计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(10):86−100 [doi: 10.15888/j.cnki.csa.009974] [CSTR: 32024.14.csa.009974]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

融合自适应双重注意力和轴向注意力 Transformer 的多光谱行人检测①

罗建国[1,2], 王燕妮[1], 韩世鹏[2], 吕　昊[2], 张耀荣[1], 吴　雪[1]

[1](西安建筑科技大学 信息与控制工程学院, 西安 710055)
[2](空军军医大学 军事生物医学工程学系, 西安 710032)
通信作者: 王燕妮, E-mail: wangyanni@xauat.edu.cn

摘　要: 针对现有的多光谱行人检测算法存在多模态相互作用不足和融合方法缺乏远程依赖性, 导致在低光照背景下小尺度行人检测性能不足的问题, 提出了一种融合自适应双重注意力和轴向注意力 Transformer 的多光谱小尺度行人检测算法 (adaptive dual attention and axial attention Transformer network, ADATNet). 采用双分支 CSPDarknet53 网络分别提取可见光和红外图像中的深度特征, 充分保留两种模态的特有信息. 设计两个特征交叉融合模块: 自适应双重注意力模块 (adaptive dual attention module, ADAM) 和轴向注意力 Transformer 特征增强 (axial attention Transformer feature enhancement, ATFE) 模块, 其中 ADAM 旨在强化模型对关键特征的关注, 同时抑制不相关或冗余的信息; ATFE 关联多模态特征的位置编码来融合增强的特征, 在确保计算效率的同时捕获长距离依赖关系. 将融合后的特征输入至检测头以输出最终检测结果. 实验结果表明, ADATNet 在 KAIST 数据集上的 MR^{-2} 降低至 7.08%, 同时在 FLIR 和 LLVIP 数据集上的 mAP50 分别达到 82.8% 和 97.6%, 较基线方法提升 4.7% 和 1.9%, 具有良好的检测性能.

关键词: 多光谱行人检测; 可见光和红外; 自适应双重注意力; 轴向注意力

引用格式: 罗建国,王燕妮,韩世鹏,吕昊,张耀荣,吴雪.融合自适应双重注意力和轴向注意力 Transformer 的多光谱行人检测.计算机系统应用, 2025,34(10):86–100. http://www.c-s-a.org.cn/1003-3254/9974.html

Fusion of Adaptive Dual Attention and Axial Attention Transformer for Multi-spectral Pedestrian Detection

LUO Jian-Guo[1,2], WANG Yan-Ni[1], HAN Shi-Peng[2], LYU Hao[2], ZHANG Yao-Rong[1], WU Xue[1]

[1](School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China)
[2](Department of Military Biomedical Engineering, Air Force Medical University, Xi'an 710032, China)

Abstract: Existing multi-spectral pedestrian detection algorithms suffer from insufficient multimodal interaction and lack of long-distance dependence of fusion algorithms, thus resulting in poor performance in small-scale pedestrian detection in low-light conditions. To this end, this study proposes an fusion adaptive dual attention and axial attention Transformer network (ADATNet) for multi-spectral small-scale pedestrian detection. The method adopts the dual-branch CSPDarknet53 network to separately extract deep features from visible and infrared images, preserving the unique information from each modality. Meanwhile, two feature fusion modules are designed, including the adaptive dual attention module (ADAM) and axial attention Transformer feature enhancement module (ATFE). ADAM aims to enhance the model's focus on critical features and suppress irrelevant or redundant information, while ATFE correlates the positional encoding of multimodal features to fuse the enhanced features, thereby both ensuring computational efficiency

① 基金项目: 陕西省自然科学基础研究项目 (2020JM499, 2020JQ684)
收稿时间: 2025-02-27; 修改时间: 2025-04-08; 采用时间: 2025-04-27; csa 在线出版时间: 2025-09-01
CNKI 网络首发时间: 2025-09-02

and capturing long-distance dependence. The fused features are then fed into the detection head to output the final detection results. Experimental results indicate that MR^{-2} of ADATNet decreases to 7.08% on the KAIST dataset, while the mAP50 for the FLIR and LLVIP datasets reaches 82.8% and 97.6% respectively. They have an improvement of 4.7% and 1.9% over the baseline methods, demonstrating excellent detection performance.

Key words: multi-spectral pedestrian detection; visible light and infrared; adaptive dual attention; axial attention

行人检测作为计算机视觉领域的一项关键任务, 在自动驾驶[1]、视频监控[2,3]等领域发挥着关键作用. 尽管过去数十年该领域已取得显著进步, 但在光照不足和复杂背景条件下, 现有行人检测技术仍然面临巨大挑战[4]. 虽然可见光图像在良好照明条件下可以提供丰富的颜色和纹理等底层信息, 但是在低光照和复杂背景下, 可见光图像很难将小尺度行人目标与背景区分, 而红外图像由于其特殊的能量辐射成像原理, 即使在低光照和复杂背景下, 依然可以提供行人轮廓特征, 但是容易出现远距离小尺度行人漏检误检问题. 因此, 借助可见光和红外的优势, 近些年, 基于可见光与红外图像融合的多光谱检测逐渐成为行人检测的主流方式.

在多光谱行人检测中, 常见做法是采用两种路线分别从可见光图像和红外图像中提取特征并进行融合[5]. 以往研究中, 基于卷积神经网络 (CNN) 的特征融合已广泛应用于当前最先进的方法中. 其中, 文献[6]探索了不同的融合架构, 并表明中期融合可以带来理想的性能表现. 在此基础上, Zhang 等人[7]对不同模态间的交互进行编码, 并实现特征的自适应融合. Wang 等人[8]提出了一种跨尺度动态卷积驱动的 YOLO 融合网络 (CDC-YOLOFusion), 该网络引入了一种新型跨尺度动态卷积融合模块, 能够根据数据分布自适应提取和融合双模态特征. 然而, 上述使用 CNN 的方法基于卷积算子的非全局感受野, 导致信息仅在局部区域融合. 尽管这些方法比单模态检测方法具有更好的性能, 但它们通常缺乏远程依赖性, 且没有充分利用模态间的互补性, 导致检测结果不理想. 为提高网络的远程依赖学习能力, Fang 等人[9]提出了一种基于 Transformer 的跨模态特征融合方法, 在特征提取阶段整合全局上下文信息并学习长程依赖性. 虽然这种方法计算量太大, 但将特征下采样至一定大小会导致信息的丢失.

尽管上述方法已取得显著进步, 但仍存在一些关键挑战. 由于可见光和红外特征在低光照条件下的稳定性和可靠性存在显著差异, 直接进行特征融合虽能弥补信息不足, 但也可能因引入噪声而降低检测效果. 因此, 如何巧妙融合这两种模态的特征并充分利用两者的互补性, 以及如何进行特征交叉融合以实现显著的性能增益, 依然是当前多光谱行人检测面临的挑战. 为解决上述问题, 本文提出了一种基于自适应双重注意力和轴向注意力 Transformer 的可见光-红外行人检测方法 (ADATNet). 核心思想在于充分利用多模态图像信息, 从而提升行人检测的精度. 该网络先是利用双流 CSPDarknet53 网络作为特征提取器, 从可见光和红外两种模态的图像中捕获深层次的特征信息. 然后, 设计自适应双重注意力模块, 运用自适应卷积 (ADC)[10]来实现卷积运算, 通过自适应调节卷积核提取全局上下文信息. 同时, 双重注意力结合可见光和红外特征, 能够更准确地定位小尺度行人所在的关键区域, 减少背景噪声的干扰. 为进一步从源图像中获取全局互补信息, 还设计了一个轴向注意力 Transformer 特征增强模块 (ATFE). 该模块能够有效地整合增强的可见光红外图像特征, 从全局上下文角度出发, 关联可见光红外特征的位置编码, 利用其强大的全局建模能力与卷积局部建模能力相结合, 捕获长距离依赖关系的同时, 有效利用局部特征信息; 最后将融合后的特征输入检测头以输出最终检测结果.

本文分别在提取可见光和红外特征阶段引入 ADC, 通过动态调制卷积核权重, 充分挖掘每种光谱特征的优势与互补潜力. 这一机制增强了每种光谱的特征表达能力, 并为后续的多光谱融合奠定了基础. 可见光在复杂光照下的细节纹理易受干扰, 而红外图像则能在低光环境下提供稳定的目标轮廓. ADC 通过独立学习每种光谱的最优特征, 专注于各模态的核心优势, 确保模态间特征的互补性. 这种独立特征学习方式通过自适应卷积调节, 使每个模态的特征得到最佳表达, 从而增强多模态融合时的互补效果.

本文的主要工作如下.

(1) 提出融合可见光-红外图像的小尺度行人检测

网络 ADATNet. 通过引入 ADC, 充分挖掘可见光和红外图像之间的互补特性, 增强多光谱图像融合检测性能.

(2) 提出了一种多模态特征融合方法, 由两个核心模块组成: 自适应双重注意力 (ADAM) 和轴向注意力 Transformer 特征增强 (ATFE). ADAM 用于准确定位行人所在关键区域, 减少背景噪声的干扰, 而 ATFE 通过全局-局部建模, 用以捕获远程依赖关系, 降低小尺度行人漏检率.

(3) 该网络在 KAIST、FLIR、LLVIP 这 3 个多光谱行人数据集上的实验结果表明, 与其他方法相比, 该方法获得了具有竞争力的检测性能.

