计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009878] [CSTR: 32024.14.csa.009878]

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

融合改进注意力的自适应双分支密集行人检测①

李建东[1,2], 焦晓光[1], 曲海成[1]

[1](辽宁工程技术大学 软件学院, 葫芦岛 125105)
[2](辽宁工程技术大学 矿业学院, 阜新 123000)
通信作者: 焦晓光, E-mail: jxg0305xg@163.com

摘　要: 为解决复杂背景干扰导致的行人检测精度低和漏检率高的问题, 本文提出一种融合改进注意力的自适应双分支密集行人检测算法 DACD-YOLO. 首先, 主干网络采用自适应融合双分支结构, 通过动态权重实现不同特征的融合, 并引入深度可分离卷积降低计算量, 有效缓解传统单分支网络中信息丢失的问题; 其次, 提出自适应视觉中心, 通过动态优化增强层内特征提取, 并重设通道数以平衡精度与计算量; 然后, 提出坐标双通道注意力机制, 结合异构卷积核设计轻量化融合模块, 降低计算复杂度并增强对关键特征的捕捉能力; 最后, 采用膨胀卷积检测头, 通过不同膨胀率卷积融合多尺度特征, 有效增强小目标和遮挡目标的特征提取能力. 实验结果表明, 与原版 YOLOv8n 相比, 改进算法在 WiderPerson 数据集上的 *mAP*@0.5 和 *mAP*@0.5:0.95 分别提高 2.3% 和 2.2%, 在 CrowdHuman 数据集上 *mAP*@0.5 和 *mAP*@0.5:0.95 分别提升 3.5% 和 4.6%. 实验证明, 改进算法在密集行人检测方面相较于原算法具有显著的精度提升.

关键词: 行人检测; 自适应; 双分支; 深度可分离卷积; 注意力机制; 异构卷积核

引用格式: 李建东,焦晓光,曲海成.融合改进注意力的自适应双分支密集行人检测.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9878.html

Adaptive Dual-branch Dense Pedestrian Detection with Improved Attention

LI Jian-Dong[1,2], JIAO Xiao-Guang[1], QU Hai-Cheng[1]

[1](Software College, Liaoning Technical University, Huludao 125105, China)
[2](College of Mining, Liaoning Technical University, Fuxin 123000, China)

Abstract: To address the low accuracy and high miss detection rates in pedestrian detection caused by complex background interference, this study proposes an adaptive dual-branch dense pedestrian detection algorithm, DACD-YOLO, incorporating improved attention mechanisms. First, the backbone network employs an adaptive dual-branch structure, which fuses different features through dynamic weighting while introducing depthwise separable convolution to reduce the computational cost, effectively mitigating the information loss present in traditional single-branch networks. Second, an adaptive vision center is proposed to enhance intra-layer feature extraction through dynamic optimization, with channel numbers reconfigured to balance accuracy and computational load. A coordinate dual-channel attention mechanism is then introduced, combining a heterogeneous convolution kernel design within a lightweight fusion module to reduce computational complexity and improve the capture of key features. Lastly, a dilation convolution detection head is utilized, fusing multi-scale features through convolutions with varying dilation rates, effectively enhancing feature extraction for small and occluded objects. Experimental results show that, compared to the original YOLOv8n, the proposed algorithm improves *mAP*@0.5 and *mAP*@0.5:0.95 by 2.3% and 2.2%, respectively, on the WiderPerson dataset, and by 3.5% and 4.6%, respectively, on the CrowdHuman dataset. The experiments demonstrate that the proposed

① 基金项目: 辽宁省教育厅基本科研项目 (JYTMS20230804); 辽宁工程技术大学学科创新团队 (LNTU20TD-23)
收稿时间: 2024-11-18; 修改时间: 2024-12-09; 采用时间: 2025-01-10; csa 在线出版时间: 2025-04-01

algorithm significantly enhances accuracy in dense pedestrian detection compared to the original method.

Key words: pedestrian detection; adaptive; dual-branch; depth separable convolution; attention mechanism; heterogeneous convolution kernel

随着人工智能技术的蓬勃发展, 深度学习在行人检测技术中的应用日益广泛[1]. 相比于传统依赖手工设计特征的算法, 深度学习算法能够自动提取数据特征, 处理复杂的非线性数据, 展现出更高的扩展性和通用性[2], 更适合现代通信和实时处理任务. 凭借其精准的识别和定位能力, 深度学习行人检测技术在智能交通、安全监控以及无人机等领域展现出重要的应用价值[3−5] 然而, 在人员密集场所, 行人姿态的多样性、尺度变化以及遮挡等问题成为影响精度检测的主要因素[6,7]. 这些复杂多变的因素要求行人检测算法必须具备出色的适应能力和高效的处理能力, 以便在各种环境下准确识别和定位.

在拥挤人群中, 行人之间的遮挡会造成目标信息缺失, 从而造成目标漏检; 在远距离识别中, 小目标行人通常伴随着低分辨率问题, 使得特征提取变得困难; 在移动设备和无人机通信等系统中, 降低复杂度可以保证算法在有限的资源下顺利运行. 针对这些问题, 研究者们提出了多种解决方案. 针对目标遮挡的影响, 周大可等人[8]提出结合空间和通道双重注意力机制的遮挡感知模型, 有效增强遮挡区域的特征提取能力. Liu 等人[9]通过融合 5 层注意力机制并引入多层双向加权特征融合结构, 有效缓解信息漏检, 但影响了时效性. 针对小目标检测的问题, 陈秀锋等人[10]引入小目标检测层, 融合异类冗余框, 有效降低了小目标漏检率, 但在一定程度上牺牲了检测速度. Zhang 等人[11]提出融合残差网络和特征金字塔的小目标行人检测算法, 通过特征选择、对齐模块和级联自动聚焦查询模块, 提高了小目标检测精度和抗干扰能力, 但该算法在大目标检测方面未展现出明显优势. 针对计算复杂度偏高问题, Dong 等人[12]在算法颈部引入轻量级 Ghost 模块, 通过分通道卷积替代普通卷积, 减少了计算负载. Shen 等人[13]选择 ShuffleNetV2 作为主干网络, 引入通道重排操作, 有效降低了计算量. 虽然 Ghost 和 ShuffleNetV2 都有效减少了算法规模, 但由于其网络结构相对较浅, 导致其在捕捉细节和小尺寸目标方面的能力略显不足.

这些行人检测算法在目标遮挡、小目标检测和减少计算复杂度的设计上各有优势, 但在如何平衡高精度检测与算法复杂度方面仍存在不足. 为此, 本文针对复杂场景对 YOLOv8n 进行改进, 提出了一种兼顾轻量和性能的密集行人检测算法 DACD-YOLO (dual-branch+ADPEVC+C2f-HC+DG-Detect-YOLO). 本文主要工作如下.

(1) 提出一种自适应融合双主干网络, 通过动态权重实现多特征的加权融合, 显著提升遮挡行人和小目标的检测性能, 同时引入深度可分离卷积, 在保证检测精度的同时降低计算量.

(2) 提出自适应视觉中心 (adaptive explicit visual center), 通过自适应调节来增强层内不同尺度的重要特征, 抑制不重要特征, 减少局部密集分布中的漏检问题, 并通过通道数重设平衡精度与计算量.

(3) 提出 CDA 注意力机制 (coordinate dual-branch attention), 通过双通道完善深、浅层特征的优势互补, 结合异构卷积核, 设计一款轻量化模块 C2f-HC (C2f-HetConv-CDA), 在保持低计算量的同时, 提升算法对空间和细粒度特征的理解和捕捉能力.

(4) 设计多尺度膨胀卷积模块 DG-Conv (dilation wise residual-GELU-Conv), 通过不同膨胀率卷积融合多尺度信息, 并结合原检测头设计膨胀卷积检测头 DG-Detect (DG-Conv-Detect), 以更好地适应输入特征的变化, 增强对小目标和遮挡目标的多尺度特征提取.

1 DACD-YOLO

YOLOv8 在行人检测中表现出色, 但在复杂背景下的密集行人检测中仍存在局限, 包括密集遮挡导致的信息丢失、层内特征不足、小目标检测能力弱, 以及精度与计算复杂度平衡欠佳. 为解决这些问题, 本文提出了一种改进的密集行人检测算法 DACD-YOLO, 旨在以较低的计算量实现更高的检测精度, 其整体架构如图 1 所示.

