DOI:10.19651/j.cnki.emt.2518797

基于改进 YOLO11n 的轻量级密集行人检测算法*

黄思禄 钟 寒
(中国人民公安大学信息网络安全学院 北京 100038)

摘 要: 在密集行人场景中,由于遮挡严重、小目标多、尺度变化大,且环境复杂,容易造成行人漏检、错检及定位不准等问题。针对以上挑战,本文提出了一种轻量化的密集行人检测算法 DC-YOLO。该算法基于 YOLO11n,在主干网络上提出了轻量级特征提取网络 EfficientNetV2S-S3,提高模型对小目标和多尺度目标的特征提取能力,降低模型参数量和计算成本;在颈部网络上提出了 P-LightNeck 模块,进一步提高了对小目标的特征融合能力,实现检测精度与效率的协同优化;引入 RepNCSPELAN4 卷积模块,通过多尺度卷积和重参数化技术,强化遮挡目标的特征提取能力,并提高推理效率;设计了动态多尺度协同注意力模块 DynaMSAttn,增强模型对不同尺度目标和复杂环境的适应性。实验结果显示,与 YOLO11n 相比,DC-YOLO 算法在 CrowdHuman 数据集上,mAP@0.5、mAP@0.5-0.95 分别提升 4.7%和 4.5%,同时参数量降低了 46.2%,通过对比实验和消融实验,验证了 DC-YOLO 算法在密集行人检测任务中具有优秀的检测效果和鲁棒性。

关键词: 密集行人检测;YOLO11;轻量化模型;多尺度卷积;动态注意力机制
中图分类号: TP391.4;TN919.8 **文献标识码:** A **国家标准学科分类代码:** 520.6040

Improved YOLO11n-based lightweight dense pedestrian detection algorithm

Huang Silu Zhong Han
(School of Information Network Security, People's Public Security University of China, Beijing 100038, China)

Abstract: In dense pedestrian scenes, severe occlusions, numerous small targets, significant scale variations, and complex environments often lead to missed detections, false detections, and inaccurate localization of pedestrians. To address these challenges, this paper proposes a lightweight dense pedestrian detection algorithm DC-YOLO. The algorithm is based on YOLO11n. In the backbone network, a lightweight feature extraction network, EfficientNetV2S-S3, is proposed to enhance the model's feature extraction capability for small and multi-scale targets while reducing model parameters and computational costs. In the neck network, the P-LightNeck module is proposed to further improve the feature fusion capability for small targets, achieving collaborative optimization of detection accuracy and efficiency. The RepNCSPELAN4 convolutional module is introduced to strengthen the feature extraction capability for occluded targets through multi-scale convolution and re-parameterization techniques, while improving inference efficiency. A dynamic multi-scale collaborative attention module, DynaMSAttn, is designed to enhance the model's adaptability to targets of varying scales and complex environments. Experimental results show that, compared to YOLO11n, the DC-YOLO algorithm achieves improvements of 4.7% in mAP@0.5 and 4.5% in mAP@0.5-0.95 on the CrowdHuman dataset, while reducing the parameter count by 46.2%. Comparative experiments and ablation experiments verify that the DC-YOLO algorithm exhibits excellent detection performance and robustness in dense pedestrian detection tasks.

Keywords: dense pedestrian detection;YOLO11;lightweight model;multi-scale convolution;dynamic attention mechanism

0 引 言

随着经济社会发展,人类活动日益频繁,公共区域如车站、景区、校园、商场和体育中心等行人密集场景的安全管理对政府,尤其是公安机关来说,是一个巨大的挑战。2024 年 11 月 11 日,珠海驾车冲撞行人案导致 35 人死亡、43 人

收稿日期:2025-05-13
* 基金项目:高等学校学科创新引智基地项目(B20087)、中国人民公安大学"双一流"创新研究项目(2023SYL07)资助

受伤。2025 年 1 月 29 日，印度北方邦踩踏事件导致超过 30 人死亡。为防止此类严重事件的发生，加强公共场所安全防范，进行密集行人检测显得尤为重要。除安防领域外，行人检测在自动驾驶、辅助驾驶和智能机器人应用等系统中也发挥着重要作用。行人检测的方法主要有基于传统机器学习和基于深度学习两种检测方法。基于传统机器学习的行人检测算法，主要是通过设计手工特征，结合级联分类器的方法。

随着深度学习的快速发展，基于深度学习技术的行人检测相关研究取得了显著进展，并且比传统机器学习方法大大提高了精度。基于深度学习的行人检测算法可以分为两阶段算法和单阶段算法。两阶段行人检测算法，主要通过分步优化行人检测任务，首先生成候选区域，然后对候选区域进行精细化分类和定位，R-CNN、Fast R-CNN[1]、Faster R-CNN[2]等系列算法是其中的主要代表。其中，Girshick 等[1]提出的目标检测模型 R-CNN，开创性地将深度卷积神经网络与自下而上的区域提议相结合，用于实现目标的准确定位与分类。单阶段目标检测算法与两阶段算法相比，核心区别在于取消了候选机制，借助网络“端到端”架构，直接从输入图像映射至目标类别与边界框预测结果，提高了检测速度。其中，Liu 等[3]提出将行人检测视为高语义特征检测任务，设计单阶段的中心和尺度预测方法，降低了训练难度，提升了检测精度。Han 等[4]通过解耦网络和记忆增强机制有效结合行人检测与重识别任务，显著提升了检测准确性和鲁棒性。

2016 年，Redmon 等[5]提出单阶段目标检测的经典算法 YOLO（you only look once），将目标检测转化为回归问题，通过单个神经网络直接从图像像素预测边界框和类别概率，实现端到端训练与实时检测，尤其适用于行人检测等对实时性要求高的场景。近几年，YOLO 算法不断迭代升级[6]，涌现出各种对 YOLO 算法改进的行人检测模型。梁天添等[7]提出改进检测头和引入高效多尺度注意力机制等方法优化 YOLOv8s，解决了恶劣天气条件下行人检测目标与背景难以区分的问题。胡伟超等[8]通过提升颈部网络的多尺度特征融合能力等方法改进 YOLOv8n，提高了复杂环境中对多尺度行人的检测能力。Liu 等[9]通过引入视觉注意力模块（VAM）和特征重建模块（FRM）改进 YOLOv7，解决复杂交通场景下行人的检测精度低和鲁棒性不足等问题。Zhang 等[10]通过优化 YOLOv8S 的 C2f 模块，利用深度可分离卷积等方法，减少了模型参数，提升了行人检测模型在边缘设备的推理速度。