1 相关研究

1.1 可见光行人检测

行人检测作为计算机视觉领域的重要任务, 近年来取得了显著进展. 现有行人检测技术通常分为两大类别: 一类是基于候选区域的两阶段模型, 一类是基于回归的一阶段模型. 两阶段模型一般先根据输入图像生成候选区域, 再对候选区域进行分类和坐标回归, 通常比一阶段模型有更高的检测精度. Girshick 等人[11]提出了 R-CNN 算法, 其使用区域提议的方法生成待测锚框. Ren 等人[12]将 Fast R-CNN[13]模型应用到了行人检测领域. Li 等人[14]在 Faster R-CNN 的基础上设计了一种尺度感知的 R-CNN 算法 (SAF R-CNN), 该网络由大型和小型子网络组成, 分别用于检测大尺度和小尺度行人目标. 尽管两阶段检测算法经过多次优化已提高了精度, 但当其应用于实际环境中时, 两阶段算法模型的计算较广泛, 因此过程也更为复杂. 一阶段模型是将目标检测问题转化成回归问题, 因此不需要生成候选区域, 可直接获得目标类别和位置信息, 在实时处理上优于两阶段算法. 与 Faster R-CNN 相比, YOLO 能更准确地区分目标对象的类别和背景信息. Gao 等人[15]对 YOLOv3 进行了优化, 并将其集成到自动驾驶系统中, 提高了网络模型的实时性. Ma 等人[16]在 YOLOv4[17]的基础上, 采用了特征金字塔网络融合方法进行特征提取, 解决了在复杂环境下小尺寸行人误检率高等问题. Lv 等人[18]针对行人检测准确性不高的问题, 提出 YOLOv5-AC 检测模型, 旨在校正位置偏差, 这一改进有效提高了检测模型的精确度.

上述方法促进了行人检测的发展. 然而, 对于实际应用, 可见图像的鲁棒性无法达到值得信赖的水平. 尤其是在低光照条件下, 其缺陷是不容忽视的.

1.2 红外行人检测

与可见光图像相比, 红外图像能够提供独特的视角信息, 从而突破可见光图像的限制, 其利用热辐射来辅助识别行人的轮廓和特征, 在低光照下更容易检测到行人. Zhou 等人[19]利用 YOLOv5 提出了一种新的红外行人检测网络, 并在颈部网络中引入了一种坐标注意力特征金字塔网络 (CA-FPN). 该方法通过坐标注意模块将位置信息集成到深度特征图中, 从而提高系统的整体性能. Wang 等人[20]设计了 ELAN 网络架构和辅助标头的训练方法, 目的是在不影响检测时间的情况下增加训练成本, 提高准确率. Teutsch 等人[21]提出一种利用最大稳定极值区域 (MSER) 对长波红外图像中的人进行实时检测的两阶段方法. Biswas 等人[22]引入了一种基于局部转向核 (local steering kernel, LSK) 描述符的多维模板形式的中级属性, 用于检测热红外图像中的行人. Chen 等人[23]引入了一种新颖的注意力引导编码器-解码器网络来增加红外图像中的上下文信息, 从而有效地突出行人.

尽管红外传感器对低光照条件具有不可比拟的优势, 但它们仍存在一些固有的局限性, 例如分辨率低、缺乏颜色信息、远距离检测效果差等, 这使得小尺度行人检测极易出现误检和漏检. 虽然上述方法一定程度上提高了行人检测的性能, 但单一模态检测方法在低光照变化下小尺度行人检测性能依然不佳.

1.3 可见光与红外融合的行人检测

为克服单一模态信息量有限和鲁棒性不足的问题, 多光谱特征融合[24]技术近年来备受关注. 该技术联合利用可见光图像与红外图像之间的互补性, 有效提升了复杂环境下的行人检测性能. 随着多模态感知技术的发展, 国内外众多学者在多光谱行人检测领域持续深入探索, 提出了多种具有代表性的检测方法. Hwang 等人[25]提出一种基于 ACF+T+THOG 的多光谱行人检测算法, 通过引入红外图像作为补充训练的模型, 在检测性能上明显优于单模态检测网络. 然而, 该方法依赖传统手工特征提取算子, 存在设计灵活性不足和鲁棒性较差等问题, 从而限制了其在复杂场景中的泛化能力. 随着深度学习的发展, 现有方法主要围绕融合阶段、特征对齐与适应性融合、注意力机制与 Transformer 建模等方面展开.

融合阶段和方式是影响多模态检测性能的关键因

素. 根据特征交互的阶段, 融合策略可分为早期融合、中期融合和晚期融合, 各自具有不同的适用场景与优势. Wagner 等人[26]基于深度卷积神经网络和 R-CNN 架构设计了两种融合网络来处理可见光和红外图像对, 并发现晚期融合策略优于早期融合和传统的聚合通道特征 (ACF)[27]方法. 在此基础上, Liu 等人[28]进一步研究了多光谱行人检测的融合策略, 提出了包含 4 种融合策略的双流 Faster R-CNN 网络, 得出中期融合策略的效果明显优于早期融合、晚期融合及置信度融合.

针对模态间存在的几何错位、特征分布不一致及环境扰动等问题, 研究者引入了多种特征对齐机制与环境感知模块. Zhou 等人[29]提出了 MBNet, 通过差分模态感知融合模块和照明感知特征对齐模块, 实现了检测精度和实时性的双重优化. Liu 等人[30]提出了一种基于拆分和聚合策略的模块, 旨在发现 RGB 图像和红外图像对之间的共享特征和模态特定特征, 以实现跨模态特征学习 (CFL). Ho 等人[31]提出了一种不确定性感知的多光谱行人检测框架, 通过引入不确定性跨模态引导 (UCG) 模块, 引导相对不确定模态的特征分布来学习更可靠模态的特征分布, 以减少模态间的差异, 从而更好地区分行人和背景. Kim 等人[32]提出了一个解决错位情况的多光谱行人检测 (MLPD) 通用框架. Liu 等人[33]提出了一种基于区域的光照与温度感知融合模块, 能够根据光照、温度等环境因素动态调整融合权重, 从而显著提升行人检测性能. 张惊雷等人[34]通过低对比度校正与多尺度增强策略, 有效提升了小尺寸目标的检测能力. 此外, 孙颖等人[35]利用门控融合网络动态调整两种特征的权重分配, 从而实现跨模态特征的有效融合.

为提高融合过程中的信息选择性与判别性, 注意力机制被广泛应用于多模态特征加权与重构. Zhang 等人[36]提出了 CIAN, 其在跨模态交互注意力的指导下融合了中层红外特征和可见光特征. 针对光照和特征模态不平衡问题, BAANet[37]引入模态感知注意力模块, 在不同阶段对两种模态的特征进行重校准, 从而增强融合表示的区分能力. Zhang 等人[38]提出 GAFF, 结合模态间与模态内注意力机制, 有效提升多光谱特征的融合表现. Shen 等人[39]提出了 ICAFusion, 通过全局特征交互提升模型的检测性能. Lee 等人[40]提出一种基于层次化和跨模态引导的模型, 通过跨引导注意力模块 (CGAM) 有效融合不同模态的特征, 从而提升了检测性能. 考虑到 Transformer 具备优秀的长距离依赖建模能力, 研究者逐步将其引入多模态融合中. 为解决现有方法缺乏对跨模态远程依赖性进行建模的能力, Fang 等人[9]设计的跨模态融合 Transformer (CFT), 深入挖掘可见光与红外模态间的深层交互, 为复杂场景下的目标检测提供了新路径.

综上所述, 现有多光谱融合方法主要聚焦于解决模态间的特征不平衡问题, 或通过光照感知、特征对齐、注意力引导及远程依赖建模等手段提升融合质量. 然而, 这些方法在充分挖掘多光谱信息互补性方面仍存在不足. 相比之下, 本文提出的方法通过促进多光谱特征之间的深度交互, 实现可见光-红外图像信息的自适应融合, 能够更准确地捕捉远距离感兴趣区域, 从而有效提升多模态小尺度行人检测的性能.

2 本文方法

2.1 网络结构

ADATNet 的整体结构如图 1 所示. 该方法由 3 个阶段构成: 单模态特征提取、双模态特征融合、颈部聚合以及检测头检测. 单模态特征提取部分由双分支 CSPDarknet53 组成, 分别提取可见光和红外特征. 双模态特征融合部分采用提出的跨模态信息融合 (cross-modality information fusion, CMIF) 模块, 将提取的可见光-红外特征进行充分融合. 颈部聚合部分中 Neck 模块对 3 个不同尺度的融合特征做进一步聚合. 最终通过 Head 模块输出检测结果.

首先, 输入一对可见光和红外图像, 分别对其进行特征提取, 得到可见光红外特征 F_{vi} 和 F_{ir}. 在特征提取阶段, 所提方法通过利用多尺度特征来捕获不同大小的行人目标, 其中 F1–F5 代表不同尺度的特征提取过程. 其次, 将提取的 F_{vi} 和 F_{ir} 进行跨模态特征融合, 得到融合的特征 F_{fused}. 一般来说, 加法运算通常作为特征融合方法, 而本文使用 ADAM 和 ATFE 作为新的融合方法. 最后, 将来自融合模块的特征图送到 ADATNet 的颈部进行多尺度特征融合, 然后传送到检测头进行分类和回归.

