中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2023)05-1287-21

论文引用格式:Bie Q, Wang X, Xu X, Zhao Q J, Wang Z, Chen J and Hu R M. 2023. Visible-infrared cross-modal pedestrian detection: a summary. Journal of Image and Graphics, 28(05):1287-1307(别倩, 王晓, 徐新, 赵启军, 王正, 陈军, 胡瑞敏. 2023. 红外—可见光跨模态的行人检测综述. 中国图象图形学报, 28(05):1287-1307)[DOI:10.11834/jig.220670]

红外—可见光跨模态的行人检测综述

别倩[1,2],王晓[1,2*],徐新[1,2],赵启军[3],王正[4],陈军[4],胡瑞敏[4]

1. 武汉科技大学计算机科学与技术学院,武汉　430065; 2. 武汉科技大学智能信息处理与实时工业系统湖北省重点实验室,武汉　430065; 3. 四川大学视觉合成图形图像技术国家级重点实验室,成都　610065; 4. 武汉大学多媒体网络通信工程湖北省重点实验室,武汉　430072

摘　要:可见光图像在光照充足的条件下可以提供一系列辅助检测行人的信息,如颜色和纹理等信息,但在低照度场景下表现并不理想。红外图像虽然不能提供颜色和纹理信息,但红外图像根据热辐射差异成像而不依赖于光照条件这一特性,使其可以在低照度场景下有效区分行人区域与背景区域并提供清晰的行人轮廓信息。由于红外和可见光两种模态之间直观的互补性,同时使用红外和可见光图像的行人检测任务被认为是一个很有前景的研究方向,受到了广泛关注,大幅促进了在安防(如安全监控和自动驾驶)和疫情防控等领域应用的发展。本文对红外—可见光跨模态的行人检测工作进行全面梳理,并对未来方向进行深入思考。首先,该课题具有独特性质。可见光图像对应三通道的颜色信息而红外图像对应单通道的温差信息,如何在两种模态存在本质差异的前提下,充分利用二者的互补性是红外—可见光跨模态行人检测领域的核心挑战和主要任务。其次,近几年红外—可见光跨模态行人检测研究针对的问题可分为两类,即模态差异大和实际应用难。针对模态差异大的问题,可分为图像未对准和融合不充分两类问题。针对实际应用难的问题,又分为标注成本、实时检测和硬件成本3类问题。本文依次对跨模态行人检测的主要研究方向展开细致且全面的描述并进行相应的总结。然后,详细地介绍与跨模态行人检测相关的数据集和评价指标,并以不同的评价指标对相关方法在不同层面上进行比较。最后,对跨模态行人检测领域存在的且尚未解决的问题进行讨论,并提出对未来相关工作方向的一些思考。

关键词:跨模态行人检测;可见光图像;红外图像;深度学习;行人检测

Visible-infrared cross-modal pedestrian detection: a summary

Bie Qian[1,2], Wang Xiao[1,2*], Xu Xin[1,2], Zhao Qijun[3], Wang Zheng[4], Chen Jun[4], Hu Ruimin[4]

1. School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan　430065, China; 2. Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan　430065, China; 3. National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu　610065, China; 4. Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan　430072, China

Abstract: The precision of pedestrian detection is focused on instances-relevant location on given input images. However, due to the perception of visible images to light changes, visible images are challenged for lower visibility conditions like

收稿日期:2022-07-07;**修回日期**:2022-12-06;**预印本日期**:2022-12-13

*通信作者:王晓　wangxiao2021@wust.edu.cn

基金项目:国家自然科学基金项目(U1803262,62071338);省级大学生创新创业训练计划项目(S202210488063);湖北省资助项目(D20211106,2022CFB018);四川大学视觉合成图形图像技术国家级重点实验室开放课题项目(2021SCUVS003)

Supported by: National Natural Science Foundation of China(U1803262, 62071338)

extreme weathers. Hence, visible images-based pedestrian detection is merely suitable for the development of temporal applications like autonomous driving and video surveillance. The infrared image can provide a clear pedestrian profile for such low-visibility scenes according to the temperature difference between the human body and the environment. Under the circumstances of sufficient light, visible images can also provide more information-lacked in infrared images like hair, face, and other related features. Visible and infrared images can provide visual information-added in common. However, the key challenges of visible and infrared images is to utilize the two modalities-between and their modality-specific noise mutually. To generate temperature information, the difference is leaked out that the visible image consists of color information in red, green, and blue (RGB) channels, while the infrared image has one channel only. And, imaging mechanism-based wavelength range of the two is different as well. The emerging deep learning technique based cross-modal pedestrian detection approaches have been developing dramatically. Our summary aims to review and analyze some popular researches on cross-modal pedestrian detection in recent years. It can be segmented into two categories: 1) the difference between two different modalities and 2) the cross-modal detectors application to the real scene. The application of cross-modal pedestrian detectors to the actual scene can be divided into three types: cost analysis-related data annotation, real-time detection, and cost-analysis of applications. The research aspects between two modalities can be divided into: the misalignment and the inadequate fusion. The misalignment of two modalities shows that the visible-infrared image pairs are required to be strictly aligned, and the features from different modalities are called to match at corresponding positions. The inadequate fusion of two modalities is required to maximize the mutual benefits between two modalities. The early research on the insufficient fusion of two-modality is related to the study of the fusion stage (when to fuse) of two-modality. The later studies on the insufficient fusion of two-modality data are focused on the study of the fusion methods (how to fuse) of two-modality. The fusion stage can be divided into three steps: image, feature, and decision. Similarly, the fusion methods can be segmented into three categories: image, feature, and detection. Subsequently, we introduce some commonly used cross-modal pedestrian detection datasets, including the Korea Advanced Institute of Science and Technology (KAIST), the forward looking infrared radiometer (FLIR), the computer vision center-14 (CVC-14), and the low-light visible-infrared parred (LLVIP). Then, we introduce some evaluation metrics method for cross-modal pedestrian detectors, including missed rate (MR), mean average precision (mAP), and a pair of visible and thermal images in temporal (speed). Finally, we summarize the challenges to be resolved in the field of cross-modal pedestrian detection and our predictions are focused on the future direction analysis of cross-modal pedestrian detection. 1) In the real world, due to the different parallax and field of view of two different sensors, the problem of misalignment of visible-infrared modality feature modules is more concerned about. However, the problem of unaligned modality features is possible to sacrifice the performance of the detector and hinder the use of unaligned data in datasets, and is not feasible to the application of dual sensors in real life to some extent. Thus, the problem of two modalities' position is to be resolved as a key research direction. 2) At present, the datasets of cross-modal pedestrian detection are all captured on sunny days, and current advanced cross-modal pedestrian detection methods are only based on all-day pedestrian detection on sunny days. However, to realize the cross-modal pedestrian detection system throughout all day and all weathers, it is required to optimize and beyond day and night data on sunny days. We also need to focus on the data under extreme weather conditions. 3) Recent studies on cross-modal pedestrian detection are focused on datasets captured by vehicle-mounted cameras. Compared to datasets captured from the monitoring perspective, the scenes of vehicle-mounted datasets are changeable, which can suppress over-fitting effectively. However, the nighttime images in the vehicle-mounted datasets may be brighter than those of the surveillance perspective datasets because of their headlight brightness at night. Therefore, we predict that multiple visual-angles datasets can be used to train the cross-modal pedestrian detector at the same time. It can not only increase the robustness of the model in darker scenes, but also suppress over-fitting at a certain scene. 4) Autonomous driving systems and robot systems are required to be quick responded for detection results. Although many models have fast inference ability on GPU (graphics processing unit), the inference speed on real devices need to be optimized, so real-time detection will be the continuous development direction of cross-modal pedestrian detection as well. 5) There is still a large gap in cross-modal pedestrian detection technology for small scale and partial or severe occluded pedestrians. However, driving systems-assisted detection and occlusion can be as a very common problem in life for small targets of pedestrians at a distance to alert drivers to slow down in advance. The cross-modal pedestrian detection technology can be forecasted and recognized for small scale targets and occlusion as the direction of future research.

Key words: cross-modal pedestrian detection; visible image; infrared image; deep learning; pedestrian detection

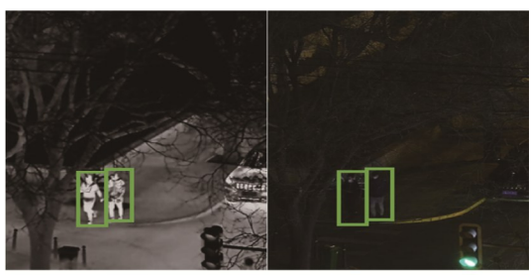
0 引 言

行人检测是计算机视觉领域最基本的任务之一，具有广泛的现实应用价值。行人检测是监控监视、监控追踪和自动驾驶等应用系统的核心技术，同时也是行人重识别、行人检索等任务的前提技术。近年来，基于深度学习方法的行人检测技术（Ouyang 和 Wang，2013；Tian 等，2015；Lin 等，2018；Li 等，2020b；Liu 等，2019；赵永强 等，2020；甑烨 等，2021）取得了很大进展，然而在光照不良条件下，可见光图像中的行人目标易与周围环境呈现一样的外观特征，因此仅基于单一可见光模态的行人检测器难以区分行人与周围背景，易造成漏检、误检等情况。

人脸识别（Lee 等，2016；He 等，2017；Wu 等，2021）、行人追踪（Farooq 等，2015）、行人重识别（Andreas 等，2013；Ye 等，2018；Vladimir 等，2019；李俊宏 等，2020；郑伟诗和吴岸聪，2018；赵才荣 等，2021；吴岸聪 等，2022；Li 等，2020a）和目标检测（Gupta 等，2014；Song 和 Xiao，2016；Deng 和 Latecki，2017；Xu 等，2018；蒋亭亭 等，2021；Fang 和 Wang，2022）等领域的许多研究表明，相较于单一可见光模态，引入一种新的模态可获得更好的效果。

在常见的可见光+深度图、可见光+雷达图以及可见光+红外图等跨模态输入组合中，雷达点云图提供的信息不足以将行人与物体区分开，深度图虽然可以提供行人轮廓信息但提供的轮廓不够准确。而热红外摄像机由于其根据温差成像的特性，相较于深度图和雷达图可以提供更清晰的行人轮廓。同时结合可见光与红外图像可以有效地解决在低照度场景（如黑夜、雾天、雨天等）下可见光图像表现较差的问题，并能在一定程度上解决行人遮挡和背景混淆等问题，如图 1 所示，其中，矩形框代表真实标注框。因此，相比于其他结合模态，结合红外—可见光两种模态在行人检测领域得到了更广泛的应用。

近几年，跨模态行人检测方面的研究依然备受关注。图 2 是以跨模态行人检测为关键词在谷歌学术搜索得到的 2015—2021 年发表的跨模态行人检测论文数量统计。可以看出，论文总体发表数量呈现上升趋势，虽然 2020 年的论文发表数量有所下降，但 2021 年呈现爆发式增长，推测是新冠疫情原因导致许多研究成果延后一年。结合可见光—红外两种模态的行人检测任务已逐渐成为并将持续成为研究者关注的热点任务。



(a) 红外图像　　(b) 可见光图像

图 1　在低照度情况下红外图像为可见光图像提供辅助信息的可视化图

Fig. 1　A visualization of an infrared image providing supplementary information to a visible image at low illumination ((a) infrared image; (b) visible image)

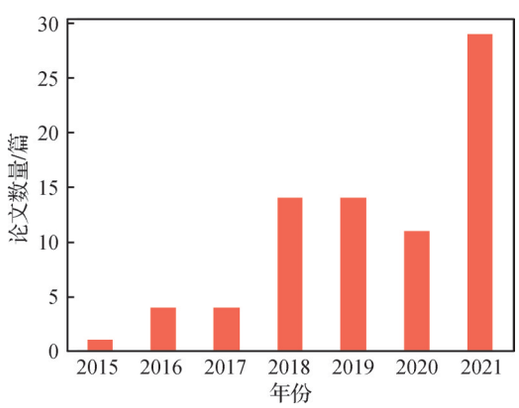


图 2　2015—2021 年跨模态行人检测论文数量统计

Fig. 2　The number of publications on cross-modal pedestrian detection from the year 2015 to 2021

本文的贡献可以总结为以下 3 点：1）与不同于 Li（2021）根据网络结构将近几年的跨模态行人检测工作分为单阶段检测和双阶段检测两类，本文根据研究针对的问题详细分类，更有利于对跨模态行人检测领域感兴趣的研究者快速了解该领域关注的问题和热点方向；2）基于不同的评价指标对近几年的跨模态行人检测方法在不同层面进行对比，提供了丰富的对比数据；3）总结了跨模态行人检测领域尚未完美解决的挑战，并提出对于未来跨模态行人检测领域方向的思考。这将有助于启发研究者实现更有价值的跨模态行人检测器。