尽管上述行人检测算法都取得了不错的进展，但密集行人检测由于严重遮挡、尺度变化大、密集边界预测模糊[11]等问题，仍然是现实世界复杂场景中最具挑战性的任务之一[12]。因此，本文针对这些问题，提出了基于 YOLO11n 的轻量化密集行人检测算法（DenseCrowd-YOLO，DC-YOLO）。本文主要工作如下：

1）针对行人检测算法实时性的要求，以轻量化主干网络 EfficientNetV2S 为基础，根据行人检测任务小目标较多的实际，提出了 EfficientNetV2S-S3 主干，通过减少主干下采样次数，避免小目标在深层特征图中信息丢失，在增强主干网络特征提取能力的同时，大幅减少模型参数，提升推理和训练效率。

2）针对小目标漏检、误检的问题，在颈部网络上提出了 P-LightNeck 模块，进一步提高了模型对小目标的特征融合能力，实现检测精度与效率的协同优化。

3）根据行人检测任务中目标尺寸变化大、背景复杂的困扰，引入了 RepNCSPELAN4 模块，有效聚合不同尺度和层次的特征，提高算法性能。

4）为解决行人相互遮挡的难题，设计了动态多尺度协同注意力模块（dynamic multi-scale collaborative attention，DynaMSAttn），提升了算法对遮挡目标和复杂背景的应对能力。

1　YOLO11 算法

YOLO11 拥有出色的检测精度、速度和效率，是目前 YOLO 系列中最适合做实时检测任务的模型。YOLO11 的主要架构和 YOLOv8 类似，主要由主干、颈部和头部 3 部分组成，其在 YOLOv8 算法基础上，采用了更高效的 C3k2 模块代替原先的 C2f 模块，提高了计算的效率。同时在主干最后部分添加了融合空间注意力机制的 C2PSA 模块，增强模型聚焦能力。此外，如图 1(a)所示还在检测头中加入了深度可分离卷积 DWConv[13]，有效降低了计算冗余并提高了效率。YOLO11 支持边缘设备、云平台和移动端等多种环境，适应性极强，并提供了 n、s、m、l、x 五个版本，其中 n 是最轻量的版本，检测速度最快。根据密集行人检测算法实时性和高效性的要求，本文选取 YOLO11n 作为基准模型。

2　DC-YOLO 算法

DC-YOLO 以 YOLO11n 为基础，聚焦密集行人检测任务中高遮挡、小目标多、尺度多变和场景复杂等挑战，通过对主干网络、颈部特征融合和注意力机制等模块进行改进，提出的轻量化的密集行人检测算法，其结构如图 1 所示。

首先提出更轻量的主干网络 EfficientNetV2S-S3 代替 YOLO11 的主干，如图 1(b)所示，通过在浅层引入 Fused-MBConv 模块，在深层引入 MBConv 模块，并减少主干下采样，捕捉更多小目标细节特征，显著降低模型复杂度和参数量，增强对小目标的特征提取能力。其次，如图 1(c)所示，在颈部网络设计了高效的特征融合模块 P-LightNeck，对主干网络输出的特征进行初步处理和融合，为后续操作提供更好的特征输入，并加入 P2 小目标特征融合网络分支，去除密集行人检测任务中冗余的 P5 大目标特征融合网络

