计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(4):329−338 [doi: 10.15888/j.cnki.csa.009022]

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

基于 YOLO 的多模态特征差分注意融合行人检测①

王　钊[1], 解文彬[1], 文　江[2]

[1](陆军工程大学 指挥控制工程学院, 南京 210007)
[2](陆军研究院 工程设计研究所, 广州 510515)
通信作者: 解文彬, E-mail: 1171801987@qq.com

摘　要: 针对可见光模态与热红外模态间的差异问题和如何充分利用多模态信息进行行人检测, 本文提出了一种基于 YOLO 的多模态特征差分注意融合行人检测方法. 该方法首先利用 YOLOv3 深度神经网络的特征提取主干分别提取多模态特征; 其次在对应多模态特征层之间嵌入模态特征差分注意模块充分挖掘模态间的差异信息, 并经过注意机制强化差异特征表示进而改善特征融合质量, 再将差异信息分别反馈到多模态特征提取主干中, 提升网络对多模态互补信息的学习融合能力; 然后对多模态特征进行分层融合得到融合后的多尺度特征; 最后在多尺度特征层上进行目标检测, 预测行人目标的概率和位置. 在 KAIST 和 LLVIP 公开多模态行人检测据集上的实验结果表明, 提出的多模态行人检测方法能有效解决模态间的差异问题, 实现多模态信息的充分利用, 具有较高的检测精度和速度, 具有实际应用价值.

关键词: 多模态; YOLOv3; 特征差分; 注意机制; 行人检测

引用格式: 王钊,解文彬,文江.基于 YOLO 的多模态特征差分注意融合行人检测.计算机系统应用,2023,32(4):329–338. http://www.c-s-a.org.cn/1003-3254/9022.html

Pedestrian Detection Based on Multimodal Feature Differential Attention Fusion and YOLO

WANG Zhao[1], XIE Wen-Bin[1], WEN Jiang[2]

[1](Command and Control Engineering Academy, Army Engineering University of PLA, Nanjing 210007, China)
[2](Engineering Design Institute, Army Research Institute, Guangzhou 510515, China)

Abstract: In order to address the difference between visible light modality and thermal infrared modality and make full use of multimodal information to perform pedestrian detection, this study proposes a multimodal feature differential attention fusion pedestrian detection method based on YOLO. The method first uses the feature extraction backbone of the YOLOv3 deep neural network to extract multimodal features respectively. Second, the differential attention module of modal features is embedded between the corresponding multimodal feature layers to fully mine the difference information between modalities, and the difference feature representation is strengthened through the attention mechanism, so as to improve the quality of feature fusion. Then, the difference information is fed back to the multimodal feature extraction backbone to improve the network's ability to learn and fuse multimodal complementary information. In addition, the multimodal features are fused in layers to obtain the multi-scale features. Finally, target detection is performed on the multi-scale feature layer to predict the probability and location of pedestrian targets. The experimental results on the public multimodal pedestrian detection datasets of KAIST and LLVIP show that the proposed multimodal pedestrian detection method can effectively address the difference between modalities and realize the full use of multimodal information. Furthermore, it has high detection accuracy and speed and is of practical application value.

Key words: multimodal; YOLOv3; feature differential; attention mechanism; pedestrian detection

① 收稿时间: 2022-08-19; 修改时间: 2022-09-22; 采用时间: 2022-10-14; csa 在线出版时间: 2022-12-23
CNKI 网络首发时间: 2022-12-27

行人检测是通用目标检测的一个特殊分支, 是计算机视觉的关键任务之一, 是视频监控系统[1,2] 和自动驾驶[3] 等各种现实应用的基础. 通常, 行人检测是基于颜色模态进行的. 然而, 颜色模态很容易受到如微光[4] 和背景杂波[5] 等具有挑战性的环境的影响. 虽然基于颜色模态的方法在行人检测方面取得了很大进展, 但在具有挑战性的环境中仍然遇到了困难.

为了解决在颜色模态中遇到的问题, 热模态作为一种附加模态被逐渐采用[6]. 许多研究表明, 通过结合不同光谱[7,8] 的视觉信息进行编码可以得到比单一模态更丰富的物体视觉表征. 由于这一优点, 许多工业应用都采用了多模态信息来解决相应问题[9]. 同时越来越多的行人检测研究也尝试采用多模态来提高检测性能[10–13].

然而, 由于多光谱图像是由两台相机根据不同的光谱波段[7] 捕获的, 因此探测到的多模态信息存在模态间的差异性. 图 1 从人类视觉的角度直观展示了不同场景下模态间的差异性. 在夜晚低照度情况下可见光图像 (图 1(a)) 中的行人目标很难从背景中区分出来, 热红外图像 (图 1(c)) 中的行人目标显著, 易于识别. 而图 1(b) 和图 1(d) 所示的情况恰好相反, 在白天场景中可见光图像中的行人信息显著易于识别, 但热红外图像中的行人目标因与背景较为相似导致目标信息不突出, 难以检测.

(a) 可见光 (夜晚) (b) 可见光 (白天)

(c) 热红外 (夜晚) (d) 热红外 (白天)

图 1 模态差异的例子

为了缓解上述多模态信息间的差异问题, Li 等人[11] 提出了光照感知网络来预测每个模态的光照得分, 并以此作为判定每个模态重要性的依据. 通过实验得出在夜间环境下热模态特性比颜色模态更重要, 而在白天则相反. 因此, 最终分类和定位的结果是依据光照得分来细化的. 虽然光照因素可以作为判定颜色模态和热模态在多模态行人检测中的重要性依据, 但将光照因素作为模态重要性的唯一评判标准是不够准确的. 在基于多模态信息的行人检测中, 根据光照感知预测的光照得分往往导致不同模态的重要性出现两级分化的情况, 一个模态的重要性要远远高于另一个模态, 这就必然导致光照得分过于低的模态的信息得不到充分利用, 这对基于多模态信息的行人检测来说是一大损失. 因此, 根据图像信息获得的光照权重不能完全代表模态的重要性, 过于依赖光照得分可能会导致性能受限.

