为训练集。在RGBT234数据集上进行测试时,将GTOT数据集(50个视频序列)作为训练集。

2.6 在线跟踪过程

在跟踪过程中,固定卷积层的所有参数,仅微调

全连接层的参数,用一个全连接层替换具有 K 个分支的特定域层。当跟踪某个视频序列中的目标时,需要先在第1帧上收集500个正样本和5000个负样本,并使用这些样本进行简单的训练,在本文方法中共训练40轮。在视频的后续帧中,每帧仅收集20个正样本和100个负样本来增加正样本和负样本的总数,以提供用于长期更新和短期更新(Nam和Han,2016)的样本源,长期跟踪和短期跟踪是为了提高跟踪器的鲁棒性。为了估计视频序列中每帧图像中的目标状态,使用高斯采样的方式在前1帧的目标物体周围采集256个候选框,然后将其作为网络的输入,以获得候选框的二分类分数。每个候选框都有一个正得分和一个负得分,分别表示为 $f^+(x_i^t)$ 和 $f^-(x_i^t)$,最后从这些候选框中选择得分最高的候选框,并将其作为当前帧的跟踪结果。同时,采用候选框回归技术来提高定位的准确性;但需要注意的是,只使用每个视频序列的第1帧来训练候选框回归器,因为后续帧中的目标状态可能不可靠。

3 实验分析

3.1 数据集和评价标准

GTOT数据集包含50个可见光—热红外视频序列,同时包含不同的场景,如水池、公共区域和校园道路等。该数据集有7种不同的挑战因素,分别是尺度变化(large scale variation,LSV)、快速运动(fast motion,FM)、低光照(low illumination,LI)、形变(deformation,DEF)、热交叉(thermal crossover,TC)、小目标(small object,SO)和遮挡(occlusion,OCC)。

RGBT234数据集是由RGBT210(red green blue-thermal 210)数据集(Li等,2017)扩展而来,由234个可见光—热红外视频序列组成。该数据集共有12种挑战因素,分别是运动模糊(motion blur,MB)、无遮挡(no occlusion,NO)、部分遮挡(partial occlusion,PO)、严重遮挡(heavy occlusion,HO)、低照度(low illumination,LI)、低分辨率(low resolution,LR)、热交叉(thermal crossover,TC)、形变(deformation,DEF)、快速移动(fast motion,FM)、比例变化(scale variation,SV)、背景杂乱(background clutter,BC)和相机移动(camera moving,CM)。最长的视频序列包含4000多帧,最短的视频序列不超过100帧。

采用精确率 (precision rate, PR) 和成功率 (success rate, SR) 作为 RGBT 目标跟踪的评价标准, 同时在 RGBT234 (Li 等, 2019) 和 GTOT (Li 等, 2016) 上比较本文方法与其他跟踪方法的性能。PR 是输出目标位置在标注真值的阈值距离内帧数所占的百分比。由于 GTOT 数据集中的大多数跟踪对象很小, 因此将其阈值距离设置为 5 像素。将 RGBT234 数据集的阈值距离设置为 20 像素。SR 是在标注的边界框和输出的边界框之间的 IoU 重叠率高于设定阈值的帧数所占的百分比。

3.2 在 GTOT 数据集上的评估

为了评估本文方法的性能, 在 GTOT 数据集上将本文方法与 10 个 RGBT 跟踪方法进行比较, 它们分别是 MDNet (Nam 和 Han, 2016) + RGBT、KCF (kernelized correlation filter) (Henriques 等, 2015) + RGBT、Struck (Hare 等, 2016) + RGBT、CN (color name) (Danelljan 等, 2014) + RGBT、SCM (sparsity-based collaborative model) (Zhong 等, 2012) + RGBT、SGT (sparse graph tracker) (Li 等, 2017)、SiamDW (deeper and wider siamese networks) (Zhang 和 Peng, 2019) + RGBT、CMRT (cross-model ranking algorithm for RGB-T tracking) (Li 等, 2018c)、DAPNet