DACD-YOLO 采用梯度分流的思想, 在 YOLOv8n 的基础上进行优化, 它由双分支主干网络 (Backbone)、颈部网络 (Neck)、头部网络 (Head) 这 3 部分组成, 在

主干中设计多层自适应融合辅助分支通过深度可分离卷积和动态融合权重, 缓解密集遮挡导致的信息丢失, 实现低计算量下的动态高效特征提取. 将原算法的 C2f 替换为低计算量的 C2f-HC, 通过异构卷积核降低模块复杂度, 并通过 CDA 注意力机制, 实现深浅层双通道的特征优势互补. Neck 位于 Backbone 和 Head 之间, 通过特征金字塔网络融合多尺度特征, 并引入 ADPEVC 模块进行自适应调节, 增强层内特征表达. Head 部分采用 3 个多尺度膨胀卷积检测头, 融合不同尺度和层级特征信息, 有效提升行人检测效果.

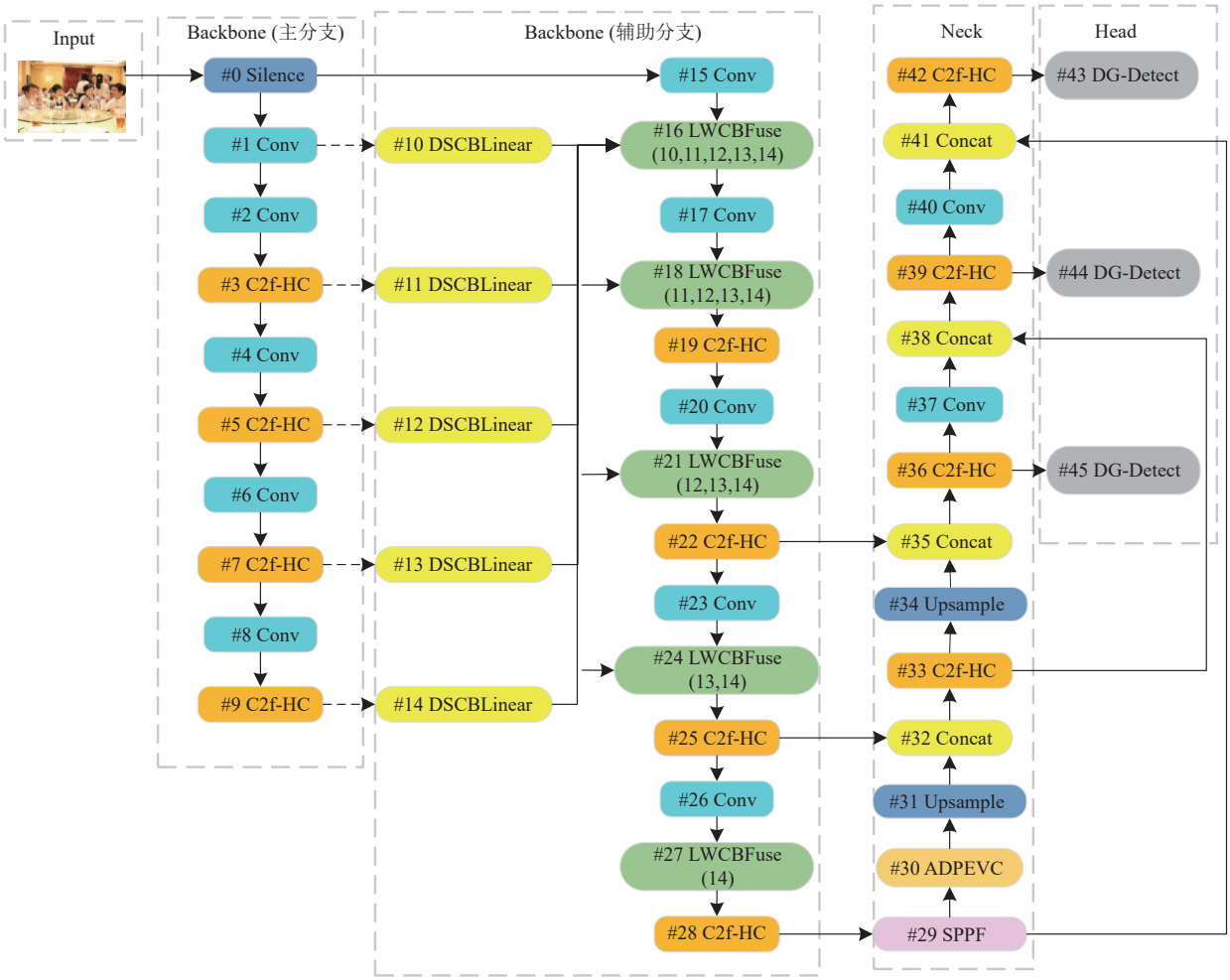


图 1 DACD-YOLO 网络结构图

1.1 自适应融合双分支结构

Wang 等人[14]提出一种可编程梯度信息 (programmable gradient information, PGI) 通过辅助可逆分支增强特征提取并缓解信息丢失, 但训练时的计算开销较大, 且缺乏对不同任务和输入的适应能力. 为此, 本文提出自适应融合双分支结构, 以低成本提升检测精度和稳定性.

在该双分支体系中, 主分支作为网络的关键路径, 专注于高效完成前向传播、特征提取及最终的目标检测推理. 与原算法相似, 主分支由 5 个 Conv 模块与 4 个 C2f-HC 模块组成, 如图 1 所示, Conv 模块通过下采样与特征增强来构建稳健的表征基础, C2f-HC 模块则通过多路径特征提取与融合提高特征复杂场景的适应能力. 然而, 随着网络深度的加深, 原始数据在逐层传递过程中面临信息持续损耗, 致使输出层所能保留的有效信息愈发匮乏, 尤其对密集场景与小目标检测的收敛性能影响显著. 为此, 本文引入了自适应融合辅助分支, 通过可逆结构和自适应融合策略在反向传播过程中为主分支提供更为完整而精确的梯度信息, 从而缓解因信息缺失导致的梯度不稳定性. 双分支体系的设

计旨在使辅助分支为主分支的参数更新注入可靠梯度信号, 而主分支则承担基础特征提取与最终预测任务, 二者形成高效协同, 实现精度的有效提升. 实验结果表明, 在 YOLOv8 框架上引入双分支网络仅以少量增加的计算开销换取了显著的精度提升, 为高性能目标检测提供了更优解.

在辅助分支中, DSCBLinear (depthwise separable convolution CBLinear) 与 LWCBFuse (learnable weights CBFuse) 紧密协同, 构建了多阶段、多特征图的处理流程, 能够高效提炼与整合特征, 为主分支在参数优化过程中提供充足且稳定的参考信息. 同时, 通过引入深度可分离卷积的手段, 有效控制了辅助分支所需的额外计算开销. 深度可分离卷积由逐深度卷积与逐点卷积两部分组成, 其中前者聚焦于空间特征的提取, 后者则专注于通道特征的整合[15]. 本文将其运用到 DSCBLinear 中, 其结构如图 2 中 A-stage 所示.

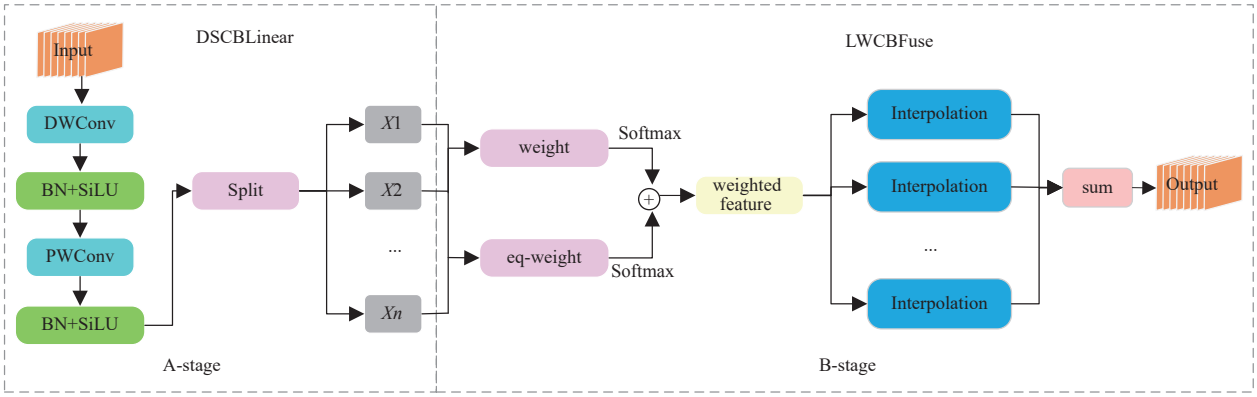


图 2 辅助分支部分结构图

假设输入通道数为C_{in}, 输出通道数为C_{out}, 卷积核大小为$k\times k$, 则普通标准卷积的参数量C_s为:

$$C_s = C_{\text{in}} \times C_{\text{out}} \times k^2 \tag{1}$$

深度可分离卷积的参数量C_{DSC}为:

$$C_{\text{DSC}} = C_{\text{DW}} + C_{\text{PW}} = C_{\text{in}} \cdot k^2 + C_{\text{in}} \cdot C_{\text{out}} \tag{2}$$

其中, C_{DW}表示逐深度卷积的参数量, C_{PW}表示逐点卷积的参数量. 深度可分离卷积与普通标准卷积的参数量之比为:

$$\frac{C_{\text{DSC}}}{C_s} = \frac{1}{C_{\text{out}}} + \frac{1}{k^2} \tag{3}$$

因此, 深度可分离卷积可以显著减少计算开销.