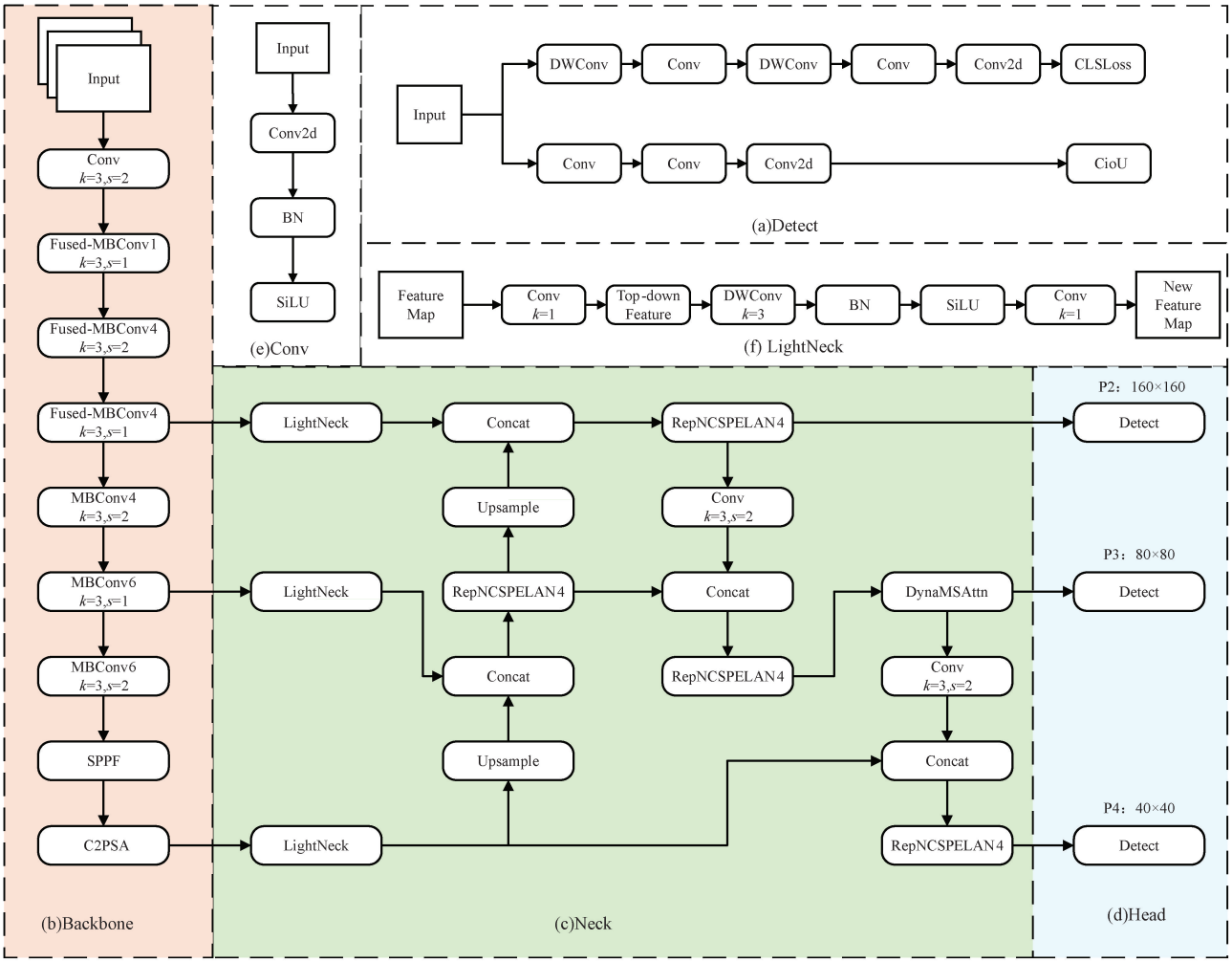


图 1 DC-YOLO 网络结构

Fig. 1 Architecture of DC-YOLO network

分支,在保证轻量化的同时,提高模型对小目标的网络特征融合能力。再次,引入 RepNCSPELAN4 模块取代 YOLO11 颈部的 C3k2 模块,巧妙融合了可重参数化、跨阶段局部网络(CSP)以及高效层聚合网络结构(ELAN)[14]的设计理念,提高模型应对复杂场景和多尺度目标的能力。最后,设计了动态多尺度协同注意力模块 DynaMSAttn,融合了多尺度空洞卷积模块[15]、动态池化通道注意力机制和空间注意力机制,增强模型对遮挡目标及复杂环境的检测能力。这些改进的有效融合,使得 DC-YOLO 在密集行人检测任务中表现优异,具有优秀的检测效果和鲁棒性。

2.1 EfficientNetV2S-S3 主干网络

EfficientNetV2 是在 2021 年提出的卷积神经网络[16],相比前代 EfficientNet[17],其通过引入 Fused-MBConv 模块和优化的网络结构设计,在训练速度和参数效率方面取得了显著提升。在密集行人检测任务中,检测的目标往往较小,部分目标甚至仅仅占图像的几个像素。在 YOLO11 中,主干网络进行多次下采样后,虽然增大了网络的感受野,但小目标可能仅仅剩下几个像素,由于网络堆叠过多层数,其特征会被弱化,产生细节失真,导致模型对小目标的检测性能不佳。据此,本文以 EfficientNetV2S 作为基础,针对密集行人检测任务进行了特定的修改和优化,减少下采样次数,保留更多的高分辨率特征信息,避免小目标在深层特征图中信息丢失,增强了主干网络对小目标的特征提取能力,提出了专门用于密集行人检测任务的 EfficientNetV2S-S3 主干网络。该主干网络在保持高精度的同时显著提升了推理和训练效率。

EfficientNetV2S-S3 采用阶段式架构,网络由 7 个阶段组成,阶段 0 通过步长为 2 的 3×3 卷积进行初始下采样和特征提取;在浅层(阶段 1～3)使用 Fused-MBConv 模块,不使用压缩-激励(squeeze-and-excitation,SE)注意力机制,其中在阶段 2 进行下采样;在深层(阶段 4～6)使用标准 MBConv 模块,并使用 SE 注意力机制[18],其中在阶段 4 和 6 进行下采样。MBConv 模块由 1×1 扩展卷积、3×3 深度可分离卷积、SE 模块和 1×1 压缩卷积组成,通过深度

可分离卷积与 SE 模块平衡参数量与特征表达能力，适用于深层阶段。Fused-MBConv 模块，将 MBConv 中的 1×1 扩展卷积与 3×3 深度可分离卷积替换为单个标准 3×3 卷积，并去掉 SE 模块，通过结构的简化提高浅层计算效率，是训练加速的关键设计（如图 2 所示）。这种混合设计充分利用了 Fused-MBConv 在浅层的计算效率和 MBConv 在深层的特征表达能力。通过上述设计，本文主干网络只进行到 16 倍下采样，得到的特征图大小为 40×40，保留了更多小目标信息；同时，为了充分利用浅层特征信息，捕捉更多小目标细节特征，将 4 倍下采样的特征图（160×160）引入特征融合网络。

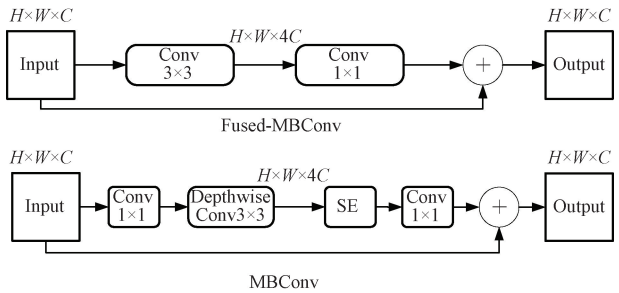


图 2　Fused-MBConv 模块和 MBConv 模块结构对比图

Fig. 2　Comparative architecture of MBConv and Fused-MBConv modules

为适配密集行人检测场景的实时性需求并优化模型计算效率，本文使用宽度因子缩减策略，将通道缩放系数调整为原始值的 75%。该设计通过压缩特征空间维度显著降低模型复杂度与推理时延，但不可避免地削弱了特征表示能力，导致检测精度出现可控性衰减。精度衰减主要是由于高频细节特征的弱化，本文通过减少下采样次数以提升特征图空间分辨率，强化了细粒度纹理信息的跨尺度传递，从而在维持实时处理能力的前提下，有效对冲了宽度因子缩减引发的表征能力损失，最终达成精度与效率的协同优化。

此外，为适配目标检测任务的特性，减少冗余计算，将原 EfficientNetV2S 网络结构中的阶段 7 的 1×1 卷积和平均池化省略，改为将阶段 2～6 输出的特征图传入颈部进行多尺度特征融合。

2.2　P-LightNeck 特征融合模块

为应对密集行人检测任务中的多尺度目标检测挑战，本文借鉴特征金字塔网络设计思想[19]，提出融合轻量化架构与小目标检测机制的 P-LightNeck 模块。

如图 1(f)所示，P-LightNeck 模块通过四阶段优化策略实现高效特征融合：1）共享投影压缩，对于每个输入特征图，采用参数共享的微型转换网络将异构特征映射至统一的通道维度，实现参数精简；2）使用深度可分离卷积（分组卷积＋点卷积）替代标准卷积，显著降低计算复杂度；3）从深层特征开始，通过上采样和横向连接逐层融合深层语义信息与浅层细节特征；4）对每个特征层级使用轻量化的 3×3 卷积进行特征平滑，提升空间感知鲁棒性。P-LightNeck 作为一个高效的特征融合模块，对主干网络输出的特征进行初步处理和融合，为后续的其他颈部操作提供更好的特征输入，这样可以在保持计算效率的同时，提高特征的表达能力，尤其是对于不同尺度目标的检测能力具有显著增益。

