Multispectral Pedestrian Detection Based on the Fusion of Infrared and Visible Image*

XIANG Jing, JIANG Mingxin*, ZHOU Qingqing, ZHANG Yuheng
(Faculty of Electronic Information Engineering, Huaiyin Institute of Technology, Huai'an Jiangsu 223003, China)

Abstract: Targeting at the problems of low accuracy of pedestrian detection in harsh environments, high computational complexity and large number of parameters in current multispectral pedestrian detection algorithms, a multispectral pedestrian detection algorithm based on infrared and visible image fusion, named CMDFT-YOLOv8 is proposed. The difference enhancement module is used to retain and enhance the features specific to different modes, and cross-modal feature interaction is proposed to effectively fuse the inputs from different modes. In addition, partial convolution is used to improve the BottleNeck structure of the double-path feature extraction network, reduce the redundant information and computational complexity of the backbone network, and extract spatial features more effectively. Compared with the benchmark method, the proposed multispectral pedestrian detection algorithm achieves 67.3%mAP in weak light environment, which is 5.8%higher than that of the benchmark model, 80.2% lower in model parameters, and 33.2% lower in model computational complexity. The proposed algorithm effectively improves the accuracy of the detection model in weak light environment and reduces the number of model parameters.

Key words: multispectral pedestrian detection; modal feature differential enhancement; cross-modal feature interaction; YOLOv8; multi-modal image fusion
EEACC: 6135 **doi**: 10.3969/j.issn.1004-1699.2026.02.021

基于红外和可见光图像融合的多光谱行人检测*

项 靖, 姜明新*, 周清清, 张宇恒
(淮阴工学院电子信息工程学院, 江苏 淮安 223003)

摘 要: 针对目前多光谱行人检测算法在恶劣环境下对行人检测精度较低, 且计算复杂度和参数量大的问题, 提出一种基于红外和可见光图像融合的多光谱行人检测算法 CMDFT-YOLOv8。利用差异增强模块保留和增强不同模态特定的特征, 并提出了跨模态特征交互来有效地融合来自不同模态的输入; 另外使用部分卷积改进双路特征提取网络的 BottleNeck 结构, 减少骨干网络的冗余信息和计算复杂度, 更有效地提取空间特征。与基准方法相比, 所提出的多光谱行人检测算法在弱光环境下的 mAP 达到了 67.3%, 相较于基准模型提升了 5.8%, 模型参数量下降 80.2%, 模型计算复杂度下降 33.2%。所提出的算法可有效提高弱光环境下检测模型的精度和降低模型参数量。

关键词: 多光谱行人检测; 模态特征差分增强; 跨模态特征交互; YOLOv8; 多模态图像融合
中图分类号: TP391.4 **文献标识码**: A **文章编号**: 1004-1699(2026)02-0377-09

红外光谱图像具有优越的图像对比度和对目标轮廓的敏感性, 被认为是计算机视觉中可见光谱图像的补充。多光谱图像特征的表示和融合是一个具有挑战性的问题, 其为各种视觉的下游任务提供服务, 如行人检测[1-2]、语义分割[3]和自动驾驶[4]等。目标检测是计算机视觉的一个重要分支, 在现实世界中有着广泛的应用。然而在现实世界中, 环境往往是开放和动态的, 例如恶劣的天气条件和不断变化的光照, 这些单模态检测方法很容易受到影响。为了提高目标检测器在全天候条件下的可感知性、可靠性和鲁棒性, 融合可见光和红外图像已成为提高暗环境下目标检测性能的典型方法。

目前, 一种常见的融合方法是将 RGB 图像与红外图像进行中期融合, 因为它们是互补的[5]。当光线充足时, RGB 相机可以获得物体的更多细节, 但在昏暗的光线条件下几乎没有帮助。另一方面, 红

项目来源: 国家自然科学基金项目(62205120); 江苏高校青蓝工程项目(202004)
收稿日期: 2024-01-08 **修改日期**: 2024-03-11

外图像可以确保在光线不足或被遮挡的情况下可以获得物体的轮廓,但缺少物体的纹理和颜色等信息。因此,设计一种融合方法来有效地利用来自红外模态和可见光模态的不同视觉信息,已经成为当下热门的研究目标。

在现有的方法中,已经利用了许多融合策略来利用可见光和红外图像的互补特征。Liu 等[6]提出了四种融合策略并进行比较,得出了与早期融合相比,多模态特征的中期融合得到的效果更好。Li 等[7]将语义信息引入到决策阶段,显著提高了多光谱行人的检测性能。Zhou 等[8]考虑了照明条件,提出了使用照明感知特征对齐模块来互补并对齐两个模态特征。Fang 等[9]首次提出了使用 Transformer 的自注意力机制在模态内和模态间进行融合特征,但直接将传统的 Transformer 应用于图像对可能难以提取有用的特征,且计算量大。此外,Chen 等[10]提出了一种后期融合方法 ProbEn。ProbEn 基于贝叶斯规则和条件独立性假设的概率集成方法,通过概率边缘化,高效地处理丢失的模态特征,但其缺点是并未充分利用不同模态之间的互补信息。Fang 等[11]提出了基于 YOLOv5 的多模态检测模型,使用跨模态注意特征融合模块来识别输入模态之间的相关性。