1 跨模态行人检测

经过众多人工智能(artificial intelligence, AI)领域研究者的共同努力,跨模态行人检测在各方面都取得了显著性的成功。本文根据2007—2021年跨模态行人检测领域的研究热点和重要数据集的提出时间整理了该课题的发展脉络,如图3所示。

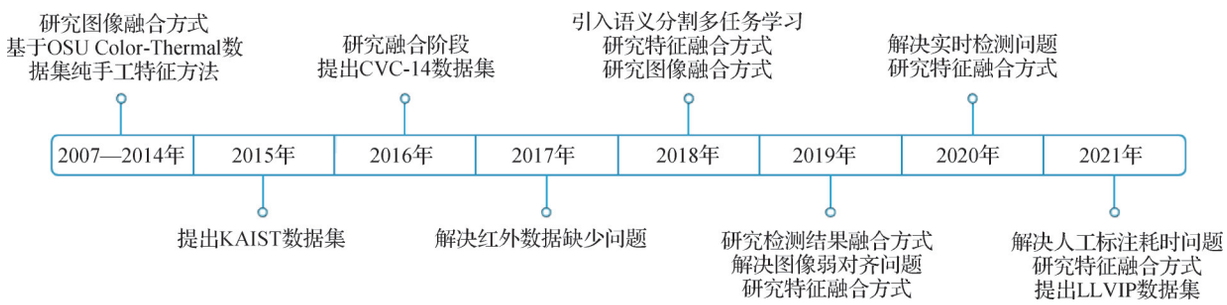


图3 跨模态行人检测领域历年的研究热点

Fig. 3 The hot research directions of cross-modal pedestrian detection in recent years

如何有效融合可见光和红外两种模态一直是备受关注的热点问题。在KAIST(Korea Advanced Institute of Science and Technology)数据集提出之前,面向红外—可见光图像的跨模态行人检测研究都是基于OSU Color-Thermal(Ohio State University Color-Thermal)数据集(Davis和Sharma,2007)。首先通过图像融合技术,其次通过背景减法在融合后的图像上产生候选区域,最后对这些候选区域进行分类判断有无行人。早期,对两模态图像融合方法的研究主要包括基于轮廓的融合方法(Davis和Sharma,2007)、基于联合双边滤波器的融合方法(Choi和Park,2010)和基于黎曼流形的融合方法(San-Biagio等,2012)等。对候选区域分类方法的研究主要包括基于周期性步态分析的分类方法(Leykin等,2007)等。2014年之前,面向红外—可见光的跨模态行人检测任务并未引起广泛关注,主要原因在于OSU Color-Thermal中的图像大多在光线较好的白天场景下收集,结合红外进行检测的优势并不明显。

Hwang等人(2015)通过以红外图像通道特征和红外图像的HOG(histogram of oriented gradient)特征作为附加通道特征扩展聚集通道特征,提出了多光谱ACF(aggregated channel features)行人检测器。实验证明,相比于仅用可见光图像训练的检测器,使用可见光—红外图像对的检测器在不同光照、不同距离和不同遮挡条件下的性能表现显著提高。随后,其提出的KAIST数据集广泛应用在跨模态行人检测领域,促进了面向红外—可见光图像的跨模态行人检测研究的再次兴起。

早期基于背景建模的方法只能检测运动的目标,对于静止的行人目标无法处理。手工特征方法可以检测静止的行人目标,但手工特征无法满足各种环境下对检测器高精度性能和鲁棒性的要求,且手工特征方法无法实现一个端到端的检测器。随着深度学习技术的推广,研究者发现基于深度卷积神经网络(convolutional neural network, CNN)的方法可以在行人检测任务中提取更深层次、更具有区分力度的行人特征。如图3所示,2016年及以后的所有关于跨模态行人检测任务的研究都是基于CNN模型上的改进。跨模态行人检测的基本流程图如图4所示,由于跨模态行人检测输入为红外—可见图像对,因此检测过程中通常需要进行融合操作。但不同的跨模态行人检测方法对应不同的融合阶段,因此不同方法对应的位置不同。所以融合操作模块并未在图4中展示,具体细节如图5所示。

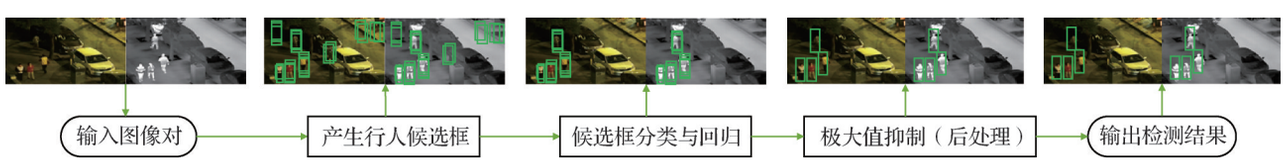
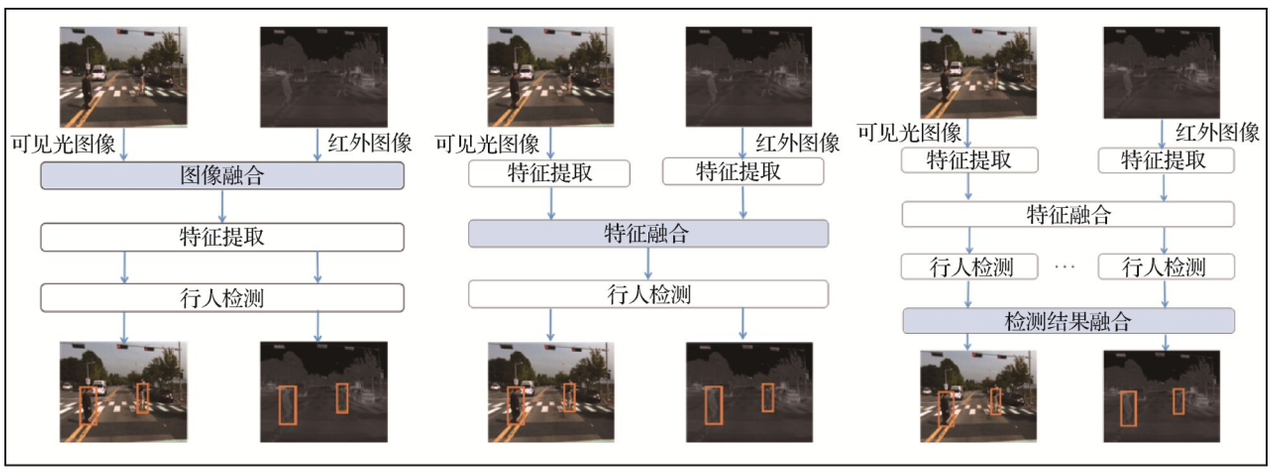


图4 跨模态行人检测的基本流程图

Fig. 4 Basic flow chart of cross-modal pedestrian detection



(a) 图像阶段融合　(b) 特征阶段融合　(c) 决策阶段融合

图 5　3 个融合阶段的可视化图

Fig. 5　A visualization of the three fusion stages ((a) image fusion stage; (b) feature fusion stage; (c) decision fusion stage)

针对近几年跨模态行人检测研究的问题，可分为模态差异大问题的研究和实际应用难问题的研究两类。针对模态差异大问题的研究旨在解决相对于单一可见光模态，新引入一个红外模态后出现的问题；针对实际应用难问题的研究旨在解决将红外—可见光跨模态行人检测器应用到现实生活场景中出现的问题。

1.1　基于模态差异大问题的研究

1.1.1　基于图像未对准问题的研究

图像未对准是指来自两种模态的图像之间存在位置偏移问题，两种图像在同一空间位置显示的信息不同。由于两模态图像是由红外—可见光传感器分别拍摄的，不同的传感器具有不同的视角和视野，并且双传感器同步拍摄较为困难，因此对于结合红外—可见光的行人检测任务，图像未对准问题是一个较为常见的挑战。在 2019 年之前，French 等人(2018)提出两模态图像融合之后进行特征提取和检测，但表现并不理想，主要是因为跨模态行人检测数据集中图像未对准的问题。Li 等人(2018)在清洗 KAIST 数据集时，发现两种模态图像之间的未对准问题会造成检测器的性能下降，因此重新标注 KAIST(Hwang 等，2015)数据集中的训练数据集。Zhang 等人(2019a)在 2019 年首次针对跨模态行人检测领域中两模态图像弱对齐问题进行系统地研究，以红外模态作为参考模态，可见光模态作为感知模态，提出区域特征对齐(region feature alignment，RFA)模块，用于解决两种模态间的图像弱对齐问题。通过两个模态真实行人标注框的中心坐标偏差算出两个模态的偏移量，根据模态间的偏移量动态地调整感知模态(可见光模态)，这个过程中参考模态(红外模态)是不被调整的。不同于 Zhang 等人(2019a)针对两种模态图像之间弱对齐问题的研究，Wanchaitanawong 等人(2021)旨在解决两模态图像之间存在较大偏差时的行人检测问题，提出基于两模态回归和两模态交并比(intersection over union，IOU)的检测方法，模型中的 RPN(region proposal network)网络和检测网络具有一个两模态的边界框回归器，用于独立调整边界框的位置。Wanchaitanawong 等人(2021)认为在两模态之间存在较大偏差时，两模态 IOU 不应仅计算一个统一的真实框和预测框的交并比，而应同时计算两种模态各自的真实框和预测框的交并比。实验表明，Wanchaitanawong 等人(2021)提出的方法在两种模态图像存在强未对齐情况下有较明显的性能提升。

1.1.2　基于融合不充分问题的研究

关于两种模态融合不充分的研究可以分为对两种模态融合阶段(何时融合)的研究和对两种模态融合方式(如何融合)的研究。早期对于两模态融合不充分的研究致力于融合阶段的研究，根据不同融合阶段所对应的融合信息不同，可将融合阶段的研究分为 3 类，分别是图像阶段融合、特征阶段融合以及决策阶段融合。同样，根据融合信息不同，可以将融合方式的研究分为 3 类，分别是对图像融合方式的研究、特征融合方式的研究以及检测结果融合方式

的研究。

1)图像阶段融合。如图5(a)所示,此时融合的信息为图像信息,融合操作在特征提取操作之前发生,这个阶段的融合称为图像阶段融合。在图像阶段进行融合操作,不需要图5(b)(c)所示的双分支网络结构,只需改变单分支网络的第1层卷积层输入通道数即可。

对于图像阶段融合方式的研究,Wagner等人(2016)提出的早期融合阶段中采取的图像融合方法是直接叠加型图像融合。对于直接叠加型图像融合,只需将单分支网络的第1层卷积层的输入通道数由可见光的三通道变为可见光+红外的四通道。Hou等人(2018)使用3种常见的图像融合方法,即Laplacian pyramid(Burt和Adelson,1983)、curvelet(Candès等,2006)和wavelet(Ranchin和Wald,1993)进行可见光和红外图像的融合。首先将RGB格式的可见光图像转为HIS(hue-intensity-saturation)格式,其次将可见光图像的I通道和热图像单通道分别使用3种图像融合方法进行融合,然后使用融合后得到的通道替换原可见光图像的I通道,最后再将经过替换后的HIS格式的可见光图像重新转化为RGB格式的可见光图像作为检测器的输入进行检测。相较于Wagner等人(2016)提出的早期、后期的融合方法,使用3种图像融合方法的行人检测器性能表现并不理想。Vandersteegen等人(2018)在基于红外图像单通道包含的信息比可见光图像的某个通道包含的信息更有价值的假设下,将RGB图像中的3个通道分别替换成热图像单通道,同时将RGB格式转化为LUV(L表示明亮程度,U、V表示图片色度)格式,将U或V两通道分别替换为热图像通道得到了5种不同的替换模型。实验结果表明,替换RGB格式中的R通道与替换LUV格式中的U通道可得到相同的效果。由于直接替换RGB中的R通道不需要格式转化,因此更推荐替换可见光图像中的R通道这一方式。French等人(2018)采用Connah等人(2015)提出的多光谱图像边缘融合算法融合红外—可见光图像,相比于Liu等人(2016a)提出的早期、中期和晚期融合算法,图像边缘融合算法性能表现并不理想。French等人(2018)分析图像边缘融合算法表现较差的原因在于数据集中的红外图像和可见光图像并没有完全对齐。本文也认为图像对未对准问题是图像融合方法检测性能较低的主要原因,因为像素级的图像融合比特征级的特征图融合对两模态图像对齐的要求更高。