在本文中, 我们受到模态差分[13] 思想的启发提出了一种基于 YOLO 的多模态特征差分注意融合行人检测方法, 以解决多模态差异性问题和充分利用多模态信息.

1 相关研究

由于低成本热传感器的日益普及, 将热成像技术用于行人检测已经引起了研究者们的极大兴趣. 在本节中, 首先简要介绍了行人检测文献中基于可见光和热传感器的单模行人检测, 然后介绍了基于两种传感器融合的多模态行人检测.

1.1 基于可见光谱的行人检测

雾、雨、尘等不利天气条件、光照变化、复杂背景等使得在可见光谱中的行人检测任务极其困难. 为了在一定程度上解决这一问题, Tian 等人[14] 提出了一种结合场景属性学习行人检测语义属性的联合优化方法. Li 等人[10] 也用到了类似的方法, 其在论文中将行人检测与其他视觉任务相结合来影响行人检测. 在另一项研究[15] 中, 采用通用的 Faster-RCNN[16] 方法来解决行人检测问题. Liu 等人[17] 提出了一种无锚框的行人检测方法, 可以显著提高检测速度.

1.2 基于红外热成像的行人检测

相对而言, 仅基于热红外图像的行人检测文献有限. Baek 等人[18] 提出了一种仅针对夜间的检测方法, 即利用定向梯度的热-位置-强度直方图 (TπHOG) 结合加性核 SVM (AKSVM) 的方法. Devaguptapu 等人[19] 利用 Cycle-GAN[20] 的图像-图像平移框架生成伪 RGB 等效图像进行行人检测. 近年来, 领域自适应技术被应用于热图像的行人检测[21,22] 中, 该技术可以充分利用现有的红外数据.

1.3 基于多模态的行人检测

从 Hwang 等人[7] 提出的多光谱行人检测基线方法 ACF+T+THOG 开始, 可见光和热红外图像融合在

行人检测中已被证明是有效的, 能够表现出比单模态行人检测更好的性能.

Liu 等人[8] 对各种融合结构进行了广泛深入的研究, 具体研究了 4 种不同的网络融合方法 (早期、中期、后期和结果融合), 并在严格配准的可见光与热红外多模态数据集上进行了行人检测实验, 结果表明在网络提取特征的中间阶段进行多模态融合得到的模型检测性能最好.

Li 等人[11] 提出了基于照明感知的 Faster R-CNNs 方法来执行多模态行人检测, 利用神经网络预测输入图像的光照值, 并将光照值作为不同模态的重要性依据指导多模态融合, 在公开的多光谱数据集上的行人检测结果表明该方法能达到很高的检测精度.

Zheng 等人[23] 使用了两个 SSD 检测器和门控融合单元 (gated fusion units) 实现了可见光和热红外模态特征的有效融合, 在多模态行人检测数据集上取得了较高的检测性能并达到了当时最快检测速度.

Zhang 等人[12] 发现了可见光-热红外图像对中的位置偏移问题, 由于多模态数据集没有严格对齐, 因此其中一种模态数据的位置偏移会影响整个多模态融合效果, 最终导致检测性能较差. 为此作者设计了一种区域对齐模块来捕获位置偏移, 自适应对齐两种模式的区域特征, 并提出了一种新的多模态融合方法, 通过特征重新加权来选择更可靠的特征, 抑制无用的特征, 在 KAIST 多模态行人检测数据集上通过大量实验表明该方法具有很高的检测精度和较强的鲁棒性.

Zhou 等人[13] 借助差分电路的思想提出了模态差分解决了多模态数据的模态不平衡问题, 即将一个模态的差异信息融入另一个模态, 实现模态间的信息交流, 并借助光照感知获得的光照权重实现两种模态的融合, 在公共数据集上达到了当时最佳检测精度和最快的检测速度.

Fu 等人[24] 在 YOLO 一阶段目标检测框架的基础上提出空间自适应像素级特征融合网络, 采用双 YOLO 目标检测框架分别用于提取可见光和热红外图像的多尺度特征, 并借助注意力机制提取像素级特征权重实现双流网络的像素级多尺度特征融合, 该方法在检测精度方面与主流方法不相上下, 但由于采用了 YOLO 检测框架, 使得检测速度有了较大提升.

Fang 等人[25] 提出了一种简单有效的跨模态特征融合方法, 通过在双流 YOLO 网络之间嵌入 Transformer 注意力机制, 指导网络学习远程依赖关系, 并在特征提取阶段集成全局上下文信息. 更重要的是, 通过利用 Transformer 的自注意机制, 网络可以同时进行模内和模间信息融合, 鲁棒捕获可见光和热红外域之间的潜在相互作用, 从而显著提高多模态目标检测性能. 在多个多模态数据集上的实验结果表明该方法可以达到很高的检测性能.