在密集行人检测任务中，目标尺度分布呈现显著的长尾特性。目标检测任务中的小目标从不同维度的标准认定，有不同的定义。有研究认为，基于像素大小来说，小目标边界框的长度和宽度像素应小于 32；基于图像占比来说，小目标的边界框应覆盖原始图像的 1%以下[20]。在公共安防场景中，前端设备通常部署于 4～6 m 的高度，对于广场、公园、步行街和交通主干道等位置，往往还会安装高点监控。这导致在监控中，单个行人目标普遍存在像素分辨率低（部分目标仅仅占数个像素）、空间分布密集等特点。所以，在密集行人检测任务里，要加强对小目标的关注，提高中小目标检测精度。

传统的 YOLO11 模型，提供了 P3、P4、P5 三种不同尺度的颈部融合网络分支，分别用于检测小、中、大目标。此种架构在密集行人目标检测任务中存在两大局限：一方面是底层特征（P2）的细节信息未得到有效利用；另一方面是高层特征（P5）对应的大目标检测分支在密集行人场景中冗余。

针对上述问题，DC-YOLO 通过引入更高分辨率的 P2 特征图（较 P5 分辨率提升 8 倍）加入特征融合网络，通过跨尺度特征融合强化细粒度空间信息提取。同时，为减少模型对低分辨率特征的过度依赖和节约计算资源，本文移除在密集行人检测任务中冗余的大目标 P5 特征融合网络分支，实现了模型复杂度与检测精度的平衡。实验结果证明，该方案能用少量计算成本换取较高的精度。

综上所述，P-LightNeck 模块通过精心的结构设计和优化策略，在大幅提高小目标特征提取能力的同时实现了计算效率和参数效率的平衡，是一种适用于密集行人检测任务的高效颈部网络。

2.3　RepNCSPELAN4 模块

密集行人检测任务存在目标尺寸差异大、遮挡严重等挑战，在基准模型 YOLO11 中，随着网络层数的增加，原始信息丢失的概率也增加，导致在输出阶段保存预测目标综合信息的能力降低。据此，DC-YOLO 在颈部引入 RepNCSPELAN4 模块，有效聚合不同尺度和层次的特征。如图 3(a)所示，RepNCSPELAN4 模块巧妙融合了可重参数化、跨阶段局部网络（CSP）以及高效层聚合网络结构（ELAN）的设计理念，有效提高了神经网络的推理性能和效率。该模块主要由 RepNCSP 和卷积模块构成。如图 3(b)～(d)所示，RepNCSP 的子模块 RepNBottleneck 将重参数化的卷积块（RepConv）[21]与 SiLU 激活函数集成在一

起,显著增强了模型的学习能力。模型进行可重参数化,在训练阶段使用多分支结构,在推理阶段则等效转化为单一标准卷积操作,有利于实现推理的高效化;CSPNet 将输入的特征图分为两部分,一部分直接进入后续的融合阶段,另一部分经过密集块进行进一步的特征提取,这种跨阶段部分连接的策略,可以解决梯度信息重复问题,缓解深层网络中梯度消失困扰,进一步提升计算效率和推理速度;ELAN 结构通过多路径特征处理和渐进式特征聚合机制,有效融合不同尺度的目标特征,能很好的处理密集行人检测任务中的多尺度目标和遮挡目标。

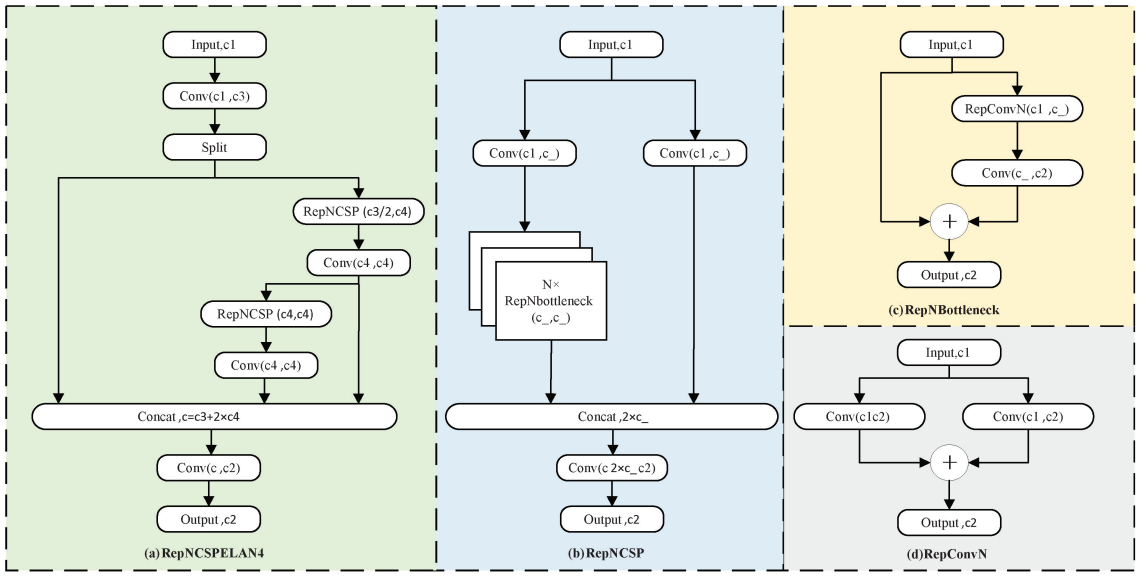


图 3 RepNCSPELAN4 模块结构

Fig. 3 Architectural of the RepNCSPELAN4 module

2.4 DynaMSAttn 注意力模块

为增强模型对不同尺度目标和复杂环境的适应性,本文设计了动态多尺度协同注意力模块 DynaMSAttn。DynaMSAttn 模块核心功能是通过动态池化策略[22]与多尺度特征增强,提升模型对遮挡目标、小目标及复杂背景的区分能力,提高定位准确度,在密集人群等复杂场景中有优异表现。DynaMSAttn 模块由 3 个关键组件构成:动态池化通道注意力机制、多尺度空洞卷积模块和空间注意力机制,其整体结构设计如图 4 所示。