2.2 跨模态信息融合

本文在特征提取结构的基础上, 提出了双模态互补信息融合方法, 由自适应双重注意力 (ADAM) 和轴向注意力 Transformer 特征增强 (ATFE) 模块组成. 所提出的 CMIF 结构如图 2 所示.

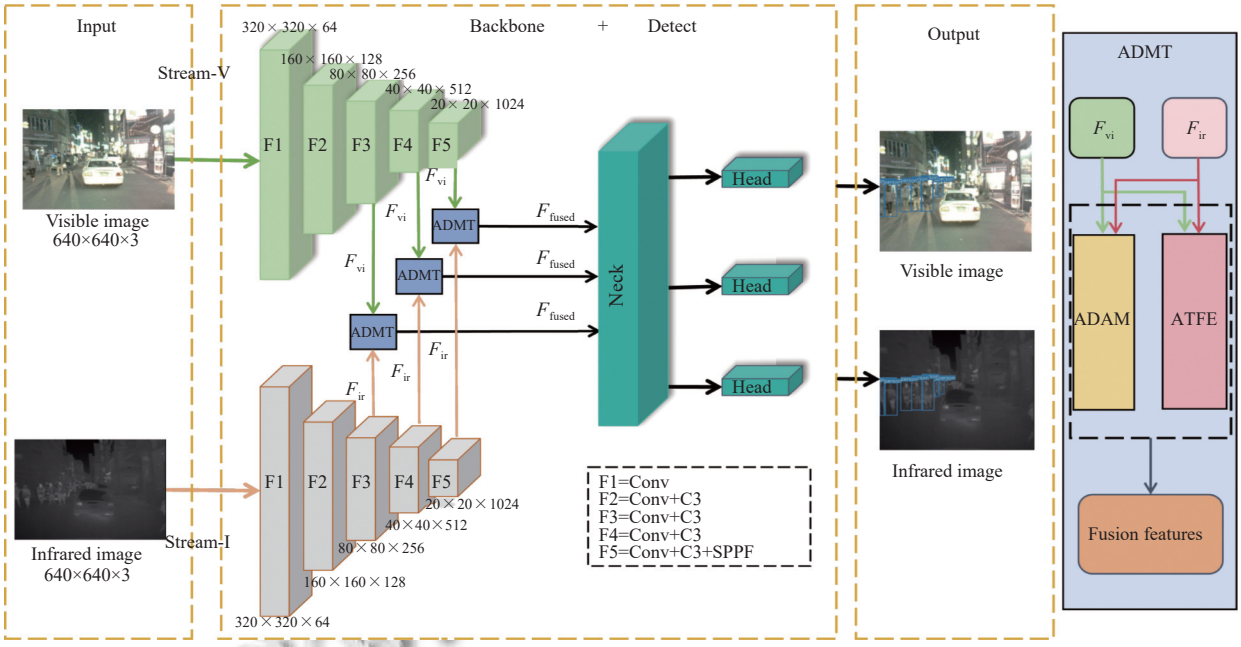


图 1 多光谱行人检测框架

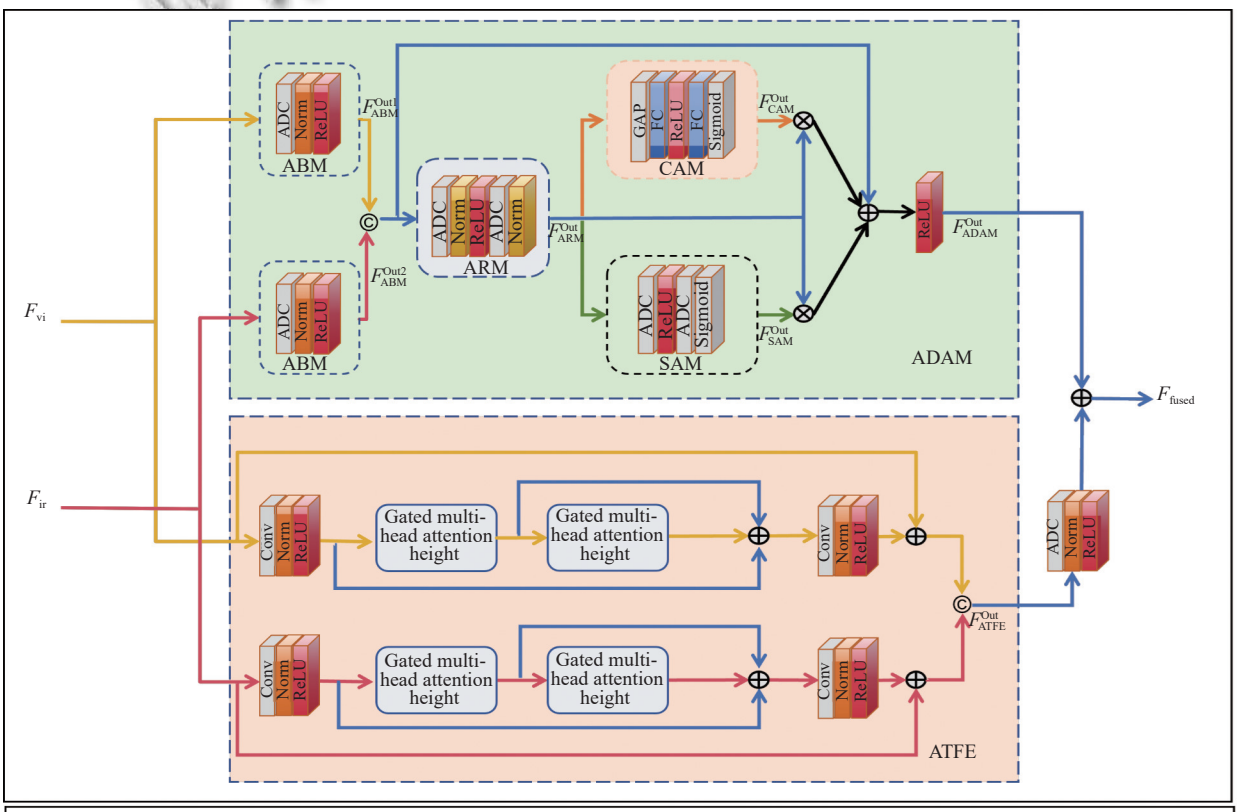


图 2 CMIF 整体结构

2.2.1 自适应双重注意力模块

ADAM 模块主要由自适应基本模块 (ABM)、自适应残差模块 (RM)、通道注意力模块 (CAM) 和自适应空间注意力模块 (SAM) 组成, 其主要作用是为了得到信息丰富的融合特征和显著的行人目标信息, 并且实现特征的降噪提纯.

由于 ADAM 是为了能够同时获得可见光图像中丰富的颜色、纹理信息以及红外图像中的语义信息, 而 Conv 的滑动窗口机制使其只能捕获局部信息, 限制了其利用全局上下文的能力, 导致可见光红外图像信息受到一定损失. 为充分保留可见光-红外图像中的重要互补信息, 在 ADAM 中引入 ADC 来提取特征. 具体来说, ADC 通过学习一系列映射, 从全局上下文特征表示中生成“门”, 从而相应地调制卷积核. 如图 3 所示, 提出的 ADC 由 3 个部分组成, 即上下文编码模块、通道交互模块和门解码模块. 首先, 上下文编码模块将全局上下文信息编码为潜在表示 K; 接着, 通道交互模块进一步处理 K, 将其转换为输出维度为 0 的中间表示 O; 最后, 门解码模块结合 K 和 O, 生成两个关键的门控信号 $G(1)$ 和 $G(2)$, 用于构建最终的注意力门 G. 通过经调制的卷积核, 使卷积层能够在全局上下文的指导下, 动态地捕获具有代表性的局部特征, 并在这一指导下组合感兴趣的局部特征. 与之前的 Conv 基于特征图的修改不同, 修改后的 ADC 可以直接调制卷积核, 并在全局上下文信息的指导下自适应地修改卷积层的权重.