基于可见光与热红外的双模态行人检测方法融合了两个模态的互补信息, 可以有效应对光照、复杂环境等不利因素, 能够实现全天候行人检测, 相比单模态行人检测具有更强的鲁棒性和可靠性. 但在多模态行人检测研究中存在以下几个问题: 一是现有检测方法中很少考虑两个模态间存在的差异性, 二是简单的融合方法不能最大限度地发掘出多模态信息的真正价值. 为解决上述问题, 充分发挥出双模态互补信息在行人检测中的作用, 提出了本文的多模态行人检测方法.

2 方法介绍

为解决模态差异问题和充分利用多模态信息, 本文提出了基于 YOLO 的多模态特征差分注意融合行人检测方法. 首先利用深度神经网络提取多模态特征; 其次在多模态特征层间嵌入模态特征差分注意模块获取模态间的差异信息并分别反馈到特征提取网络中, 提升网络对多模态信息的融合能力; 然后对多模态特征分层融合得到融合多尺度特征; 最后在多尺度特征层上进行目标检测, 预测行人目标的概率和位置.

2.1 网络主体架构

网络总体架构如图 2 所示, 基于 YOLOv3[26] 目标检测网络搭建而成, 主要由两条特征提取主干、模态特征差分注意、模态特征融合、多尺度特征金字塔和检测层组成. 随着 YOLO 系列一阶段目标检测网络的提出, 目标检测的精度和速度都得到了很大提升, 其强大的特征提取能力和近乎实时的检测速度是我们选用 YOLO 框架构建多模态行人检测网络的依据.

在多模态行人检测网络中, 首先使用两条 YOLOv3 网络中提出的 Darknet53 特征提取主干分别提取可见光模态和热红外模态的多尺度特征, V_1、V_2、V_3、V_4、V_5 和 T_1、T_2、T_3、T_4、T_5 分别是两个模态的原始图像经过卷积操作和 2 倍、4 倍、8 倍、16 倍、32 倍下采样得到的 5 层多尺度特征图. 多尺度特征的作用主要是为了解决行人检测中的行人目标存在大、中、小等多尺度问题.

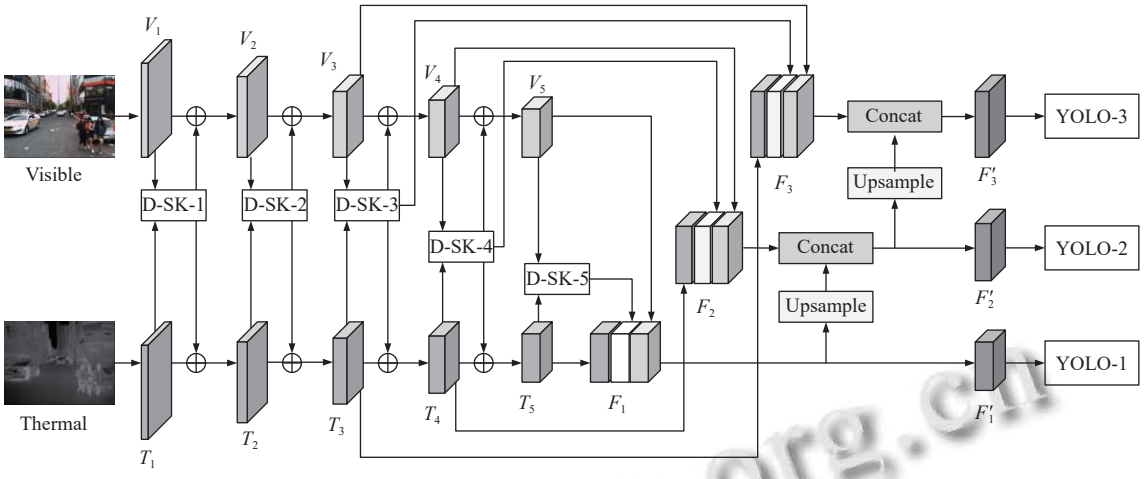

图 2 基于 YOLO 的多模态特征差分注意融合行人检测网络

其次, 在两个模态对应的特征层之间插入模态特征差分注意模块 (modal feature differential selective kernel convolution module, D-SK), 获取两个模态对应特征层间的差异信息, 并将差异信息以加和的方式反馈到特征提取过程中, 实现模态间的信息交流, 该模块的详细信息将在后文进行说明.

然后, 选取可见光模态的 V_3、V_4、V_5 层特征, 模态特征差分注意模块的 D-SK-3、D-SK-4、D-SK-5 层差分特征和热红外模态的 T_3、T_4、T_5 层特征, 采用拼接融合的方式分别将对应特征层的特征和差分特征进行融合得到 F_1、F_2、F_3 三个多尺度融合特征.

最后, 将融合特征 F_1 经过相关卷积操作得到特征 F_1'将融合特征 F_1 经过上采样后与融合特征 F_2 融合, 再经过相关卷积操作得到特征F_2'; 将融合特征 F_1 经过上采样后与融合特征 F_2 融合后的特征进过上采样与融合特征 F_3 融合, 再经过相关卷积操作得到特征F_3'; 在多尺度特征F_1'、F_2'、F_3'之后分别接上 YOLOv3 检测头进行最终的行人目标预测.

2.2 模态特征差分注意

(1) 模态特征差分思想

在多模态特征融合阶段, 目前经常采用在特征通道方向上的拼接融合、对应特征逐元素相加融合等跨模态特征融合方法, 从最终的实验结果来看这些方法都非常有效, 相比基于单模态行人检测来说检测性能提升了很多. 但不同模态之间特征差异较大而这些特征融合方法又较为直接, 没有考虑到模态特征之间的信息交流, 盲目的特征融合使得模态间信息交流不充分进而阻碍了模型性能的进一步提升. 为了解决模态间的信息交流以准确得到跨模态特征信息, 本文通过借鉴 MB-Net[13] 中提出的模态差分思想构建图 3 所示的模态特征差分注意模块来挖掘可见光模态和热红外模态之间的差异信息, 由于通过模态差分得到的差异信息中既包含有用信息也包含了无用的噪声信息, 因此为了尽可能减少噪声影响, 选择对模态差分信息进行注意力机制加强, 强化有用信息的同时抑制噪声.