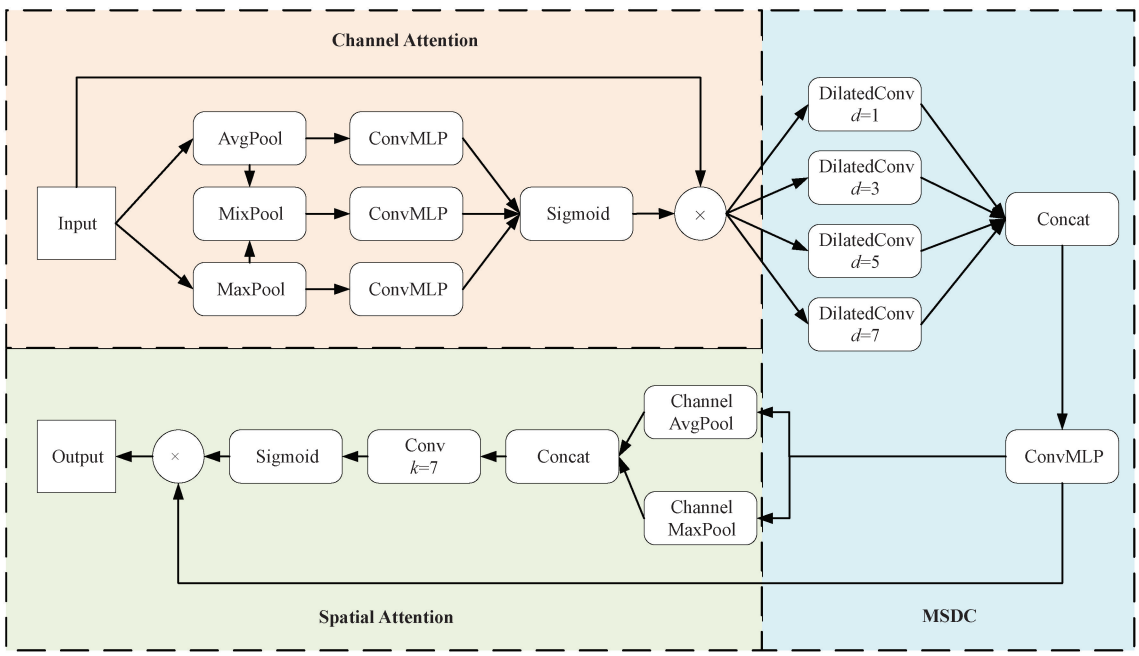


图 4 DynaMSAttn 模块结构图

Fig. 4 Architecture of the DynaMSAttn module

首先，通道注意力机制采用创新的动态池化策略，结合了平均池化利于提取背景和整体特征以及最大池化利于定位目标边界的优势，并引入自适应调整的混合池化策略，用来自适应捕捉目标的关键信息。然后通过多分支池化的特征生成通道注意力权重，增强关键通道响应。动态池化策略，可以增强对密集行人检测任务中被遮挡目标和小目标的通道敏感度。对于给定的输入特征 $X \in R^{C\times H\times W}$，动态池化可以表示为：

$$F_{avg} = \mathrm{Avg}Pool(X) \in R^{C\times 1\times 1} \tag{1}$$

$$F_{max} = \mathrm{Max}Pool(X) \in R^{C\times 1\times 1} \tag{2}$$

$$F_{mix} = \alpha \cdot F_{avg} + (1-\alpha) \cdot F_{max} \tag{3}$$

式中：F_{avg} 是平均池化，用来获取特征图每个通道的全局信息；F_{max} 是最大池化，用来获取特征图每个通道的显著特征；F_{mix} 是动态混合池化，$\alpha \in [0,1]$ 是混合参数，根据训练数据自动调整参数，以找到平均池化与最大池化的最佳混合比例。

通道注意力的计算过程可以表示为：

$$M_{avg} = \sigma(W_2 \cdot \delta(W_1 \cdot F_{avg})) \tag{4}$$

$$M_{max} = \sigma(W_2 \cdot \delta(W_1 \cdot F_{max})) \tag{5}$$

$$M_{mix} = \sigma(W_2 \cdot \delta(W_1 \cdot F_{mix})) \tag{6}$$

$$M_c = \sigma(M_{avg} + M_{max} + M_{mix}) \tag{7}$$

$$X' = X \odot M_c \tag{8}$$

式中：$W_1 \in R^{\frac{C}{r}\times C}$ 和 $W_2 \in R^{C\times\frac{C}{r}}$ 是降维和升维的权重矩阵；R 是降维比率；δ 表示 ReLU 激活函数；σ 表示 Sigmoid 激活函数；M_c 表示通道注意力权重；$\odot$ 表示逐元素乘法（通道注意力图与特征图的相乘）。

其次，通过多尺度空洞卷积模块（MSDC），扩大感受野，提高不同尺度的上下文信息捕获能力，有效应对密集行人检测任务中目标尺寸跨度大的问题。多尺度空洞卷积 MSDC 可以表示为：

$$F_k = \delta(BN(Conv_{d=k}(X'))), k \in \{1,3,5,7\} \tag{9}$$

$$X'' = MLP(Concat(F_1, F_3, F_5, F_7)) \tag{10}$$

式中：$Conv_{d=k}$ 表示膨胀率为 k 的 3×3 卷积操作；BN 表示 Batch Normalization（批归一化）；$Concat$ 表示沿通道维度拼接；MLP 表示 ConvMlp 操作，通过 1×1 卷积实现通道压缩。

最后，通过动态生成空间权重图，加强特征图中的关键区域关注度，抑制背景噪声，提升密集行人检测中的被遮挡目标和小目标定位精度。空间注意力的计算过程可以表示为：

$$F^s_{avg} = \frac{1}{C}\sum_{i=1}^{C} X''_i \tag{11}$$

$$F^s_{max} = \max_{i\in\{1,2,\cdots,c\}} X''_i \tag{12}$$

$$M_s = \sigma(Conv_{7\times 7}(Concat(F^s_{avg}, F^s_{max}))) \tag{13}$$

$$X_{out} = X'' \odot M_s \tag{14}$$

式中：X''_i 表示输入特征图的第 i 个通道；M_s 表示空间注意力权重；X_{out} 为 DynaMSAttn 模块的最终输出。

综合上述过程，DynaMSAttn 模块的整体数学表达式为：

$$DynaMSAttn(X) = MSDC(X \odot M_c) \odot M_s \tag{15}$$

式中：MSDC 是多尺度空洞卷积模块。

3　实验过程和结果分析

3.1　实验数据集

本文使用旷世科技于 2018 年发布的 CrowdHuman 数据集[23]，这是评估密集行人检测算法的基准数据集，是专为解决密集行人检测任务的挑战而设计的。该数据集共有训练图像 15 000 张、验证图像 4 370 张和测试图像 5 000 张，包含了高达 47 万个人类实例，平均每张图包含 22.6 个人，重叠率高达 50%，远超其他同类数据集。该数据集标注包含了头部边界框和人类全身边界框，其场景涵盖全球 40 余个不同城市的街道、广场和景区等，既有监控视角，又有水平视角，场景光照多样，很好的体现了现实世界的复杂场景。此外，本文还使用了 WiderPerson 数据集[24]验证本文所提出的算法在密集行人检测任务上的有效性。WiderPerson 数据集适用于野外场合的密集行人检测任务，共有训练图像 8 000 张、验证图像 1 000 张、测试图像 4 382 张，共有 399 786 个标注，即平均每张图像有 29.87 个标注，其场景和遮挡情形具有多样性，极具挑战性。