Navon 等人[16]提出基于对称性的神经网络设计方法, 利用权重空间中的对称性, 显著提升了模型的泛化能力, 并为理解神经网络的权重结构提供了新的视角. 基于此思想, 本文设计 LWCBFuse 结构, 如图 2 中 B-stage 所示, 通过融合多输入特征图, 结合各特征图的重要性及其相互关系, 实现了高效的特征整合. 该结构中的可学习权重使网络能够动态调整每个特征图的贡献, 而等变性权重矩阵有效地捕捉了不同特征之间的复杂关系. 通过矩阵乘法整合这些权重, 模块生成了自适应的融合权重, 增强了特征表示能力.

首先输入 n 个特征图$X_i \in \mathbb{R}^{C_i\times H_i\times W_i}$, C_i是通道数, H_i和W_i是高度和宽度, 对原始权重进行 Softmax 归一化:

$$\alpha_i = \frac{\mathrm{e}^{w_i}}{\sum_{j=1}^{n} \mathrm{e}^{w_j}} \tag{4}$$

其中, α_i表示第i个输入特征图的归一化权重, w_i表示第i个输入特征图的原始权重, j 表示索引变量, 用于遍历所有特征图的原始权重w_j. 通过 Softmax 对等变性权重矩阵的每行进行归一化, 以计算不同特征图间的相互关系:

$$\beta_{ij} = \frac{\mathrm{e}^{w_{ij}}}{\sum_{k=1}^{n} \mathrm{e}^{w_{ik}}} \tag{5}$$

其中, w_{ij}表示等变性权重矩阵中的第i行第j列的元素, k用于遍历当前行中所有列的索引.

可学习权重α_i和等变权重β_j通过矩阵乘法得到组合权重向量γ_i. 然后, 对每个特征图进行双线性插值调整大小, 使其统一为指定尺寸, 将组合权重应用于每个特征图以获得加权特征, 最后将所有加权特征图求和, 从而得到最终的输出特征图Y:

$$Y=\sum_{i=1}^{n}\gamma_i\cdot Interpolate(X_i,size(H,W)) \tag{6}$$

其中, *Interpolate* 表示插值操作, 用于将输入特征图调整到目标尺寸 *size*, 以便与其他特征图对齐, 确保后续融合操作可以顺利进行.

结合自适应融合权重和等变性权重矩阵, 模块生成依赖于特征图重要性及其相互关系的加权特征. 通过多次训练迭代, 自适应机制逐步优化输入特征图的贡献, 使得模型在密集行人检测中能够精准分配特征权重, 更有效地整合小目标和遮挡场景下的特征信息. 同时, 自适应融合权重能够有效关联深层与浅层特征, 避免因“平均融合”导致的特征损失, 从而显著提升融合效果, 实现更精确且稳定的检测性能.

1.2 自适应视觉中心块

显式视觉中心块 (explicit visual center, EVC)[17]是一种广义的层内的特征调节方法, 由多层感知机 (multi-layer perceptron, MLP) 和可学习的视觉中心 (learnable visual center, LVC) 两个模块组成, 其中轻量级 MLP 用于捕获特征图中的全局长距离依赖关系, LVC 专注于局部特征聚合. 本文引入改进的 EVC 模块, 并设计通道重标定机制来动态调整通道权重, 使网络能够有效地捕捉局部和全局特征, 提升网络对视觉中心和关键特征的关注度. 通道重标定机制使用了类似 SE 注意力机制[18]的策略, 通过全局平均池化提取每个通道的全局特征, 然后使用轻量化的 1×1 卷积生成通道的动态权重, 有效提升了网络的表达能力, 结构图如图 3 所示.

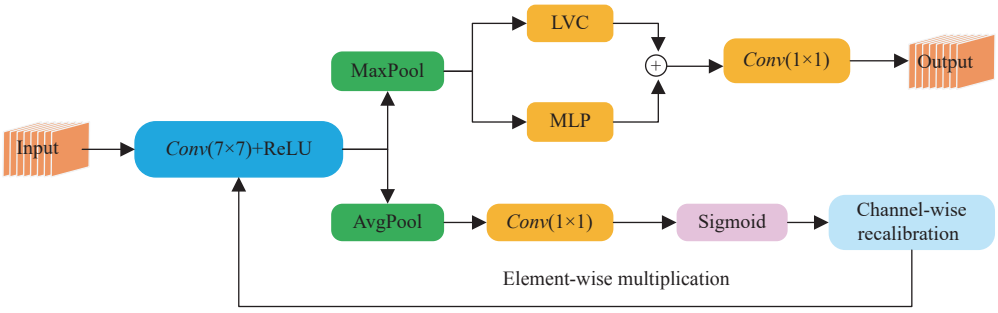


图 3 自适应视觉中心块

首先对输入进行卷积操作, 并通过激活函数 (ReLU) 提取初步特征, 随后通过最大池化缩小特征图的空间尺寸. 接着, 将池化后的特征分别传入 LVC 和 MLP 进行进一步处理, 从而同时捕获局部细节和全局信息, 增强特征的多样性及模型整体的感知能力. 这些不同模块提取的特征会在通道维度上进行拼接, 并通过 1×1 的卷积融合输出. 同时, 采用通道重标定机制对通道进行调整, 通过全局平均池化获取全局特征得到X_{avg}:

$$X_{avg}=AvgPool(X_{conv}) \tag{7}$$

通过1×1卷积和 Sigmoid 激活函数生成权重, 并将这些权重应用于缩放初步提取的特征, 从而增强模型对重要特征的关注. 随后将权重w归一化到 (0, 1) 的范围, 通过w体现不同通道的重要程度:

$$w=\sigma(Conv_{1\times1}(X_{avg})) \tag{8}$$

其中, σ表示 Sigmoid 激活函数.

最后, 将这些权重w应用到原始卷积后的特征图上, 从而实现了通道级的自适应调节:

$$X_{out}=X\cdot w \tag{9}$$

不同的输入特征会生成不同的通道权重, 使得模型能够对重要特征给予更高的响应, 而减少不重要特征的影响. 在整个网络训练过程中, 权重是通过反向传播自适应更新的. 当输入数据的特征模式发生变化时, 网络中的卷积核和通道重标定的权重会根据损失函数进行自适应地优化和更新, 使得网络可以学习到更适合于不同任务或数据模式的特征表达方式.

原模型中的 SPPF 模块提供了多尺度特征提取的全局和局部视野, 而 ADPEVC 模块进一步强化了层内的特征调节, 两者在全局和细节特征提取上进行互补, 使得整个网络能够更好地适应复杂环境中的行人检测需求. 增加通道数可以增加网络容量, 捕获更丰富的特征, 但会出现增加计算成本和过拟合现象, 为在特征丰富性与计算成本之间权衡, 基于大量实验验证的结果, 本文将 ADPEVC 和 SPPF 的通道数设为 256.

1.3 坐标双通道注意力机制

注意力机制借鉴人类注意力的选择性, 有选择地关注其中的关键部分, 可以减少卷积操作带来的信息丢失, 提升目标识别能力. SE 注意力机制以低成本计算通道注意力, 有效提升了性能, 但它只考虑通道间信息而忽略了位置信息. CBAM 注意力机制通过通道和空间注意力机制相结合可增强特征表示能力[19], 但在捕捉全局上下文信息和处理复杂场景方面可能略有不足. CA 注意力机制结合了通道和空间注意力, 通过引入坐标信息来增强特征图的表示能力[20]. 它通过对输入特征图在水平和垂直两个方向进行全局平均池化, 生成两个方向的注意力向量, 并通过一系列卷积操作将这些向量融入特征图中, 从而保留位置信息并突出重要特征, 但它只考虑了一个统一的特征通道. 本文提出的 CDA 注意力机制将生成过程分为“浅层”和“深层”两部分, 使得注意力机制充分利用两层特征进行优势互补, 并引入 SiLU 激活函数, 从而更好地提取不同尺度下的特征信息, 结构如图 4 所示.