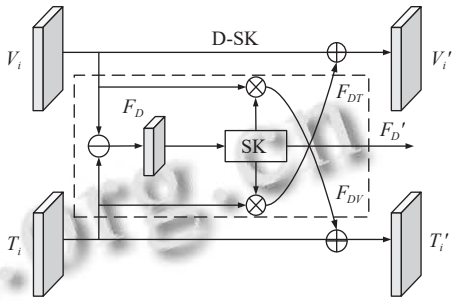

图 3 模态特征差分注意模块

模态差分思想主要受到差分放大器的启发, 在差分放大器中为了突出差异性, 选择抑制共模信号并放大差模信号. 模态差分的主要优点在于即保留了原始特征信息又引入了互补模态的差异信息, 加强了多模态特征信息之间的融合与交流. 在模态差分思想中, 用每个通道上的公共特征和差分特征的组合来表示可见光模态图像特征 (V_i) 和热红外模态图像特征 (T_i), 具体表示形式如式 (1) 和式 (2) 所示:

$$V_i = \frac{V_i + V_i}{2} + \frac{T_i - T_i}{2} = \frac{V_i + T_i}{2} + \frac{V_i - T_i}{2} \tag{1}$$

$$T_i = \frac{T_i + T_i}{2} + \frac{V_i - V_i}{2} = \frac{T_i + V_i}{2} + \frac{T_i - V_i}{2} \tag{2}$$

其中, $(V_i+T_i)/2$ 表示模态公共特征, $(V_i-T_i)/2$ 和 $(T_i-V_i)/2$ 分别表示可见光模态和热红外模态捕获的独有特征. 通过在每个模态的特征中融入互补模态的特征, 达到两个模态间的信息交流, 增强模态间的依赖关系, 使得每一条主干网络在学习过程中不是单独地学习各自模态的特征, 而是同时对多个模态中的信息进行学习, 丰富学习内容, 这正是深度网络所需要的.

(2) 模态特征差分注意模块构建

在构建多模态行人检测网络时, 选择将模态特征差分注意模块 (图 3) 插入到主干网络的每一个特征提取层中. 如图 2 中所示, 将 5 个 D-SK 插入到主干网络的 5 个卷积层之间, 这样做是基于两个考虑: 1) 在主干中的每个特征提取层之间加入 D-SK 可以起到充分利用、学习多模态的互补信息; 2) 由于在 D-SK 模块中加入了选择性核卷积 (selective kernel convolution, SK)[27], 其通过大量实验研究表明将 SK 放置在特征提取的浅层网络位置处效果最显著, 因此在本文中选择从主干网的第 1 层特征提取层开始使用添加了 SK 的 D-SK 模块, 并在第 5 层特征提取层中只进行模态特征差分注意获取差异信息表示而不进行差分特征反馈.

模态特征差分注意模块的构建, 首先将两条特征提取主干网络中对应特征层的特征 V_i 和 T_i 相减得到相同层的模态差异特征 F_D; 其次对 F_D 进行图 4 所示的 SK 操作得到 F_D', 使其能自适应调整感受野的大小, 产生不同的有效感受, 加强对多尺度信息的敏感性; 然后将 F_D'分别与可见光模态特征 V_i 和热红外模态特征 T_i 通过逐元素相乘的方式得到每个模态的差分注意特征 F_{DV} 和 F_{DT}; 最后将差分特征 F_{DV} 和 F_{DT} 作为互补信息通过逐元素相加的方式分别融合进热红外特征和可见光特征中并得到新的热红外特征 T_i'和新的可见光特征 V_i'. 该过程的具体表达如式 (3) 和式 (4) 所示:

$$V_i' = V_i \oplus F_{DT} = V_i \oplus T_i \otimes f(F_D) \tag{3}$$

$$T_i' = T_i \oplus F_{DV} = T_i \oplus V_i \otimes f(F_D) \tag{4}$$

其中, V_i'和 T_i'代表经过模态特征差分注意后的可见光模态和热红外模态特征; F_{DV} 为可见光互补特征, F_{DT} 为热红外互补特征; F_D 为初始模态差分特征; $f(x)$函数表示 SK 操作; $\oplus$代表逐元素相加, $\otimes$代表逐元素相乘.

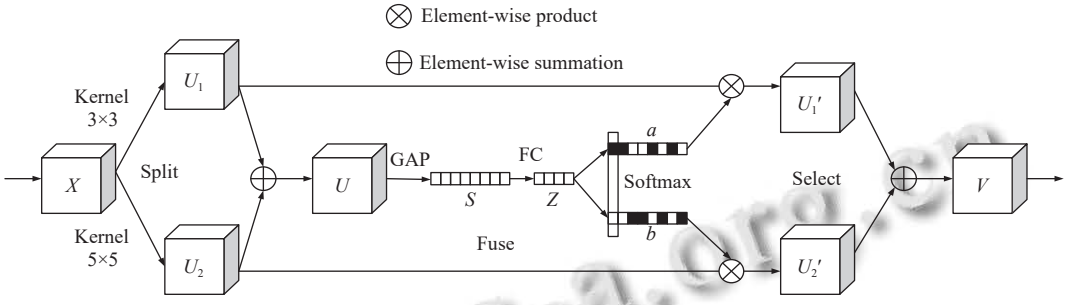

图 4 选择性核卷积

(3) 选择性核卷积

选择性核卷积 (SK) 是 attention 机制中的重要一员, 其能自适应调整感受野的大小, 产生不同的有效感受, 加强对多尺度信息的敏感性, 使得网络能够同时学习到多尺度目标信息. SK 的构建如图 4 所示, 从整体上看可以分为 3 个部分, 即 Split、Fuse 和 Select.