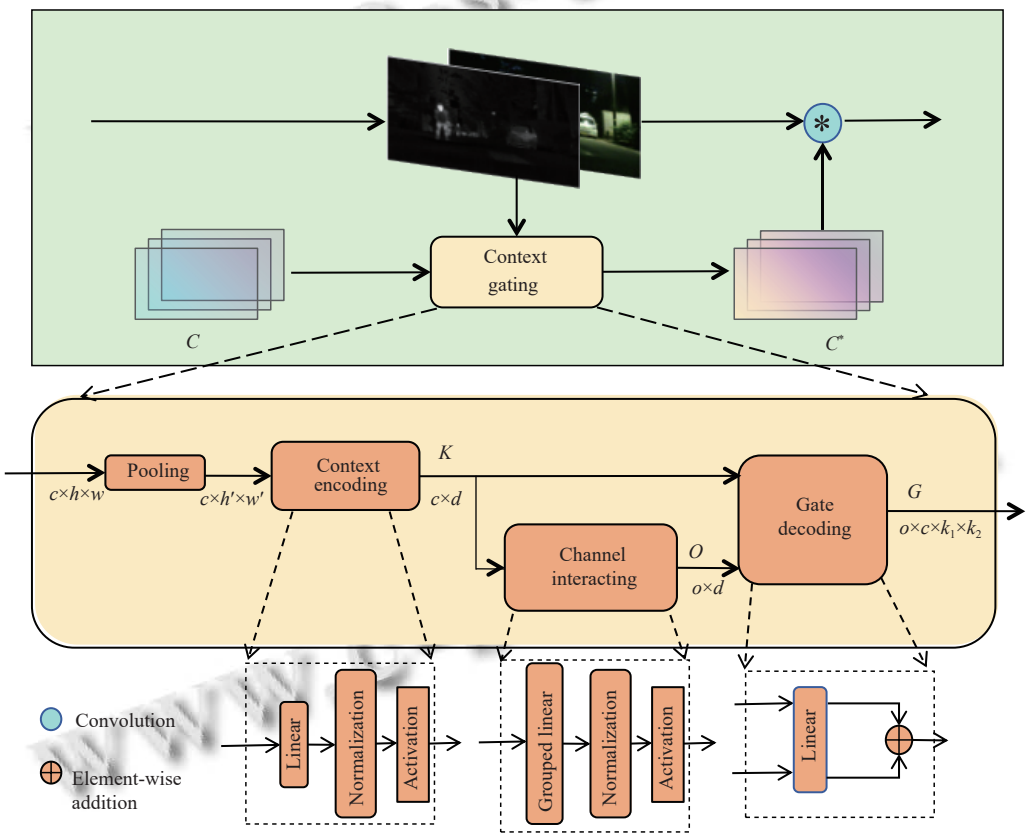


图 3 自适应卷积

将提取的可见光特征F_{vi}和红外特征F_{ir}分别输入到由自适应卷积 (ADC)、批量归一化层 (Norm) 和修正线性单元 (ReLU) 组成的自适应基本模块 (ABM) 中, 自适应的提取可见光-红外图像的特征, 其过程可以表示为:

$$F_{\text{ABM}}^{\text{Out1}} = Re(Norm(ADC(F_{\text{vi}}))) \tag{1}$$

$$F_{\text{ABM}}^{\text{Out2}} = Re(Norm(ADC(F_{\text{ir}}))) \tag{2}$$

其中, $F_{\text{ABM}}^{\text{Out1}}$、$F_{\text{ABM}}^{\text{Out2}}$表示可见光-红外经过 ABM 提取的输出特征, $Norm(\cdot)$ 和 $Re(\cdot)$ 分别表示批量归一化和 ReLU 操作. 将经过 ABM 的可见光红外特征$F_{\text{ABM}}^{\text{Out1}}$、$F_{\text{ABM}}^{\text{Out2}}$经过拼接后, 得到初步融合的特征$F_{\text{RM}}^{\text{In}}$作为 RM 的输入, 从可见光图像中获取纹理特征, 并从红外图像

中获取热分布. RM 由两个自适应卷积层组成:

$$F_{\mathrm{RM}}^{\mathrm{Out}} = Norm(ADC_3^{C,C}(Re(Norm(ADC_3^{C,C}(F_{\mathrm{RM}}^{\mathrm{In}}))))) \quad (3)$$

其中, $ADC_3^{C,C}(\cdot)$ 表示输入和输出通道均为 C 的 3×3 的自适应卷积.

为避免噪声的影响, 同时引入两种注意力机制, 采用并行的 CAM 和 SAM 来处理 RM 的输出 $F_{\mathrm{RM}}^{\mathrm{Out}}$, 从而更充分地捕捉可见光和红外中的显著特征, 降低背景噪声对重要特征的影响. 其中, 为确保实用性和有效性, CAM 的设计基于挤压激励 (SE) 模块. 具体来说, CAM 有 5 层: 第 1 层是全局平均池化 (GAP) 层, 用于生成通道描述符; 第 2 层是全连接 (FC) 层, 用于降低维度; 第 3 层是 ReLU 层; 第 4 层是另一个 FC 层, 用于提高维度. 最后一层是 Sigmoid 激活层. SAM 由 2 个自适应卷积层、1 个 ReLU 和 1 个 Sigmoid 组成, 用于突出空间中的特征. SAM 的特征提取过程可以描述为:

$$F_{\mathrm{SAM}}^{\mathrm{Out}} = Sig(ADC_3^{\frac{C}{R},1}(Re(ADC_3^{C,\frac{C}{R}}(F_{\mathrm{RM}}^{\mathrm{Out}})))) \quad (4)$$

其中, $ADC_3^{C,\frac{C}{R}}(\cdot)$ 表示输入通道为 C、输出通道为 $\frac{C}{R}$ 的 3×3 自适应卷积. $ADC_3^{\frac{C}{R},1}(\cdot)$ 表示输入和输出通道分别为 $\frac{C}{R}$ 和 1 的 3×3 自适应卷积, $Sig(\cdot)$ 是 Sigmoid 激活函数.

尽管注意力机制可以提高可见光-红外融合行人检测的检测能力, 但传统的基于注意力的行人检测方法在通道级特性和空间级特性上被极大压缩, 这不可避免地导致了大量信息丢失. 为在保持或增强注意力优势的同时, 减少信息损耗, 将通过 RM 的特征 $F_{\mathrm{RM}}^{\mathrm{Out}}$ 分别与经过 CAM 处理后的特征 $F_{\mathrm{CAM}}^{\mathrm{Out}}$、SAM 处理后的特征 $F_{\mathrm{SAM}}^{\mathrm{Out}}$ 相乘, 使浅层特征和深层特征得到更充分的交互, 其过程可以表示为:

$$F_{\mathrm{RC}} = F_{\mathrm{CAM}}^{\mathrm{Out}} \cdot F_{\mathrm{RM}}^{\mathrm{Out}} \quad (5)$$

$$F_{\mathrm{RS}} = F_{\mathrm{SAM}}^{\mathrm{Out}} \cdot F_{\mathrm{RM}}^{\mathrm{Out}} \quad (6)$$

其中, F_{RC} 和 F_{RS} 是相乘后的特征. 在 F_{RC}、F_{RS} 和 $F_{\mathrm{RM}}^{\mathrm{In}}$ 之间构建残差连接, 以充分提取卷积运算期间可能丢失的互补信息, 使特征在网络中更有效地传播, 其可以表示为:

$$F_A = F_{\mathrm{RM}}^{\mathrm{In}} + F_{\mathrm{RC}} + F_{\mathrm{RS}} \quad (7)$$

其中, F_A 表示经过残差结构后的融合特征.

ADAM 模块中对于通过 RM、CAM 分支和 SAM 分支的特征进行相乘和残差连接后, 产生两个不同维度的融合特征, 并将其经过 ReLU 激活函数后生成包含丰富的可见光-红外融合特征 $F_{\mathrm{ADAM}}^{\mathrm{Out}}$.

2.2.2 轴向注意力 Transformer 特征增强模块

为了更充分的融合可见光和红外特征, 提出一种基于轴向注意力的 Transformer 特征增强模块 ATFE, 如图 2 所示. 该模块将 Transformer 强大的全局建模能力与卷积局部建模能力相结合, 既能够捕获长距离依赖关系, 又能够有效利用局部特征信息, 从而在复杂背景下精准定位目标区域. ATFE 模块分别对初步提取的可见光和红外特征进行增强, 通过横向和纵向两个轴向的全局信息建模, 实现跨模态特征的空间对齐与语义互补, 提升了目标区域的显著性表达能力.

为降低全局建模的计算复杂度, 在 ATFE 中引入计算上更有效的轴向注意力机制, 以在融合全局特征的同时能够降低计算复杂度. 具体而言, 首先在特征图的高度轴上执行自注意力建模, 随后在宽度轴上进行建模, 从而在两个空间方向上分别建模长距离依赖关系, 达到空间信息的全局建模效果.

门控轴向注意力的前向传播如图 4 所示. 在自注意力机制的基础上添加了相对位置编码 R, 以编码可见光与红外图像的空间结构关系, 从而在特征表示中引入精确的位置信息, 增强对目标区域的感知能力. 与此同时, 为进一步提升特征增强的选择性, 引入可学习的门控机制 G, 对相对位置编码在注意力计算中的作用进行调节, 从而控制非局部上下文信息对当前特征的影响. 改进后的自注意力机制在高度轴上的形式为:

$$y_{ij} = \sum_{h=1}^{H} Softmax\left(q_{ij}^{\mathrm{T}}k_{ih} + G_Q q_{ij}^{\mathrm{T}}R_{ih}^{q} + G_K k_{ij}^{\mathrm{T}}R_{ih}^{k}\right)\left(G_V v_{ih} + G_V R_{ih}^{v}\right) \quad (8)$$

其中, R^q、R^k、$R^v \in R_{H\times H}$ 是高度轴的位置嵌入, G_Q、G_K、$G_V \in R$ 是可学习的参数, 它们共同构成门控机制. 当一个相对位置编码被准确学习时, 门控机制会赋予其更高的权重. 对查询向量 Q, 键向量 K 和值向量 V 引入门控机制, 确保了在特征融合中更高效地过滤和强化有用信息. 具体而言, G_Q 动态过滤特征需求, 使模型只关注当前模态中对行人目标有贡献的区域, 抑制无关区域的干扰. G_K 用于调节不同模态间的特征相关性, 提升跨模态匹配的准确性. G_V 在特征中控制权重分配, 放大对检测任务有用的信息特征, 抑制背景噪声的干扰. 这种基于门控机制的改进有效克服了传统注意力机制在多模态特征处理中存在的特征混淆与对无关区

域响应较强的问题, 进一步增强了可见光与红外特征之间的互补关系, 提高了小目标在复杂场景下的检测鲁棒性.