(dense feature aggregation and pruning network) (Zhu 等, 2019) 和 RT-MDNet (real-time multi-domain convolutional neural networks) (Jung 等, 2018) + RGBT。其中 MDNet + RGBT、CMRT、SiamDW + RGBT、SGT 和 DAPNet 是 RGBT 目标跟踪器, 其他跟踪器是拓展的 RGBT 目标跟踪器。通过将两种不同模态的特征拼接成单个向量或将热红外特征视为可见光特征的额外通道, 把原本基于 RGB 的跟踪器扩展为 RGBT 目标跟踪器。根据图 4 的结果, 在 GTOT 数据集上, 本文方法 (86.1%, 70.9%) 的 PR 和 SR 分别比基准跟踪器 MDNet + RGBT (80.0%, 63.7%) 高 6.1% 和 7.2%。这些实验数据证明了本文方法的有效性。

从图 4 中可以看出, 本文方法 PR 比 DAPNet (Zhu 等, 2019) 低 2.1%。同时将本文方法和 DAPNet 在 GTOT 数据集的 7 种挑战因素上进行了比较。本文方法在形变 (DEF) 挑战下 PR、SR 分别为 88.4%、73.7%, 而 DAPNet 在形变挑战下的 PR、SR 分别为 91.9%、77.1%。这表明本文方法在形变 (DEF) 这个挑战因素下跟踪性能低于 DAPNet。原因可能是 GTOT 数据集上的目标相对较小, 本文方法不能很好地处理小目标的尺度变化。

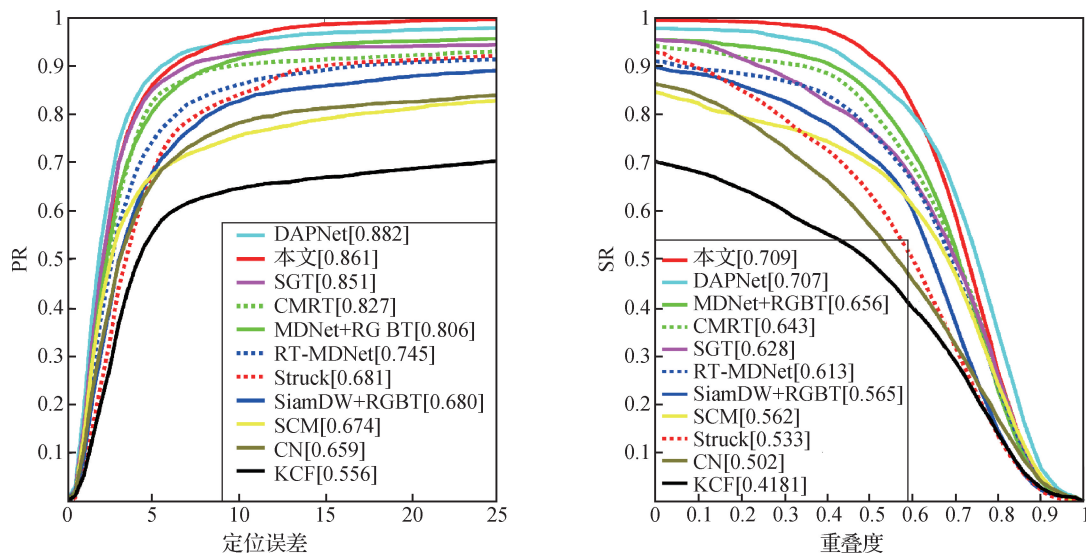


图 4 GTOT 数据集上不同跟踪器的评估曲线

Fig. 4 Evaluation curves of different trackers on GTOT dataset

3.3 在 RGBT234 数据集上的评估

为了评估本文方法的整体性能, 在 RGBT234 数据集上将本文方法与一些 RGBT 方法进行比较, 包括 RT-MDNet (Jung 等, 2018) + RGBT、SiamDW

(Zhang 和 Peng, 2019) + RGBT、MDNet (Nan 和 Ham, 2016) + RGBT、SOWP (spatially ordered and weighted patch) (Kim 等, 2015) + RGBT、DAPNet (Zhu 等, 2019)、CFnet (correlation filter networks)

(Valmadre 等, 2017) + RGBT、CSR-DCF (discriminative correlation filter with channel and spatial reliability) (Lukežić 等, 2017) + RGBT、CMRT (Zhu 等, 2019)、MEEM (multi-expert entropy minimization tracking network) (Zhang 等, 2014) + RGBT、KCF (Henriques 等, 2015) + RGBT, 其中 RT-MDNet + RGBT、MDNet + RGBT、DAPNet、CMRT 和 SGT 是