3.2　实验环境与评价指标

本文实验使用 NVIDIA RTX 4090D 显卡，显存为 24 GB，CPU 使用 18 核、60 GB 的 AMD EPYC 9754 128-Core Processor，操作系统为 ubuntu22.04，开发环境为 python 3.12.3、PyTorch 2.5.1、CUDA12.1 的深度学习框架。模型训练轮次均 200 轮，图片输入尺寸统一为 640×640，优化器采用 SGD，批量大小设置为 8，初始学习率设为 0.01，动量参数为 0.937，最后 10 轮统一关闭 mosaic 增强，且不使用预训练权重，如表 1 所示。

表 1　模型训练参数

Table 1　Model training configuration

参数	值
epochs	200
imgsz	640
optimizer	SGD
batch	8
lr0	0.01
momentum	0.937
pretrained	False

为更好的评价算法在密集行人检测任务中的效果，本文采用精确率（precision，P）、召回率（recall，R）、均值平均精度（mean average precision，mAP）、F1 分数以及模型参

数量作为评价指标。

$$P = \frac{TP}{TP + FP} \times 100\% \tag{16}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{17}$$

$$AP = \int_0^1 P(R)\mathrm{d}R \tag{18}$$

$$mAP = \frac{\sum_{i=1}^{K} AP_i}{K} \times 100\% \tag{19}$$

$$F_1 = \frac{2P \cdot R}{P + R} \tag{20}$$

式中:TP 表示被正确预测的真阳性,FP 表示被错误预测为阳性的假阳性,FN 表示被错误预测为阴性的假阴性,K 表示样本的种类数量。mAP@0.5 是指预测框与真实框之间交并比(intersection over union,IoU)阈值为 0.5 时的平均精确率的平均值,mAP@0.5∶0.95 是指 IoU 阈值从 0.5～0.95(步长为 0.05)下计算得到的平均精确率的平均值,F1 得分可以衡量模型在误检和漏检之间的平衡能力。

3.3 主干网络融合改进的对比实验

为验证本文提出的 EfficientNetV2S-S3 主干网络改进在密集行人检测任务上的优势,将轻量化的 EfficientNetV2S、宽度因子分别为 75% 和 100% 的 EfficientNetV2S-S3 三个主干网络依次融入 YOLO11n 中进行对比实验。实验结果如表 2 所示,相较于基线 YOLO11n 模型,轻量化 EfficientNetV2S 主干的融入导致性能轻微下降,但参数量减少至 2.1 M,表明其轻量化特性牺牲了部分检测精度。相比之下,EfficientNetV2S-S3 主干显著提升了性能。其中,宽度因子 75% 的版本 mAP@0.5 提高了 2%,mAP@0.5-0.95 提高 2.2%,同时参数量大幅降低至 1.0 M;宽度因子 100% 的版本比宽度因子 75% 的版本检测精度略有提高,但参数量增至 1.2 M。综上,宽度因子 75% 的 EfficientNetV2S-S3 主干网络在参数量仅 1.0 M 的条件下实现了最优的精度-效率权衡。

表 2 主干网络融合改进比较

Table 2 Comparison of improved backbone networks

Mode	mAP@0.5/%	mAP@0.5-0.95/%	F1-score	Parameters/10^6
YOLO11n(none)	78.7	48.5	0.76	2.6
EfficientNetV2S	77.3	47.0	0.75	2.1
EfficientNetV2S-S3(宽度因子 75%)	80.7	50.7	0.78	1.0
EfficientNetV2S-S3(宽度因子 100%)	81.1	50.9	0.78	1.2

3.4 注意力模块的对比实验

为验证本文提出的 DynaMSAttn 注意力机制在密集行人检测任务上的优势,将 DynaMSAttn 注意力机制和近年优秀的 CAA、LSKA、SCSA 三个注意力机制分别集成到 YOLO11n 模型中进行对比实验。CAA 注意力机制是 Cai 等[25]于 2024 年提出的加强模型对复杂上下文和多尺度目标适应能力的注意力机制;LSKA 注意力机制是 Lau 等[26]在 2023 年创新设计通过一维卷积分解策略解决大核卷积计算负担问题的注意力机制;SCSA 注意力是 Si 等[27]于 2024 年提出的缓解语义差异增强特征表达的协同空间通道注意力。不同注意力机制的实验结果如表 3 所示。

表 3 不同注意力机制比较

Table 3 Comparative analysis of attention mechanisms

注意力机制	mAP@0.5/%	mAP@0.5-0.95/%	F1-score	Parameters/10^6
YOLO11n(none)	78.7	48.5	0.76	2.6
CAA	78.1	48.0	0.76	2.8
LSKA	79.0	48.8	0.76	2.7
SCSA	78.8	48.7	0.76	2.6
DynaMSAttn	80.0	49.5	0.78	2.8

由实验结果可知,CAA 注意力机制的平均精度最差,这是因为其使用固定的平均池化层,不足以提取复杂的密集场景特征,此外,CAA 机制专注于空间注意力,但在密集人群这类遮挡严重的场景中,通道注意力同样重要。LSKA 注意力机制的 mAP 值有少量提升,主要得益于其使用了空洞卷积来增大感受野,在高遮挡的密集行人检测任务中能帮助算法更好的区分紧密相邻的人物,减少漏检和误检。SCSA 注意力融合了分组空间注意力和通道注意力等,对目标检测效果有一定改善,但效果有限。本文提出的 DynaMSAttn 注意力机制虽然参数量比最低的 SCSA

注意力机制增加了 0.2×10^6，但无论是 mAP 还是 F1 得分，均是最高的，这是因为其融合了动态池化通道注意力机制、多尺度空洞卷积模块和空间注意力机制，能更好的捕获密集行人检测任务中不同尺度和被遮挡的目标。

3.5　实验结果分析

为验证 DC-YOLO 算法在密集行人检测任务中性能的优越性，将其与当前主流轻量化目标检测模型，包括 RT-DETR-l[28]、YOLOv5n、YOLOv8n、YOLOv9t[29]、YOLOv10n、YOLO11s、YOLO11n 等模型进行了系统的对比。从 mAP@0.5、mAP@0.5-0.95、F1 得分及模型参数量 4 个指标综合评估各模型的性能。实验结果如表 4 所示，结果表明，本文的方法在显著减少参数量的同时，成功实现了检测精度与计算效率的协同优化。