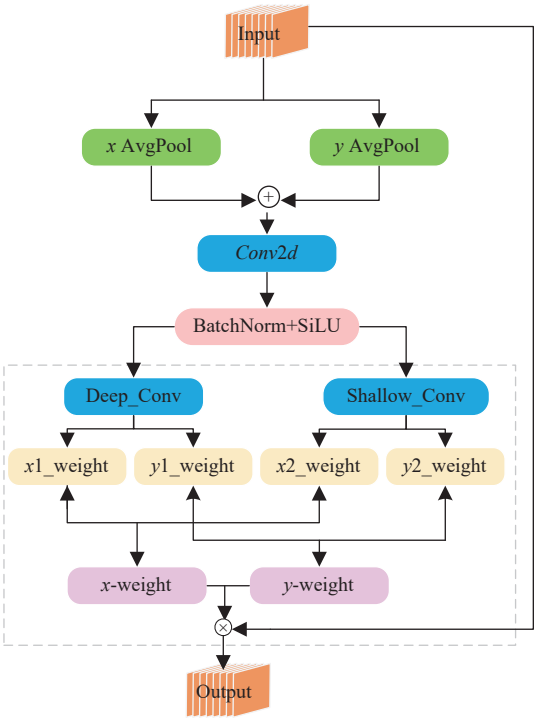


图 4 CDA 注意力机制结构图

首先, 对输入特征图X分别在水平和垂直两个方向进行平局池化, 得到特征X_x和X_y. 其次, 将这两个方向的特征映射进行拼接, 进入1×1卷积层进一步处理, 紧接着经过批归一化和 SiLU 激活函数, 进一步提取空间特征, 有利于模型的反向传播. 接着, 处理后的特征被分为两路, 一路通过逐深度卷积提取深层次的 x 和 y 方向特征权重, 可用式 (10) 表示:

$$\begin{cases}\alpha_x^{\text{deep}} = \sigma(Conv_{1\times1}Conv_{3\times3}(X_x))\\ \alpha_y^{\text{deep}} = \sigma(Conv_{1\times1}Conv_{3\times3}(X_y))\end{cases} \tag{10}$$

深层路径通过3×3的逐深度卷积提取复杂的上下文信息, 再经过1×1卷积生成两个方向的深层注意力权重α_x^{deep}和α_y^{deep}, σ代表 Sigmoid 激活函数. 另一路通过浅层卷积提取浅层次的 x 和 y 方向特征权重. 可用式 (11) 表示:

$$\begin{cases}\beta_x^{\text{shallow}} = \sigma(Conv_{1\times1}(X_x))\\ \beta_y^{\text{shallow}} = \sigma(Conv_{1\times1}(X_y))\end{cases} \tag{11}$$

浅层路径使用1×1卷积直接处理以保留细节特征, 生成两个方向的浅层注意力权重β_x^{shallow}和β_y^{shallow}. 最后, 深层和浅层的特征权重分别融合生成最终的 x 方向和 y 方向的注意力权重, 这些权重作用于输入特征上, 输出增强后的特征映射, 以提升模型在空间维度上捕捉特征的能力.

在深层路径中, 为获得较大感受野, 增强全局信息捕获, 增加3×3卷积, 随后通过1×1卷积进行降维, 形成一倒瓶颈结构, 在降低计算成本的同时, 保持丰富的特征表达能力. 在浅层网络中仅使用1×1卷积, 通过小感受野保持更多局部细节, 并减少计算复杂度.

1.4 融合 CDA 的轻量化模块

降低计算复杂度能够在资源受限的情况下实现更少的资源消耗. 本文设计 C2f-HC 模块来替换原算法的 C2f 模块, 通过异构卷积核和 CDA 注意力机制的结合, 保持算法的轻量化和精度, 适用于多种检测任务.

模型压缩可以有效降低复杂度, Li 等人[21]通过网络架构优化、剪枝等方法减少行人检测算法的参数量, 尽管剪枝操作在减少参数量方面有显著优势, 但增加了开发和调试的时间成本. Singh 等人[22]提出一种延迟为零的高效异构内核卷积 (HetConv), 使用不同类型的卷积核来处理输入特征图, 在保持精度的同时减少计算量. 使用高效的卷积滤波器设计能避免连接剪枝和过滤器剪枝中繁琐训练和修剪过程. 在本文中 HetConv 使用 1×1 和 3×3 卷积核的组合, 这种组合相比标准卷积能够在减少计算量的同时保持特征提取能力. 本文结合 HetConv 和 CDA 注意力机制思想, 改进原算法的 C2f, 设计出 C2f-HC, 其结构图如图 5 所示.

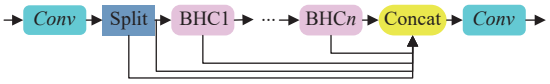


图 5 C2f-HC 结构图

输入特征图为$X \in \mathbb{R}^{H\times W\times C}$, 其中$H$是高度, W是宽度, C是通道数, X经过 *Conv* 层进行 Split 拆分, 一部分直接传递形成Y_i, 另一部分经过 BHC (Bottleneck-HetConv-CDA) 模块对卷积核进行异构化, 形成Y_j, 其中的 CDA 注意力机制增强了算法对重要特征的捕捉能力, 然后将Y_i和Y_j的特征图进行融合, 得到融合后的特征图Z. 最后融合后的特征图Z经过一个1×1卷积层, 输出最终的特征图O. 在 C2f-HC 中, BHC 中每个输入通道数都是上一级的 0.5 倍, 这可以有效降低计算量. 与原 C2f 中的 Bottleneck 相比, BHC 通过使用 1×1 和 3×3 卷积核的组合, 减少了总参数量, 而 Bottleneck 通常包含多个标准卷积层, 每个卷积层都需要大量的参数. C2f-HC 在保持高性能的同时, 实现计算复杂度的显著降低. 将 CDA 置于 Bottleneck 内部, 可在每次特征经过瓶颈处理后实现注意力机制的增强. 这样做能够在每个局部模块中对特征进行逐步筛选和强化, 从而提升特征表示的质量, 确保网络在处理过程中更加精准地关注关键信息, 最终提升模型的整体性能.

1.5 膨胀卷积检测头

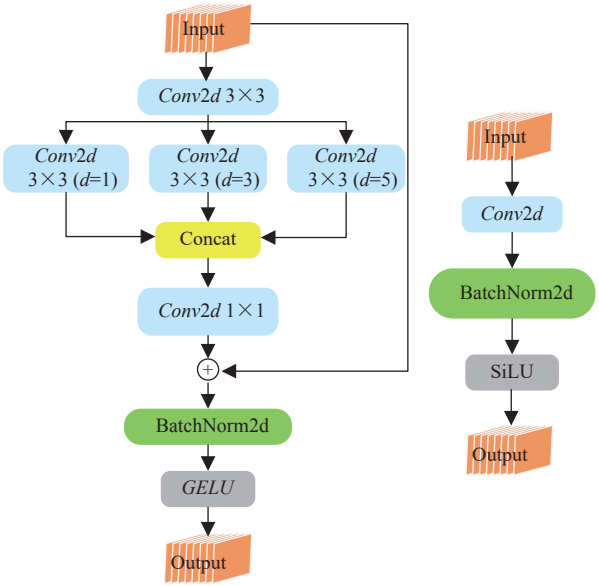
当前许多网络通过结合多尺度感受野来增强特征提取, 但这种无差别地获取上下文信息的做法可能会增加计算负担, 从而导致效率低下. 如 ESPNetV2[23]高低阶段采用相同的大感受野, 导致低阶的特征提取效率低下. Wei 等人[24]根据两步残差特征提取方法, 提出一种新的高效架构 DWRSeg, 显著提高捕获多尺度信息的效率与准确率.

在密集行人检测中, 单一尺度的卷积核难以适应复杂背景下的图像数据. 为增强检测的灵敏度, 设计了一种新的卷积结构 DG-Conv, 通过使用不同膨胀率的卷积核来捕捉多尺度特征, 如图 6(a) 所示. 相比于原始的 *Conv* 结构, DG-Conv 在处理不同尺度和复杂背景的目标时更加灵活, 能够捕捉更细致的特征, 这对遮挡和小目标行人的检测尤为有利. 图 6(b) 展示了原始 *Conv* 结构.