RGBT 目标跟踪器, 其余的是拓展的 RGBT 目标跟踪器。从图 5 中可以看出, 本文方法 PR 比基准跟踪器 MDNet + RGBT (72.2%) 高 7.0%, SR 比基准跟踪器 MDNet + RGBT (49.5%) 高 6.3%。本文方法的 PR 和 SR (79.2% 和 55.8%) 分别比 DAPNet (76.6% 和 53.7%) 高 2.6% 和 2.1%。

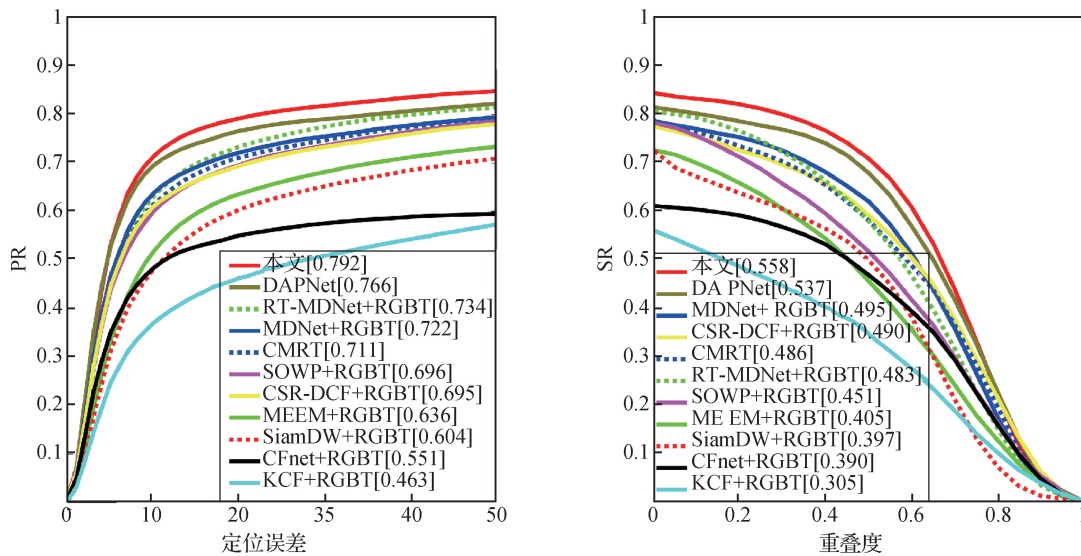


图5 RGBT234 数据集上不同跟踪器的评估曲线

Fig. 5 Evaluation curves of different trackers on RGBT234 dataset

在 RGBT234 数据集不同的挑战因素下, 将本文跟踪器与其他主流跟踪器进行比较, 分别是 MDNet + RGBT, DAPNet, RT-MDNet + RGBT, SiamDW + RGBT, CMRT, SOWP + RGBT, M⁵L, CSR-DCF + RGBT。在表 1 中详细展示了实验结果。从表 1 中可以看出, 本文方法几乎在所有挑战下均优于其他 RGBT 跟踪器, 尤其是在背景杂乱 (BC), 遮挡 (HO) 和形变 (DEF) 等挑战下的跟踪效果尤其突出。表明本文跟踪器在处理物体外观变化和不利条件方面具有更好的优势, 但是在热交叉的情况下, 本文跟踪器性能不如 DAPNet, 这是热交叉情况下的热红外模态的信息不可靠所致, 由于本文方法是将互补特征和模态特征直接进行元素级相加, 因此在热交叉情况下引入了一些噪声。但是在其他挑战中, 本文跟踪器性能优于 DAPNet。

3.4 定性分析

在 RGBT 目标跟踪数据集中选取了 4 个视频序列, 将本文方法和 4 个 RGBT 跟踪方法进行了定性结果比较, 其分别是 SiamDW (Zhang 和 Peng, 2019) +

RGBT, C-COT (continuous convolution operator tracker) (Danelljan 等, 2016), RT-MDNet (Jung 等, 2018) + RGBT 和 MDNet (Nam 和 Han, 2016) + RGBT。可视化的结果展示在图 6 中。本文跟踪器能够有效应对遮挡、背景杂乱和外观变化等挑战。图 6(a)(c) 所示的背景混乱和部分遮挡的情况下, 本文方法很好地区分目标和背景, 从而准确地跟踪目标。图 6(b) 所示的高强度照明的条件下, 可见光图像几乎无法定位目标物体的位置, 而热红外图像提供了额外的信息, 本文方法充分利用了热红外图像所提供的信息来对目标的位置进行准确定位。如图 6(d) 所示, 在摄像机强烈抖动和强光照射的情况下, 只有本文方法可以准确地跟踪到目标, 而其他方法都跟踪失败。