近年来,国内外大多数关于多模态行人检测的方法均采用并行的骨干网络处理红外和可见光图像,因此额外的骨干网络也会使得模型需要学习更多参数。其次,为了发现不同模态之间的关联信息,许多研究人员设计了复杂的融合模块,如使用 Transformer 进行多模态融合[12]。虽然 Transformer 在多模态融合领域取得了极大的发展,例如图像和音频频谱中的空间区域可以自由地进行聚合。但与视听学习任务不同的是红外图像和彩色图像对有更多的共同特征。由于存在冗余信息,直接将传统的 Transformer 应用于图像对可能难以提取有用的特征。其次,这些复杂的模块导致更高的计算复杂性,从而限制了模型的实际应用。最后,先前的大多数研究只关注于进行特征融合和利用融合后的特征,并未考虑到每个模态的独有特征。融合信息主要依靠物体的纹理,使目标与背景杂波不同。在没有充分考虑特定特征的情况下,融合的信息成为主要的判别线索,这限制了判别能力的上限。

本文的主要贡献如下:①为了更好地融合可见光模态和红外模态特征,本文提出了一种基于跨模态特征差分 Transformer 的多光谱行人检测算法 CMDFT-YOLOv8,通过改进的差分结构来充分地获取每个模态图像的差异信息,交叉这些差异信息,并高效地计算两个模态间自注意力,来增强和互补融合前每个模态的独有信息,有效地融合多模态图像特征。②针对目前在多模态图像目标检测任务中,使用双路骨干网络所导致的计算量大与推理速度慢的问题,通过引入 PConv 来构建 PConv-GRNBlock 模块,以改进双路骨干网络的 BottleNeck 结构,提高了不同模态的特征表达能力和感受野的覆盖范围,同时降低双路骨干网络的计算量与参数量。

1 本文方法介绍

近红外和可见光图像融合的多光谱行人检测分为 3 个步骤,首先对数据集进行预处理操作,以适应设备和网络的需求;随后搭建多光谱行人检测网络模型,提取不同模态下的行人区域特征,实现精准检测;最后将图像数据可视化并对比前沿算法验证本文算法性能。

1.1 总体框架

本文总体框架如图 1 所示,所提出融合方案的目标是探索红外图像和可见光图像模态的特有特征,并使用所提出的跨模态特征差分 Transformer 对特有特征进行融合,来进行多模态行人检测。本文将 YOLOv8 主干网络重新设计为双路独立的特征提取网络,分别提取可见光和红外图片对的特征。并将不同阶段的特征图输入到所提出的跨模态特征差分 Transformer 中,以增强不同模态之间的互补信息特征。其次,本文使用轻量级网络 FasterNet[13]中的部分卷积 PConv,提出了 PConv-GRNBlock 模块以改进双路特征提取网络中 C2f 的 BottleNeck 结构,以降低骨干网络的计算量的同时更有效地提取空间特征。

1.2 多模态图像融合策略

根据融合策略,多模态目标检测可分为三类,包括早期融合、晚期融合和中期融合[14-16]。早期融合是最直观的融合方法,其中红外和可见光图像被连接以生成 4 通道图像,然后将其输入到检测架构中[16]。先前的研究表明,融合早期阶段的两个输入会放弃每个输入特有的特征,从而降低检测精度。在后期融合中,将两种模态的输入分别输入到两个单模态目标检测模型中以生成边界框,融合不同结合策略的输出结果。大多数关于多模态对象检测的研究都集中在中期融合,因为这种策略在设计特定于输入的融合模块时提供了更大的灵活性,从而使检测模型能够探索输入模态之间更深的相关性。本文提出的中期融合方案实现多模态的多尺度融合,能够探索潜在的模态特异性特征,丰富图像特征表示。

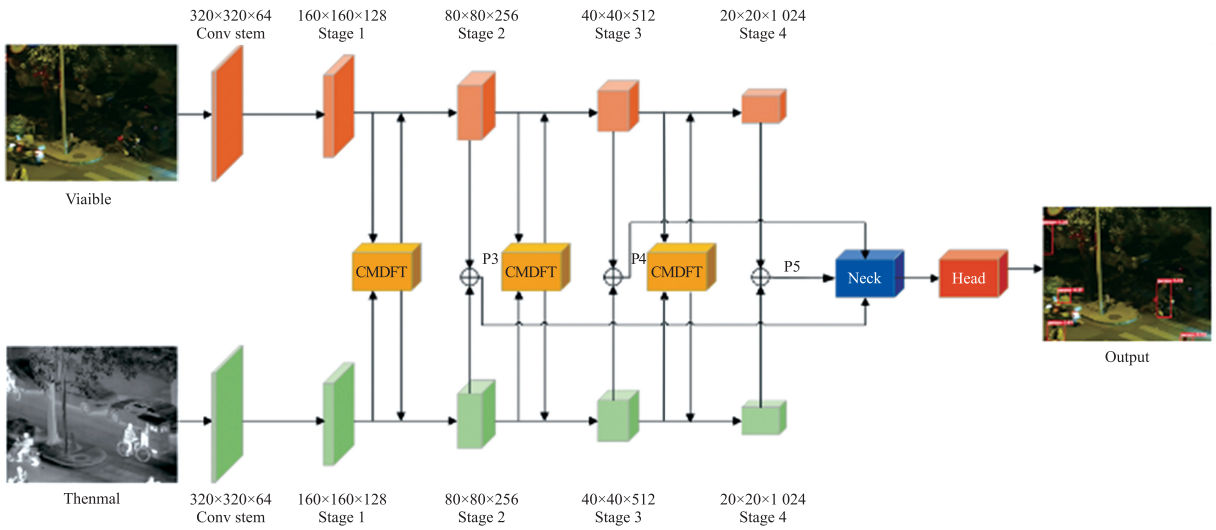


图 1　CMDFT-YOLOv8 网络总体框架图

1.3　跨模态特征差分 Transformer

在先前的研究中,已经提出了许多融合方法来以适当的方式融合每个模态的属性,但并未关注到融合之前的各个模态所拥有的特定特征。因此,本文提出了跨模态特征差分 Transformer 来增强和互补融合前模态的特有信息。