如图 2 所示, 可见光特征 F_{vi} 与红外特征 F_{ir} 分别输入至 ATFE 的两个分支, 进行轴向注意力建模和特征增强. 提取的可见光-红外特征进入 ATFE 后经过 1×1 的卷积、批量标准化和 ReLU 激活函数处理, 调整通道维度并增强非线性表达能力. 随后依次通过两个门控多头自注意力模块, 分别在特征图的高度轴和宽度轴上进行轴向注意力建模. 特征经过高度轴处理后与初始输入进行残差连接, 再与宽度轴建模后的结果进行第 2 次残差连接, 最大程度地保留了原始信息并融合不同维度的上下文信息. 最终, 融合后的特征再次经过 1×1 卷积、归一化与激活处理, 分别得到增强后的可见光与红外特征. 两个模态的增强特征在通道维度上进行拼接融合, 输出融合特征 $F_{\text{ATFE}}^{\text{Out}}$.

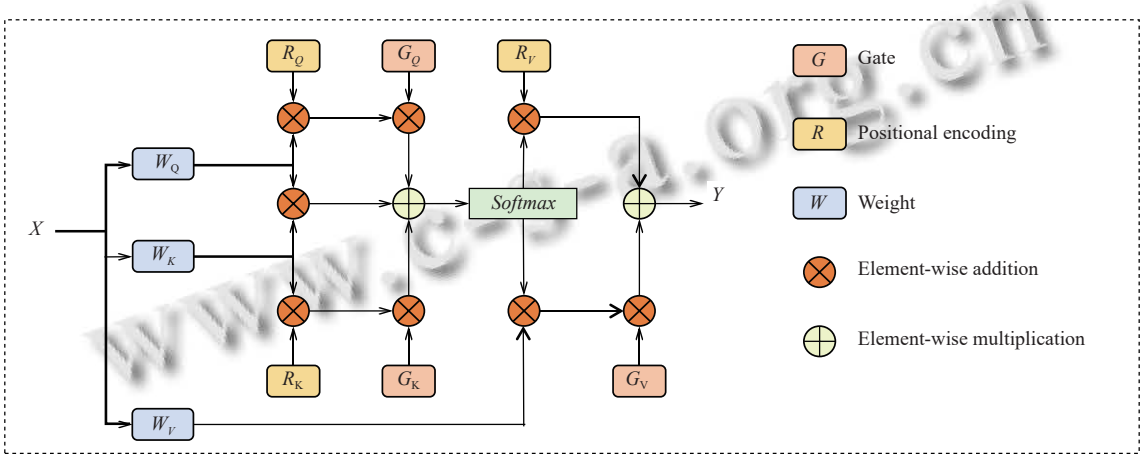


图 4 门控轴向注意力

融合后的特征 $F_{\text{ATFE}}^{\text{Out}}$ 经过 ABM, 由 ADC 自适应提取融合的可见光红外特征, 与通过 ADAM 模块提取的深度融合特征 $F_{\text{ADAM}}^{\text{Out}}$ 进行元素相加, 得到最终融合结果 F_{fused}, 并将其输入至后续的颈部网络与检测头中, 完成多模态行人检测任务.

3 实验结果

3.1 数据集介绍及评价指标

为了验证本文所提方法的有效性, 在 3 个不同规模的可见光-红外图像数据集上进行实验.

KAIST 数据集[25]: 是一个移动视角交通场景行人数据集, 包含白天和夜晚时校园、街道和乡村的各种的交通路段场景. 该数据集共有 95 328 张可见光图像和红外图像, 共有 103 128 个密集注释. 为与其他方法进行公平比较, 在实验中采用由 AR-CNN[41]改进的注释进行训练, 并使用了 2 252 张由 Liu 等人[28]改进的注释的图像进行测试. 具体来说, 其中 1 455 张是在白天拍摄的, 另外 797 张是在夜晚拍摄的.

FLIR 数据集[42]: 是一个质量较高的多光谱目标检测数据集, 包含白天和夜间场景, 并提供了对齐版本的可见光和红外图像. 该数据集包含 5 142 个多光谱图像对, 其中 4 129 对用于训练, 1 013 对用于测试, 涵盖 3 个对象类别: “人”“汽车”和“自行车”.

LLVIP 数据集[43]: 是最近发布的用于弱光视觉的可见红外配对行人数据集. 该数据集包含 30 976 张图像, 即 15 488 对, 其中大部分是在黑暗场景中拍摄的. 所有图像在时间和空间上都严格对齐, 并且仅包括行人作为检测类别.

对数平均漏检率 (MR^{-2}): KAIST 数据集使用该指标进行评估, 该指标通过对在 10^{-2}–100 范围内采样的每个图像误报 (FPPI) 的缺失率进行平均来计算[44]. MR^{-2} 的值越低, 性能越好.

平均精度均值 (mAP): FLIR 和 LLVIP 数据集使用 MS-COCO[45]种引入的常用目标检测指标 mAP 进行评估, mAP 数值越高表示性能越好. mAP50 是指 *IoU* 阈值为 0.5 时的 mAP; mAP75 是指 *IoU* 阈值为 0.75 时的 mAP; mAP 是指 *IoU* 阈值在 0.5–0.95 之间, 间隔为 0.05 的 mAP 平均值.

3.2 实验环境及设置

本实验通过在 CSPDarknet53 中添加红外特征提

取分支, 构建了一个双模态检测网络作为基础模型, 且仅使用加法作为特征融合方法. 为保证实验的公平性, 在训练时对所提方法和基准模型设置相同的参数. 具体来说, 本文算法是在具有 Ubuntu 20.04 系统的服务器上使用 PyTorch 1.13.0 框架实现的, 并在具有 4 张 NVIDIA TITAN X GPU 的显卡上进行实验. 训练阶段需要 70 个 epoch, 批量大小为 8. 模型采用 SGD 优化器, 初始学习率和初始动量分别为 0.01 和 0.937, 权重衰减因子为 0.000 5, 学习率衰减方法为余弦退火. 训练时图像的输入尺寸为 640×640, 测试时图像的输入尺寸为 640×512.

3.3 实验验证及分析

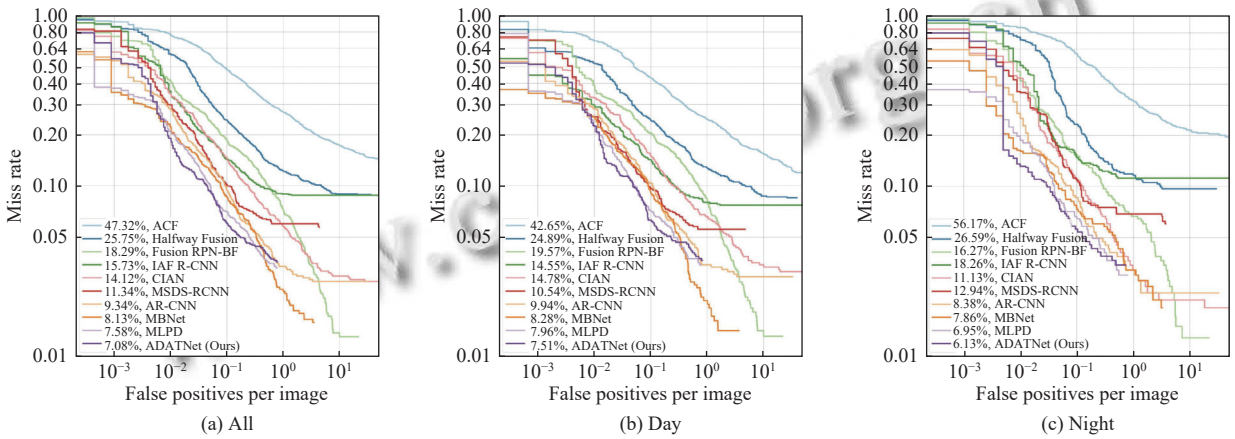
为验证所提方法的性能, 将其与 KAIST、FLIR 和 LLVIP 这 3 个数据集上的最先进方法进行对比, 并从定量和定性方面展示所提方法的优越性.

在 KAIST 训练集上训练本文模型, 并在 KAIST 测试集上进行评估. 与其他主流多光谱行人检测算法 ACF[25]、Halfway Fusion[28]、Fusion RPN-BF[46]、IAF R-CNN[6]、IATDNN-IASS[47]、CIAN[36]、MSDS-RCNN[48]、AR-CNN[41]、CMPD[49]、MBNet[29]、BAANet[37]、RITANet[33]和 MLPD[32]进行对比, 实验结果见表 1. 可以看出, ADATNet 在全天 (All)、白天 (Day) 和夜间 (Night) 这 3 种场景下的 MR^{-2} 值分别为 7.08%、7.51% 和 6.14%, 均优于现有方法, 表现出更高的检测精度和更强的场景适应能力. 尤其是与现有性能最优的 MLPD 方法相比, 在 3 个评估条件下分别提高了 0.50%、0.44% 和 0.81%. 此外, ADATNet 在保持高精度的同时, 也具备较快的检测速度, 具有良好的实用性.