在 DG-Conv 卷积中, 输入特征图$X \in \mathbb{R}^{C\times H\times W}$, 其中$C$、$H$和$W$分别为输入通道数、高度和宽度. 首先通过一个3×3卷积层提取局部特征, 记为X_1. 然后将X_1传递给 3 个并行的膨胀卷积分支, 这些分支的膨胀率分别为 1、3 和 5. 通过使用不同膨胀率的卷积, 可以扩大感受野并捕获到不同尺度的特征. 不同膨胀率卷积可用式 (12) 表示为:

$$\begin{cases} X_2 = Conv2d_{3\times3,d=1}(X_1) \\ X_3 = Conv2d_{3\times3,d=3}(X_1) \\ X_4 = Conv2d_{3\times3,d=5}(X_1) \end{cases} \tag{12}$$

其中, d 表示膨胀率.



(a) DG-Conv结构图 (b) *Conv*结构

图 6 DG-Conv 和 *Conv* 结构对比图

随后, 来自各个分支的特征被拼接在一起, 形成更加丰富的多尺度特征表示, 并通过1×1卷积层进行通道融合与维度压缩, 从而减少计算复杂度. 在此过程中, 输入特征图还通过残差连接跳过卷积路径, 直接与经过卷积操作后的特征相加以保留原始输入的信息, 从而帮助梯度更好地流动, 减轻梯度消失问题. 之后, 特征经过批归一化来标准化数据分布, 并通过 *GELU* 激活函数对特征进行非线性变换, 使得网络能够更好地捕捉复杂特征. 可用公式表示为:

$$X_{\text{final}} = GELU(BN(Conv_{1\times1}(cat(X_2, X_3, X_4)))) \tag{13}$$

其中, *GELU* 是一种激活函数, 它将输入值映射到一个非线性的输出. 其数学表达式为:

$$GELU(x) = x\cdot\Phi(x) = \frac{x}{2}\left(1 + erf\left(\frac{x}{\sqrt{2}}\right)\right) \tag{14}$$

其中, Φ表示标准正态分布的累积分布函数, *erf* 表示误差函数:

$$erf(x)=\frac{2}{\sqrt{\pi}}\int_0^x \mathrm{e}^{-t^2}\mathrm{d}t \tag{15}$$

GELU 函数的特点是在接近 0 的位置保持了连续性, 在较大的输入值时保持了较大的梯度. 与传统的 ReLU 激活函数相比, GELU 可以带来更好的性能和收敛速度.

将原版算法检测头的 Conv 更换为 DG-Conv 设计出一款新的检测头 DG-Detect, DG-Conv 的引用使得卷积操作能够更好地适应多尺度特征的变化, 从而捕捉更细致的特征. 其结构图如图 7 所示.

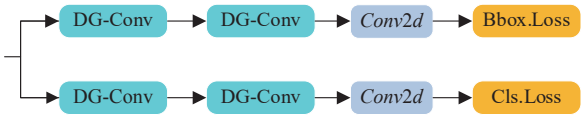


图 7 DG-Detect 结构图

DG-Detect 结合了 YOLOv8 原检测头 (Detect) 结构和 DWR 模块的理念, 通过多尺度膨胀卷积增强特征提取能力. 输入特征图通过 DG-Conv 模块后, 进入最终的 Conv2d 层进行处理, 生成用于分类和边界框回归的特征图, 这些设计能够在前向传播过程中有效完成密集行人检测任务, 实现对小目标和遮挡目标的高效、精确检测.

1.6 损失函数

边界框回归损失用于衡量预测边界框与真实边界框之间的偏差. 本文采用 YOLOv8n 默认的 CIoU (complete *IoU*) 损失函数, 以确保预测框与真实框在位置、尺度及宽高比上的匹配, 其定义如下:

$$L_{\mathrm{CIoU}}=1-IoU+\frac{\rho^2}{c^2}+\alpha v \tag{16}$$

其中, IoU 为预测框与真实框之间的交并比, 用于衡量两者重叠程度; ρ 表示预测框与真实框中心点之间的欧氏距离; c 表示同时包围预测框与真实框的最小外接矩形的对角线长度; α 和 v 用于量化预测框与真实框在宽高比例上的一致性.

CIoU 综合考虑了中心点距离与宽高比的一致性, 从而为边界框回归提供了更全面的优化指标. CIoU 的使用有效提升了目标定位的精度与检测的稳定性.

2 实验结果与分析

2.1 实验配置

本文实验在 Ubuntu 18.04 操作系统下执行, 显卡为 RTX 3090 (24 GB 显存), 采用深度学习框架 PyTorch 2.0.0, Python 版本为 3.8, CUDA 版本为 11.8. 根据实验观察算法的性能变化来确定超参数配置, 关键超参数配置见表 1.

表 1 关键超参数配置

名称	参数
epoch	300
batch	16
imgsz	640
lr0	0.01
weight_decay	0.000 5
momentum	0.937

2.2 实验数据集

行人检测数据集呈现出高度复杂且多变的特性, 在现实场景中, 诸如遮挡现象、杂乱背景以及光照变化等因素都会对检测结果造成影响. 因此, 本研究中的密集行人检测任务需要依赖内容丰富、能够有效模拟真实世界行人检测场景的数据集.

当前, 行人检测数据集在规模、标注精度、场景复杂性和任务覆盖上呈现出多样化与专业化的发展趋势, 但也存在一些不足. 例如 INRIA Person 数据集[25]和 JAAD 数据集[26]的规模较小. Caltech Pedestrian 数据集[27]规模较大, 但场景相对单一, 可能导致在其他环境下的泛化能力不足. COCO 数据集[28]中的 Person 子集规模较大、场景丰富, 但单张图像平均人类实例数量不足. 为客观验证改进对算法的影响, 本文选取两款复杂场景下的密集行人检测数据集 WiderPerson[29]和 CrowdHuman[30]. WiderPerson 数据集包含步行、马拉松、广场舞等多种户外特征, 其中有 236 073 个行人实例, 是现有拥挤度最高的公开数据集之一, 平均人类实例数量 29.51, 本文选取此数据集进行训练和验证, 选取 9 000 张图像, 将其按照 8:2 的比例进行训练集与验证集的划分. 在验证阶段, 本文采用 CrowdHuman 数据集进行性能评估. 该数据集以复杂的严重遮挡场景为特征, 每幅图像平均包含约 22.6 个人类实例. 为了从多个维度评估算法的泛化能力, 直观展示其在不同形态和状态下的提升效果, 同时避免单一形态目标提升对实验结果的干扰, 与 WiderPerson 数据集的二分类处理方式不同, 本文针对 CrowdHuman 数据集设置了 3 个差异化标签. 此外, 为进一步丰富数据多样性, 本文对样本进行了随机旋转、翻转、调整对比度等数据增强操作. 最终, 从中选取了 15 000 张图像, 并按 8:2 的比例

划分为训练集和验证集. 在这种标签配置下, 一个数据集即可实现多个二分类数据集的验证效果, 从而为算法性能提供了更全面而细致的评估依据.

2.3 评价指标

实验通过平均精度 (average precision, *AP*)、平均精度均值 (mean average precision, *mAP*)、每秒浮点运算次数 (FLOPs)、参数量 (parameters) 来评估算法性能, 同时通过准确率 (precision, *P*)、召回率 (recall, *R*) 观察算法的误检情况, F_1 分数综合考虑了精确率和召回率. 相关计算公式如下:

$$P = \frac{TP}{TP + FP} \tag{17}$$

$$R = \frac{TP}{TP + FN} \tag{18}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{19}$$

其中, TP、FP分别为识别正确、错误的正样本数量, FN为错误的识别负样本, 实际为正样本的数量.

$$AP = \int_0^1 P \mathrm{d}R \tag{20}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{21}$$

其中, N为类别个数, *mAP*@0.5 是交并比IoU为 0.5 的平均精度, *mAP*@0.5:0.95 是$IoU \in [0.5, 0.95]$区间阈值的平均精度.

AP 用于评估单个类别的检测性能, *mAP* 通过综合所有类别的 *AP* 来衡量算法整体性能, 是目标检测质量的关键指标. parameters 是参数量, FLOPs 反映了算法在执行浮点运算时的计算量, 两者是衡量算法复杂度的重要因素. F_1 分数综合精确率和召回率, 平衡评估算法的精准识别与全面覆盖能力, 尤其适用于不平衡数据场景.