2)特征阶段融合。如图5(b)所示,此时融合的信息为两种模态的特征信息,融合操作在特征提取过程中发生,这个阶段的融合称为特征阶段融合。由于图像阶段融合的行人检测效果并不理想,近几年的研究大都致力于两种模态特征阶段的融合。

对于单层特征图融合阶段的研究,Wagner等人(2016)提出晚期融合方法,两个子网络分别提取两种模态的特征,并在最后的全连接层融合两个分支提取出的不同特征。实验表明,其提出的属于特征阶段融合的晚期融合检测效果优于其同时提出的属于图像阶段融合的早期融合检测效果。Liu等人(2016a)对特征融合阶段进一步研究,提出4个不同的特征融合阶段,分别是早期融合、中期融合、后期融合和置信度融合。4个融合阶段都基于双分支的网络结构。其中置信度融合属于决策融合阶段的研究。早期融合是两个分支经过一个卷积层后进行特征融合;中期融合是两个分支经过4个卷积层后进行特征融合;后期融合是两个分支经过最后一个卷积层后进行特征融合。实验结果表明,中期融合阶段优于其他3个融合阶段,主要原因在于早期的特征图包含较为丰富的细节信息,后期特征图包含较为丰富的语义信息,而中期特征图不仅包含丰富的语义信息同时也包含细节信息。König等人(2017)为找到RPN中可见光和红外特征的最佳融合阶段,以产生更有效的候选区域,在Liu等人(2016a)提出的早期融合、中期融合和晚期融合的基础上,分别在第2层和第3层卷积层后对两模态特征进行融合,并通过实验对比不同特征阶段融合检测器的性能表现,实验结果表明,在第3个卷积层之后进行两模态特征图融合效果最好。

对于多层特征图融合阶段的研究,Choi等人(2016)认为仅融合单个特征层是不够的,浅层的特征分辨率高,但对于行人特征识别能力差。然而随着卷积层的加深,特征图对行人特征的识别能力增加,特征分辨率随之下降。因此,Choi等人(2016)提出通过多尺寸特征图融合来平衡不同卷积层之间特征识别能力和分辨率。不同于基于R-CNN(region convolutional neural network)网络系列的检测器,Zheng等人(2019)设计了基于双SSD(single shot detector)的跨模态行人检测器(Liu等,2016b),采用

SSD模型中6个不同尺寸特征图检测不同尺度目标的思想，分别将两分支的特征图在conv4_3，conv7，conv8_2，conv9_2，conv10_2，conv11_2共6个不同尺寸的卷积层进行融合。在此网络模型基础上，Zhuang等人(2022)对在输入通道融合两模态特征、conv8_2层之后融合两模态特征和Zheng等人(2019)提出的6层融合3种阶段下检测器的表现进行比较，实验结果证明，对于双分支SSD结构而言，6层融合检测的效果最好。

常见的特征融合方式包括4种方式，分别是平均融合、级联融合、增强融合和加权融合，如图6所示。

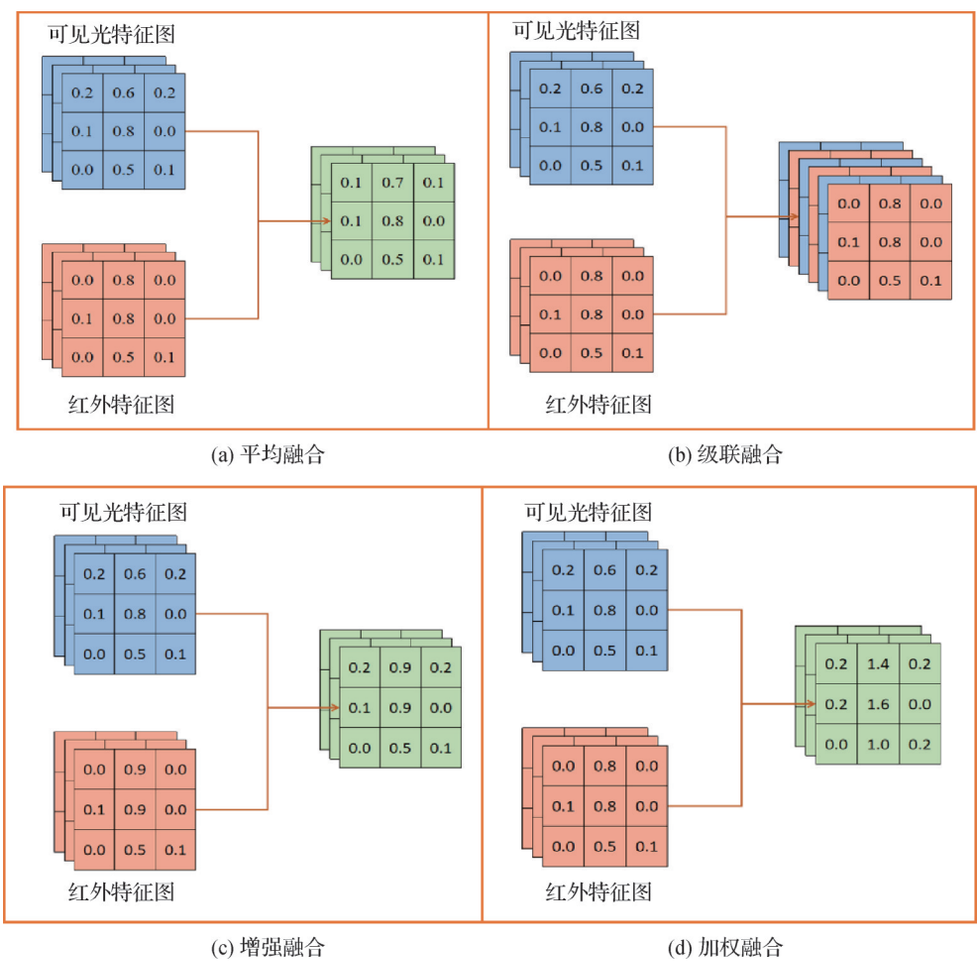


图6　4种常见融合方式的可视化图

Fig. 6　A visualization of the four fusion methods

((a) average fusion; (b) concatenation fusion; (c) maximization fusion; (d) addition fusion)

Liu等人(2016a)提出使用级联的方式进行两种模态特征图的融合，如图6(b)所示，但两种模态特征图的直接级联操作会导致通道数加倍，为能够继续使用预训练的网络模型参数，在连接层后面引入NiN(network-in-network)层以减少通道数，这种级联融合方式在之后的研究中得到广泛使用。不同于单独使用级联融合方式，Lee等人(2018)引入特征加强的操作，使红外特征图和可见光特征图中强度较高的部分会在联系特征图中也得到强调。首先利用Hadamard积方法融合输入的可见光特征图和红外特征图得到两种模态的关联特征图，其次将两种模态的特征图和关联特征图级联得到最终的融合特征图。Guan等人(2018)组织对比实验，以比较级联融合(图6(b))、增强融合(图6(c))和加权融合(图6(d))3种特征融合方法的性能，经实验表明，加权融合方法的效果最佳，增强融合方法其次，而最常用的级联融合方法检测效果最差。

为充分利用两种模态之间的互补性，Zhang等人(2019b)首次将注意力机制应用于跨模态行人检测，提出跨模态交互式注意力网络(cross-modality

interactive attention network, CIAN)。首先,利用可见光和红外两个分支分别提取两种模态的特征。其次,通过获取两个模态的全局特征将两个模态之间的相关性编码于注意力模块。最后,以编码得到的重要信息自适应地调整两种模态的融合权重,融合后特征图中的有效信息会更加突出,而无效信息将被抑制。Zheng等人(2019)提出以门控融合单元(gated fusion unit, GFU)的思想融合来自于两种模态的特征图。GFU基于门控机制的运行方式,在保证输出融合特征图通道数与输入特征图通道数相同的前提下,可以保留两种模态的有效特征,同时降低无效特征的影响。Zhang等人(2019a)以红外模态为参考模态,可见光模态为感知模态,提出基于置信度融合的方法。对于某一行人实例,若感知模态预测该候选区域为行人的概率与参考模态预测其为行人的概率相差较大,则抑制感知模态的特征。Zhang等人(2020a)为保持不同光谱特征之间的一致性,提出以循环融合方式动态调整模态特征的融合过程(cyclic fuse-and-refine, CFR),认为融合后的双光谱特征通常比单光谱特征更具有判别性,使用融合后的双光谱特征优化单光谱特征,可使经过优化的单光谱特征相较于原光谱特征更具有判别性。但由于连续多次使用融合的多光谱特征来调整单光谱特征,两种模态之间的互补性会随着一致性增加而降低,所以通过控制循环优化的次数来实现两种模态之间一致性与互补性的平衡,实验结果表明,经过3次循环优化的检测效果最好。Zhou等人(2020)发现可见光—红外两种模态都有其内在特质,两种模态中都存在有效信息及噪音信息(可见光图像在低照度条件下难以区分周围背景和行人,红外图像在高温条件下难以区分周围背景和行人),仅使用简单的线性融合策略未能充分利用两模态之间的互补性。因此,Zhou等人(2020)提出使用差分模态感知(differential modality aware fusion, DMAF)模块,DMAF充分利用两种模态之间的互补性,从而可以产生更有效的融合特征表示。Zhang等人(2021a)提出由语义信息引导的类内和类间注意力机制特征融合模块(guided attentive feature fusion, GAFF)。在两模态特征融合阶段,两种注意力机制将由行人的语义信息引导,类内注意力机制更关注对应模态内的有效信息,因此可见光和红外特征图中含有行人的区域会加强;类间注意力机制更关注两个模态间的有效信息,因此两种模态中预测行人信息更准确的模态将得到更多的信任。Yang等人(2022)提出基于注意力机制的双向自适应注意力融合模块(bi-directional adaptive attention gate, BAA-Gate),注意力机制用于抑制两个模态中的噪声信息同时挑选两个模态间的有效信息。经过注意力机制挑选后的红外模态特征再经过光照加权以用于调整可见光模态特征,同时可见光模态特征也以相同的方式用于调整红外模态特征。与其他基于注意力机制的融合方法不同,Yang等人(2022)将光照权重与注意力机制相结合,光照权重不仅用于两种模态检测结果的融合,而且用于适应性地调整两个模态特征之间校准和融合的强度。Kim等人(2022a)提出基于感兴趣区域(region of interest, RoI)不确定性和预测不确定性的跨模态行人检测框架。针对RoI的不确定性,在结合两种模态特征时,设计不确定性感知特征融合(uncertainty-aware feature fusion, UFF)模块,以减少融合过程对不确定性RoI的特征影响;针对模态差异,提出不确定性感知的跨模态引导(uncertainty-aware cross-modal guiding, UCG)模块,使用预测不确定性表示各模态中的RoI预测的可靠性,引导高不确定性模态的特征分布靠近低不确定性模态的特征分布。

3)决策阶段融合。如图5(c)所示,决策阶段融合一般发生在检测结果产生之后,决策融合阶段融合的信息为不同分支的检测结果。

常见的检测结果融合方式为平均融合(两模态检测结果的权重都为0.5)。Liu等人(2016a)提出的置信度融合是两个分支网络的级联,首先,两个子网络分别生成候选区域和置信度。其次,可见光模态网络的输出会作为输入送入红外模态网络,红外模态分支网络的输出也会作为输入送入可见光模态网络以重新计算置信度。最后,通过平均融合方式合并两模态网络得到检测结果。Li等人(2019)提出置信度级的非级联融合,候选区域来自于两种模态的第5个卷积层后的融合特征图,然后两个子网络分别以生成的候选区域作为输入生成分类和边界框预测的结果,最后将两个子网络产生的结果平均融合得到最终的检测结果。

但在不同的环境下,两种模态检测结果的可靠性不一致,此时应根据不同的场景自适应地调整不同模态的检测结果权重。引入辅助因素融合检测结

果是较为有效的融合方式。引入辅助因素融合检测结果是指网络根据引入的光照、温度等条件计算出不同分支的检测结果所对应的权重参数，将计算出的权重参数用于不同分支检测结果的融合。Li等人(2019)研究发现，在光照良好条件下，相于比单一模态的行人检测，可见光图像和红外图像融合检测的效果较好；在光照不良条件下，相比于使用两种模态融合检测，单独使用红外图像检测的效果更好。因此，提出采用光照感知网络(illumination aware network，IAN)预测图像光照，从而计算得到两种模态在不同光照下的权重以适应性地调整对每一种模态检测结果的信任程度。与Li等人(2019)将光照权重用于两种模态检测结果的融合不同，Guan等人(2019a)提出将光照权重用于白天和夜晚两个子网络检测结果的融合。两者的区别在于Li等人(2019)认为在白天场景下可见光网络的检测结果应该赋予更大的权重，在夜晚场景下红外网络的检测结果应该赋予更大的权重。而Guan等人(2019a)认为在白天场景下白天分支网络的检测结果应该有更大的权重，在夜晚场景下夜晚分支网络的检测结果应该有更大的权重。在引入光照条件因素的基础上，Zhuang等人(2022)提出在外界温度高(如夏季的白天)的场景下，人体与外界环境温差不大，红外图像并不能提供清晰的行人轮廓。应根据温度和光照条件共同决定更信任哪一模态检测结果，从而进一步引入温度条件结合光照条件进行两模态检测结果的融合权重计算。