表 4　不同算法对比实验表

Table 4　Comparative analysis of object detection algorithms

算法	mAP@0.5/%	mAP@0.5-0.95/%	F1-score	Parameters/10^6
RT-DETR-l	76.4	43.7	0.74	32.0
YOLOv5n	78.4	47.9	0.76	2.5
YOLOv8n	79.2	48.9	0.77	3.0
YOLOv9t	79.1	49.0	0.77	2.1
YOLOv10n	78.5	48.8	0.75	2.7
YOLO11s	83.0	53.6	0.80	9.4
YOLO11n(Baseline)	78.7	48.5	0.76	2.6
DC-YOLO(Ours)	83.4	53.0	0.80	1.4

在检测精度方面，当处于宽松定位要求(mAP@0.5)的场景时，本文的方法展现出了卓越的性能，以 83.4% 的精度领先于所有对比模型，mAP@0.5 与基线模型 YOLO11n(78.7%)相比，高出了 4.7%；相较于参数量第二少的 YOLOv10n(78.5%)，更是提升了 4.9%。虽然 YOLO11s 凭借 9.4 M 的参数量获得了 83% 的 mAP@0.5，但它的参数量是本文方法的 6.71 倍，本文方法仅使用了极少的参数量，就实现了更高的检测精度，这充分体现了模型结构优化的显著优势。当处于严格定位要求(mAP@0.5-0.95)的场景中，本文方法的 mAP@0.5-0.95 为 53%，非常接近最优的 YOLO11s 的 53.6%，但其参数量却仅为 YOLO11s 的 14.9%。与 YOLOv5n 的 47.9%、YOLOv8n 的 48.9%、YOLOv9t 的 49%、YOLO11n 的 48.5%等轻量化模型相比，本文方法的 mAP@0.5-0.95 提升了 4%～5.1%。本文方法借助高效轻量化主干与颈部特征网络提取优化，在大幅削减参数量的同时，依然保留了对密集行人检测任务中细节的捕捉能力。在 F1 得分上，本文方法达到了 0.80，与 YOLO11s 一同处于最优水平，明显优于 YOLOv8n(0.77)、YOLOv9t(0.77)、YOLOv10n(0.75)和 YOLO11n(0.76)。F1 得分较高，得益于动态多尺度协同注意力机制提高了模型对多尺度目标和复杂环境的适应，使其在精确率和召回率之间取得了较好平衡。值得注意的是，RT-DETR-l 模型，虽然参数量和运算量远高于各个 YOLO 版本，但在密集行人检测的各项指标上并未取得好的结果。

为验证 DC-YOLO 算法的泛化性和鲁棒性，测试本文方法在不同环境不同场景下的密集行人检测效果，在 WiderPerson 公开数据集上进行了对比实验，实验结果如表 5 所示。从表中可以看出，在 WiderPerson 数据集上的效果与 CrowdHuman 数据集上的效果相似，均有效提高了检测质量。同时，由于 Widerperson 数据集的标注强度不如 CrowdHuman 数据集的标注强度，所以 DC-YOLO 算法在 Widerperson 数据集中的指标提升稍微低于在 Crowdhuman 数据集中的提升。此外，因为 WiderPerson 数据集上没有对行人头部进行标注，所以相对于 CrowdHuman 数据集来说，微小目标的数量会有所减少，检测难度与 Crowdhuman 数据集对比，稍有降低，所以 DC-YOLO 算法在 WiderPerson 数据集上，各项指标均高于 CrowdHuman 数据集上的指标。

表 5　YOLO11n 和 DC-YOLO 在不同数据集上的对比实验表

Table 5　Comparative evaluation of YOLO11n and DC-YOLO on benchmark datasets

数据集	算法	mAP@0.5/%	mAP@0.5-0.95/%	F1-score	Parameters/10^6
CrowdHuman	YOLO11n	78.7	48.5	0.76	2.6
	DC-YOLO	83.4	53.0	0.80	1.4
WiderPerson	YOLO11n	88.4	62.6	0.83	2.6
	DC-YOLO	90.3	64.9	0.84	1.4

3.6 消融实验

为验证所提改进模块对密集行人检测任务的有效性,以 YOLO11n 为基线模型,使用 CrowdHuman 数据集,设计消融实验,对比主干网络 EfficientNetv2S-S3、颈部网络优化 P-LightNeck、RepNCSPELAN4、注意力机制 DynaMSAttn 四个模块对 DC-YOLO 算法的影响。消融实验结果如表 6 所示。

表 6 DC-YOLO 消融实验

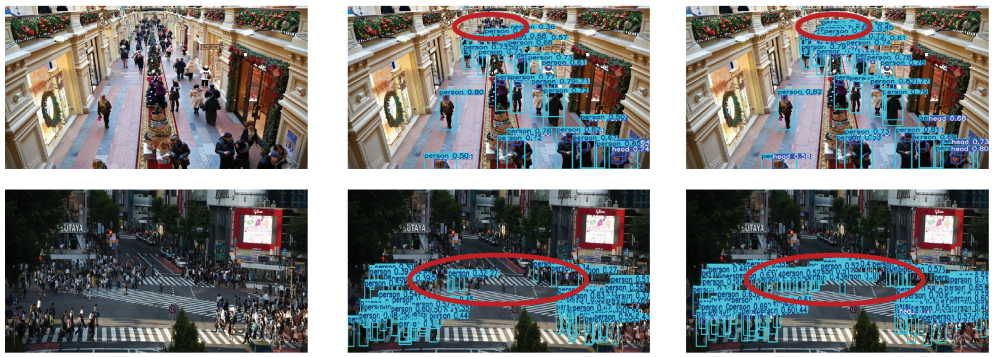
Table 6 DC-YOLO ablation experiment

实验次数	Efficient Netv2S-S3	Light NeckP2	RepNCSPELAN4	DynaMSAttn	mAP@0.5/%	mAP@0.5-0.95/%	F1-score	Parameters/10^6
1					78.7	48.5	0.76	2.6
2	√				80.7	50.7	0.78	1.0
3		√			81.4	51.1	0.78	1.9
4			√		79.8	49.8	0.77	3.2
5				√	80.0	49.5	0.78	2.8
6	√	√			82.3	52.0	0.79	1.0
7	√	√	√		82.9	52.9	0.79	1.2
8	√	√	√	√	83.4	53.0	0.8	1.4