如图 2 所示,我们使用差分操作来获取每个模态的独有特征 F_D,并增强差异特征的通道间关系,通过跨模态特征交互模块(Cross-Modality Feature Interaction,CMFI)增强模态间的差分和互补性。这种设计背后的主要思想是利用交叉计算来自不同模态特征的自注意力,让网络更加关注另外模态独有的特征。

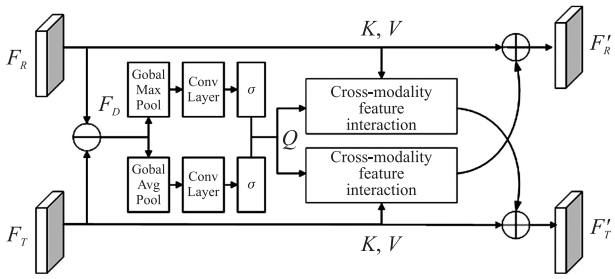


图 2　跨模态特征差分 Transformer(CMDFT)

1.3.1　差异增强模块

模态特征差分结构通常是将两个特征图进行减法来直接计算不同模态的差异信息,然后用于特征提取。然而,直接获取不同模态的差异信息可能会导致噪声和差异信息的丢失。其次,简单的相减可能需要对特征进行归一化、标准化或其他形式的预处理,以便更好地捕捉到模态之间的差异。

为了减少噪声并获得高质量的互补信息,本文借鉴了模态特征差分思想。通过充分利用双模态图像来学习差异信息,原始的双模态图像被馈送到骨干网络中作为差分增强模块的输入。差分增强模块如图 2 所示,首先,在双模态特征图之间进行减法来学习差异特征图 F_D,该过程的具体表达如式(1)所示:

$$F_D^T = |F_T - F_R| \qquad F_D^R = |F_R - F_T| \tag{1}$$

$|\cdot|$表示绝对值运算,F_D 可以被看作是两个特征图的相对差异程度。其中绝对值运算确保了所获得的差分特征是有意义的。

其次,为了有效地利用所获得差异特征的通道间关系,进行了通道注意力操作。具体地,差分特征的空间维度同时被平均池化和最大池化挤压,以生成不同的空间上下文信息。然后,空间上下文信息被送入 1×1 卷积层,实现局部跨通道之间的交互,该过程的具体表达如式(2)所示:

$$F'_D = \sigma(\mathrm{Conv}(\mathrm{GMP}(F_D)) + \mathrm{Conv}(\mathrm{GAP}(F_D))) \tag{2}$$

式中:F'_D表示通道注意力运算后获得的注意力图,Conv 表示 1×1 的卷积,GMP 和 GAP 分别表示全局平均池化和全局最大池化运算,σ 表示 Sigmoid 函数。

差异增强模块旨在通过差分和通道注意力操作让网络在融合的过程中学习更多的差异信息和消除不同模态中的噪声。

1.3.2　跨模态特征交互模块

由于普通的融合特征很容易被来自其中一个模态的强特征表示所主导,通常无法充分利用特定模态的信息。为了探索每种模态的特定特征,本文提出了一种新的跨模态特征交互结构来高效的运用跨模态信息增强特征提取。

跨模态特征差分 Transformer 使用两个跨模态特

征交互模块(Cross-Modality Feature Interaction,CMFI)来分别处理来自差分增强模块增强后的差异特征 F_D^R 和 F_D^T。不同于传统 ViT 中的 Transformer 将输入特征图线性映射成为 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$,跨模态特征差分 Transformer 通过交叉来自模态差异特征 F_D^R 和 F_D^T 实现全局特征交互,其计算原理如图 3 所示。

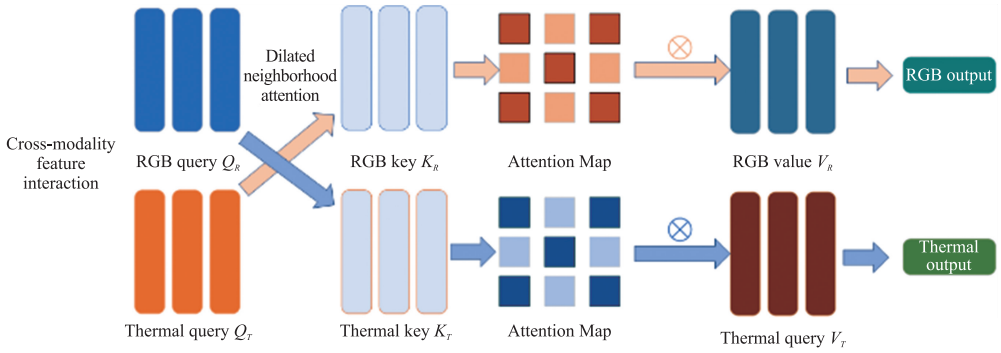


图 3 跨模态特征交互过程

将可见光模态的差分特征 F_D^R 与来自红外模态特征 F_T 经过独立的线性层进行线性映射后作为 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$,使用空洞邻近注意力进行多模态特征交互。通过空洞邻近注意力来计算可见光模态差异特征 F_D^R 与红外模态特征 F_T 之间的相似度,并计算注意力得分,通过 Softmax 函数进行归一化处理,对 F_T 执行注意力加权,来更加关注红外模态特征中的可见光模态的差分特征 F_D^R 所表达的部分。该过程的具体表达如式(3)和式(4)所示。对于红外模态来说,将红外模态的差分特征 F_D^T 与来自可见光的模态特征 K_R 和 V_R 进行线性映射后交互,其过程与可见光模态的计算过程相似,该过程的具体表达如式(5)和式(6)所示:

$$Q_R = F_D^R \cdot W_Q^R, K_T = F_T \cdot W_K^{\mathrm{T}}, V_T = F_T \cdot W_V^{\mathrm{T}} \tag{3}$$

$$\mathrm{Attention}(Q_R, K_T, V_T) = \mathrm{softmax}\left(\frac{Q_R \cdot K_T}{\sqrt{d_k}}\right) \cdot V_T \tag{4}$$

$$Q_T = F_D^T \cdot W_Q^{\mathrm{T}}, K_R = F_R \cdot W_K^R, V_R = F_R \cdot W_V^R \tag{5}$$

$$\mathrm{Attention}(Q_T, K_R, V_R) = \mathrm{softmax}\left(\frac{Q_T \cdot K_R}{\sqrt{d_k}}\right) \cdot V_R \tag{6}$$

式中:$W_Q^R, W_K^R, W_V^R \in \mathbf{R}^{d_x \times \frac{d_x}{N_h}}$ 和 $W_Q^T, W_R^T, W_V^T \in \mathbf{R}^{d_x \times \frac{d_x}{N_h}}$ 表示可见光模态和红外模态下查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 投影的权重矩阵,N_h 是注意力头的数量。根据头部的个数 N_h 进行划分后,进行多头空洞邻近注意力机制计算得到注意力得分。

每个 CMFI 是由多个空洞邻近注意力(Dilated Neighborhood Attention)[17] 组成,形成多头注意力。如图 4 所示,通过为不同的头部设置不同的扩张系数 Exp,可以实现不同尺度的特征提取,有效聚合长短距离的局部特征和扩大全局感受野的同时增强了多尺度学习的能力。空洞邻近注意力有效地减小了跨模态特征交互模块的参数量和计算复杂性,在计算复杂度和特征图的计算感受野大小之间寻求了适合的折中方案。其结构如图 3 所示。

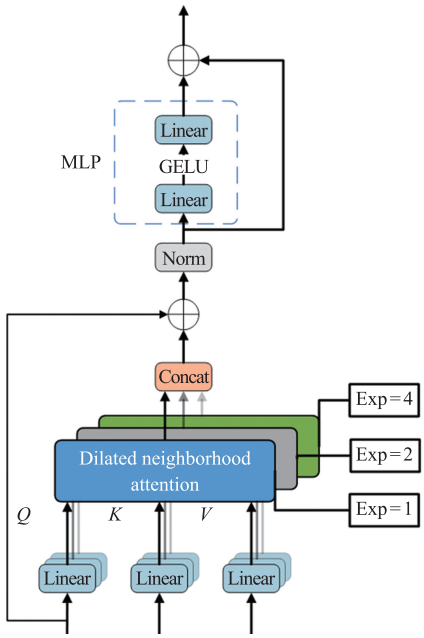


图 4 跨模态特征交互模块

首先跨模态特征交互模块通过层归一化(Layer Normalization)减少每层输入的变化范围,从而减轻内部协变量偏移的问题,有助于而提升网络的泛化能力。再通过线性层将不同模态的查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 进行线性变换,将查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 通过空洞邻近注意力计算不同头部的注意力得分,并将结果连接在一起作为输入特征图 Y。最后,将特征图 Y 经过一个层归一化和 MLP,MLP 由两个线性层和一个 GELU 激活函数组成,最后经过残差链接得到输出 Z。该过程的具体表达式见式(7)、式(8),分别表示红外和可见光模态下的计算过程。

$$\begin{aligned} &h_i^{\mathrm{T}} = \mathrm{DNA}(Q_R, K_T, V_T), 1 \leqslant i \leqslant N_h \\ &Y = \mathrm{Concat}[h_1^R, \cdots, h_n^R] + Q_R \\ &Z = \mathrm{MLP}(\mathrm{Norm}(Y)) + Y \end{aligned} \tag{7}$$

$$h_i^{\mathrm{T}}=\mathrm{DNA}(Q_R,K_T,V_T),1\leqslant i\leqslant N_h$$
$$Y=\mathrm{Concat}[h_1^{\mathrm{T}},\cdots,h_n^{\mathrm{T}}]+Q_R \tag{8}$$
$$Z=\mathrm{MLP}(\mathrm{Norm}(Y))+Y$$

如式(7)所示,DNA 表示空洞邻近注意力。对于来自可见光模态的 Q_R,它通过对来自红外模态的 K_T 和 V_T 执行注意力加权来互补多模态信息。

1.4　轻量的特征提取网络设计

由于 YOLOv8 的特征提取网络采用的是 CSP-DarkNet-53,其网络结构复杂,导致了计算量较大,对于一些计算资源有限或者实时性要求高的场景,可会存在计算瓶颈。

在轻量化特征提取网络的设计中,大多数设计都会采用 MobileNet[18-19] 系列或者 GhostConv[20] 卷积块来构建网络,其通过深度可分离卷积(DW-Conv)[21]来有效地减少计算量,但是这些操作会导致内存访问增加,使得模型运行效率降低。此外,MobileNet、ShuffleNet[22] 和 GhostNet 等网络在处理小模型时,经常需要额外的数据操作,如连接、洗牌(Shuffle)和池化,这些操作的运行时间可能对模型推理速度有着相当大的影响。