1.2 基于实际应用难问题的研究

1.2.1 基于标注成本问题的研究

在实际应用中，将跨模态行人检测器应用到某一新场景时，其性能可能由于原训练场景和当前目标场景之间的差异而显著下降。为使检测器在不同的场景下都保持较好的性能，需要将检测器应用到目标场景的数据集上进行训练使其学习目标场景的特有信息。然而目标场景是多变的，如自动驾驶的汽车一天可能会经过各种各样的场景，而不同场景光照不同、行人数量不同、行人遮挡程度不同、空旷程度不同以及天气不同等。对各种目标场景的数据集标注是一个耗费人力的过程，且为了训练跨模态行人检测器在不同光照条件下的鲁棒性，跨模态行人数据集一般是在全天候环境下拍摄的，对于数据集中较暗的场景下拍摄的可见光图像，人眼也难以区别图像中是否存在行人，这会造成许多误标和漏标的情况。为解决数据标注成本问题，Cao等人(2019a)提出将训练好的可见光行人检测器用于为跨模态行人数据集标注边界框，并将自动标注过后的跨模态数据用于跨模态行人检测器中进行训练，从而实现无监督的跨模态行人检测而无需人工标注跨模态行人数据。不同于Cao等人(2019a)使用训练好的可见光检测器，Guan等人(2019b)使用训练好的跨模态检测器，先在一种跨模态数据中学习，再将检测器迁移到另一种跨模态数据中，迁移的过程不需要人工的标注。与上述仅使用可见光检测器或使用跨模态行人检测器不同，Lyu等人(2021)首先将两种检测器分别在可见光和红外数据集上进行训练，其次利用训练好的两种检测器分别为跨模态行人数据集中的可见光和红外图像生成伪标签，最后为使生成的伪标签更加可靠，利用光照权重来确定两种模态检测器生成的伪标签在不同光照条件下的优先级。以上工作都是以无监督迁移学习的方法解决人工标注耗时的问题，Zhang等人(2021b)提出使用弱监督学习的方法解决此类问题，即采用少量人工标注的标签。首次将主动学习策略应用于跨模态行人检测领域，首先，随机挑选一部分样本进行人工标注后作为输入数据送入检测器进行训练；其次，检测器利用两个模态图像间的互补性以选取有效的样本；最后，将检测器选取的有效的样本进行人工标注，并将标注好的样本继续送入检测器训练。重复上述过程，直至当前检测器可以达到与全监督的跨模态行人检测器相同的性能后结束循环。

1.2.2 基于硬件成本问题的研究

在监控领域中，道路监控大多是可见光监控。若将原有道路监控都替换为可见光—红外一体监控所需成本较昂贵，且获取严格对齐的双光谱图像通常需要特殊校准板对双光谱图像进行校准，这意味着更高昂的成本。目前主流的方法尝试采用已有的跨模态数据训练检测器，而测试检测器仅在单一模态上进行。如果检测器的表现可达到与可见光和红外图像数据共同测试的效果，跨模态行人检测器就可利用现有的可见光或红外摄像机而不用使用特定的可见光—红外一体摄像机，这将大幅节省成本并且不用考虑双光谱数据的未对准问题。Xu等人(2017)提出采用生成伪红外的思想解决此类问题。首先，第1个网络输入为可见光图像和红外图像，该

网络学习从可见光图像到红外图像的非线性映射关系;然后,将第1个网络学习到的参数作为第2个网络中的第2个分支的参数。第2个网络中的第1个分支学习可见光图像的特征,第2个分支学习可见光图像生成的伪红外图像的特征;最后,将两个分支分别学习到的特征融合用于行人检测。Kruthiventi等人(2017)采用教师—学生网络结构解决此类问题。首先,教师网络由来自于两种模态的数据共同训练,学习来自于两种模态的特征;然后,由训练好的教师网络指导输入仅为可见光图像的学生网络,既学习可见光图像特征,又学习可见光图像中类似于热图像的特征。Liu等人(2022)同样采用教师—学生网络结构,但不同于Kruthiventi等人(2017)设计的教师网络简单的将两种模态进行融合学习,为更明确地探索两种模态之间特有和共有的联系,提出基于通道分离和融合的跨模态特征学习(cross-modal feature learning, CFL)模块,同时,为使学生网络能更有效地从训练好的教师网络中学习跨模态知识,从特征、检测和分割3个角度设计不同的蒸馏损失项。以上研究都使用红外与可见光图像进行训练,测试仅使用可见光图像。Kim等人(2022b)在以上研究的基础上,构建可分别使用可见光或红外图像测试的多功能行人检测器,设计多传感器匹配对比损失以引导网络在不考虑输入模态的情况下能产生相似地视觉表示,并设计多光谱回忆(multi spectral recalling, MSR)存储器用于储存两种模态的视觉特征表示,根据输入模态的特征与MSR中的键相似度,从MSR中取出对应的值,同时设计多光谱回忆损失以引导MSR有效记住两个模态的上下文信息。然而,仅基于单一模态图像测试的方法,与同时使用红外与可见光数据测试的方法仍然存在较大差距。对于分辨率较高的红外热像仪,其与多架无人机结合使用,成本是无法估计的。相比于车载和监控行人检测领域尝试使用单一模态通过某种变换代替双模态,机载行人检测领域意图使用分辨率较低的红外热像仪与可见光相机结合,并通过特定处理以达到与使用高分辨率红外热像仪相同的效果。de Oliveira和Wehrmeister等人(2016)将低成本的热像仪图像与用于分类任务的CNN相结合,以检测候选对象,需要的红外图像仅为227×227像素。实验结果表明,低分辨率的热图像在大多数情况下可以产生较为准确的行人候选框,并且能够在一定程度上提升检测器速度。在此研究基础上,de Oliveira和Wehrmeister等人(2018)将Haar级联(基于Haar特征的级联分类器)、LBP(local binary pattern)级联、HOG + SVM(support vector machine)和卷积神经网络(CNN)4种机器学习技术用于特征提取和分类,将显著图技术(saliency map, SM)和低分辨率热图像处理技术(thermal image processing, TIP)用于产生行人候选框,在8种自由组合中,经过实验证明,通过将CNN与TIP结合,可以达到最好的行人检测性能。Beleznai等人(2018)认为红外图像在周围环境温度较高时提供的行人轮廓并不清晰,而深度图像中偶尔存在无效像素出现于无纹理区域。因此提出使用热红外和深度图数据生成行人候选框,再通过两个分别训练的基于CNN的分类器评估生成的候选框,最后将保留较好的候选框对象。经实验结果表明,以此互补方式快速生成行人候选框的方法能够在有限的成本下提供较好的行人检测结果。

1.2.3 基于实时检测问题的研究

自动驾驶系统需要在检测到行人后快速做出一系列反应从而避免意外事故的发生,所以不论是一般的行人检测领域还是跨模态行人检测领域对行人检测器的速度都有较高的要求。虽然使用两阶段网络的跨模态行人检测器精度较高,但自动驾驶系统不仅要有较高的精度同时也需要较快的检测速度,所以目前对实时检测问题的研究都致力于采用单阶段网络结构设计跨模态行人检测器,旨在保持较高的检测速度条件下达到与两阶段检测器相同或更高的精度。单阶段检测器省略生成候选区域的过程,将行人定位任务和分类任务在同一个网络中完成,从而在速度上实现优于两级检测器的性能。Vandersteegen等人(2018)提出在不改变检测速度较高的YOLOV2(you only look once V2)(Redmon和Farhadi, 2017)网络结构下,直接将可见光图像三通道中的某一个通道替换为红外图像的单通道,采用替换后的图像作为YOLOV2检测器的输入。替换图像通道的操作耗时可忽略不计,基于此方法的跨模态行人检测器可以达到与传统可见光行人检测器一样的速度。Zheng等人(2019)使用速度和精度都较好的SSD(Liu等, 2016b)网络设计基于双分支SSD的跨模态行人检测网络。实验证明,双分支SSD检测网络速度相比于两阶段检测网络速度有较大提升。之后,双分支SSD网络结构在许多研究中广泛

使用。比如Zhou等人(2020)设计的MBNet(modality balance network)，Zhang等人(2020a)设计的CFR(cyclic fuse-and-refines)等。Cao等人(2019b)发现同一检测器处理低分率输入图像的速度相较于处理高分辨输入图像的速度会有较大提升。但使用边界框定位行人的跨模态行人检测器在对分辨率较低的输入图像检测时，很难产生分类为正样本的边界框。为在提高跨模态行人检测器检测速度的同时保证检测精度，提出采用更适用于处理低分辨率输入图像的盒级分割技术定位行人，如图7所示。盒级分割技术可以分割出大致的行人轮廓，从而减少边界框中的无效区域，增多正样本的数量。本文认为，以使检测器可以有效处理低分辨输入图像为出发点，提高检测器的速度，该研究提供了解决实时检测问题的另一种思路。

图7　盒级分割技术的可视化图

Fig. 7　A visualization of the box-level segmentation technique

由于检测耗时时间较长将阻碍行人检测方法在无人机系统上的应用，为实现机载图像中的人员快速识别，de Oliveira和Wehrmeister等人(2016)提出首先使用热红外图像产生较为数量较少且较为准确的候选框，以解决滑动窗口方法产生行人候选框较多且耗时较长的问题，其次将生成的候选框映射到热红外图像对应的可见光图像中，最后使用CNN分类器对候选框进行分类和回归。采用热图像生成候选框，能够减少搜索空间，在保证准确率的同时具有可接受的计算性能。Golcarenarenji等人(2021)提出修改YOLOV3(Redmon和Farhadi，2018)模型结构，使用扩展卷积替换原始模型在自底向上路径中的最后两个池化层，以提高获得的特征图分辨率，从而获得更多关于小目标行人的信息，提高远距离小目标行人检测结果的精度。实验结果表明，其提出的模型与YOLOV3模型检测准确率相当，但速度加倍。Shao等人(2022)认为模型庞大、内存有限和计算能力有限3种因素限制了空中行人检测算法在无人机平台上的部署，因此提出将网络修剪算法与YOLOV3检测算法相结合，以获得修剪后的YOLOV3模型。实验结果表明，与原始的YOLOV3模型相比，修剪后的模型体积缩接近95.5%，计算量减少约2/3，而准确率仅降低1.7%。

2　数据集和评价指标

2.1　数据集

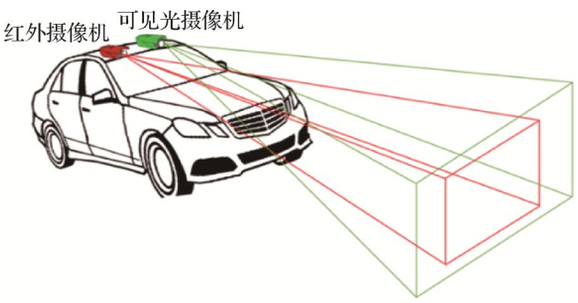
相比于单一模态的行人检测数据集，跨模态行人检测数据集由可见光—红外图像对组成，每一对图像由可见光摄像机和红外摄像机同时在同一地点分别拍摄的图像组成。

OSU Color-Thermal(Davis和Sharma，2007)数据集是在2007年收集并发表的数据集。此数据集在大学校园中一个繁忙的十字路口拍摄，包含6个序列，前3个序列在同一地点拍摄，后3个序列在另一地点拍摄，图像中包含大量行人。数据集中图像为白天场景下采集，红外图像的优势并不突出。