1) Split: 分别使用3×3和5×5卷积核对输入特征 X 进行卷积得到两组具有不同感受野的特征 U_1 和 U_2.

2) Fuse: 组合并聚合来自多个路径的信息, 以获得选择权重的全局和综合表示. 首先将不同感受野特征 U_1 和 U_2 通过逐元素相加得到融合特征 U, 接着对特征 U 进行全局平均池化得到全局特征表示 S, 最后通过简单全连接 (fully connected layers, FC) 操作得到压缩特征 Z, 降低特征尺寸提高计算效率.

3) Select: 通过跨通道注意力实现对不同空间尺度信息的自适应选择. 首先经过 Softmax 层获得通道上的不同空间尺度的权重表示 a 和 b (其中 $a+b=1$), 权重矩阵 a 和 b 的具体含义: 即表示每个特征通道中分别以3×3和5×5卷积得到的不同空间尺度信息的权重占比, 并将所有通道上的不同权重分别组合为权重矩阵 a 和 b; 然后将特征 U_1、U_2 分别与权重矩阵 a、b 通过逐元素相乘得到特征矩阵 U_1'、U_2'实现特征加权; 最后通过特征矩阵 U_1'、U_2'逐元素相加得到融合特征矩阵 V, 实现空间多尺度信息自适应融合.

3 实验结果与分析

3.1 数据集与评价指标

KAIST: KAIST[7] 是行人检测常用的多模态数据集, 由 95 000 对色-热帧、103 128 个注释边框和 1 182 不同的行人组成, 这些图像分别捕获于白天和夜间场景. Zhang 等人[12] 修复了原始数据集中的对齐问题, 而 Liu 等人[8] 对测试集进行了重新标注. 我们使用上述经过修复与重新标注的数据集进行实验.

LLVIP: LLVIP[28] 是一个高分辨率可见光-热红外配对行人检测数据集, 共包含 16 836 对经过配准的红外与可见光图像对, 12 025 对用于训练, 3 463 对用于测试, 其中大多数是在非常黑暗的场景中拍摄的, 所有的图像在时间和空间上都是严格对齐的.

对于行人检测的评价, 使用 KAIST 数据集常用的对数平均漏检率 (miss rate, MR) 来计算误差, MR 越低越好. 在 LLVIP 上使用平均检测精度 (mean average precision, mAP) 进行评价, mAP 越高越好. 所有行人检测数据集的评价都是基于行人检测的合理设置[29] 进行的, 即将测试集中的严重遮挡行人和高度低于 50 像素的行人标注剔除重新构建合理测试子集.

3.2 实验参数设置

在 PyTorch 深度学习框架上进行本文提出的多光谱行人检测实验. 实验其他参数设置如表 1 所示.

针对 KAIST 数据集中的多尺度行人目标检测采用文献 [30] 中通过聚类得到的一组 anchor 模板作为目标检测框, 尺寸分别为 [48, 157]、[34, 104]、[84, 50]、[27, 80]、[26, 63]、[25, 40]、[18, 54]、[16, 44]、[13, 24]. LLVIP 数据集上的行人目标多为中大尺度目标, 使用 YOLOv3 网络中的原始 anchor 模板即可.

表 1 实验参数设置

参数名	参数设置
显存	11 GB
GPU型号	GTX 1080 Ti
优化算法	随机梯度下降 (SGD)
初始学习率	0.001
训练次数 (epoch)	50
批处理大小 (batch size)	4
数据增广	裁剪、缩放、翻转

3.3 实验结果与分析

3.3.1 在 KAIST 数据集上的结果与比较

(1) 实验结果

为了说明本文所提出的多模态行人检测算法在 KAIST 多模态数据集上具有性能优势, 选择与 ACF+T+THOG[7]、Halfway Fusion[8]、GFD-SSD[23]、模态自适应[31] 等方法进行比较, 结果如表 2 所示.

表 2 在 KAIST 数据集上的 MR 与检测速度

方法	全体 (%)	白天 (%)	夜晚 (%)	时间 (s)
ACF+T+THOG	47.32	42.57	56.17	2.73
Halfway Fusion	37.19	37.12	35.33	0.43
GFD-SSD	28.00	25.80	30.03	0.04
模态自适应	26.96	—	—	—
本文	**20.26**	**21.34**	**23.07**	**0.05**

表 2 中展示了在 KAIST 数据集上的不同方法的 MR 值和检测速度比较, MR 值越低、检测速度越快表示模型性能越好. 在全天候、白天和夜晚 3 个合理测试子集上本文所提方法的 MR 值分别为 20.26%、21.34% 和 23.07%, 相比基线方法 ACF+T+THOG (47.32%、42.57% 和 56.17%) 来说 MR 值分别降低了将近 27%、21% 和 33%, 与其他几种方法相比本文方法在检测性能上也有很大提升, 而且本文方法的检测速度达到了每 0.05 s 检测一幅图像, 但略低于 Halfway Fusion 的 0.043 s/幅和 GFD-SSD 的 0.04 s/幅.