FasterNet 是一个新型的网络,它的主要特点是运行速度快,同时具有更低的延迟和更高的吞吐量。其提出了一种新的部分卷积 PConv,通过分析输入特征图的冗余性,对特征图进行部分卷积以提高计算效率。相较于常规的卷积操作,PConv 在保持相同计算精度的前提下,可以显著减少浮点运算次数(FLOPs)和降低内存访问的频率,从而加快计算速度。如图 5 所示,PConv 仅在部分输入通道上进行常规卷积,而其余通道保持不变。其将第一个或最后一个连续的 c_p 通道进行计算,而输入和输出特征图具有相同数量的通道。因此,PConv 的 FLOP 计算量如式(9)所示:

$$h\times w\times k^2\times c_p^2 \tag{9}$$

在部分输入通道比 $\frac{c_p}{c}=\frac{1}{4}$ 的情况下,PConv 的 FLOP 仅为常规 Conv 的 1/16。

针对目前在多模态目标检测任务中,使用双路特征提取网络所导致的计算量大与推理速度慢的问题,通过引入 PConv 来构建 PConv-GRNBlock 模块,以改进 YOLOv8 的 BottleNeck 结构,进一步降低双路骨干网络的计算量与参数量。

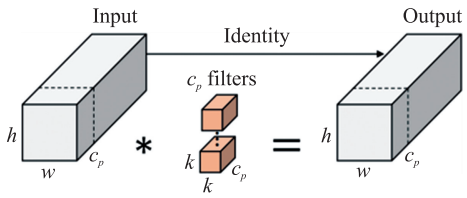


图 5　PConv 卷积结构

PConv-GRNBlock 模块结构是参考 ConvNeXt V2[23] 网络设计方法,如图 6 所示,首先使用 7×7 的 PConv 大卷积核进行特征提取,大尺寸的卷积核可以在图像上覆盖更大的区域,从而捕获更丰富的上下文信息。其次,引入全局响应归一化(GRN)在每个通道上进行归一化,以提升通道之间对比度和选择性。GRN 与 YOLOv8 中 BottleNeck 的 BN 层相比具有两大优势:首先,它无需引入额外参数,因为其操作仅限于对特征图的归一化;其次,GRN 能够适应任意大小的 batch_size,不同于 BN 层是对 batch_size 敏感。中间部分由两个 1×1 的卷积块和 GELU 激活函数后接一个 GRN 层组成,第一个卷积块的输出通道数是 4 倍的 PConv,第二个卷积块的通道数则和第一个相同,组成逆瓶颈层结构,能够提供更多的信息传递路径和计算资源,从而增强了模型的表达能力。最后通过残差结构,将输入与输出进行连接,学习输入和输出之间的映射关系,从而提高模型的性能。

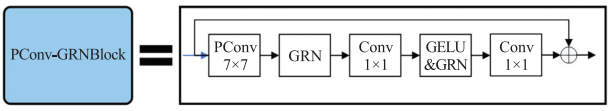


图 6　PConv-GRNBlock 模块结构图

C2f_PConv-GRNBlock 模块的结构是根据 C2f 模块改进而来,如图 7 所示,对于输入特征图 X,经过卷积和特征图切分后,被送入 m 个依次串联的 PConv-GRNBlock 模块中,最后每个模块的结果进行拼接卷积后输出,该过程的具体表达式见式(9):

$$Y^0=\mathrm{Split}(\mathrm{Conv}(X))$$
$$Y^i=\mathrm{PConv\text{-}GRNBlock}(Y^{i-1})\quad 1\leqslant i\leqslant m$$
$$Z=\mathrm{Conv}(\mathrm{Concat}(Y^0,\cdots,Y^m)+Y^0) \tag{10}$$

C2f_PConvBottleNeck 模块采用部分卷积 PConv 改进 C2f 的 BottleNeck 结构,降低了模块减少冗余计算和内存访问,可以更有效地提取空间特征。并通过设置更大的卷积核来增强感受野,在计算复杂度不变的前提下,进一步提升网络的检测精度。

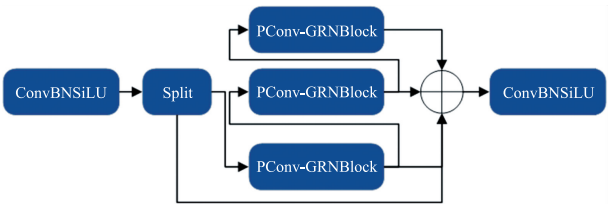


图 7　C2f_PConv-GRNBlock 模块结构图

2 实验过程与分析

2.1 数据集与评价指标

本实验使用 LLVIP[24] 数据集和 M3FD[25] 数据集评估本文所提出的算法,LLVIP 包含由 26 个不同位置的监控摄像头拍摄的可见光-红外图像对,大部分图片是在昏暗的光线下收集的,且时间和空间严格对齐。该数据集包括 15 488 个手动对齐的图片对,输入图像尺寸大小为 1 280×1 024,其中 12 025 对用于训练,3 463 对用于测试。“Pedestrian”是数据集中唯一的对象类别。

M3FD 数据集是一个用于红外和可见光图像融合与目标检测的数据集。该数据集涵盖了多种场景、环境、光照和天气条件下的高分辨率红外和可见光图像,包含了 4 177 组对齐的红外和可见光图像对以及 23 635 个标注的目标,输入图像尺寸为 1 024×768。我们去掉了除了‘Person’以外的所有标签,因为它们不是我们训练的目标。

为了更全面地对比多光谱行人检测,本文采用了 MS-COCO[26] 提出的三种目标检测评价指标进行评估:AP50、AP75、mAP。为了评价模型的性能,本文使用模型参数量(Param)、计算量 GFLOPs 作为评价指标。其中 AP50 表示 IOU 阈值为 0.5 时的 mAP,mAP 表示 IOU 从 0.5 到 0.95,步长为 0.05 时 mAP 的平均值。AP50、AP75 和 mAP 越高越好。如式(10)所示,TP 表示正确识别行人的数量,即检测器预测的 box 和 Ground Truth 满足交并比阈值;否则,将被判定为错误识别(FP)。FN 表示错误识别非行人的数量,意味着存在真正的目标,但检测器并未找到它。