KAIST(Hwang等，2015)数据集是目前跨模态行人检测领域最受欢迎并且应用最广泛的数据集。该数据集由在校园、街道以及乡下等各种常规交通场景中捕获的全天可见光—红外热图像对组成。共包括95 328幅图像和1 182个行人，每幅图像都由可见光图像和红外图像对组成。包含person、people和cyclist共3个类别，比较容易区分的个体标注为person，不易区分的个体标注为people，骑行者标注为cyclist。此外，根据行人遮挡情况，KAIST数据集可分为3类，即无遮挡、部分遮挡(≤50%行人区域被遮挡)和严重遮挡(>50%的行人区域被遮挡)。随着KAIST数据集的广泛应用，KAIST数据集的修改提升版逐渐提出。Liu等人(2016a)提出消除错误边界框标注后的提升版测试数据集。Li等人(2018)提出净化版本的训练数据集，消除了原始训练数据集中的一些错误注释。Zhang等人(2019a)为解决可见光—热图像对之间的不对齐问题，分别为每个模态重新标记行人并建立联系，提出了新的配对版KAIST数据集。

CVC-14(computer vision center-14)(González等，2016)数据集是由车载双摄像机昼夜在各种场景下拍摄的，该数据集中的可见光图像为灰度图。根据拍摄时间段，该数据集可分成白天和夜晚两个子序

列。对于训练数据集，白天和夜间子序列分别包含3 695和3 390个可见光—红外图像对。对于测试数据集，白天和夜晚子序列分别包含706和727个可见光—红外图像对。如图8所示，可见光和红外摄像机的视野和分辨率存在不一致问题。红外图像分辨率为640 × 512像素，可见光图像分辨率为1 280 × 1 024像素。可见光摄像机的视野更广，拍摄到的范围更大，可以捕捉位于车两旁的行人。红外摄像机的视野较小，主要集中于车的正前方，无法捕捉位于车两旁的行人。因此，González等人(2016)对两种模态图像都单独提供注释，使用CVC-14的数据集时，需要手动进行空间位置对齐和裁剪。



(a) 双摄像机示意图

(b) 一对示例图像

图8　CVC-14数据集的相关图像

Fig. 8　Related images of the CVC-14 dataset ((a) fixed position diagram of two cameras; (b) a pair of sample images)

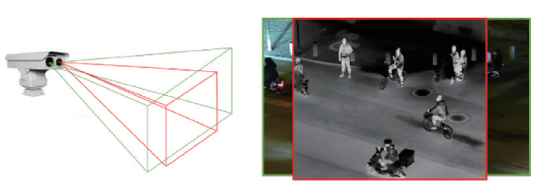
Utokyo(University of Tokyo)(Takumi等，2017)数据集包括7 512组图像，其中白天图像3 740组，夜晚图像3 772组，由可见光、远红外、中红外和近红外4种摄像机在大学校园中拍摄，包含6 066组图像未对齐的训练图像和1 466组图像对齐的测试图像。目前使用UTokyo数据集的跨模态行人检测方法大都只使用对齐的测试集图像。

FLIR(forward looking infrared radiometer)(F. A. Group，2018)数据集是由车载可见光和红外相机捕获，包含10 000个可见光—红外图像对，只有红外图像经过标注。Zhang等人(2020a)对FLIR数据集进行清洗，保留了FLIR数据集中最常用的人、汽车、自行车3个类别，移除了未对齐的可见光—红外图像对。清洗后的FLIR训练数据集和测试数据集分别包含4 129和1 013个对齐的可见光—红外图像对。

DIML(digital image media laboratory)(Park等，2018)数据集是在2018年提出的基于室内监控场景的双光谱数据集，包含1 003个对齐的可见光—红外图像对以及1 792个行人注释。红外图像和可见光图像分辨率都是640 × 480像素。

LLVIP(low-light visible-infrared paired)(Jia等，2021)数据集是在2021年提出的基于弱光场景的可见光—红外图像对数据集，包含30 976幅图像(15 488对)，每幅图像都包含行人。由于数据集的拍摄时间段在18时—22时之间，因此大部分图像都处于较暗的场景。如图9所示，该数据集是由可见—红外一体双目摄像机所拍摄的，原始拍摄的图像也存在视野和分辨率不同的问题。但LLVIP数据集已经过裁剪和校准处理，因此数据集中的每个图像对严格意义上在空间和时间都是对齐的。

本文对现有数据集在不同层面上进行了详细对比，如表1所示。



(a) 双目摄像机示意图　(b) 校准前的一对示例图像

图9　LLVIP数据集的相关图像

Fig. 9　Related images of the LLVIP dataset ((a) schematic diagram of the binocular camera; (b) a pair of sample images before calibration)

2.2　评价指标

错误的检测结果可能会使控制决策系统发出错误的信号。漏检率(miss rate，MR)是跨模态行人检测领域最常用评价指标之一，计算为

$$MR = 1 - \frac{N_{tp}}{N_g} \tag{1}$$

式中，N_{tp}表示预测为正样本且预测结果为正确的边界框数量，N_g表示真实边界框的数量。MR表示正样本未被检测到的概率，可表示为

表 1　跨模态行人检测数据集对比

Table 1　Comparison of datasets for cross-modal pedestrian detection

数据集	分辨率/像素	拍摄时间段	拍摄场景	摄像机视角	是否对齐	提出年份
OSU Color-Thermal	320×240	白天	室外	道路监控	是	2007
KAIST	640×512	全天	室外	车载摄像头	是	2015
CVC-14	640×512	全天	室外	车载摄像头	否	2016
UTokyo	320×256	全天	室外	车载摄像头	测试集对齐	2017
FLIR	640×512	全天	室外	车载摄像头	否	2018
DIML	640×480	白天	室内	房间监控	是	2018
LLVIP	1 080×720	18:00—22:00	室外	道路监控	是	2021

$$MR = \frac{N_{\mathrm{fn}}}{N_{\mathrm{g}}} \tag{2}$$

式中，N_{fn}表示预测为负样本但真实为正样本的边界框数量。

通过设定不同的置信度阈值，可以得到以每幅图像的平均误检个数（false positives per image，FPPI）为 X 轴，MR 为 Y 轴的 FPPI 曲线。其中，MR 为固定 FPPI 值对应的最小 MR。FPPI 计算为

$$FPPI = \frac{N_{\mathrm{fp}}}{N_{\mathrm{p}}} \tag{3}$$

式中，N_{fp}代表被预测为正样本但预测结果错误的边界框数量，即误检或虚检个数。N_{p}代表总图像的个数。

采用 $FPPI$ 曲线，均匀选取$[10^{-2}, 10^{0}]$范围内的 9 个 $FPPI$，得到对应的 9 个 MR 的值，则平均对数漏检率 MR^{-2} 为

$$MR^{-2} = \mathrm{e}^{\frac{1}{9}\sum_{i=1}^{9}\ln(MR_i)} \tag{4}$$

除跨模态行人检测器模型的准确度外，检测器的速度（speed）也是一个重要的评价指标。这里的速度指的是检测器处理一个可见光—红外图像对所需要的时间。一般来说，参数越多，模型复杂度越高。所以，除使用推断时间评估检测器模型的速度外，模型参数数量是另一个较为有效的评估指标。

2.3　前沿方法比较和分析

以 MR^{-2} 为评价指标时，MR^{-2} 值越低越好。为保证公平地比较，检测器都使用 KAIST 数据集中的原始训练集进行训练，在 Liu 等人（2016a）提出的提升版 KAIST 测试集进行测试。

表 2 为国内外前沿方法以 MR^{-2} 为指标在 KAIST 数据集中的无遮挡、部分遮挡（行人身体部分 ≤ 50% 被遮挡）和严重遮挡（行人身体部分 > 50% 被遮挡）以及近距离（车与人距离 ≤ 11 m）、中等距离（车与人距离在 11～28 m 之间）和远距离（车与人距离 ≥ 28 m）6 个测试数据集上的比较结果。可以看出，AR-CNN（area-region convolutional neural network）、MBNet 和 BAANet（bi-directional adaptive attention network）解决了可见光—红外图像对之间的弱对齐问题，并且采用较好的融合方式，在近距离的行人检测任务上 MR^{-2} 达到 0.00%。CIAN、MBNet 和 BAANet 等 3 种方法借鉴 SSD 多尺度检测的思想，将不同尺寸的特征图的检测结果融合，更有助于小目标行人的检测，在 KAIST 的远距离测试数据集上性能有较大提升。

表 3 为国内外前沿方法以 MR^{-2} 为指标在 CVC-14 数据集中的全天、白天、黑夜 3 个测试数据集上的比较结果。由于 CVC-14 数据集中图像未对准问题较为严重，相比于 KAIST 数据集，跨模态行人检测器在 CVC-14 数据集上性能表现较低。

表 4 为 19 种先进的跨模态行人检测方法以 MR^{-2} 为指标在 KAIST 数据集中的全天、白天、黑夜 3 个测试数据集上，仅考虑在无遮挡或部分遮挡且身高像素大于 50 像素的行人的比较结果。从表 4 可以看出，每种方法对于夜间图像和白天图像的行人检测结果都存在着一定的差异，有的方法对于白天图像的检测效果比较好，而有的方法相反。本文发现，CMT-CNN（cross-modality transfer convolutional neural network）、Halfway Fusion、IAF R-CNN（illumination-aware faster region convolutional neural network）、IATDNN+ IASS（illumination-aware two-stream deep convolutional neutral networks + illumination-aware semantic segmentation）和 MSDS-RCNN（multispectral simultaneous detection and segmentation-region convo-

表2　前沿方法在KAIST无遮挡、部分遮挡、严重遮挡以及近距离、中等距离、远距离测试数据集上的MR^{-2}性能比较

Table 2　Comparison of log-averaged miss-rate of the state-of-the-art methods on no-occlusion, partial-occlusion, heavy-occlusion, near-scale, medium-scale, and far-scale test sets of KAIST dataset

/%

方法	无遮挡	部分遮挡	严重遮挡	近距离	中等距离	远距离
ACF+T+THOG(Hwang等,2015)	62.94	81.40	88.08	28.74	53.67	88.20
Halfway Fusion(Liu等,2016a)	43.13	65.21	74.36	8.13	30.34	75.70
Fusion RPN+BDT(König等,2017)	47.45	56.10	72.20	0.04	30.87	88.86
IAF R-CNN(Li等,2019)	40.17	48.40	69.70	0.96	25.54	77.84
IATDNN+IASS(Guan等,2019a)	45.43	46.25	64.57	0.04	28.55	83.42
CIAN(Zhang等,2019b)	30.31	41.57	62.48	3.71	19.04	55.82
MSDS-RCNN(Li等,2018)	29.86	38.71	63.37	1.29	16.19	63.73
AR-CNN(Zhang等,2019a)	31.40	38.63	**55.73**	**0.00**	16.08	69.00
MBNet(Zhou等,2020)	27.74	35.43	59.14	**0.00**	16.07	55.99
BAANet(Yang等,2022)	**25.15**	**34.07**	57.92	**0.00**	**13.72**	**51.25**

注:加粗字体表示各列最优结果。

表3　前沿方法在CVC-14全天、白天、黑夜测试数据集上的MR^{-2}性能比较

Table 3　Comparison of log-averaged miss-rate of the state-of-the-art methods on all, day, and night test sets of CVC-14 dataset

/%

方法	全天	白天	黑夜
ACF + T + THOG(Hwang等,2015)	60.10	61.30	48.20
CNN + SVR(Choi等,2016)	47.30	49.30	43.80
Halfway Fusion(Liu等,2016a)	37.00	38.10	34.40
CWF + APF(Park等,2018)	31.40	31.80	30.80
AR-CNN(Zhang等,2019a)	22.10	24.70	18.10
MBNet(Zhou等,2020)	21.10	24.70	13.50
TVPD(Kim等,2022b)	19.88	**23.69**	12.35
UFF+UCG(Kim等,2022a)	**18.70**	23.87	**11.08**

注:加粗字体表示各列最优结果。

表4　前沿方法在KAIST全天、白天、夜间测试数据集上的MR^{-2}性能比较

Table 4　Comparison of log-averaged miss-rate of the state-of-the-art methods on all, day, and night test sets of KAIST dataset

/%

方法	全天	白天	黑夜
ACF + T + THOG(Hwang等,2015)	47.24	42.44	56.17
CMT-CNN(Xu等,2017)	36.83	34.56	41.82
CWF+APF(Park等,2018)	31.36	31.79	30.82
Halfway Fusion(Liu等,2016a)	26.15	24.85	27.59
Fusion RPN + BD(König等,2017)	16.53	16.39	18.16
IAF R-CNN(Li等,2019)	16.22	13.94	18.28
IATDNN+IASS(Guan等,2019a)	15.78	15.08	17.22
RFA(Zhang等,2019a)	14.61	16.78	10.21
CIAN(Zhang等,2019b)	14.12	14.77	11.13
DCRD(刘万军 等,2021)	12.58	13.12	11.65
MSDS-RCNN(Li等,2018)	11.63	10.60	13.73
CS-RCNN(Zhang等,2020b)	11.43	11.86	8.82
GAFF(Zhang等,2021a)	10.62	10.82	10.14
CFR_3(Zhang等,2020a)	10.05	9.72	10.80
AR-CNN(Zhang等,2019a)	9.34	9.94	8.38
DCRL(刘万军 等,2021)	9.16	9.86	8.18
MBNet(Zhou等,2020)	8.13	8.28	7.86
BAANet(Yang等,2022)	7.92	8.37	6.98
UFF+UCG(Kim等,2022a)	**7.89**	**8.18**	**6.96**