3.5 消融分析

为了进一步验证本文方法的各个组成部分的有效性, 在 GTOT 和 RGBT234 数据集上进行消融实验。使用“Baseline”来表示基准方法, 即 MDNet + RGBT, 是直接在通道方向上级联两个模态的特定特征, 并在训练期间使用二分类损失函数训练模型。

表 1 在 RGB234 数据集上,不同跟踪器在不同挑战上的 PR/SR 分数比较
Table 1 Challenge-based PR/SR score comparison of different trackers on RGB234 dataset

挑战	RT-MDNet + RGBT	SOWP + RGBT	MDNet + RGBT	SiamDW + RGBT	CSR-DCF + RGBT	CMRT	DAPNet	M ⁵ L	本文
FM	70.3/43.5	70.3/43.5	58.6/36.3	59.7/36.5	52.9/35.8	61.3/38.4	67.0/44.3	65.9/42.0	74.5/48.5
LI	73.7/47.4	72.3/46.8	67.0/45.5	60.0/39.9	69.1/47.4	74.2/49.8	77.5/53.0	76.1/49.5	84.0/57.3
CM	66.4/45.5	65.2/43.0	64.0/45.4	56.2/38.2	61.1/44.5	62.9/44.7	66.8/47.4	71.6/50.0	71.3/51.2
HO	61.8/40.4	57.0/37.9	61.9/42.1	52.0/33.7	59.3/40.9	56.3/37.7	66.0/45.7	66.2/45.7	68.0/47.2
BC	72.5/45.5	64.7/41.9	64.4/43.2	51.9/32.3	61.8/41.0	63.1/39.8	71.7/48.4	76.6/49.4	76.9/51.4
DEF	67.0/46.6	65.0/46.0	66.8/47.3	55.8/39.0	63.0/46.2	66.7/47.8	71.7/57.8	72.7/50.0	72.7/53.0
NO	89.3/58.6	86.8/53.7	86.2/61.1	78.3/53.4	82.6/60.0	89.5/61.6	90.0/64.4	90.4/61.9	90.9/65.9
TC	78.6/51.3	70.1/44.2	75.6/51.7	56.9/36.8	66.8/46.2	67.5/44.3	76.8/53.8	78.1/54.3	70.7/50.9
LR	78.6/48.3	72.5/46.2	75.9/49.3	60.5/37.0	72.0/47.6	68.7/42.0	75.0/51.0	76.2/49.6	70.7/51.2
SV	73.5/48.2	66.4/40.4	73.5/50.5	60.9/40.5	70.7/49.9	71.0/49.3	78.0/54.2	78.0/54.2	79.1/56.8
PO	78.0/51.7	74.7/48.4	76.1/51.8	60.8/39.6	73.7/52.2	77.2/53.6	82.1/57.4	82.1/57.4	85.2/59.8
MB	61.2/42.9	63.9/42.1	65.4/46.3	49.4/34.0	58.0/42.5	60.0/42.7	65.3/46.7	67.0/47.2	72.4/51.8
全部	73.4/48.3	69.6/45.1	72.2/49.5	60.4/39.7	69.5/49.0	71.1/48.6	76.6/53.7	77.0/52.1	79.2/55.8