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \\ \text{mAP} &= \frac{1}{n}\sum_{i=0}^{n} P_i(r)\,\mathrm{d}r \end{aligned} \tag{10}$$

2.2 实验参数设置

本文实验配置:操作系统为 Ubuntu18.04 LST,深度学习框架为 PyTorch2.1,CPU 为 i7-12700,内存为 128 GB,GPU 为 NVIDIA GeForce RTX 3090,显存为 24 GB。我们使用随机梯度下降(SGD)优化器,其初始学习率为 1e-2,动量为 0.937,权重衰减为 0.000 5,为了追求更好的性能,我们采用在 COCO 数据集上预先训练的 YOLOv8 模型作为权重初始化。由于 COCO 只有 RGB 图像,对红外输入的预训练模型进行微调的情况下需要对红外图像进行仔细的预处理。

2.3 不同检测模型对比实验

为了验证本文所提改进算模型的总体有效性,我们进一步将 CMDFT 模块与其他多光谱行人检测模型进行对比,包括 CFR[27]、GAFF[28]、ProbEn[10]、CFT[9],并选用 LLVIP 数据集进行实验,实验结果如表 1 所示。

表 1 相似模型大小的不同算法性能对比

Method	Modality	Param	GFLOPs	AP50	AP75	mAP
CFR	RGB+T	-	-	93.4	50.2	51.5
GAFF	RGB+T	31M	-	94.0	60.2	55.8
ProbEn	RGB+T	-	-	93.4	50.2	51.5
CFT	RGB+T	206.03M	224.40	97.5	72.9	63.6
CMDFT	**RGB+T**	**40.83M**	**150.1**	**97.3**	**77.2**	**67.3**

由表 1 可知,CMDFT 表示为本文提出的多光谱行人检测模型,我们使用 CFT 作为基线,相较于其他算法,本文所提出的 CMDFT 算法在 AP50、AP75 和 mAP 都取得了较高的检测精度,且参数量和计算复杂度都优于上述模型。此外,我们的模型在 AP75 和 mAP 上比基线模型(CFT)提高了 5.9%和 5.8%,参数量 Param 和模型计算量 GFLOPs 下降了 80.18%和 33.11%。

2.4 消融实验

为了进一步验证本文提出模型的可行性和部分改进对网络的性能提升效果,将改进方法分为两个主要的部分:第一部分是骨干网络采用单模态的 CSPDarkNet53、双路 CSPDarkNet53 和 FasterNet 改进的双路 CSPDarkNet53 进行对比,第二部分是采用最基本的融合方法和本文所提出的 CMDFT 进行对比,消融实验结果如表 2 所示。

首先我们在 LLVIP 数据集上验证 YOLOv8 的性能,在输入单模态图片情况下,YOLOv8 验证得到的 mAP 值分别为 49.1%和 60.1%,红外模态的验证结果超过了可见光模态。这可能是由于 LLVIP 中存在大量的低光场景,导致有效目标区域的损失。我们将输入模态为红外的 YOLOv8 与双路 YOLOv8 进行比较发现,简单的双流网络无法充分利用不同模式之间的内在互补性,可能会增加网络学习的难度,加剧模式的不平衡,从而导致性能下降。

为了展示 DFMDA 模块的高效性,我们将简单的 Concat 融合方法和 CMDFT 模块作比较,在参数提升 4.24M 的情况下,mAP 提升了 4.6 个百分点,网络模型的整体性能得到了有效提升。在骨干网络采用 PConv 改进的情况下,模型的参数量和计算量显著下降,且 mAP 得到了很好的保持。实验表明我们的模型可以有效地学习不同模态的互补信息,提高检测结果。

表 2　不同骨干网络和融合方法下的模型性能

Model	Backbone	Fusion Method	Param	GFLOPs	mAP
YOLOv8	CSPD53(RGB)	–	43M	165.2	49.1
YOLOv8	CSPD53(Thermal)	–	43M	165.2	60.1
YOLOv8	Two-way CSPD53	Concat	64M	254.1	62.3
YOLOv8	Two-way CSPD53	CMDFT	68.24M	266.4	66.9
Ours	Improved CSPD53	**CMDFT**	**40.83M**	**150.1**	**67.3**