注:加粗字体表示各列最优结果,加下划线方法表示仅用可见光图像测试。

lutional neutral networks)等直接使用级联方式融合红外—可见模态的方法,不能充分利用红外图像在夜间场景下的优势,因此相比于夜间图像,对于白天场景下的行人检测效果更好。而引入注意力机制的CIAN、GAFF、BAANet方法以及引入差分融合模块的MBNet方法,以自注意力、差分融合的方式调整红外特征图在特征融合过程中的权重,更充分地利用红外图像在夜间场景下的优势,因此在夜间场景下行人检测效果更好。IAF R-CNN和IATDNN+IASS引入光照权重网络,根据光照条件自适应的调整红

外—可见光图像在融合过程中的权重，但由于权重只用于两分支的检测结果融合而未在特征融合过程中使用，所以 MR^{-2} 并无明显提升。虽然 MSDS R-CNN 并未引入光照权重，但是其 MR^{-2} 提升至 11.63%，这表明将语义分割和目标检测任务结合学习有助于提高检测性能。

表 5 以速度为指标对国内外方法在 KAIST 测试集上进行比较，速度的值越低表示检测器效率越高，但速度同时也取决于机器。如表 5 所示，由于 CFR_3 方法采用双模态特征循环融合的思想而未引入其他模块，模型所需要的参数量较小，因此相比于运行在 1080Ti 的其他方法，速度较快。基于两阶段检测网络的 Halfway Fusion、Fusion RPN + BDT、IAF R-CNN、IATDNN + IASS 等方法，虽然检测精度上有较大提升，但速度较慢，不符合现实场景对于实时检测的要求。

表 5　前沿方法在 KAIST 数据集的速度比较

Table 5　Comparison of speed of the state-of-the-art methods on KAIST dataset

方法	运算平台	速度/s
ACF + T + THOG（Hwang 等，2015）	MATLAB	2.73
Halfway Fusion（Liu 等，2016a）	TITANX	0.43
Fusion RPN + BDT（König 等，2017）	MATLAB	0.80
IAF R-CNN（Li 等，2019）	TITANX	0.21
IATDNN + IASS（Guan 等，2019a）	TITANX	0.25
MSDS-RCNN（Li 等，2018）	TITANX	0.22
CIAN（Zhang 等，2019b）	1080Ti	0.07
AR-CNN（Zhang 等，2019a）	1080Ti	0.12
CFR_3（Zhang 等，2020a）	1080Ti	**0.05**
MBNet（Zhou 等，2020）	1080Ti	0.07
BAANet（Yang 等，2022）	1080Ti	0.07

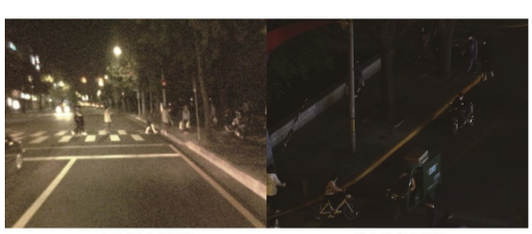
注：加粗字体表示最优结果。

3　未来方向展望

尽管近年来提出了一系列基于深度神经网络的跨模态行人检测方法并取得相对优异的性能，对比 Li 等人（2018 年）为 KAIST 数据集创建的人眼检测基线，检测器的表现仍然与人眼检测的表现存在较大差距，跨模态行人检测仍然是一个需要深入研究的方向。本文对跨模态行人检测领域尚未解决的问题进行讨论，并提出对于未来可行方向的思考。

1）模态间数据未对准问题。如图 8 和图 9 所示，CVC-14 数据集采用的双摄像机与 LLVIP 数据集采用的双目单摄像机都由于不同传感器的视野和视角不同，造成可见光—红外图像在空间上的未对准问题。在现实生活中，设备老化或外界干扰等原因将导致两传感器拍摄不同步，造成可见光—红外图像在时间上的未对准问题。当两模态图像间存在未对准问题时，两种模态图像（包括其产生的特征图）在同一空间位置所表示的行人信息不一致，从而产生错误的信息干扰检测器的优化。所以，解决两模态图像未对准问题将是一个重要的研究方向，不仅可以促进未对齐跨模态数据集（如 CVC-14，FLIR 等）的使用，还可以促进双传感器在现实生活中的应用。

2）不同视角下的跨模态行人检测问题。目前对于跨模态行人检测领域的研究大都基于车载数据集 KAIST、CVC-14 和 FLIR，对于监控视角下的跨模态行人检测研究比较缺少。车载数据集相较于监控视角拍摄的数据集场景多变，可以有效防止检测模型对某一特定环境过拟合。图 10 为 KAIST 数据集和 LLVIP 数据集的可见光图像的对比。从图 10（a）可以看出，由于夜间车辆行驶时通常会开启前照灯，相比于监控视角数据集，车载数据集中的夜间图像可能较为明亮。因此，如果同时将车载—道路监控两种视角摄像头拍摄的数据集用于跨模态行人检测器的训练，既可以增加模型在更暗场景下的鲁棒性，又可以防止模型对某一场景过拟合。



(a) KAIST　　(b) LLVIP

图 10　KAIST 数据集和 LLVIP 数据集的可见光图像对比

Fig. 10　Comparison of a visible image between KAIST dataset and LLVIP dataset ((a) KAIST; (b) LLVIP)

3）机载场景下的跨模态行人检测问题。在过去几年中，无人机在监视、搜索、救援以及快递等民用领域中的应用不断增加。针对大面积流行性疫情（如新冠疫情）造成的社会形式，Shao 等人（2021）提出使用无人机进行行人之间社交距离检测。首先检测器通过实时人头检测以准确定位行人，其次利用

获得的无人机图像(垂直图像和倾斜图像)到真实世界坐标的变换矩阵,计算出无人机图像上行人之间的社交距离,以实现可靠的行人个体之间的社交距离监控,既可以节省人力也可以保证行人之间的距离,但此研究仅针对白天场景。本文认为,若引入红外模态辅助,可实现全天候的行人距离监测,将更具实用价值。表6展示了目前先进的无人机目标检测数据集的对比。目前机载场景下没有公开的跨模态行人检测数据集,近几年关于跨模态行人检测的研究(de Oliveira 和 Wehrmeister, 2016, 2018; Beleznai等,2018; Wang等,2022)都是基于私有的数据集。因此,提出机载场景下的跨模态行人数据集是具有较大贡献价值的一项工作,且机载场景下的跨模态行人检测研究具有非常重要的现实意义。

表6　目前先进的无人机目标检测数据集对比

Table 6　Comparison of the current state-of-the-art UAV object detection datasets

数据集	任务	分辨率/像素	模态	提出年份
GMVRT-v2(Blondel等,2014)	行人检测	1 280×720	RGB	2014
CARPK(Hsieh等,2017)	目标检测	1 280×720	RGB	2017
UAVDT(Du等,2018)	目标检测	1 080×540	RGB	2018
VisDrone-2018(Zhu等,2018)	目标检测	3 840×2 160	RGB	2018
DAC-SDC(Xu等,2019)	目标检测	640×360	RGB	2019
VisDrone-2019(Zhu等,2020)	目标检测	3 840×2160	RGB	2019
Drone Vehicle(Sun等,2022)	目标检测	840×712	RGB+Thermal	2020
AU-AIR(Bozcan和Kayacan,2020)	目标检测	1 920×1 080	RGB	2020
BIRDSAI(Bondi等,2020)	目标检测	640×480	Thermal	2020
UVSD(Yang等,2019)	目标检测	960×540～5 280×2 970	RGB	2020
MOHR(Zhang等,2021c)	目标检测	5 482×3 078或7 360×4 912	RGB	2021
P-DESTRE(Kumar等,2021)	行人检测及追踪	3 840×2 160	RGB	2021

4)遮挡条件下的跨模态行人检测问题。根据Dollar等人(2012)的调查,在实际应用中,行人遮挡情况在繁华拥挤的街道、大型商场、机场等场景是很常见的,70%左右的行人都存在遮挡情况。从表2可以看出,目前跨模态行人检测的研究大多关注于检测部分遮挡或没有遮挡情况下的行人,跨模态行人检测器在严重遮挡情况下的表现并不理想。但Hwang等人(2015)的实验表明,不论是在无遮挡、部分遮挡还是严重遮挡情况下,在可见光图像没有足够信息识别遮挡的行人时,红外图像由于可以提供清晰的轮廓而具有一定的辅助检测作用。因此,结合红外图像和可见光图像用于进行遮挡条件下的行人检测是未来值得关注的一个研究方向。

5)小目标行人检测问题。对比于大目标行人包含的丰富检测信息,小目标行人的外表和轮廓信息都较模糊,检测器很难将小目标行人对象与周围环境区分开来。如表2所示,即使最先进的跨模态行人检测方法对于检测小目标行人的效果仍不理想,跨模态行人检测领域中对于小目标行人检测技术的研究仍然存在着很大空白。因此,将红外图像与可见光图像结合起来进行小目标行人检测也是未来值得关注的一个研究方向。

6)恶劣天气条件下的行人检测问题。雾天由于雾的遮挡导致可见光图像中行人能见度低,雨天由于摄像头前的雨滴导致可见光图像中的行人不可见。而在上述情况下,红外图像都能提供较为清晰的行人轮廓,所以本文提出应将跨模态行人检测器应用于雨、雾等恶劣天气条件。目前的跨模态行人检测的数据集都是在良好天气场景下拍摄的,目前先进的跨模态行人检测方法也仅适用于良好天气下的全天行人检测。要实现全天候的跨模态行人检测系统,仅关注良好天气下的白天和黑夜数据是不够的,也需要关注如雾天、雨天等恶劣天气下的数据。因此,提出基于雾天、雨天等天气下的多光谱行人数

据集是有实际意义的，并且针对恶劣天气下的跨模态行人检测任务也是一个可行的方向。

7）实时行人检测的问题。自动驾驶系统和机器人系统需要快速返回行人检测结果，如表5所示，大多跨模态行人检测器在GPU（graphics processing unit）上处理一对输入图像耗时较短，但在实际应用设备上跨模态行人检测器在GPU上处理一对输入图像的耗时并不理想。行人检测系统的检测错误将导致控制决策模块发出错误的决策信号，行人检测系统检测耗时长将导致决策模块发出的决策信号不及时。对于自动驾驶系统，错误或者延迟的决策信号都极有可能导致汽车与行人的碰撞。因此，准确高效的行人检测算法对于保障行人安全具有非常重要的现实意义，实时检测也将是跨模态行人检测持续发展的方向之一，并应同时保证行人检测的精度。

4 结 语

本文首先回顾了基于背景建模和手工特征的跨模态行人检测研究，并根据基于深度学习方法的跨模态行人检测研究针对的问题进行仔细分类并逐一展开介绍。其次，本文总结了7个可用的数据集和3个常用于评估跨模态行人检测器性能的评价指标。根据7个数据集的不同使用场景、拍摄时段等进行比较说明，并在不同的数据集上以不同的评价指标对近几年前沿的跨模态行人检测方法进行比较分析。最后，本文讨论了目前跨模态行人检测领域仍然存在的问题并提出未来可行的研究方向。

尽管基于深度学习的跨模态行人检测近些年来取得了巨大的研究进展，但是相关技术仍然难以满足需求，仍然需要深入研究。希望通过本文能够使更多读者了解当前工作，进而提出更有价值的跨模态行人检测方法。

参考文献（References）

Andreas M, Chris B, Thomas B M, Albert C and Sergio E. 2013. Trimodal person re-identification with RGB, depth and thermal features//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. Portland, USA: IEEE: 301-307 [DOI: 10.1109/cvprw.2013.52]

Beleznai C, Steininger D, Croonen G and Broneder E. 2018. Multimodal human detection from aerial views by fast shape-aware clustering and classification//Proceedings of the 10th IAPR Workshop on Pattern Recognition in Remote Sensing. Beijing, China: IEEE: 1-6 [DOI: 10.1109/prrs.2018.8486236]

Blondel P, Potelle A, Pégard C and Lozano R. 2014. Fast and viewpoint robust human detection in uncluttered environments//Proceedings of 2014 IEEE Visual Communications and Image Processing Conference. Valletta, Malta: IEEE: 522-525 [DOI: 10.1109/vcip.2014.7051621]