2.4 对比实验

2.4.1 不同算法对比

为验证算法的性能, 选取当前不同算法在 WiderPerson 数据集下进行实验对比, 实验结果见表 2. DACD-YOLO 相比于原算法 (YOLOv8n) *mAP*@0.5 和 *mAP*@0.5:0.95 分别提升 2.3% 和 2.2%, FLOPs 和 parameters 略有增加, 分别增加 4.4G 和 0.8M, 对计算成本的影响较小. 相比 YOLOv8s *mAP*@50 仅增加 0.2%, 但 parameters 和 FLOPs 分别降低 7.3M 和 16.1G. 相比 YOLOv10s, 改进算法的计算量也明显低于 YOLOv10s, 其中 parameters 不足 YOLOv10s 的一半, 充分表明该改进方法在性能与计算量之间达成了良好平衡. DACD-YOLO 在所示表中精度最高, 相比于 YOLO-Worldv2n (ultralytics 框架集成版本)、Gold-YOLOn、YOLOv5n、YOLOv7tiny, *mAP*@0.5 分别高出 2.1%、2.6%、4.4% 和 0.6%. 综上所述, DACD-YOLO 在密集行人检测场景中兼顾轻量化和精度, 有着良好的性能表现.

表 2 不同算法对比

算法	*mAP*@0.5 (%)	*mAP*@0.5:0.95 (%)	FLOPs (G)	参数量 (M)
SSD[31]	68.3	41.5	87.7	26.3
YOLOv5n	73.4	44.8	**4.2**	**1.8**
YOLOv5s	75.5	46.9	15.8	7.0
YOLOv7tiny[32]	77.2	46.1	13.2	6.0
YOLOv8n	75.5	48.6	8.2	3.0
YOLOv8s	77.6	50.4	28.7	11.1
RT-DETR-l[33]	73.9	44.3	56.9	23.8
Gold-YOLOn[34]	75.2	48.7	12.1	5.6
YOLO-Worldv2n[35]	75.7	48.9	9.4	3.5
YOLOv10n[36]	74.2	48.3	8.4	2.7
YOLOv10s	77.2	50.6	24.8	8.1
DACD-YOLO	**77.8**	**50.8**	12.6	3.8

2.4.2 注意力机制对比

注意力机制可以在处理信息时有选择性关注输入数据, 更有效地捕捉重要特征. 本文在 YOLOv8n 原算法的基础上引入 HetConv, 通过对比几种常见的注意力机制, 分析它们对 YOLOv8n 算法性能的影响. 实验采用 WiderPerson 数据集作为测试基准, 通过表 3 实验结果分析, CoT 和 CDA 注意力机制结合了卷积操作和自注意力机制后, *mAP*@0.5 提升最大为 0.3%, 但 CoT 的 FLOPs 和 parameters 增加也较大, 分别为 3.7G 和 1.9M. 添加 CA 注意力后, 参数量略微提升, 但 *mAP*@0.5:0.95 分别提升了 0.2% 和 0.1%. 综合表现最好的是 CDA, 在 FLOPs 和 parameters 变化不大的情况下, *mAP*@0.5 和 *mAP*@0.5:0.95 分别提升 0.3% 和 0.2%, 这一结果主要归因于 CDA 注意力机制结合深、浅层双通道设计, 能够同时捕捉长程依赖和空间信息, 有效缓解浅层特征丢失.

2.4.3 通道数重设

通道数对算法的性能和效率有显著影响. 增加通道数可以捕获更丰富的特征, 但可能导致冗余特征的

学习, 从而加剧过拟合, 并增加计算量和内存开销. 相反, 减少通道数可以降低计算成本和资源占用, 但可能削弱特征提取能力. 通道数的选择应综合考虑任务复杂度、数据集规模和硬件资源, 寻求性能与效率的最佳平衡. 本文重设 SPPF 和 ADPEVC 通道数, 表 4 数据是在 WiderPerson 数据集下测试的结果, 由表 4 可知, 两者通道数为 128 时, FLOPs 和参数量最小, 分别为 14.3G 和 4.7M, 但 *mAP*@0.5 和 *mAP*@0.5:0.95 最低, 分别为 77.6% 和 50.5%, 当通道数为 256、512 和 1 024 时, 三者的 *mAP*@0.5、*mAP*@0.5:0.95 相近, 但当通道数 256 时计算量更小, 此时的 FLOPs 和参数量分别为 14.4G 和 4.8M. 为综合算法的性能与轻量化, 本文选取 256 作为 SPPF 和 ADPEVC 的通道数.

表 3 注意力机制对比

注意力	*mAP*@0.5 (%)	*mAP*@0.5:0.95 (%)	FLOPs (G)	参数量 (M)
无	75.0	48.2	**6.5**	**2.3**
+ SE	75.1	48.3	6.5	2.5
+ ECA[37]	75.1	48.2	6.5	2.5
+ CBMA	75.2	48.2	6.6	2.5
+ NAM[38]	75.1	48.3	6.5	2.3
+ GAM[39]	75.2	48.4	9.3	3.6
+ CoT[40]	75.3	48.4	10.2	4.2
+ CA	75.2	48.3	6.5	2.4
+ CDA	**75.3**	**48.4**	6.6	2.4

表 4 SPPF 和 ADPEVC 不同通道数对比

YOLOv8n	辅助分支	ADPEVC	SPPF和ADPEVC通道数	*mAP*@0.5 (%)	*mAP*@0.5:0.95 (%)	FLOPs (G)	参数量 (M)
√	√	√	128	77.6	50.5	**14.3**	**4.7**
√	√	√	256	**77.7**	50.7	14.4	4.8
√	√	√	512	77.7	50.7	14.8	4.9
√	√	√	1 024	77.7	**50.8**	16.3	5.5

2.4.4 CrowdHuma 数据集验证

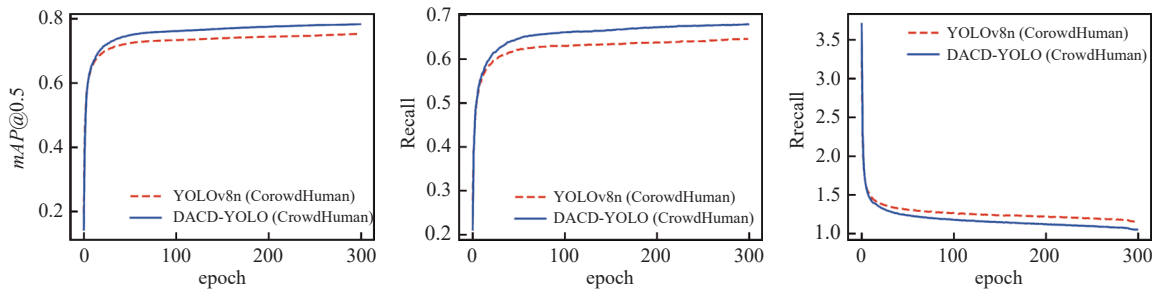
为评估所提方法的通用性和稳健性, 本文选取了 CrowdHuman 数据集进行验证, 检测结果如表 5 所示. 不同于 WiderPerson 数据集, CrowdHuman 数据集设 3 类标签: 头部 (head)、可见身体部分 (visible body) 和全身 (full body). 当存在部分信息遮挡时, 算法需要对 full body 标签进行遮挡信息的推测. 其中 *mAP* 表示总体平局精度均值, *P*、*R* 分别表示 3 种标签总体的平均准确率、平均召回率. 由表 5 可以看出, 改进算法在 CrowdHuman 数据集上精度表现较好, 其中 *mAP*@0.5 相比 YOLOv5n、YOLOv5s、YOLOv7tiny、YOLOv8n、YOLOv10n 分别提升 8.2%、4.1%、1.4%、3.5%、3.3%, 与原算法 (YOLOv8n) 相比, *AP*@0.5 (visible body) 提升 4.0%, *mAP*@0.5:0.95 提升 4.6%, *P* 和 *R* 分别提升 1.2% 和 3.8%, F_1 提升 2.8%, FLOPs 升高 4.4G. 改进算法与 YOLOv8s 相比各项精度相近, 但 FLOPs 远低于 YOLOv8s, 两者相差 16.1G, 综上所述, 改进算法相比原版 YOLOv8n 各精度指标有明显提升, 在所示算法中综合表现最优.

如图 8(a)、(b) 所示, 在 CrowdHuman 数据集上, DACD-YOLO 相比 YOLOv8n 在平均精度和召回率方面均显示出显著的提升. 从图 8(c) 中可观, DACD-YOLO 在训练过程中展示出更优的收敛效果和更低的损失, 这表明改进后的算法能够在训练过程中更精准地匹配预测边界矩与目标边界矩, 有效提升了检测效能和训练稳定性.