表 1 KAIST 数据集上不同算法对比结果

Method	MR^{-2} (*IoU*=0.5) (%)			Platform	Speed (s)
	All	Day	Night		
ACF[25]	47.32	42.57	56.17	Matlab	2.73
Halfway Fusion[28]	25.75	24.88	26.59	TITAN X	0.43
Fusion RPN-BF[46]	18.29	19.57	16.27	Matlab	0.80
IAF R-CNN[6]	15.73	14.55	18.26	TITAN X	0.25
IATDNN-IASS[47]	14.95	14.67	15.72	—	—
CIAN[36]	14.12	14.77	11.13	1080Ti	0.07
MSDS-RCNN[48]	11.34	10.53	12.94	TITAN X	0.22
AR-CNN[41]	9.34	9.94	8.38	1080 Ti	0.12
CMPD[49]	8.16	8.77	7.31	—	—
MBNet[29]	8.13	8.28	7.86	1080 Ti	0.07
BAANet[37]	7.92	8.37	6.98	TITAN X	0.07
RITANet[33]	7.64	7.73	7.11	—	—
MLPD[32]	7.58	7.95	6.95	2080 Ti	**0.012**
Ours	**7.08**	**7.51**	**6.14**	TITAN X	0.05

为进一步验证 ADATNet 的检测性能, 图 5 展示了 KAIST 数据集上与其他方法在全天、白天和夜间 3 种条件下的 MR-FPPI 曲线.



(a) All (b) Day (c) Night

图 5 MR-FPPI 曲线图

从图 5 可以看出, ADATNet 的曲线在各场景中始终处于最低水平, 表明其在不同误报率 (false positives per image) 下均具有更低的漏检率 (miss rate). 与 MLPD 相比, ADATNet 在中高误检率区间的优势尤为明显, 反映出其在复杂背景或遮挡条件下的检测鲁棒性更强. 此外, 图 5 显示出 ADATNet 在夜间低光条件下的表现尤为突出, 曲线下降更快, 水平更低, 体现了自适应双重注意力和轴向注意力 Transformer 融合策略对于可见光与红外模态特征建模的有效性.

KAIST 数据集上的部分可视化检测结果如图 6 所示. 其中, 绿色边界框表示 TPs, 指真的正样本, 即正样本被正确识别为正样本; 蓝色边界框表示 FPs, 指假的

正样本, 即负样本被错误识别为正样本; 红色边界框表示 FNs, 指假的负样本, 即正样本被错误识别为负样本. 定性分析如下: 与基准模型相比, 所提方法显著减少了漏检和误检. 如图 6 前两行, 当行人目标在可见光图像和红外图像中不明显时, 使用简单融合特征的基准模型无法突出显示目标特征, 导致漏检误检. 而 ADATNet 通过 CMIF 在两种模态的信息间进行了深度交互, 具有更高的信息融合程度, 从而极大地提高了检测性能.

图 6 KAIST 检测结果的可视化

FLIR 数据集上的实验结果见表 2. 与其他主流的多光谱行人检测算法 Halfway Fusion[28], IV-CRN[50], GAFF[38], CFT[9], ICAFusion[39], CrossFormer[40]相比, 本文方法可以取得更好结果. 可以看出, ADATNet 超越了当前最先进的 CrossFormer 方法. CrossFormer 在引入 Transformer 架构后, 有效提升了多模态信息间的特征融合质量, 实现了较高的检测精度, 其 mAP 达到 42.1%. 与之相比, ADATNet 在 mAP75 和 mAP 指标上分别提升了 0.3% 和 1.4%, 尤其在 mAP50 上提升了 3.5%, 表现出更强的检测能力. 这一性能提升主要归功于 ADATNet 中设计的自适应双重注意力模块与轴向注意力特征增强模块, 能更有效地建模局部与全局的模态间语义依赖关系, 从而进一步增强模型对目标的识别能力. 此外, 与基准模型相比, ADATNet 的 mAP50 提高了 4.7%, mAP75 提高了 4.2%, mAP 提高了 3.3%. 这一提升验证了其在捕捉图像丰富上下文信息方面的优势. 与现有多模态行人检测方法主要依赖加权机制对单模态特征进行增强或抑制不同, ADATNet 更注重建模不同层级间的跨模态依赖关系, 并融合多尺度目标特征, 从而在复杂场景中实现更优的检测性能.

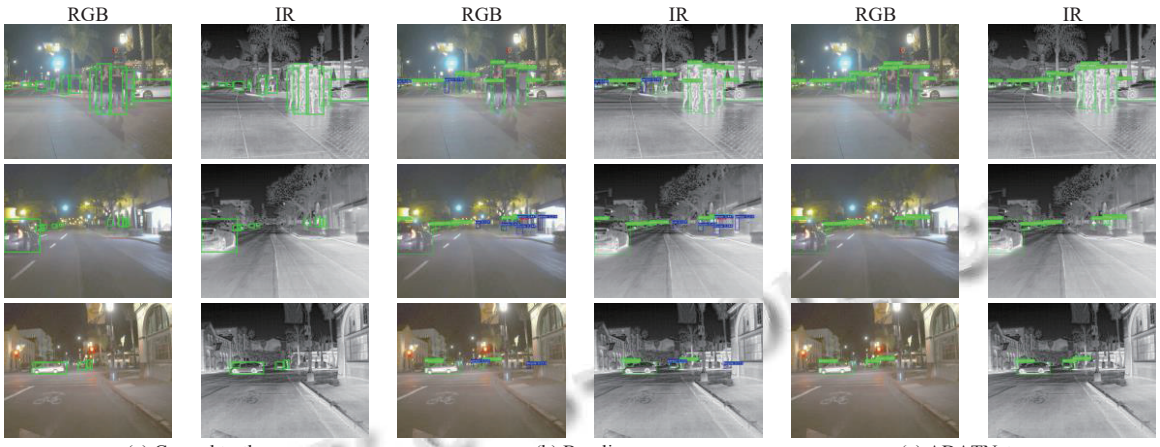
图 7 展示了 FLIR 数据集上的部分可视化检测结果, 其中绿色、蓝色、红色边界框的含义与图 6 相同. 可以看出, ADATNet 与基准模型相比表现更佳, 有效解决了误检和漏检的问题. 具体来说, 该方法在复杂背景下对小尺寸行人的检测效果表现出色. 这是因为该方法有效融合了浅层到深层的特征, 有效增强了目标的特征, 从而提高了网络的检测性能.

表 2 FLIR 数据集上不同算法对比结果 (%)

Method	Dataset type	Backbone	mAP50	mAP75	mAP
Halfway Fusion[28]	RGB+IR	VGG16	71.2	—	—
IV-CRN[50]	RGB+IR	VGG16	72.3	—	—
GAFF[38]	RGB+IR	ResNet18	72.9	32.9	37.5
GAFF[38]	RGB+IR	VGG16	72.7	30.9	37.3
CFT[9]	RGB+IR	CFB	78.7	35.5	40.2
ICAFusion[39]	RGB+IR	CSPDarknet53	79.2	36.9	41.4
CrossFormer[40]	RGB+IR	CSPDarknet53	79.3	38.5	42.1
Baseline	RGB+IR	CSPDarknet53	78.1	34.6	40.2
Ours	RGB+IR	CSPDarknet53	**82.8**	**38.8**	**43.5**

LLVIP 数据集上的实验结果如表 3 所示, 与其他方法相比, ADATNet 在此数据集上也实现了最先进的性能, mAP50 达到 97.6%, mAP75 达到 74.2%, mAP 达到 65.2%. 在不同 *IoU* 阈值下, mAP 均高于单模态检测方法以及多模态的 DRF-SGA[34]、CFT[9]、IACMDF[51]、CrossFormer[40]和 CCIFNet[52]方法. 与首次在多光谱检测中使用 Transformer 的 CFT 模型相比, ADATNet 的 mAP50、mAP75 以及 mAP 分别提高了 0.1%、1.3% 和 1.6%. 总的来说, 其在不同 *IoU* 阈值下, 都能获得最优检测结果, 表明 ADATNet 的优越性. 部分可视化检测结果如图 8 所示, 图中第 1 行的行人被树木遮挡, 第 2 和 3 行存在行人重叠, 此时基准模型存在误检和漏

检, 但是 ADATNet 通过融合两种模态的信息有效解决了误检和漏检问题. 特别要指出的是, 图 8 前两行为两种模态差异过大, 可见光信息严重不足, 甚至红外信息的有效性也降低的情况, 该方法不像基准模型那样平均对待两种模态信息, 而是通过深度交互融合两种模态信息.