Bondi E, Jain R, Aggrawal P, Anand S, Hannaford R, Kapoor A, Piavis J, Shah S, Joppa L, Dilkina B and Tambe M. 2020. BIRDSAI: a dataset for detection and tracking in aerial thermal infrared videos//Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass, USA: IEEE: 1736-1745 [DOI: 10.1109/wacv45572.2020.9093284]

Bozcan I and Kayacan E. 2020. AU-AIR: a multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance//Proceedings of 2020 IEEE International Conference on Robotics and Automation. Paris, France: IEEE: 8504-8510 [DOI: 10.1109/ICRA40945.2020.9196845]

Burt P J and Adelson E H. 1983. The Laplacian pyramid as a compact image code//Fischler M A and Firschein O, eds. Readings in Computer Vision. Los Altos, USA: Morgan Kaufmann: 671-679 [DOI: 10.1016/B978-0-08-051581-6.50065-9]

Candès E, Demanet L, Donoho D and Ying L X. 2006. Fast discrete curvelet transforms. Multiscale Modeling and Simulation, 5(3): 861-899 [DOI: 10.1137/05064182X]

Cao Y P, Guan D Y, Huang W L, Yang J X, Cao Y L and Qiao Y. 2019a. Pedestrian detection with unsupervised multispectral feature learning using deep neural networks. Information Fusion, 46: 206-217 [DOI: 10.1016/j.inffus.2018.06.005]

Cao Y P, Guan D Y, Wu Y L, Yang J X, Cao Y L and Yang M Y. 2019b. Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection. ISPRS Journal of Photogrammetry and Remote Sensing, 150: 70-79 [DOI: 10.1016/j.isprsjprs.2019.02.005]

Choi E J and Park D J. 2010. Human detection using image fusion of thermal and visible image with new joint bilateral filter//Proceedings of the 5th International Conference on Computer Sciences and Convergence Information Technology. Seoul, Korea (South): IEEE: 882-885 [DOI: 10.1109/iccit.2010.5711182]

Choi H, Kim S, Park K and Sohn K. 2016. Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks//Proceedings of the 23rd International Conference on Pattern Recognition. Cancun, Mexico: IEEE: 621-626 [DOI: 10.1109/icpr.2016.7899703]

Connah D, Drew M S and Finlayson G D. 2015. Spectral edge: gradient-preserving spectral mapping for image fusion. Journal of the Optical Society of America A, 32(12): 2384-2396 [DOI: 10.1364/josaa.32.002384]

Davis J W and Sharma V. 2007. Background-subtraction using contour-based fusion of thermal and visible imagery. Computer Vision and Image Understanding, 106(2/3): 162-182 [DOI: 10.1016/j.cviu.2006.06.010]

de Oliveira D C and Wehrmeister M A. 2016. Towards real-time people recognition on aerial imagery using convolutional neural networks// The 19th IEEE International Symposium on Real-Time Distributed Computing. York, UK: IEEE: 27-34 [DOI: 10.1109/isorc.2016.14]

de Oliveira D C and Wehrmeister M A. 2018. Using deep learning and low-cost RGB and thermal cameras to detect pedestrians in aerial images captured by multirotor UAV. Sensors, 18(7): #2244 [DOI: 10.3390/s18072244]

Deng Z and Latecki L J. 2017. Amodal detection of 3D objects: inferring 3D bounding boxes from 2D ones in RGB-depth images//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: 398-406 [DOI: 10.1109/cvpr.2017.50]

Dollar P, Wojek C, Schiele B and Perona P. 2012. Pedestrian detection: an evaluation of the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(4): 743-761 [DOI: 10.1109/tpami.2011.155]

Du D W, Qi Y K, Yu H Y, Yang Y F, Duan K W, Li G R, Zhang W G, Huang Q M and Tian Q. 2018. The unmanned aerial vehicle benchmark: object detection and tracking//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 375-391 [DOI: 10.1007/978-3-030-01249-6_23]

Fang Q Y and Wang Z K. 2022. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. Pattern Recognition, 130: #108786 [DOI: 10.1016/j. patcog. 2022. 108786]

Farooq A, Jalal A and Kamal S. 2015. Dense RGB-D map-based human tracking and activity recognition using skin joints features and self-organizing map. KSII Transactions on Internet and Information Systems, 9(5): 1856-1869 [DOI: 10.3837/tiis.2015.05.017]

French G, Finlayson G and Mackiewicz M. 2018. Multi-spectral pedestrian detection via image fusion and deep neural networks. Journal of Imaging Science and Technology, 62(5): #50406 [DOI: 10.2352/J.ImagingSci.Technol.2018.62.5.050406]

F. A. Group. 2018. Flir Thermal Dataset for Algorithm Training [EB/OL]. [2023-02-26]. https://www.flir.in/oem/adas/adas-dataset-form/.

Golcarenarenji G, Martinez-Alpiste I, Wang Q and Alcaraz-Calero J M. 2021. Efficient real-time human detection using unmanned aerial vehicles optical imagery. International Journal of Remote Sensing, 42(7): 2440-2462 [DOI: 10.1080/01431161.2020.1862435]

González A, Fang Z J, Socarras Y, Serrat J, Vázquez D, Xu J L and López A M. 2016. Pedestrian detection at day/night time with visible and fir cameras: a comparison. Sensors, 16(6): #820 [DOI: 10.3390/s16060820]

Guan D Y, Cao Y P, Yang J X, Cao Y L and Tisse C L. 2018. Exploiting fusion architectures for multispectral pedestrian detection and segmentation. Applied Optics, 57(18): 108-116 [DOI: 10.1364/ao.57.00d108]

Guan D Y, Cao Y P, Yang J X, Cao Y L and Yang M Y. 2019a. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. Information Fusion, 50: 148-157 [DOI: 10.1016/j.inffus.2018.11.017]

Guan D Y, Luo X, Cao Y P, Yang J X, Cao Y L, Vosselman G and Yang M Y. 2019b. Unsupervised domain adaptation for multispectral pedestrian detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA: IEEE: 434-443 [DOI: 10.1109/cvprw.2019.00057]

Gupta S, Girshick R, Arbeláez P and Malik J. 2014. Learning rich features from RGB-D images for object detection and segmentation// Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 345-360 [DOI: 10.1007/978-3-319-10584-0_23]

He R, Wu X, Sun Z A and Tan T N. 2017. Learning invariant deep representation for NIR-VIS face recognition//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI: 2000-2006 [DOI: 10.5555/3298483.3298529]

Hou Y L, Song Y Y, Hao X L, Shen Y, Qian M Y and Chen H J. 2018. Multispectral pedestrian detection based on deep convolutional neural networks. Infrared Physics and Technology, 94: 69-77 [DOI: 10.1016/j.infrared.2018.08.029]

Hsieh M R, Lin Y L and Hsu W H. 2017. Drone-based object counting by spatially regularized regional proposal network//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4165-4173 [DOI: 10.1109/iccv.2017.446]

Hwang S, Park J, Kim N, Choi Y and Kweon I S. 2015. Multispectral pedestrian detection: benchmark dataset and baseline//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1037-1045 [DOI: 10.1109/CVPR.2015.7298706]

Jia X Y, Zhu C, Li M Z, Tang W Q and Zhou W L. 2021. LLVIP: a visible-infrared paired dataset for low-light vision//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal, Canada: IEEE: 3489-3497 [DOI: 10.1109/iccvw54120.2021.00389]

Jiang T T, Liu Y, Ma X and Sun J L. 2021. Multi-path collaborative salient object detection based on RGB-T images. Journal of Image and Graphics, 26(10): 2388-2399 (蒋亭亭, 刘昱, 马欣, 孙景林. 2021. 多支路协同的RGB-T图像显著性目标检测. 中国图象图形学报, 26(10): 2388-2399) [DOI: 10.11834/jig.200317]

Kim J U, Park S and Ro Y M. 2022a. Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. IEEE Transactions on Circuits and Systems for Video Technology, 32(3): 1510-1523 [DOI: 10.1109/tcsvt.2021.3076466]

Kim J U, Park S and Ro Y M. 2022b. Towards versatile pedestrian

detector with multisensory-matching and multispectral recalling memory//Proceedings of the 36th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 1157-1165 [DOI: 10.1609/aaai.v36i1.20001]

König D, Adam M, Jarvers C, Layher G, Neumann H and Teutsch M. 2017. Fully convolutional region proposal networks for multispectral person detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA: IEEE: 243-250 [DOI: 10.1109/cvprw.2017.36]

Kruthiventi S S S, Sahay P and Biswal R. 2017. Low-light pedestrian detection from RGB images using multi-modal knowledge distillation//Proceedings of 2017 IEEE International Conference on Image Processing. Beijing, China: IEEE: 4207-4211 [DOI: 10.1109/icip.2017.8297075]

Kumar S V A, Yaghoubi E, Das A, Harish B S and Proença H. 2021. The P-DESTRE: a fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices. IEEE Transactions on Information Forensics and Security, 16: 1696-1708 [DOI: 10.1109/tifs.2020.3040881]

Lee Y, Bui T D and Shin J. 2018. Pedestrian detection based on deep fusion network using feature correlation//Proceedings of 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Honolulu, USA: IEEE: 694-699 [DOI: 10.23919/apsipa.2018.8659688]

Lee Y, Chen J C, Tseng C W and Lai S H. 2016. Accurate and robust face recognition from RGB-D images with a deep learning approach//Proceedings of 2016 British Machine Vision Conference. York, UK: BMVA Press: #3 [DOI: 10.5244/c.30.123]

Leykin A, Ran Y and Hammoud R. 2007. Thermal-visible video fusion for moving target tracking and pedestrian classification//Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA: IEEE: 1-8 [DOI: 10.1109/cvpr.2007.383444]

Li C Y, Song D, Tong R F and Tang M. 2018. Multispectral pedestrian detection via simultaneous detection and segmentation//Proceedings of British Machine Vision Conference 2018. Newcastle, UK: BMVA Press: #225

Li C Y, Song D, Tong R F and Tang M. 2019. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. Pattern Recognition, 85: 161-171 [DOI: 10.1016/j.patcog.2018.08.005]

Li D G, Wei X, Hong X P and Gong Y H. 2020a. Infrared-visible cross-modal person re-identification with an X modality//Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI: 4610-4617 [DOI: 10.1609/aaai.v34i04.5891]

Li G F, Yang Y F and Qu X D. 2020b. Deep learning approaches on pedestrian detection in hazy weather. IEEE Transactions on Industrial Electronics, 67(10): 8889-8899 [DOI: 10.1109/tie.2019.2945295]

Li J H, Zhang P, Wang X W and Huang S Z. 2020. Infrared small-target detection algorithms: a survey. Journal of Image and Graphics, 25(9): 1739-1753 (李俊宏, 张萍, 王晓玮, 黄世泽. 2020. 红外弱小目标检测算法综述. 中国图象图形学报, 25(9): 1739-1753) [DOI: 10.11834/jig.190574]

Li Y T. 2021. Multispectral pedestrian detection in autonomous driving: a review. IEIE Transactions on Smart Processing and Computing, 10(1): 10-16 [DOI: 10.5573/ieiespc.2021.10.1.010]

Lin C Z, Lu J W, Wang G and Zhou J. 2018. Graininess-aware deep feature learning for pedestrian detection//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 745-761 [DOI: 10.1007/978-3-030-01240-3_45]

Liu J J, Zhang S T, Wang S and Metaxas N D. 2016a. Multispectral deep neural networks for pedestrian detection//Proceedings of British Machine Vision Conference 2016. York, UK: BMVA Press: 73.1-73.13

Liu T S, Lam K M, Zhao R and Qiu G P. 2022. Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection. IEEE Transactions on Circuits and Systems for Video Technology, 32(1): 315-329 [DOI: 10.1109/tcsvt.2021.3060162]

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A C. 2016b. SSD: single shot multiBox detector//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 21-37 [DOI: 10.1007/978-3-319-46448-0_2]

Liu W, Liao S C, Ren W Q, Hu W D and Yu Y N. 2019. High-level semantic feature detection: a new perspective for pedestrian detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5182-5191 [DOI: 10.1109/cvpr.2019.00533]

Liu W J, Dong L B and Qu H C. 2021. Small-scale pedestrian detection based on improved R-FCN model. Journal of Image and Graphics, 26(10): 2400-2410 (刘万军, 董利兵, 曲海成. 2021. 改进 R-FCN 模型的小尺度行人检测. 中国图象图形学报, 26(10): 2400-2410) [DOI: 10.11834/jig.200287]