注:加粗字体为每行最优值。

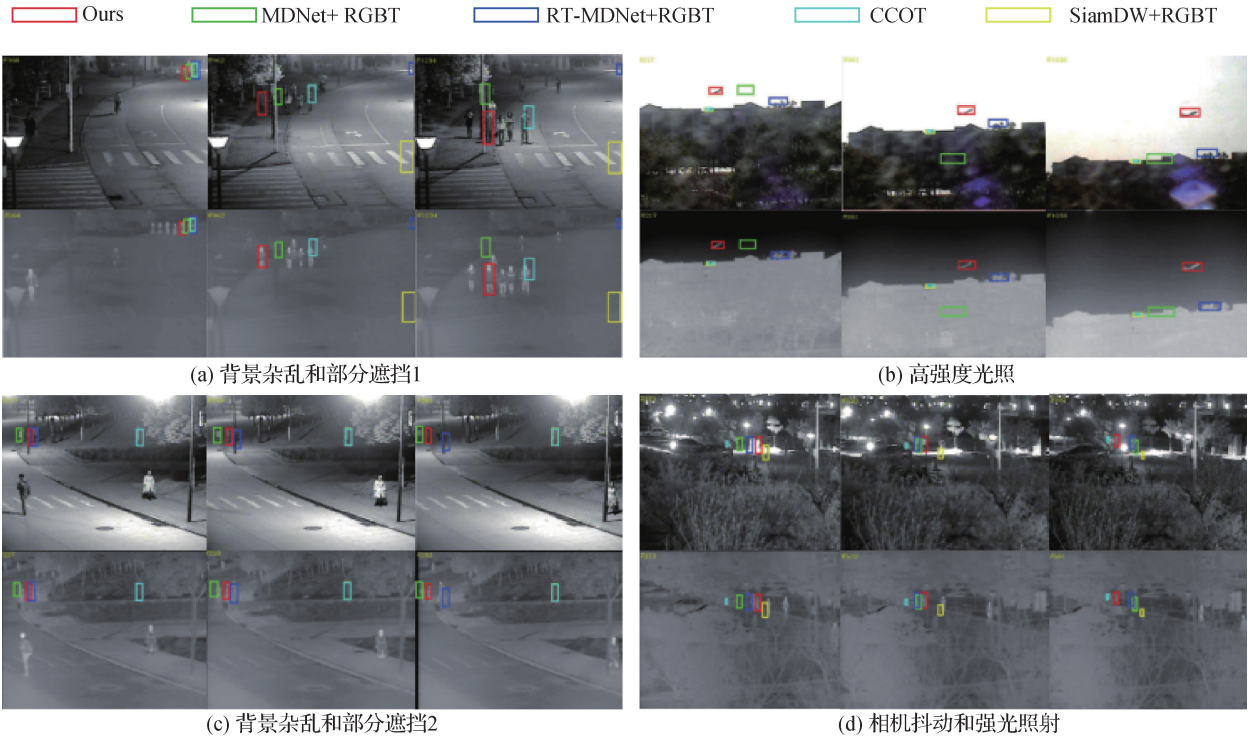


图 6 不同跟踪方法的实例

Fig. 6 Sample results of different tracking method((a) background clutter and partial occlusion 1; (b) high illumination; (c) background clutter and partial occlusion 2; (d) camera moving and high illumination)

实验包括:1) Baseline + DWL, 在通道方向上级联两个模态的特定特征, 同时在训练期间使用动态权重损失函数来优化整个网络;2) Baseline + DWL + MIFL, 使用跨模态交互和融合模块来融合不同层的多模态特征, 在训练期间使用动态损失函数训练模型;3) Baseline + DWL + MIFL + CFL, 使用跨模态交互和融合模块融合不同层的多模态特征, 同时利用互补特征学习模块来计算不同模态的互补特征, 将模态特定特征和互补特征进行融合, 最后使用动态损失函数训练网络模型。

表2展示了在 GTOT 和 RGBT234 数据集上各个组成成分的实验结果。从表2可以得出:1) Baseline + DWL 的结果优于 Baseline, 在相同设置的测试环境下, 在 RGBT234 数据集上 PR 和 SR 分别提升了 3.4% 和 3.2%, 在 GTOT 数据集上 PR 和 SR 分别提升了 3.5% 和 3.9%。这表明在训练过程中使用动态权重损失函数以自适应优化网络中的参数, 使得训练的模型的鲁棒性更高。2) Baseline + DWL + MIFL 的结果优于 Baseline + DWL, 这表明通过跨模态交互和融合模块后得到的多模态特征比直接在通道上级联两种模态的特征所获得的多模态特征更具有判别性。3) Baseline + DWL + MIFL + CFL 优于 Baseline + DWL + MIFL, 这表明互补特征学习模块有效探索了两种模态之间的互补性, 模态互补和特定特征的融合之后能更加鲁棒地表示目标对象。从实验结果得出, 每一个模块的加入都会使跟踪精度有进一步的提升。最后, 本文方法和 Baseline (MDNet + RGBT) 的帧速率分别为 1.67 帧/s 和 2.1 帧/s。