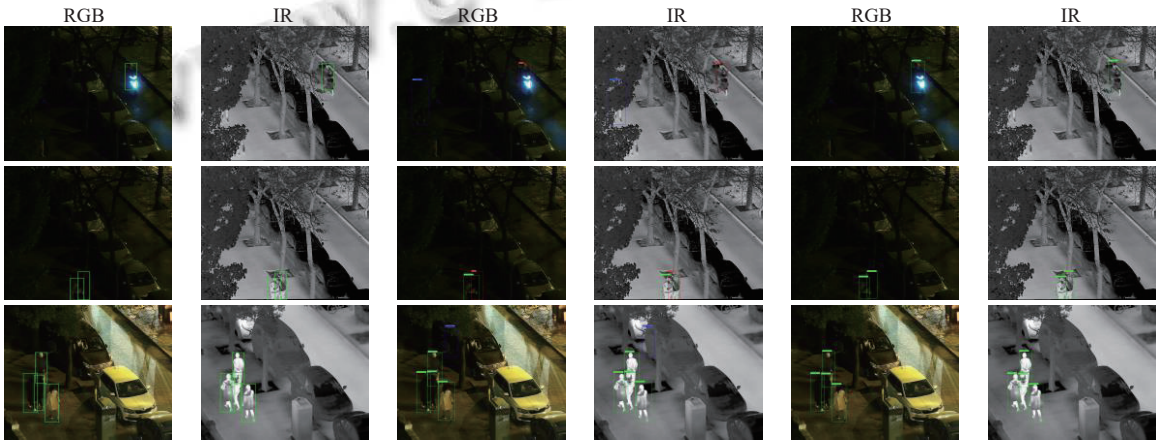


(a) Ground truth (b) Baseline (c) ADATNet

图 7 FLIR 检测结果的可视化

表 3 LLVIP 数据集上不同算法对比结果 (%)

Method	Dataset type	Backbone	mAP50	mAP75	mAP
Faster R-CNN	RGB	ResNet50	91.4	48.0	49.2
Faster R-CNN	IR	ResNet50	96.1	68.5	61.1
SSD[9]	RGB	VGG16	82.6	31.8	39.8
SSD[9]	IR	VGG16	90.2	57.9	53.5
YOLOv5[43]	RGB	CSPDarknet53	90.8	51.9	50.5
YOLOv5[43]	IR	CSPDarknet53	94.6	72.2	61.9
DRF-SGA[34]	RGB+IR	CSPDarknet53	96.4	—	63.9
CFT[9]	RGB+IR	CFB	97.5	72.9	63.6
IACMDF[51]	RGB+IR	ResNet50	97.3	74.1	65.1
CrossFormer[40]	RGB+IR	CSPDarknet53	97.4	—	65.1
CCIFNet[52]	RGB+IR	ResNet50	97.6	72.6	64.1
Baseline	RGB+IR	CSPDarknet53	95.7	72.5	62.3
Ours	RGB+IR	CSPDarknet53	**97.6**	**74.2**	**65.2**



(a) Ground truth (b) Baseline (c) ADATNet

图 8 LLVIP 检测结果的可视化

3.4 热图可视化结果分析

图 9 展示了基准模型和 ADATNet 在 3 个数据集上实际检测时的特征热图, 图 9(a), (b) 为 RGB 和红外图像的真实框. 可以观察到, ADATNet 所提取的激活特征更集中于目标, 而基准模型提取的激活特征较为分散, 且还存在将背景特征错误判定为目标类激活特征的情况. 该可视化结果进一步证明本文方法相比基准模型在提取可见光红外互补性特征方面更具优势.

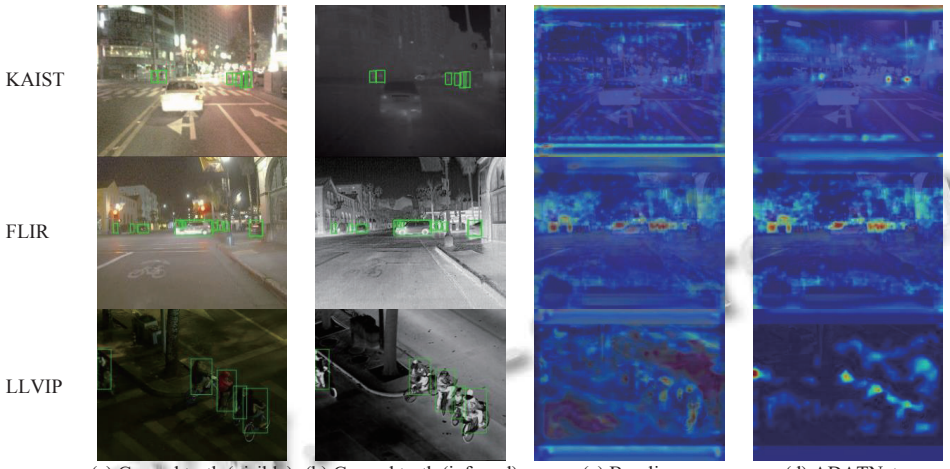


图 9 热图的可视化结果

3.5 消融实验

不同模块的加入对 ADATNet 性能的提升不同, 为了验证所提出方法中每个模块的有效性, 在 FLIR 数据集上进行全面的消融实验, 通过实验数据客观证明不同组件加入 CMIF 方法产生的性能变化. 评估指标有精确率、召回率、mAP50、mAP75 和 mAP.

(1) CMIF 的必要性

为进行比较, 使用双分支并行的 CSPDarknet53 和逐元素相加作为融合操作的网络作为基准模型, 然后, 分别将 ADAM 和 ATFE 引入基准模型中, 最后, 将由 ADAM 和 ATFE 构成的 CMIF 引入基准模型中. 表 4 显示了 4 种模型的性能指标表现, 即没有使用 ADAM 和 ATFE、仅使用 ADAM、仅使用 ATFE 和同时使用 ADAM 和 ATFE.

表 4 加入不同模块的消融实验结果 (%)

ADAM	ATFE	Precision	Recall	mAP50	mAP75	mAP
—	—	79.2	71.6	78.1	34.6	40.2
√	—	80.0 (+0.8)	73.8 (+2.2)	80.9 (+2.8)	37.0 (+2.4)	41.7 (+1.5)
—	√	81.5 (+2.3)	75.5 (+3.9)	81.3 (+3.2)	35.1 (+0.5)	40.4 (+0.2)
√	√	**83.6 (+4.4)**	**74.6 (+3.0)**	**82.8 (+4.7)**	**38.8 (+4.2)**	**43.5 (+3.3)**

注: 括号中的数值为较基本骨干网络的变化值

从表 4 可以看出, 仅将 ADAM 模块嵌入基准模型, 整个模型开始准确关注源图像的显著特征, 使模型可以提取更多目标特征. 与基准模型相比, 准确率和召回率都有所提升, mAP50 提高 2.8%, mAP75 提高 2.4%, mAP 提高 1.5%. 仅将 ATFE 模块添加到基准模型, 整个模型很好地构建了远程依赖关系, 与基准模型相比, mAP50 提高 3.2%, mAP75 提高 0.5%, mAP 提高 0.2%. 当 ADAM 与 ATFE 相结合时, 可以自适应的同时从源图像中捕获局部和全局信息, 与基准模型相比, mAP50 提高 4.7%, mAP75 提高 4.2%, mAP 提高 3.3%.

从实验结果看出, 在基准模型中同时添加 ADAM 和 ATFE 给模型性能带来最大的提升, 准确率、召回率、mAP50、mAP75 和 mAP 都取得最高效果, 表明改进方法是有效的.

(2) ADC 的影响

为证明 CMIF 中采用的 ADC 有效性, 在 CMIF 中使用 Conv 替换所有 ADC 进行实验. 如表 5 所示, 当 ADAM 中的卷积为 Conv 时, 整个网络的 mAP50 只提高 3.3%, mAP75 只提高 0.5%, mAP 只提高 0.5%; 当 ADAM 中卷积为 ADC 时, 整个网络的 mAP50 提高了 4.7%, mAP75 提高了 4.2%, mAP 提高了 3.3%. 模型的精度、召回率、mAP50、mAP75、mAP 均达到最高值. 可以明显看出, 使用 ADC 时, 有效获取和利用了源图像中的全局上下文信息, 其效果最优.

表 5 Conv 与自适应卷积的比较 (%)

Module	Precision	Recall	mAP50	mAP75	mAP
Baseline	79.2	71.6	78.1	34.6	40.2
Conv	82.8 (+3.6)	73.6 (+2.0)	81.4 (+3.3)	35.1 (+0.5)	40.7 (+0.5)
ADC	**83.6 (+4.4)**	**74.6 (+3.0)**	**82.8 (+4.7)**	**38.8 (+4.2)**	**43.5 (+3.3)**

4 结论

针对多光谱行人检测任务中存在的多模态相互作用不足和融合方法缺乏远程依赖性的问题, 本文提出了一种跨模态互补信息融合的行人检测网络. 具体而言, 提出了一种基于自适应双重注意力和轴向注意力 Transformer 的多光谱小尺度行人检测算法, 充分利用可见光和红外模态之间的互补特性. 其中 ADAM 模块强化模型对关键特征的关注, 同时抑制不相关或冗余的信息, 为后续的检测提供了更有效的融合特征信息, 减少了低光照条件下的噪声信息干扰. ATFE 模块关联多模态特征的位置编码来融合增强的特征. 实验结果表明, 所提出的方法在 KAIST、FLIR、LLVIP 数据集上均表现出最佳检测性能. 在未来的工作中, 将进一步优化网络结构, 探索更高效的特征融合策略, 在保证精度的情况下提高算法的推理速度, 以更好地满足实际应用需求.