(2) 特征图可视化

图 5 展示一组夜间图像在经过本文方法融合前后的特征可视化结果, 从直观的角度展示了多模态行人检测的优势. 从图中可以发现, 融合前特征可视化结果相比原图中的目标来说不够准确, 对于近距离的目标特征显示还算清晰, 但对远距离小目标来说出现了多余和遗漏, 会造成目标的误检和漏检影响性能, 在经过本文特征融合方法之后, 目标特征得到了明显改善, 显著突出了原图中的目标信息, 这也证明了本文提出的方法能有效融合多模态信息.

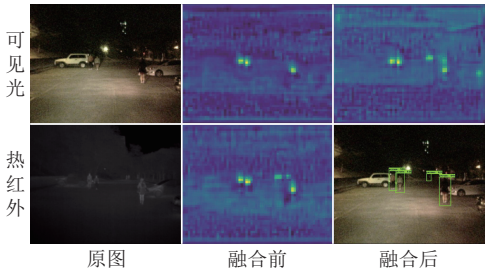

图 5 KAIST 数据集上的案例融合前后特征图可视化结果

3.3.2 在 LLVIP 数据集上的结果与比较

(1) 实验结果

在实际应用中行人数据具有丰富性、多样性的特

点, 为了验证本文方法在不同行人数据集上同样具有鲁棒性, 选择在 LLVIP 多模态数据集上进行相同的实验验证, 检测结果与对比见表 3.

表 3 在 LLVIP 数据集上的平均检测精度与速度

方法	数据	mAP0.5 (%)	mAP0.5:0.95 (%)	速度 (fps)
YOLOv3[28]	可见光	85.9	43.3	—
YOLOv3[28]	热红外	89.7	52.8	—
YOLOv5[28]	可见光	90.8	50.0	—
YOLOv5[28]	热红外	94.6	61.9	—
本文	双模态	**93.7**	**56.1**	**18**

从表 3 中可知, 在 LLVIP 数据集上, 当阈值取 0.5 时本文所提方法的平均检测精度 (mAP) 为 93.7%, 当阈值在 [0.5, 0.95] 之间变动时本文方法在阈值区间的 mAP 为 56.1%, 且本文方法达到了 18 fps 的检测速度.

在 LLVIP 数据集的可见光子集上使用 YOLOv3 方法和 YOLOv5 方法取得的 mAP 明显低于本文方法的 mAP, 从直观的角度来看本文方法比基于单模态可见光的检测方法更优、性能更好.

在热红外子集上, 本文方法的 mAP 要明显高于使用 YOLOv3 方法取得的 mAP, 但略低于使用 YOLOv5 方法取得的 mAP. 分析出现该结果的原因: ① 该数据集为高清数据集, 行人目标中包含较少小尺寸行人, 且热红外图像中的行人目标比可见光图像中的行人目标更为突出更易于检测, 这种判断可以从表 3 中的结果得到验证, 即用相同方法对不同模态数据进行检测, 在热红外模态上的 mAP 高于可见光模态上的 mAP; ② YOLOv5 网络强大的特征提取能力, 能提取出比 YOLOv3 网络更强、更全面的行人特征, 结合目标更显著的热红外单模态数据集, 使得 YOLOv5 网络在热红外单模态数据集上能达到更高的检测性能.

(2) 特征图可视化

对 LLVIP 数据集上的一个案例融合前后特征图可视化结果如图 6 所示, 从上到下依次展示了双模态原图、融合前特征、融合后特征和检测结果, 对比融合前后的特征图可以发现经过本文方法融合后目标区域的特征更加显著, 与背景之间的差异性更明显, 证明本文方法对模态间的信息能起到有效融合的作用.

3.4 消融实验

为了验证本文提出的方法中每个模块的加入都能促进模型性能的提升, 于是在相同的实验条件下对本文提出的方法分别在 KAIST 和 LLVIP 数据集上进行消融实验.

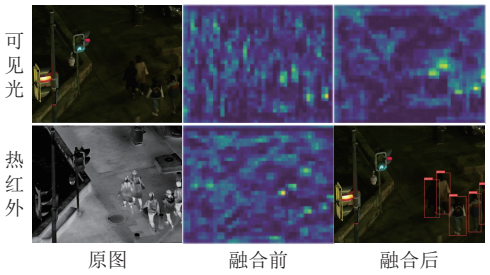

图 6 LLVIP 数据集上的案例融合前后特征图可视化结果

本文方法的主要创新在于: ① 在提取完不同模态的差异信息后加入了选择性核卷积注意模块 (SK), 该模块的加入是为了发挥注意力机制的作用, 提升对有用信息的关注并抑制无用信息的表达; ② 选择特征拼接的方式完成不同模态间的信息融合, 并将提取到的模态间的差异信息作为融合特征的重要组成部分与两个模态的特征一起完成特征融合.

因此针对本章的创新点, 提出以下消融实验: ① 在提取完模态差异特征后不进行选择性核卷积操作, 并保持其他操作不变, 进行消融实验; ② 在融合特征中不加入模态差异信息, 只对两个模态的特征进行拼接完成特征融合, 并保持其他操作不变, 进行消融实验. ③ 同时为了验证不同的特征融合方法对模型性能的影响, 本章采用不同融合方法进行了实验, 在保持其他操作不变的情况下将特征拼接替换为特征相加, 进行消融实验. 下面展示了在不同数据集上的消融实验结果.

(1) 在 KAIST 数据集上的消融结果

在 KAIST 数据集上的消融实验结果见表 4, 从上到下依次展示了 5 种不同的实验结果.