表2 在 RGBT234 和 GTOT 数据集上的成分分析结果

Table 2 Component analysis results on RGBT234 and GTOT datasets

Baseline	DWL	MIFL	CFL	RGBT234		GTOT	
				PR	SR	PR	SR
✓	-	-	-	0.722	0.495	0.800	0.637
✓	✓	-	-	0.756	0.527	0.835	0.676
✓	✓	✓	-	0.770	0.538	0.838	0.686
✓	✓	✓	✓	0.792	0.558	0.861	0.709

注:DWL(dynamic weighted learning)表示动态损失函数,MIFL(modal interaction and fusion learning)表示跨模态交互和融合模块,CFL(complementary feature learning)表示互补特征学习模块,✓表示采用,-表示未采用。

4 结 论

本文方法通过门机制对可见光和热红外特征之间的互补性进行了有效的挖掘。提出了一个动态权重损失函数,通过对两个特定分支的预测结果进行一致性和不确定性的约束来优化整个训练模型。在 RGBT234 和 GTOT 这两个基准数据集上将本文方法与其他跟踪方法进行对比,实验结果表明本文方法优于其他方法。

然而本文方法也存在不足,相对于实时的 RGBT 跟踪方法,本文方法在实时性方面存在很大的优化空间,未来会考虑加入改进的 ROI-align(region of interest-align)模块来提高跟踪器的速度。同时本文方法通过门机制获取的互补特征,在获取的特征中可能仍然有一些噪声,未来也会将采取一些措施来进一步抑制噪声,以获得更加鲁棒的特征表示。

参考文献 (References)

- Chen Y P, Kalantidis Y, Li J S, Yan S C and Feng J S. 2018. A²-nets: double attention networks [EB/OL]. [2021-04-21]. <https://arxiv.org/pdf/1810.11579.pdf>
- Cheng Y H, Cai R, Li Z W, Zhao X and Huang K Q. 2017. Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1475-1483 [DOI: 10.1109/CVPR.2017.161]
- Danelljan M, Khan F S, Felsberg M and Van de Weijer J. 2014. Adaptive color attributes for real-time visual tracking//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 1090-1097 [DOI: 10.1109/CVPR.2014.143]
- 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]
- Fu J, Liu J, Tian H J, Li Y, Bao Y J, Fang Z W and Lu H Q. 2019. Dual attention network for scene segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3141-3149 [DOI: 10.1109/CVPR.2019.00326]
- 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//Pro-

- ceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea (South); IEEE: 91-99 [DOI: 10.1109/ICCVW.2019.00017]
- Hare S, Golodetz S, Saffari A, Vineet V, Cheng M M, Hicks S L and Torr P H S. 2016. Struck: structured output tracking with kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10): 2096-2109 [DOI: 10.1109/TPAMI.2015.2509974]
- Henriques J F, Caseiro R, Martins P and Batista J. 2015. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3): 583-596 [DOI: 10.1109/TPAMI.2014.2345390]
- Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]
- 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]
- 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 2015 IEEE International Conference on Computer Vision. Santiago, Chile; IEEE: 3011-3019 [DOI: 10.1109/ICCV.2015.345]
- Lan X Y, Ye M, Zhang S P and Yuen P. 2018. Robust collaborative discriminative learning for RGB-infrared tracking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1): 7008-7015
- Li C L, Cheng H, Hu S Y, Liu X B, Tang J and Lin L. 2016. 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, Liang X Y, Lu Y J, Zhao N and Tang J. 2019. RGB-T object tracking: benchmark and baseline [EB/OL]. [2021-04-21]. <https://arxiv.org/pdf/1805.08982.pdf>
- Li C L, Lin L, Zuo W M, Tang J and Yang M H. 2019a. Visual tracking via dynamic graph learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11): 2770-2782 [DOI: 10.1109/TPAMI.2018.2864965]
- Li C L, Liu L, Lu A D, Ji Q and Tang J. 2020a. 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, 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, Zhao N, Lu Y J, Zhu C L and Tang J. 2017. Weighted sparse representation regularized graph learning for RGB-T object tracking//Proceedings of the 25th ACM International Conference on Multimedia. California, 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 X T, Zhao H L, Han L, Tong Y H and Yang K Y. 2020b. GFF: gated fully fusion for semantic segmentation[DB/OL]. [2021-04-21]. <https://arxiv.org/pdf/1904.01803.pdf>
- 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]
- Lukežić A, Vojir T, Zajc L C, Matas J and Kristan M. 2017. Discriminative correlation filter with channel and spatial reliability//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA; IEEE: 4847-4856 [DOI: 10.1109/CVPR.2017.515]
- 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]
- Ruan W J, Chen J, Wu Y, Wang J Q, Liang C, Hu R M and Jiang J J. 2019. Multi-correlation filters with triangle-structure constraints for object tracking. *IEEE Transactions on Multimedia*, 21(5): 1122-1134 [DOI: 10.1109/TMM.2018.2872897]
- Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2021-04-21]. <https://arxiv.org/pdf/1409.1556.pdf>
- Tu Z Z, Lin C, Li C L, Tang J and Luo B. 2020. M⁵L: multi-modal multi-margin metric learning for RGBT tracking [EB/OL]. [2021-04-21]. <https://arxiv.org/pdf/2003.07650.pdf>
- 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 Q L, Wu B G, Zhu P F, Li P H, Zuo W M and Hu Q H. 2020. ECA-net: efficient channel attention for deep convolutional neural networks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 11531-11539 [DOI: 10.1109/CVPR42600.2020.01155]
- 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
- Xu N W, Xu G, Zhang X C and Bavirisetti D P. 2018. Relative object tracking algorithm based on convolutional neural network for visible and infrared video sequences//Proceedings of the 4th International Conference on Virtual Reality. Hong Kong, China; ACM; 44-49 [DOI: 10.1145/3198910.3198918]
- 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]
- Yao R, Xia S X, Zhang Z and Zhang Y N. 2017. Real-time correlation filter tracking by efficient dense belief propagation with structure pre-serving. IEEE Transactions on Multimedia, 19 (4): 772-784 [DOI: 10.1109/TMM.2016.2631727]
- Yuan Y, Yang H, Fang Y M and Lin W S. 2015. Visual object tracking by structure complexity coefficients. IEEE Transactions on Multimedia, 17(8): 1125-1136 [DOI: 10.1109/TMM.2015.2440996]
- Zhang J M, Ma S G and Sclaroff S. 2014. MEEM: robust tracking via multiple experts using entropy minimization//Proceedings of 2014 IEEE Conference on European Conference on Computer Vision. Zurich, Switzerland; Springer; 188-203 [DOI: 10.1007/978-3-319-10599-4_13]
- Zhang S L, Yu X, Sui Y, Zhao S C and Zhang L. 2015. Object tracking with multi-view support vector machines. IEEE Transactions on Multimedia, 17(3): 265-278 [DOI: 10.1109/TMM.2015.2390044]
- Zhang X M, Zhang X H, Du X D, Zhou X M and Yin J. 2018. Learning multi-domain convolutional network for RGB-T visual tracking//Proceedings of the 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics. Beijing, China; IEEE; 1-6 [DOI: 10.1109/CISP-BMEI.2018.8633180]
- 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]
- Zhong W, Lu H C and Yang M H. 2012. Robust object tracking via sparsity-based collaborative model//Proceedings of 2012 IEEE Conference on Computer vision and Pattern Recognition. Providence, USA; IEEE; 1838-1845 [DOI: 10.1109/CVPR.2012.6247882]
- Zhu Y B, Li C L, Tang J and Luo B. 2020. Quality-aware feature aggregation network for robust rgbt tracking//IEEE Transactions on Intelligent Vehicles, 99: # 2980735 [DOI: 10.1109/TIV.2020.2980735]
- Zhu Y B, Li C L, Luo B, Tang J and Wang X. 2019. Dense feature aggregation and pruning for RGBT tracking//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France; ACM; 465-472 [DOI: 10.1145/3343031.3350928]

作者简介

王福田,男,讲师,主要研究方向为计算机视觉、语义分割和边缘计算。E-mail:wft@ahu.edu.cn

李成龙,通信作者,男,副教授,主要研究方向为视觉计算、深度学习和智能系统与应用。E-mail:lcl1314@foxmail.com

张淑云,女,硕士研究生,主要研究方向为 RGBT 标跟踪。E-mail:shyunzhang@163.com

罗斌,男,教授,主要研究方向为数字图像处理、模式识别和计算机应用技术。E-mail:ahu_lb@163.com