由此可见, DACD-YOLO 在保持高精度的同时, 显著降低了计算复杂度, 有着良好的性能表现.

表 5 CrowdHuman 数据集结果对比

算法	*AP*@0.5 (%)			*mAP*@0.5 (%)	*mAP*@0.5:0.95 (%)	*P* (%)	*R* (%)	F_1	FLOPs (G)
	head	visible body	full body						
YOLOv5n	71.0	67.0	73.7	70.6	39.7	81.1	61.6	70.0	**4.2**
YOLOv5s	75.7	71.7	76.8	74.7	44.3	83.2	66.6	74.0	15.8
YOLOv7tiny	76.9	73.7	81.6	77.4	44.0	84.3	**70.2**	**76.6**	13.2
YOLOv10n	76.6	72.1	77.9	75.5	47.7	82.4	65.6	73.7	8.4
YOLO-Worldv2n	75.6	70.2	79.0	74.9	47.6	84.2	64.3	72.9	9.2
YOLOv8n	75.4	70.5	79.9	75.3	47.6	84.1	64.6	73.1	8.2
YOLOv8s	79.2	**74.7**	**82.8**	**78.9**	52.1	**85.4**	68.8	76.2	28.7
DACD-YOLO	**79.2**	74.5	82.7	78.8	**52.2**	85.3	68.4	75.9	12.6



(a) *mAP*@50对比图 (b) 召回率对比图 (c) loss对比图

图 8 对比曲线图

2.5 消融实验

为了验证本文所提出的模块对算法改进的效果, 选取 WiderPerson 和 CrowdHuman 两个数据集进行消融实验评估, 用“√”表示添加模块. 表 6 为 WiderPerson 消融实验结果, 由表 6 可知, 使用双分支网络后, *mAP*@0.5 和 *mAP*@0.5:0.95 有巨大提升, 相比原算法分别提升 2.0 % 和 1.4%, FLOPs 增加 6.3G, *P* 和 *R* 分别提升 0.4% 和 2.3%. 引入优化后的 ADPEVC 模块后因重设 SPPF 和 ADPEVC 通道数, 在计算量减少的情况下, *mAP*@0.5 增加 0.2%. 为继续降低算法计算量, 将原算法的 C2f 替换为 C2f-HC, 从表中可以看到, 该模块的引入使算法计算量有所降低, 其中 FLOPs 降低 3.5G, 参数量降低 1M, 但 *mAP*@0.5 仅降低 0.2%, 有效降低算法复杂度. 最后引入 DG-Detect 检测头, 增加少量的参数量和 FLOPs 提升了 0.3% 的 *mAP*@0.5. 表 7 为在 CrowdHuman 数据集下的消融实验结果, 使用双分支网络后, *mAP*@0.5 提升 3.2%, 加入 ADPEVC 并重设通道数后, 计算量有所降低, 但 *mAP*@0.5 提升 0.3%, 使用轻量化模块 C2f-HC 后, *mAP*@0.5 降低 0.3%, 最后引入 DG-Detect 检测头, FLOPs 和参数量分别增加 1.8G 和 0.1M, *mAP*@0.5 和 *mAP*@0.5:0.95 分别提升 0.3% 和 0.4%.

表 6 消融实验结果 (WiderPerson 数据集)

YOLOv8n	辅助分支	ADPEVC	C2f-HC	DG-Detect	*mAP*@0.5 (%)	*mAP*@0.5:0.95 (%)	*P* (%)	*R* (%)	FLOPs (G)	参数量 (M)
√	—	—	—	—	75.5	48.6	81.6	63.6	**8.2**	**3.0**
√	√	—	—	—	77.5	50.0	82.0	65.9	14.5	4.9
√	√	√	—	—	77.7	50.5	81.9	**66.7**	14.3	4.7
√	√	√	√	—	77.5	50.2	82.2	66.2	10.8	3.7
√	√	√	√	√	**77.8**	**50.8**	**82.4**	66.6	12.6	3.8

表 7 消融实验结果 (CrowdHuman 数据集)

YOLOv8n	辅助分支	ADPEVC	C2f-HC	DG-Detect	*mAP*@0.5 (%)	*mAP*@0.5:0.95 (%)	*P* (%)	*R* (%)	FLOPs (G)	参数量 (M)
√	—	—	—	—	75.3	47.6	84.1	64.6	**8.2**	**3.0**
√	√	—	—	—	78.5	51.6	84.9	68.3	14.5	4.9
√	√	√	—	—	78.8	52.1	85.0	**68.5**	14.3	4.7
√	√	√	√		78.5	51.8	85.0	68.1	10.8	3.7
√	√	√	√	√	**78.8**	**52.2**	**85.3**	68.4	12.6	3.8

由此可见, 改进算法在行人检测中兼顾轻量与性能, 相比原算法精度显著提升, 验证了改进方案的可行性.

2.6 检测效果对比

为验证算法的准确性, 随机选取部分未经训练的图片进行推理, 图 9 展示了使用 WiderPerson 训练权重的检测结果, 其中 YOLOv8n 的漏检情况已在图中用圆形圈出, 其余对比结果用方框显示. 从图 9(a) 可见, 在无遮挡等良好条件下, 原算法和改进算法均能准确检测出行人, 但改进算法的精度明显高于原算法. 图 9(b) 中, 在光线昏暗且人员严重遮挡的情况下, 原算法左后方有部分人物存在漏检情况, 而改进算法能够成功检测出. 图 9(c) 为地铁中行人检测结果, 改进算法也成功检测出了左侧被遮挡的行人. 在检测小目标群体时, 改进算法同样表现出色. 例如在图 9(d) 中, 改进算法成功检测出原算法漏检的右侧黄色轿车前方的行人. 与其

他 3 个算法对比, DACD-YOLO 算法表现也较为优秀, 例如图 9(b) 左后方的 YOLOv8n 漏检的行人, YOLOv8s 和 YOLO-Worldv2n 虽然也检测出, 但数量略少于改进算法. 由此可见, 改进算法对于遮挡及小目标行人有很好的检测效果, 检测精度明显提升, 漏检率明显降低, 是处理复杂场景下密集行人检测的有效解决方案.