2.5　检测结果可视化展示

为了更直观地对比改进前后网络的模型的性能,本文分别选择复杂背景下夜间行人的可见光和红外图片,并在基础的 YOLOv8、CFT 和本文提出的算法进行可视化比较。

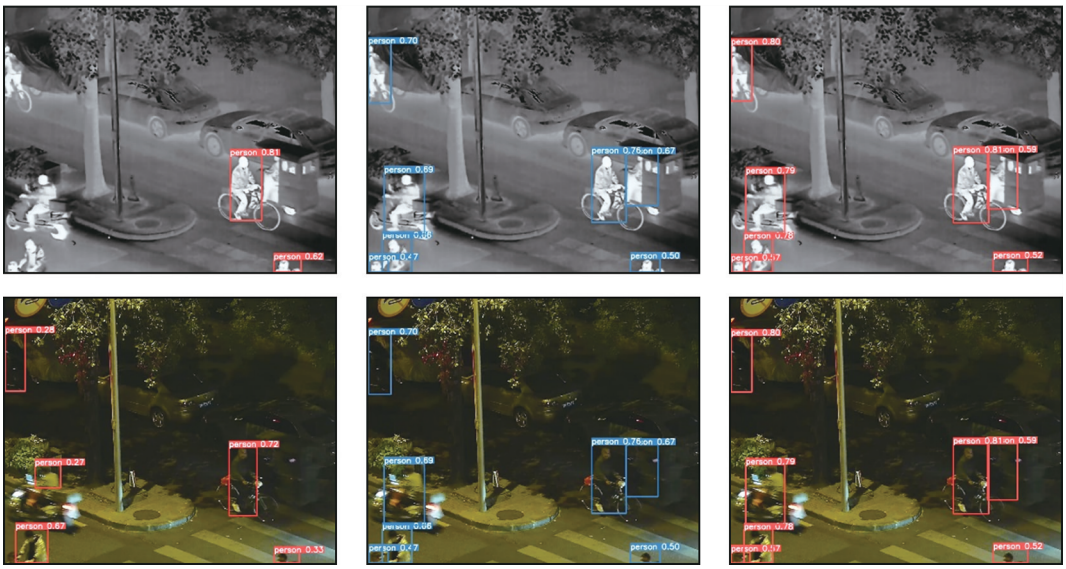
①在 LLVIP 数据集上的检测结果展示

在图 8(a) YOLOv8 的红外和可见光图中,可以看到可见光模态下从复杂背景中检测得到行人的置信度得分很低,甚至出现漏检情况。对于图 8(a) YOLOv8 的红外模态下来说,由于红外特征图提供的图像纹理信息不是很充分,但是能清晰地捕捉到行人的纹理信息,从而能检测出可见光模态下关注不到的物体。YOLOv8 的可见光和红外检测结果,证明了红外图像和可见光图像之间的互补性,所以实现跨模态特征互补是很有必要的,能够增强多光谱目标检测中的特征表示能力。

图 8(b)表明 CFT 方法利用 Transformer 网络可以有效地获取模态之间的潜在特征,从而有效地提高了多光谱物体检测的性能。图 8(c)表明所提出的模型与 CFT,均能够检出复杂背景下的行人目标,但是所提出的方法检测得到的置信度分数更高,在低光照环境下检测更加准确。

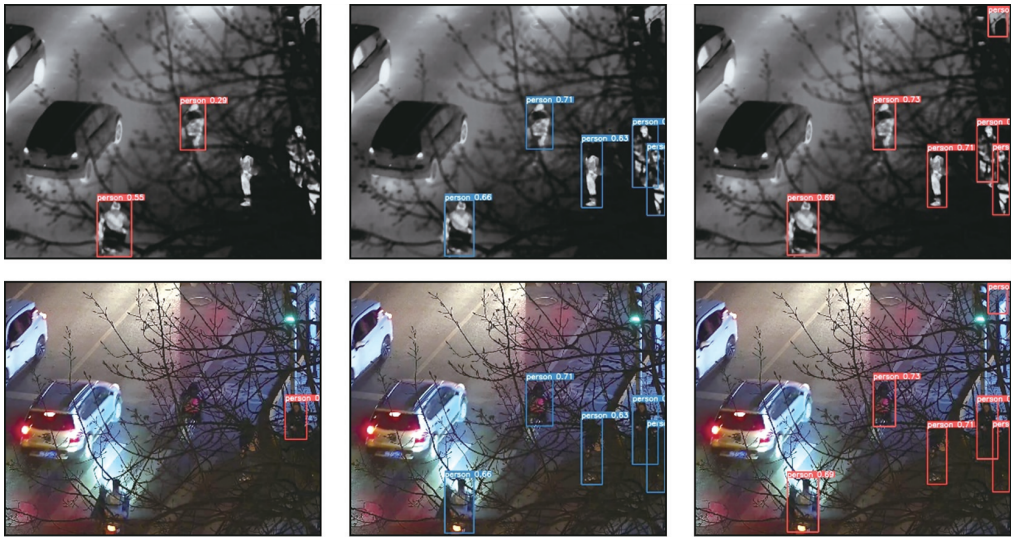
②在 M3FD 数据集上的检测结果展示

图 9 所示为在 M3FD 数据集检测结果的可视化。在图 9(a)中,在复杂背景下存在着目标遮挡,YOLOv8 的检测效果不是很好,出现漏检情况,而所提出的方法相较于基线方法并未出现漏检,且取得了最好的检测结果。



(a) YOLOv8　　(b) CFT　　(c) 本文算法

图 8　在 LLVIP 数据集上的检测结果对比



(a) YOLOv8 (b) CFT (c) 本文算法

图 9 在 M3FD 数据集上的检测结果对比

3 结论

经过实验验证的结果可以看出,基于跨模态特征差分 Transformer 的多光谱行人检测算法 CMDFT-YOLOv8,主要由两个改进方式组成:①针对现如今多模态目标检测网络参数量大、计算量复杂度高的问题,构建 PConv-GRNBlock 模块,以改进 YOLOv8 的 BottleNeck 结构,进一步降低双路骨干网络的计算量与参数量。在保证网络检测精度不下降的情况下,有效地减少了网络整体参数量和计算复杂度,实现了网络检测推理速度的大幅提升。②为了更好地融合可见光模态和红外模态特征。本文提出了一种基于跨模态特征差分 Transformer 的多光谱行人检测算法 CMDFT-YOLOv8,通过差异增强模块获取不同模态之间的差异信息并进行增强,在跨模态特征交互模块中与来自其他模态的补充信息进行交叉计算,以充分挖掘潜模态特定特征中提高检测性能的潜在能力;跨模态特征交互模块通过空洞临近注意力机制以学习长程依赖关系并集成全局上下文信息,且有效地降低自我注意力的二次复杂,从而增强了双流 CNN 在多光谱目标检测中的表示能力。

综上所述,本文方法能显著提升行人检测精度,且有效降低模型参数量和计算复杂度,由于我们的方法简单而有效,它可以应用于其他计算机视觉融合任务,如 RGB-激光雷达、RGB-D 任务等,具有实际应用价值

参考文献:

[1] 张国立,常帅,宋延嵩,等. 基于可变形卷积和多尺度残差注意力的多光谱行人检测方法[J]. 激光与光电子学进展,2024,61(10):351-358.