参考文献

1 Yuan W, Yang M, Wang CX, *et al*. VRDriving: A virtual-to-real autonomous driving framework based on adversarial learning. IEEE Transactions on Cognitive and Developmental Systems, 2021, 13(4): 912–921. [doi: 10.1109/TCDS.2020.3006621]

2 Li XD, Ye M, Liu YG, *et al*. Accurate object detection using memory-based models in surveillance scenes. Pattern Recognition, 2017, 67: 73–84. [doi: 10.1016/j.patcog.2017.01.030]

3 别倩, 王晓, 徐新, 等. 红外-可见光跨模态的行人检测综述. 中国图象图形学报, 2023, 28(5): 1287–1307. [doi: 10.11834/jig.220670]

4 Cao JL, Pang YW, Xie J, *et al*. From handcrafted to deep features for pedestrian detection: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(9): 4913–4934. [doi: 10.1109/TPAMI.2021.3076733]

5 Ranjbarzadeh R, Dorosti S, Jafarzadeh Ghoushchi S, *et al*. Nerve optic segmentation in CT images using a deep learning model and a texture descriptor. Complex & Intelligent Systems, 2022, 8(4): 3543–3557.

6 Li CY, Song D, Tong RF, *et al*. Illumination-aware Faster R-CNN for robust multispectral pedestrian detection. Pattern Recognition, 2019, 85: 161–171. [doi: 10.1016/j.patcog.2018.08.005]

7 Zhang L, Zhu XY, Chen XY, *et al*. Weakly aligned cross-modal learning for multispectral pedestrian detection. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019. 5126–5136.

8 Wang ZA, Liao XH, Yuan J, *et al*. CDC-YOLOFusion: Leveraging cross-scale dynamic convolution fusion for visible-infrared object detection. IEEE Transactions on Intelligent Vehicles, 2024. [doi: 10.1109/TIV.2024.3443264]

9 Fang QY, Han DP, Wang ZK. Cross-modality fusion transformer for multispectral object detection. arXiv:2111.00273, 2021.

10 Lin XD, Ma L, Liu W, *et al*. Context-gated convolution. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 701–718.

11 Girshick R, Donahue J, Darell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 580–587.

12 Ren S, He K, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149.

13 Girshick R. Fast R-CNN. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 1440–1448.

14 Li JN, Liang XD, Shen SM, *et al*. Scale-aware fast R-CNN for pedestrian detection. IEEE Transactions on Multimedia, 2018, 20(4): 985–996.

15 Gao XL, Ge DY, Chen ZF. The research on autopilot system based on lightweight YOLO-V3 target detection algorithm. Journal of Physics: Conference Series, 2020, 1486(3): 032028. [doi: 10.1088/1742-6596/1486/3/032028]

16 Ma LZ, Chen Y, Zhang JL. Vehicle and pedestrian detection based on improved YOLOv4-tiny model. Journal of Physics: Conference Series, 2021, 1920(1): 012034. [doi: 10.1088/1742-6596/1920/1/012034]

17 Wang CY, Bochkovskiy A, Liao HYM. Scaled-YOLOv4: Scaling cross stage partial network. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 13024–13033.

18 Lv HH, Yan HB, Liu KY, *et al*. YOLOv5-AC: Attention mechanism-based lightweight YOLOv5 for track pedestrian detection. Sensors, 2022, 22(15): 5903. [doi: 10.3390/s22155903]

19 Zhou L, Gao S, Wang SM, *et al*. IPD-Net: Infrared pedestrian detection network via adaptive feature extraction and coordinate information fusion. Sensors, 2022, 22(22): 8966. [doi: 10.3390/s22228966]

20 Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 7464–7475.

21 Teutsch M, Mueller T, Huber M, *et al*. Low resolution person detection with a moving thermal infrared camera by hot spot classification. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Columbus: IEEE, 2014. 209–216.

22 Biswas SK, Milanfar P. Linear support tensor machine with LSK channels: Pedestrian detection in thermal infrared images. IEEE Transactions on Image Processing, 2017, 26(9): 4229–4242. [doi: 10.1109/TIP.2017.2705426]

23 Chen YF, Shin H. Pedestrian detection at night in infrared images using an attention-guided encoder-decoder convolutional neural network. Applied Sciences, 2020, 10(3): 809. [doi: 10.3390/app10030809]

24 Kim JU, Park S, Ro YM. Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(3): 1510–1523. [doi: 10.1109/TCSVT.2021.3076466]

25 Hwang S, Park J, Kim N, *et al*. Multispectral pedestrian detection: Benchmark dataset and baseline. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 1037–1045.

26 Wagner J, Fischer V, Herman M, *et al*. Multispectral pedestrian detection using deep fusion convolutional neural networks. Proceedings of the 24th European Symposium on Artificial Neural Networks. Bruges: ESANN, 2016. 509–514.

27 Dollár P, Appel R, Belongie S, *et al*. Fast feature pyramids for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(8): 1532–1545. [doi: 10.1109/TPAMI.2014.2300479]

28 Liu JJ, Zhang ST, Wang S, *et al*. Multispectral deep neural networks for pedestrian detection. arXiv:1611.02644, 2016.

29 Zhou KL, Chen LS, Cao X. Improving multispectral pedestrian detection by addressing modality imbalance problems. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 787–803.

30 Liu TS, Lam KM, Zhao R, *et al*. Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(1): 315–329. [doi: 10.1109/TCSVT.2021.3060162]

31 Ho J, Kalchbrenner N, Weissenborn D, *et al*. Axial attention in multidimensional transformers. arXiv:1912.12180, 2019.

32 Kim J, Kim H, Kim T, *et al*. MLPD: Multi-label pedestrian detector in multispectral domain. IEEE Robotics and Automation Letters, 2021, 6(4): 7846–7853. [doi: 10.1109/LRA.2021.3099870]

33 Liu YH, Hu C, Zhao BX, *et al*. Region-based illumination-temperature awareness and cross-modality enhancement for multispectral pedestrian detection. IEEE Transactions on Intelligent Vehicles, 2024. [doi: 10.1109/TIV.2024.3367688]

34 张惊雷, 宫文浩, 贾鑫. 基于自引导注意力的双模态校准融合目标检测算法. 模式识别与人工智能, 2023, 36(9): 793–805. [doi: 10.16451/j.cnki.issn1003-6059.202309003]

35 孙颖, 侯志强, 杨晨, 等. 基于双模态融合网络的目标检测算法. 光子学报, 2023, 52(1): 0110002.

36 Zhang L, Liu ZY, Zhang SF, *et al*. Cross-modality interactive attention network for multispectral pedestrian detection. Information Fusion, 2019, 50: 20–29. [doi: 10.1016/j.inffus.2018.09.015]

37 Yang XX, Qiang YQ, Zhu HJ, *et al*. BAANet: Learning bi-directional adaptive attention gates for multispectral pedestrian detection. Proceedings of the 2022 International Conference on Robotics and Automation. Philadelphia: IEEE, 2022. 2920–2926.

38 Zhang H, Fromont E, Lefèvre S, *et al*. Guided attentive feature fusion for multispectral pedestrian detection. IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2021. 72–80.

39 Shen JF, Chen YF, Liu Y, *et al*. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. Pattern Recognition, 2024, 145: 109913. [doi: 10.1016/j.patcog.2023.109913]

40 Lee S, Park J, Park J. CrossFormer: Cross-guided attention for multi-modal object detection. Pattern Recognition Letters, 2024, 179: 144–150. [doi: 10.1016/j.patrec.2024.02.012]

41 Zhang L, Liu ZY, Zhu XY, *et al*. Weakly aligned feature fusion for multimodal object detection. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(3): 4145–4159. [doi: 10.1109/TNNLS.2021.3105143]

42 FLIR ADA Team. 用于算法训练的 Teledyne FLIR 数据集. https://www.flir.cn/oem/adas/adas-dataset-form/. [2025-02-01].

43 Jia XY, Zhu C, Li MZ, *et al*. LLVIP: A visible-infrared paired dataset for low-light vision. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE, 2021. 3489–3497.

44 Dollar P, Wojek C, Schiele B, *et al*. Pedestrian detection: An evaluation of the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(4): 743–761. [doi: 10.1109/TPAMI.2011.155]

45 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer, 2014. 740–755.

46 König D, Adam M, Jarvers C, *et al*. Fully convolutional region proposal networks for multispectral person detection. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 243–250.

47 Guan DY, Cao YP, Yang JX, *et al*. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. Information Fusion, 2019, 50: 148–157. [doi: 10.1016/j.inffus.2018.11.017]

48 Li CY, Song D, Tong RF, *et al*. Multispectral pedestrian detection via simultaneous detection and segmentation. arXiv:1808.04818, 2018.

49 Li Q, Zhang CQ, Hu QH, *et al*. Confidence-aware fusion using Dempster-Shafer theory for multispectral pedestrian detection. IEEE Transactions on Multimedia, 2023, 25: 3420–3431. [doi: 10.1109/TMM.2022.3160589]

50 Tang YY, Jiang B. The infrared-visible complementary recognition network based on context information. Proceedings of the 14th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI). Shanghai: IEEE, 2021. 1–6.

51 Wang CS, Qian JJ, Wang J, *et al*. Illumination-aware cross-modality differential fusion multispectral pedestrian detection. Electronics, 2023, 12(17): 3576. [doi: 10.3390/electronics12173576]

52 Yan CQ, Zhang H, Li XL, *et al*. Cross-modality complementary information fusion for multispectral pedestrian detection. Neural Computing and Applications, 2023, 35(14): 10361–10386. [doi: 10.1007/s00521-023-08239-z]

(校对责编: 王欣欣)