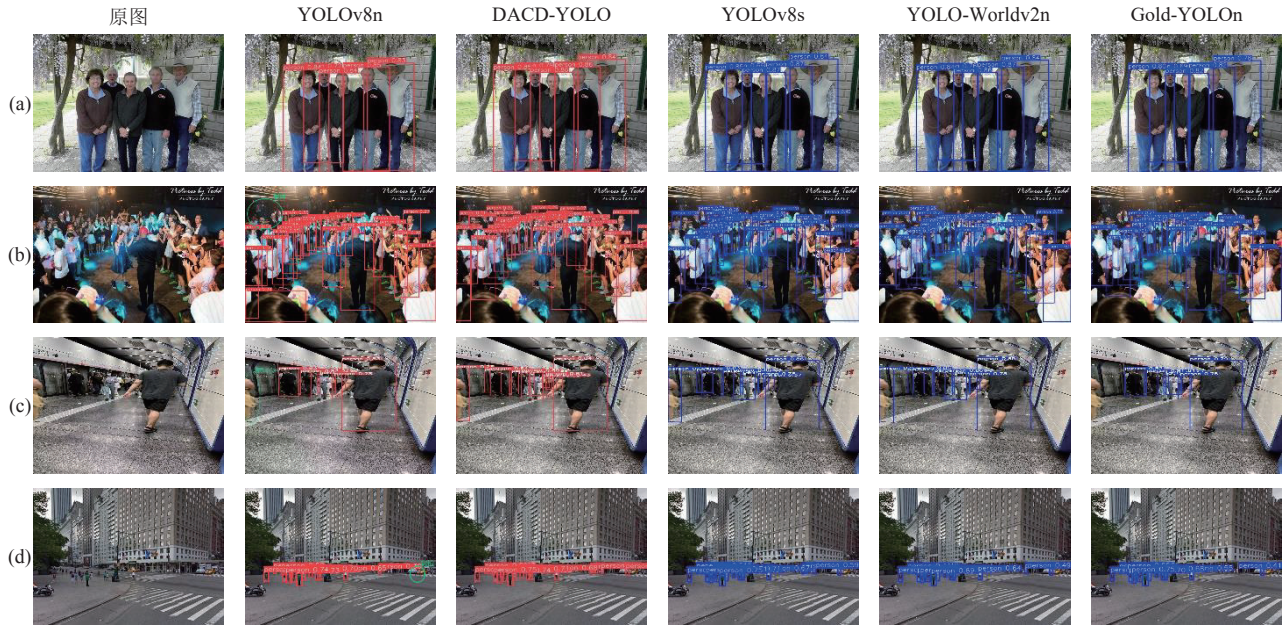


图 9 检测效果对比图

3 结论与展望

本文针对复杂场景下的密集行人检测, 对 YOLOv8n 算法进行改进, 提出一种融合改进注意力的自适应双分支密集行人检测算法, 通过动态加权特征融合和深度可分离卷积, 有效缓解信息丢失. 为缓解原算法层内交互不足的问题, 提出自适应视觉中心模块, 有效减少局部区域密集分布中的漏检, 并通过通道数重设以平衡精度与计算量. 为降低计算复杂度, 设计轻量化 C2f-HC 模块, 替换原 C2f 模块, 结合异构卷积和 CDA 注意力机制, 实现深浅层特征的互补, 提升空间关系和细粒度特征的捕捉能力. 为提高对小目标及遮挡目标的多尺度检测精度, 引入具有不同膨胀率的膨胀卷积检测头, 以更有效地捕获多尺度上下文信息. 实验证明, DACD-YOLO 在 WiderPerson 和 CrowdHuman 数据集上的 *mAP*@0.5 分别提升了 2.3% 和 3.5%, 相较于 YOLOv8n 精度取得显著提升, DACD-YOLO 与 YOLOv8s 的精度相近, 但前者的计算量远小于后者, 充分体现了改进算法在复杂场景下的优越性.

尽管改进的算法在测试集上表现出色, 但实际应用中的复杂性和多样性可能带来新的挑战. 未来研究将着重提升算法在噪声、光照变化和动态背景中的鲁棒性, 并通过剪枝、稀疏化等轻量化技术, 以实现性能与计算量的平衡与稳定.

参考文献

1 Huang LC, Wang ZW, Fu XB. Pedestrian detection using RetinaNet with multi-branch structure and double pooling attention mechanism. Multimedia Tools and Applications, 2024, 83(2): 6051–6075. [doi: 10.1007/s11042-023-15862-4]

2 Verma R, Ukkusuri SV. Crosswalk detection from satellite imagery for pedestrian network completion. Transportation Research Record, 2024, 2678(7): 845–856. [doi: 10.1177/03611981231210545]

3 娄翔飞, 吕文涛, 叶冬, 等. 基于计算机视觉的行人检测方法研究进展. 浙江理工大学学报, 2023, 49(3): 318–330.

4 Gong LX, Huang X, Chen JL, *et al*. Reparameterized dilated architecture: A wider field of view for pedestrian detection. Applied Intelligence, 2024, 54(2): 1525–1544. [doi: 10.1007/s10489-023-05255-3]

5 张嘉辉, 赵威, 王子琛, 等. 基于检测和重识别的无人机行人跟踪算法. 北京航空航天大学学报, 2024, 50(8): 2538–2546. [doi: 10.13700/j.bh.1001-5965.2022.0675]

6 Huan RH, Zhang J, Xie CJ, *et al*. MLFFCSP: A new anti-

occlusion pedestrian detection network with multi-level feature fusion for small targets. Multimedia Tools and Applications, 2023, 82(19): 29405–29430. [doi: 10.1007/s11042-023-14721-6]

7 He YZ, He N, Yu HG, *et al*. From macro to micro: Rethinking multi-scale pedestrian detection. Multimedia Systems, 2023, 29(3): 1417–1429. [doi: 10.1007/s00530-023-01058-1]

8 周大可, 宋荣, 杨欣. 结合双重注意力机制的遮挡感知行人检测. 哈尔滨工业大学学报, 2021, 53(9): 156–163. [doi: 10.11918/201904144]

9 Liu QL, Ye HX, Wang SM, *et al*. YOLOv8-CB: Dense pedestrian detection algorithm based on in-vehicle camera. Electronics, 2024, 13(1): 236. [doi: 10.3390/electronics13010236]

10 陈秀锋, 王成鑫, 吴阅晨, 等. 改进 YOLOv5s 算法的车辆目标实时检测方法. 哈尔滨理工大学学报, 2024, 29(1): 107–114. [doi: 10.15938/j.jhust.2024.01.012]

11 Zhang Y, Zhang SF, Xin DR, *et al*. A small target pedestrian detection model based on autonomous driving. Journal of Advanced Transportation, 2023, 2023(1): 5349965. [doi: 10.1155/2023/5349965]

12 Dong XD, Yan S, Duan CQ. A lightweight vehicles detection network model based on YOLOv5. Engineering Applications of Artificial Intelligence, 2022, 113: 104914. [doi: 10.1016/j.engappai.2022.104914]

13 Shen X, Wei XK. A real-time subway driver action sensoring and detection based on lightweight ShuffleNetV2 network. Sensors, 2023, 23(23): 9503. [doi: 10.3390/s23239503]

14 Wang CY, Yeh IH, Liao HYM. YOLOv9: Learning what you want to learn using programmable gradient information. Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2024. 1–21.

15 谢国波, 林松泽, 林志毅, 等. 基于改进 YOLOv7-tiny 的道路病害检测算法. 图学学报, 2024, 45(5): 987–997.

16 Navon A, Shamsian A, Achituve I, *et al*. Equivariant architectures for learning in deep weight spaces. Proceedings of the 40th International Conference on Machine Learning. Honolulu: JMLR.org, 2023. 1073.

17 Quan Y, Zhang D, Zhang LY, *et al*. Centralized feature pyramid for object detection. IEEE Transactions on Image Processing, 2023, 32: 4341–4354. [doi: 10.1109/TIP.2023.3297408]

18 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141. [doi: 10.1109/CVPR.2018.00745]

19 Wang YF, Wang W, Li Y, *et al*. An attention mechanism module with spatial perception and channel information interaction. Complex & Intelligent Systems, 2024, 10(4): 5427–5444. [doi: 10.1007/s40747-024-01445-9]

20 Hou QB, Zhou DQ, Feng JS. Coordinate attention for efficient mobile network design. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 13708–13717. [doi: 10.1109/CVPR46437.2021.01350]

21 Li MJ, Chen S, Sun C, *et al*. An improved lightweight dense pedestrian detection algorithm. Applied Sciences, 2023, 13(15): 8757. [doi: 10.3390/app13158757]

22 Singh P, Verma VK, Rai P, *et al*. HetConv: Heterogeneous kernel-based convolutions for deep CNNs. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4830–4839.

23 Mehta S, Rastegari M, Shapiro L, *et al*. ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9182–9192.

24 Wei HR, Liu X, Xu SC, *et al*. DWRSeg: Rethinking efficient acquisition of multi-scale contextual information for real-time semantic segmentation. arXiv:2212.01173, 2023.

25 Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 886–893.

26 Rasouli A, Kotseruba I, Kunic T, *et al*. PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 6261–6270. [doi: 10.1109/ICCV.2019.00636]

27 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]

28 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.

29 Zhang SF, Xie YL, Wan J, *et al*. Widerperson: A diverse dataset for dense pedestrian detection in the wild. IEEE Transactions on Multimedia, 2020, 22(2): 380–393. [doi: 10.1109/TMM.2019.2929005]

30 Shao S, Zhao ZJ, Li BX, *et al*. CrowdHuman: A benchmark for detecting human in a crowd. arXiv:1805.00123, 2018.

31 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 21–37.

32 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. [doi: 10.1109/CVPR52729.2023.00721]

33 Zhao Y, Lv WY, Xu SL, *et al*. DETRs beat YOLOs on real-time object detection. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 16965–16974.

34 Wang CC, He W, Nie Y, *et al*. Gold-YOLO: Efficient object detector via gather-and-distribute mechanism. Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2024. 2224.

35 Cheng TH, Song L, Ge YX, *et al*. YOLO-World: Real-time open-vocabulary object detection. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 16901–16911.

36 Wang A, Chen H, Liu LH, *et al*. YOLOv10: Real-time end-to-end object detection. Proceedings of the 38th Conference on Neural Information Processing Systems. Vancouver: NeurIPS, 2024. 14458.

37 Wang QL, Wu BG, Zhu PF, *et al*. ECA-Net: Efficient channel attention for deep convolutional neural networks. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 11531–11539.

38 Liu YC, Shao ZR, Teng YY, *et al*. NAM: Normalization-based attention module. arXiv:2111.12419, 2021.

39 Cheng JF, Zhong YR, Dai YC, *et al*. Global attention mechanism: Retain information with a global perspective. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.

40 Li YH, Yao T, Pan YW, *et al*. Contextual Transformer networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 1489–1500. [doi: 10.1109/TPAMI.2022.3164083]

(校对责编: 张重毅)