表 4 在 KAIST 数据集上的消融结果 (IoU=0.5)

方法	MR (%)
KAIST_yolov3_concat_no_sknet_and_difference	23.21
KAIST_yolov3_concat_no_sknet_but_difference	22.75
KAIST_yolov3_concat_sknet_no_difference	21.60
KAIST_yolov3_concat_sknet_difference	20.26
KAIST_yolov3_add_sknet_difference	19.80

1) 以不使用选择性核卷积和模态特征差异信息的方法作为基线, 初始基线的 MR 值为 23.21%.

2) 在基线方法的基础上将模态特征差异信息加入融合特征中, 该方法的 MR 值为 22.75%, 相比基线的 MR 值降低了 0.5 个百分点.

3) 在基线方法的基础上加入选择性核卷积操作, 并保持其他操作不变, 该方法的 MR 值为 21.60%, 相比基线 MR 值降低了 1.6 个百分点, 效果显著.

4) 在基线方法的基础上同时使用选择性核卷积和

模态特征差异信息, 该方法的 MR 值为 20.26%, 相比基线和前两种方法, MR 值分别降低了约 3%、2.5% 和 1.3%, 从结果来看两个模块的同时使用对模型性能的提升最为明显.

5) 最后对特征融合方式进行了消融分析, 将特征拼接的方式替换为特征相加的方式完成模态特征融合, 并保持其他操作不变, 取得的 MR 值相比使用特征拼接的融合方法进一步降低了约 0.5 个百分点.

通过上述消融实验进一步证明了本文提出的创新方法中每个模块的使用都能提升模型性能, 尤其当两个模块同时使用时模型性能提升最为显著, 同时也验证了在特征融合方法中特征相加比特征拼接更有效.

(2) 在 LLVIP 数据集上的消融结果

图 7 所示的 P-R 曲线展示了在 LLVIP 数据集上的消融实验结果. 图中共有 5 条检测精度-召回率 (P-R) 曲线, 每条曲线代表消融实验中的一种方法, 按精度从低到高依次排列如下.

1) 蓝色曲线代表不使用选择性核卷积和模态特征差异信息 (mAP 为 89.8%).

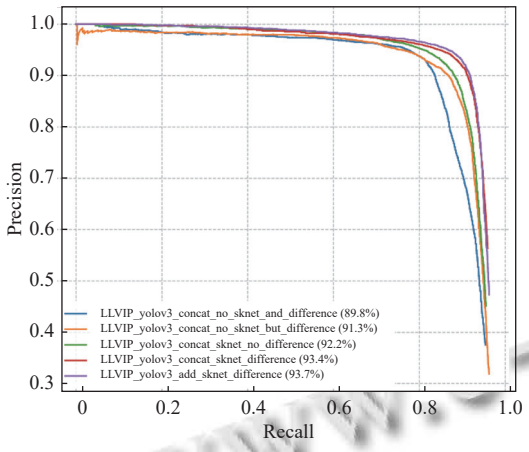

图 7 在 LLVIP 数据集上的消融实验 P-R 曲线

2) 橙色曲线代表不使用选择性核卷积但使用模态特征差异信息 (mAP 为 91.3%).

3) 绿色曲线代表在融合特征中使用选择性核卷积但不使用模态特征差异信息 (mAP 为 92.2%).

4) 红色曲线代表使用选择性核卷积和模态特征差异信息 (mAP 为 93.4%).

5) 紫色曲线代表使用特征相加的融合方式 (mAP 为 93.7%).

将蓝色曲线代表的方法作为对比基线, 在此基础上加入模态特征差异信息得到橙色曲线所示结果, mAP 提升了 1.5 个百分点; 在蓝色曲线代表的方法的基础上加入选择性核卷积得到绿色曲线所示结果, mAP 提升了 2.4 个百分点; 在蓝色曲线方法的基础上同时使用选择性核卷积和模态特征差异信息得到红色曲线所示结果, mAP 提升了 3.6 个百分点; 用特征相加的融合方法替换红色曲线表示的方法中的特征拼接融合方法得到紫色曲线所示结果, 在此基础上 mAP 提升了 0.3 个百分点.

(3) 消融实验总结

通过上述消融实验可得如下结论: 1) 选择性核卷积和模态特征差异信息的使用都能提升模型的检测精度, 相比之下选择性核卷积的使用提升效果更好. 2) 从结果来看两个模块的同时使用具有叠加效果, 比仅使用单个模块获得的性能提升更强, 从模型中分析可知, 在初步获取到的模态特征差异信息后使用选择性核卷积进行注意机制加强能够显著增强有用信息表达并抑制无用信息, 再将经过注意机制加强的模态差异信息融入不同模态的特征提取中使得模态间的信息交流更加有效, 最后将模态特征差异信息加入融合特征中以得到更鲁棒的特征表示. 3) 观察实验结果, 特征相加依然比特征拼接有效, 但在本实验中使用两种特征融合方法得到的实验结果相差并不大.

3.5 检测结果展示

(1) 在 KAIST 数据集上的检测结果展示

本文方法在 KAIST 数据集上的部分检测结果如图 8 所示. 分别展示了白天和夜间两种不同光照环境下的行人检测结果, 第 1 列为原始可见光、热红外图像, 第 2 列为对应的行人检测结果. 从图 8 中结果可知, 无论面对何种光照情况、何种尺寸行人, 本文算法都能正确检测出行人目标. 从对 KAIST 数据集的已有信息可知, 该数据集中行人尺寸整体偏小、环境因素较为丰富、图像清晰度整体较差, 使得行人检测难度较高, 且该数据集为目前常用的行人检测数据集, 因此本文提出的方法在 KAIST 数据集上的良好表现使得该算法被应用于视频监控等实际应用领域成为可能.