Lyu C, Heyer P, Munir A, Platisa L, Micheloni C, Goossens B and Philips W. 2021. Visible-thermal pedestrian detection via unsupervised transfer learning//Proceedings of the 5th International Conference on Innovation in Artificial Intelligence. Xiamen, China: ACM: 158-163 [DOI: 10.1145/3461353.3461369]

Ouyang W L and Wang X G. 2013. Joint deep learning for pedestrian detection//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE: 2056-2063 [DOI: 10.1109/iccv.2013.257]

Park K, Kim S and Sohn K. 2018. Unified multi-spectral pedestrian detection based on probabilistic fusion networks. Pattern Recognition, 80: 143-155 [DOI: 10.1016/j.patcog.2018.03.007]

Ranchin T and Wald L. 1993. The wavelet transform for the analysis of remotely sensed images. International Journal of Remote Sensing,

14(3): 615-619 [DOI: 10.1080/01431169308904362]

Redmon J and Farhadi A. 2017. YOLO9000: better, faster, stronger//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6517-6525 [DOI: 10.1109/cvpr.2017.690]

Redmon J and Farhadi A. 2018. YOLOv3: an incremental improvement [EB/OL]. [2018-04-08]. https://arxiv.org/pdf/1804.02767.pdf

San-Biagio M, Crocco M, Cristani M, Martelli S and Murino V. 2012. Low-level multimodal integration on Riemannian manifolds for automatic pedestrian detection//Proceedings of the 15th International Conference on Information Fusion. Singapore , Singapore: IEEE: 2223-2229

Shao Y H, Zhang X P, Chu H Y, Zhang X Q, Zhang D and Rao Y B. 2022. AIR-YOLOv3: aerial infrared pedestrian detection via an improved YOLOv3 with network pruning. Applied Sciences, 12(7): #3627 [DOI: 10.3390/app12073627]

Shao Z F, Cheng G, Ma J Y, Wang Z Y, Wang J M and Li D R. 2021. Real-time and accurate UAV pedestrian detection for social distancing monitoring in COVID-19 pandemic. IEEE Transactions on Multimedia, 24: 2069-2083 [DOI: 10.1109/TMM.2021.3075566]

Song S R and Xiao J X. 2016. Deep sliding shapes for Amodal 3D object detection in RGB-D images//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 808-816 [DOI: 10.1109/cvpr.2016.94]

Sun Y M, Cao B, Zhu P F and Hu Q H. 2022. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. IEEE Transactions on Circuits sand Systems for Video Technology, 32(10): 6700-6713 [DOI: 10.1109/TCSVT.2022.3168279]

Takumi K, Watanabe K, Ha Q S, Tejero-De-Pablos A, Ushiku Y and Harada T. 2017. Multispectral object detection for autonomous vehicles//Proceedings of on Thematic Workshops of ACM Multimedia 2017. California, USA: ACM: 35-43 [DOI: 10.1145/3126686.3126727]

Tian Y L, Luo P, Wang X G and Tang X O. 2015. Deep learning strong parts for pedestrian detection//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, USA: IEEE: 1904-1912 [DOI: 10.1109/iccv.2015.221]

Vandersteegen M, van Beeck K and Goedemé T. 2018. Real-time multispectral pedestrian detection with a single-pass deep neural network//Proceedings of the 15th International Conference on Image Analysis and Recognition. Póvoa de Varzim, Portugal: Springer: 419-426 [DOI: 10.1007/978-3-319-93000-8_47]

Vladimir V K, Vladimir A K, Jiri H, Walter G K and Vladimir M. 2019. Thermal GAN: multimodal color-to-thermal image translation for person re-identification in multispectral dataset. Computer Vision ECCV 2018 Workshops, 606-624 [DOI: 10.1007/978-3-030-11024-6_46]

Wagner J, Fischer V, Herman M and Behnke S. 2016. Multispectral pedestrian detection using deep fusion convolutional neural networks//The 24th European Symposium on Artificial Neural Networks. Bruges, Belgium: ESANN: 509-514

Wanchaitanawong N, Tanaka M, Shibata T and Okutomi M. 2021. Multi-modal pedestrian detection with large misalignment based on modal-wise regression and multi-modal IoU//Proceedings of the 17th International Conference on Machine Vision and Applications. Aichi, Japan: IEEE: 1-6 [DOI: 10.23919/mva51890.2021.9511366]

Wang C Q, Luo D, Liu Y, Xu B and Zhou Y J. 2022. Near-surface pedestrian detection method based on deep learning for UAVs in low illumination environments. Optical Engineering, 61(2): #023103 [DOI: 10.1117/1.OE.61.2.023103]

Wu A C, Lin C Z and Zheng W S. 2022. Single-modality self-supervised information mining for cross-modality person re-identification. Journal of Image and Graphics, 27(10): 2843-2859 (吴岸聪, 林城栋, 郑伟诗. 2022. 面向跨模态行人重识别的单模态自监督信息挖掘. 中国图象图形学报, 27(10): 2843-2859) [DOI: 10.11834/jig.211050]

Wu F, Jing X Y, Feng Y J, Ji Y M and Wang R C. 2021. Spectrum-aware discriminative deep feature learning for multi-spectral face recognition. Pattern Recognition, 111: #107632 [DOI: 10.1016/j.patcog.2020.107632]

Xu D, Ouyang W L, Ricci E, Wang X G and Sebe N. 2017. Learning cross-modal deep representations for robust pedestrian detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 4236-4244 [DOI: 10.1109/cvpr.2017.451]

Xu D F, Anguelov D and Jain A. 2018. Pointfusion: deep sensor fusion for 3D bounding box estimation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 244-253 [DOI: 10.1109/cvpr.2018.00033]

Xu M, Yu X S, Chen D Y, Wu C D, Jia T and Ru J Y. 2018. Pedestrian detection in complex thermal infrared surveillance scene. Journal of Image and Graphics, 23(12): 1829-1837 (许茗, 于晓升, 陈东岳, 吴成东, 贾同, 茹敬雨. 2018. 复杂热红外监控场景下行人检测. 中国图象图形学报, 23(12): 1829-1837) [DOI: 10.11834/jig.180299]

Xu X W, Zhang X Y, Yu B, Hu X S, Rowen C, Hu J T and Shi Y Y. 2019. DAC-SDC low power object detection challenge for UAV applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(2): 392-403 [DOI: 10.1109/tpami.2019.2932429]

Yang F, Fan H, Chu P, Blasch E and Ling H B. 2019. Clustered object detection in aerial images//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 8311-8320 [DOI: 10.1109/iccv.2019.00840]

Yang X X, Qian Y Q, Zhu H J, Wang C X and Yang M. 2022. BAANet: learning bi-directional adaptive attention gates for multi-

spectral pedestrian detection//Proceedings of 2022 International Conference on Robotics and Automation. Philadelphia, USA: IEEE: 2920-2926 [DOI: 10.1109/ICRA46639.2022.9811999]

Ye M, Lan X Y, Li J W and Yuen P. 2018. Hierarchical discriminative learning for visible thermal person re-identification//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI: 7501-7508 [DOI: 10.1609/aaai.v32i1.12293]

Zhang H, Fromont E, Lefevre S and Avignon B. 2020a. Multispectral fusion for object detection with cyclic fuse-and-refine blocks//Proceedings of 2020 IEEE International Conference on Image Processing. Abu Dhabi, United Arab Emirates: IEEE: 276-280 [DOI: 10.1109/icip40778.2020.9191080]

Zhang H, Fromont E, Lefevre S and Avignon B. 2021a. Guided attentive feature fusion for multispectral pedestrian detection//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 72-80 [DOI: 10.1109/wacv48630.2021.00012]

Zhang H, Fromont E, Lefevre S and Avignon B. 2021b. Deep active learning from multispectral data through cross-modality prediction inconsistency//Proceedings of 2021 IEEE International Conference on Image Processing. Anchorage, USA: IEEE: 449-453 [DOI: 10.1109/icip42928.2021.9506322]

Zhang H J, Sun M S, Li Q, Liu L L, Liu M and Ji Y Z. 2021c. An empirical study of multi-scale object detection in high resolution UAV images. Neurocomputing, 421: 173-182 [DOI: 10.1016/j.neucom.2020.08.074]

Zhang L, Liu Z Y, Zhang S F, Yang X, Qiao H, Huang K Z and Hussain A. 2019b. Cross-modality interactive attention network for multispectral pedestrian detection. Information Fusion, 50: 20-29 [DOI: 10.1016/j.inffus.2018.09.015]

Zhang L, Zhu X Y, Chen X Y, Yang X, Lei Z and Liu Z Y. 2019a. Weakly aligned cross-modal learning for multispectral pedestrian detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5126-5136 [DOI: 10.1109/ICCV.2019.00523]

Zhang Y T, Yin Z S, Nie L Z and Huang S. 2020b. Attention based multi-layer fusion of multispectral images for pedestrian detection. IEEE Access, 8: 165071-165084 [DOI: 10.1109/access.2020.3022623]

Zhao C R, Qi D, Dou S G, Tu Y P, Sun T L, Bai S, Jiang X Y, Bai X and Miao D Q. 2021. Key technology for intelligent video surveillance: a review of person re-identification. Science in China (Information Sciences), 51(12): 1979-2015 (赵才荣, 齐鼎, 窦曙光, 涂远鹏, 孙添力, 柏松, 蒋忻洋, 白翔, 苗夺谦. 2021. 智能视频监控关键技术: 行人再识别研究综述. 中国科学: 信息科学), 51(12): 1979-2015 [DOI: 10.1360/SSI-2021-021]

Zhao Y Q, Rao Y, Dong S P and Zhang J Y. 2020. Survey on deep learning object detection. Journal of Image and Graphics, 25(4): 629-654 (赵永强, 饶元, 董世鹏, 张君毅. 2020. 深度学习目标检测方法综述. 中国图象图形学报, 25(4): 629-654) [DOI: 10.11834/jig.190307]

Zhen Y, Wang Z L and Wu F. 2021. Pedestrian detection method based on density and score refinement. Journal of Image and Graphics, 26(2): 425-437 (甄烨, 王子磊, 吴枫. 2021. 融合密度和精细分数的行人检测. 中国图象图形学报, 26(2): 425-437) [DOI: 10.11834/jig.200060]

Zheng W S and Wu A C. 2018. Asymmetric person re-identification: cross-view person tracking in a large camera network. Scientia Sinica Informationis, 48(5): 545-563 (郑伟诗, 吴岸聪. 2018. 非对称行人重识别: 跨摄像机持续行人追踪. 中国科学: 信息科学), 48(5): 545-563 [DOI: 10.1360/N112018-00017]

Zheng Y, Izzat I H and Ziaee S. 2019. GFD-SSD: gated fusion double SSD for multispectral pedestrian detection [EB/OL]. [2019-03-16]. https://arxiv.org/ftp/arxiv/papers/1903/1903.06999.pdf

Zhou K L, Chen L S and Cao X. 2020. Improving multispectral pedestrian detection by addressing modality imbalance problems//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 787-803 [DOI: 10.1007/978-3-030-58523-5_46]

Zhu P F, Wen L Y, Bian X, Ling H B and Hu Q H. 2018. Vision meets drones: a challenge [EB/OL]. [2020-04-20]. https://arxiv.org/pdf/1804.07437.pdf

Zhu P F, Wen L Y, Du D W, Bian X, Hu Q H and Ling H B. 2020. Vision meets drones: past, present and future [EB/OL]. [2020-01-16]. https://arxiv.org/pdf/2001.06303v1.pdf

Zhuang Y F, Pu Z Y, Hu J and Wang Y H. 2022. Illumination and temperature-aware multispectral networks for edge-computing-enabled pedestrian detection. IEEE Transactions on Network Science and Engineering, 9(3): 1282-1295 [DOI: 10.1109/tnse.2021.3139335]

作者简介

别倩，女，本科生，主要研究方向为跨模态行人检测、计算机视觉。E-mail: bieqian@wust.edu.cn

王晓，女，通信作者，特任副教授，主要研究方向为行人检测。E-mail: wangxiao2021@wust.edu.cn

徐新，男，教授，主要研究方向为模式识别。E-mail: xuxin@wust.edu.cn

赵启军，男，教授，主要研究方向为视频分析与处理。E-mail: qjzhao@scu.edu.cn

王正，男，教授，主要研究方向为多源行为分析。E-mail: wangzwhu@whu.edu.cn

陈军，男，教授，主要研究方向为安防应急处理。E-mail: chenj@whu.edu.cn

胡瑞敏，男，教授，俄罗斯国家工程院外籍院士，主要研究方向为社会安全。E-mail: hrm@whu.ed