[2] 刘珂琪,董绵绵,郜辉,等. 基于光照感知权重融合的多模态行人检测算法[J]. 激光与光电子学进展,2023,60(16):145-155.

[3] 赵什陆,张强. 深度学习多模态图像语义分割前沿进展[J]. 中国图像图形学报,2023,28(11):3320-3341.

[4] 钱多,殷俊. 基于俯视角融合的多模态三维目标检测[J]. 南京大学学报(自然科学),2023,59(6):996-1002.

[5] 唐霖峰,张浩,徐涵,等. 基于深度学习的图像融合方法综述[J]. 中国图像图形学报,2023,28(1):3-36.

[6] Liu J J, Zhang S T, Wang S, et al. Multispectral Deep Neural Networks for Pedestrian Detection[C]//2016 27th British Machine Vision Conference(BMVC), York, UK, 2016;73.1-73.13.

[7] Li C, Song D, Tong R, et al. Multispectral Pedestrian Detection via Simultaneous Detection and Segmentation [C]//2018 British Machine Vision Conference(BMVC), Newcastle, UK, 2018:5126-5136.

[8] Zhou K, Chen L, Cao X. Improving Multispectral Pedestrian Detection by Addressing Modality Imbalance Problems [C]//ECCV 2020:16th European Conference on Computer Vision, Glasgow, UK, Springer, 2020:787-803.

[9] Fang Q Y, Han D P, Wang Z K. Cross-Modality Fusion Transformer for Multispectral Object Detection[EB/OL]. (2022-10-04). https://arxiv.org/abs/2111.00273.

[10] Chen Y T, Shi J, Ye Z, et al. Multimodal Object Detection via Probabilistic Ensembling[C]//17th European Conference on Computer Vision, Tel Aviv, Israel, 2022:139-158.

[11] Fan Q Y, Wang Z K. Cross-Modality Attentive Feature Fusion for Object Detection in Multispectral Remote Sensing Imagery[J]. Pattern Recognition, 2022, 130:108786.

[12] 金宇锋,陶重犇. 基于 Transformer 的融合信息增强 3D 目标检测算法[J]. 仪器仪表学报,2023,44(12):

297-306.

[13] Chen J R, Kao S H, He H, et al. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 2023: 12021-12031.

[14] Choi H, Kim S, Park K, et al. Multi-Spectral Pedestrian Detection Based on Accumulated Object Proposal with Fully Convolutional Networks [C]//2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, IEEE, 2016: 621-626.

[15] Kieu M, Bagdanov A D, Bertini M, et al. Task-Conditioned Domain Adaptation for Pedestrian Detection in Thermal Imagery [C]//16th European Conference on Computer Vision, Online, 2020: 546-562.

[16] Wagner J, Fischer V, Herman M, et al. Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks [C]//27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (FSANN), Bruges, Belgium, 2016, 587: 509-514.

[17] Hassani A, Walton S, Li J, et al. Neighborhood Attention Transformer [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 2023: 6185-6194.

[18] Howard A, Sandler M, Chu G, et al. Searching for MobileNetv3 [C]//2019 IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 2019: 1314-1324.

[19] Sandler M, Howard A, Zhu M L, et al. MobileNetv2: Inverted Residuals and Linear Bottlenecks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018: 4510-4520.

[20] Han K, Wang Y H, Tian Q, et al. GhostNet: More Features from Cheap Operations [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 2020: 1580-1589.

[21] Zhang P, Lo E, Lu B. High Performance Depthwise and Pointwise Convolutions on Mobile Devices [C]//2020 AAAI Conference on Artificial Intelligence, New York, NY, USA, 2020, 34(4): 6795-6802.

[22] Ma N N, Zhang X Y, Zheng H T, et al. Shufflenet v2: Practical Guidelines for Efficient CNN Architecture Design [C]//15th European Conference on Computer Vision (ECCV), Munich, Germany, 2018: 122-138.

[23] Woo S, Debnath S, Hu R, et al. Convnext v2: Co-Designing and Scaling Convnets with Masked Autoencoders [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 2023: 16133-16142.

[24] Jia X Y, Zhu C, Li M Z, et al. LLVIP: A Visible-Infrared Paired Dataset for Low-Light Vision [C]//2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 2021: 3496-3504.

[25] Liu J, Fan X, Huang Z, et al. Target-Aware Dual Adversarial Learning and a Multi-Scenario Multi-Modality Benchmark to Fuse Infrared and Visible for Object Detection [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 2022: 5802-5811.

[26] Lin T Y, Maire M, Belongie S, et al. Microsoft Coco: Common Objects in Context [C]//13th European Conference on Computer Vision, Zurich, Switzerland, Springer, 2014: 740-755.

[27] Zhang H, Fromont E, Lefevre S, et al. Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks [C]//2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 2020: 276-280.

[28] Zhang H, Fromont E, Lefèvre S, et al. Guided Attentive Feature Fusion for Multispectral Pedestrian Detection [C]//2021 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2021: 72-80.

项　靖(1998—),男,现为淮阴工学院电子信息工程学院 22 级硕士研究生,研究方向为多模态目标检测,1217128095@qq.com;

姜明新(1979—),女,博士,教授,硕士生导师,研究方向为计算机视觉和深度学习,jiangmingxin@126.com。