(2) 在 LLVIP 数据集上的检测结果展示

图 9 展示了在 LLVIP 数据集上的部分案例检测结果, 夜间检测结果为图 9(a) 所示, 白天检测结果为图 9(b) 所示, 在不同情况下目标均能被正确检测到, 且目标框大小合适、与行人目标贴合密切. 通过实验表明, 在 LLVIP 数据集上, 本文算法同样具有良好的检测能力.

图 8 在 KAIST 数据集上的部分检测结果

(a) 夜晚

(b) 白天

图 9 在 LLVIP 数据集上的部分案例检测结果

4 总结

针对模态差异问题和如何充分利用多模态信息进行行人检测, 本文提出了一种基于 YOLO 的多模态特征差分注意融合行人检测方法. 该方法借鉴差分电路的思想, 通过模态特征相减获取模态间差异信息, 并将差异信息反馈至不同模态实现模态间的信息交流与模态特征互补融合, 有效地解决了模态差异问题, 实现了多模态信息的充分利用. 在两种多模态数据集上对本文提出的方法进行实验验证, 结果均表明本文方法能显著提升行人检测精度, 并且能够实现较高的检测速度, 具有实际应用价值.

参考文献

1 Bilal M, Khan A, Khan M U K, *et al*. A low-complexity pedestrian detection framework for smart video surveillance systems. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(10): 2260–2273. [doi: 10.1109/TCSVT.2016.2581660]

2 Zhang SZ, Cheng D, Gong YH, *et al*. Pedestrian search in surveillance videos by learning discriminative deep features. Neurocomputing, 2018, 283: 120–128. [doi: 10.1016/j.neucom.2017.12.042]

3 Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the KITTI vision benchmark suite. Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 3354–3361.

4 Lee WJ, Lee SW. Improved spatiotemporal noise reduction for very low-light environments. IEEE Transactions on Circuits and Systems II: Express Briefs, 2016, 63(9): 888–892. [doi: 10.1109/TCSII.2016.2536218]

5 Song WF, Li S, Chang T, *et al*. Context-interactive CNN for person re-identification. IEEE Transactions on Image Processing, 2020, 29: 2860–2874. [doi: 10.1109/TIP.2019.2953587]

6 Chen SJ, Shen HL. Multispectral image out-of-focus deblurring using interchannel correlation. IEEE Transactions on Image Processing, 2015, 24(11): 4433–4445. [doi: 10.1109/TIP.2015.2465162]

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

8 Liu JJ, Zhang ST, Wang S, *et al*. Multispectral deep neural networks for pedestrian detection. Proceedings of the British Machine Vision Conference 2016. York: BMVC, 2016.

9 Choi Y, Kim N, Hwang S, *et al*. KAIST multi-spectral day/night data set for autonomous and assisted driving. IEEE

Transactions on Intelligent Transportation Systems, 2018, 19(3): 934–948. [doi: 10.1109/TITS.2018.2791533]

10 Li CY, Song D, Tong RF, *et al*. Multispectral pedestrian detection via simultaneous detection and segmentation. Proceedings of the 2018 British Machine Vision Conference. Newcastle: BMVC, 2018. 225.

11 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]

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

13 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.

14 Tian YL, Luo P, Wang XG, *et al*. Pedestrian detection aided by deep learning semantic tasks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 5079–5087.

15 Zhang LL, Lin L, Liang XD, *et al*. Is faster R-CNN doing well for pedestrian detection? Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 443–457.

16 Ren SQ, He KM, Girshick RB, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: ACM, 2015. 91–99.

17 Liu W, Liao SC, Ren WQ, *et al*. High-level semantic feature detection: A new perspective for pedestrian detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5182–5191.

18 Baek J, Hong S, Kim J, *et al*. Efficient pedestrian detection at nighttime using a thermal camera. Sensors, 2017, 17(8): 1850. [doi: 10.3390/s17081850]

19 Devaguptapu C, Akolekar N, Sharma MM, *et al*. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach: IEEE, 2019. 1029–1038.

20 Zhu JY, Park T, Isola P, *et al*. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2242–2251.

21 Guo TT, Huynh CP, Solh M. Domain-adaptive pedestrian detection in thermal images. Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP). Taipei: IEEE, 2019. 1660–1664.

22 Kieu M, Bagdanov AD, Bertini M, *et al*. Task-conditioned domain adaptation for pedestrian detection in thermal imagery. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 546–562.

23 Zheng Y, Izzat IH, Ziaee S. GFD-SSD: Gated fusion double SSD for multispectral pedestrian detection. arXiv:1903.06999, 2019.

24 Fu L, Gu WB, Ai YB, *et al*. Adaptive spatial pixel-level feature fusion network for multispectral pedestrian detection. Infrared Physics & Technology, 2021, 116: 103770.

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

26 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv:1804.02767, 2018.

27 Li X, Wang WH, Hu XL, *et al*. Selective kernel networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 510–519.

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

29 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]

30 施政, 毛力, 孙俊. 基于 YOLO 的多模态加权融合行人检测算法. 计算机工程, 2021, 47(8): 234–242. [doi: 10.19678/j.issn.1000-3428.0058745]

31 陈莹, 朱宇. 模态自适应权值学习机制下的多光谱行人检测网络. 光学 精密工程, 2020, 28(12): 2700–2709.

(校对责编: 孙君艳)