从实验结果看,4 个改进模块对于密集行人检测性能均有提升。主干网络 EfficientNetv2S-S3 实现了在较低参数量的情况下,有效提高了检测精度,这既得益于该网络结构继承了 EfficientNetv2 的高效特征提取及 MBConv 和 Fused-MBConv 的有效融合,又受益于其专门针对密集行人特征降低了主干网络下采样,提高了模型对小目标的特征提取能力。在颈部采用轻量化特征金字塔融合小目标检测网络分支 P2 的 P-LightNeck 模块,对精度的提升最大,其中 mAP@0.5 提高了 2.7%,mAP@0.5-0.95 提高了 2.6%。这主要是因为 CrowdHuman 数据集有大量的小目标,而引入小目标检测网络分支十分有效的增强了对小目标的关注度,提高了检测精度,同时,对特征金字塔轻量化和移除大目标检测网络分支 P5 降低了参数量。在颈部引入 RepNCSPELAN4 卷积模块,虽然使参数增加了 0.6 M,但使得 mAP@0.5-0.95 比基线提高了 1.3%,表明聚合不同尺度和层次的特征,能有效应对密集行人检测任务中的多尺度行人目标。DynaMSAttn 注意力机制模块,通过增加少许参数量的代价,使 mAP@0.5 比基线提高了 1.3%,主要是因为多尺度空洞卷积和动态池化的策略,加强了对密集行人检测任务中被遮挡目标和小目标的通道敏感度。把 4 个模块逐一融合,包括 mAP@0.5、mAP@0.5-0.95、F1-score 在内的各项性能均能稳步提高,表明各个模块之间能较好的进行融合,能很好的适应密集行人检测任务。

3.7 可视化分析

为更直观地展示本文算法的有效性,随机选取了不同场景的图片进行检测结果可视化,将 DC-YOLO 算法和基准算法 YOLO11n 进行比较。在图 5 中,使用椭圆型框将两种算法检测结果的主要差异处圈起来。图 5 第 1 行为在夜间场景下,YOLO11n 算法将左下角的半个人和远处户



(a) 原始图像
(a) Original images　　(b) YOLO11n　　(c) DC-YOLO

图 5　YOLO11n 和 DC-YOLO 在不同场景下的检测结果可视化

Fig. 5　Multi-scenario detection visualization of YOLO11n and DC-YOLO

外显示屏幕前经过的行人漏检了，而 DC-YOLO 算法准确的检出了此两处行人；图 5 第 2 行为在有障碍物遮挡的服装店场景下，左边遮挡区域，YOLO11n 算法仅仅检测出 1 个行人，而 DC-YOLO 算法检测出 5 名行人，余 1 名行人漏检；图 5 第 3 行为在小目标场景中，DC-YOLO 算法比 YOLO11n 算法检测出的目标更多，如图中椭圆线框圈出部分；图 5 第 4 行为在商场中的多尺度目标场景下，图像前部行人占据的像素点位较多，而图像后部的行人仅仅占据几个像素点，YOLO11n 算法把该图后部的行人当成背景导致未检出，DC-YOLO 算法则是正确检测出了图像后部仅有几个像素的行人；图 5 第 5 行为在行人密集的十字路口场景中，YOLO11n 算法仅仅检测出 85 个目标，而 DC-YOLO 算法检测出了 145 个目标，DC-YOLO 算法检出的目标数量是 YOLO11n 算法检出目标数量的 170.59%。从图 5 的不同场景检测结果可以看出，DC-YOLO 算法能有效提高检测性能，降低漏检率，同时定位更精准，证明了本文算法的有效性。

4　结　　论

本文针对密集行人检测任务中遮挡严重、小目标多、尺度变化大和环境复杂等挑战，提出了 DC-YOLO 轻量化算法。该算法通过提出轻量化主干网络 EfficientNetV2S-S3，显著降低模型复杂性和参数量，并增强了对小目标的特征提取能力；设计高效的特征融合模块 P-LightNeck，提高颈部网络对小目标的网络特征融合能力；引入 RepNCSPELAN4 模块，强化模型应对复杂场景和多尺度目标的能力；设计动态多尺度协同注意力模块 DynaMSAttn，增强模型对遮挡目标及复杂环境的检测能力。

实验结果表明 DC-YOLO 较基准模型 YOLO11n，在 CrowdHuman 公开数据集上，mAP@0.5、mAP@0.5-0.95 分别提升了 4.7%和 4.5%，且参数量降低了 46.2%。同时，该算法在 WiderPerson 公开数据集上表现也不俗。实验证明，DC-YOLO 在密集行人检测任务中有明显优势。

DC-YOLO 算法虽然在密集行人检测任务中具有突出优势，但在行人极度密集、遮挡异常严重、夜间光照条件差、极端雨雪、沙尘等极端环境下，仍然存在漏检、错检及定位精度差等问题。下一步，本文将继续研究算法改进，提高算法在各种极端环境下的鲁棒性和检测效果。

参考文献

[1]　GIRSHICK R. Fast R-CNN[C]. IEEE International Conference on Computer Vision, 2015: 1440-1448.

[2]　REN SH Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[3]　LIU W, LIAO SH C, REN W, et al. High-level semantic feature detection: A new perspective for pedestrian detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 5187-5196.

[4]　HAN CH CH, ZHENG ZH D, SU K, et al. DMRNet++: Learning discriminative features with decoupled networks and enriched pairs for one-step person search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 7319-7337.

[5]　REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[6]　WANG AO, CHEN H, LIU L H, et al. YOLOv10: Real-time end-to-end object detection[C]. Advances in Neural Information Processing Systems, 2024, 37: 107984-108011.

[7]　梁天添，杨淞淇，钱振明. 基于改进 YOLOv8s 的恶劣天气车辆行人检测方法[J]. 电子测量技术. 2024, 47(9): 112-119.

LIANG T T, YANG S Q, QIAN ZH M. Improved YOLOv8s method for vehicle and pedestrian detection in adverse weather [J]. Electronic Measurement Technology, 2024, 47(9): 112-119.

[8] 胡伟超,皮建勇,胡倩,等. 面向复杂场景密集行人检测的 YOLOv8 改进模型[J]. 电子测量技术. 2024, 47(14): 159-169.
HU W CH, PI J Y, HU Q, et al. Improved YOLOv8 model for dense pedestrian detection in complex scenes[J]. Electronic Measurement Technology, 2024, 47(14): 159-169.

[9] LIU W B, QIAO X Y, ZHAO CH Y, et al. VP-YOLO: A human visual perception-inspired robust vehicle-pedestrian detection model for complex traffic scenarios. [J]. Expert Systems with Applications, 2025, 274: 126837.

[10] ZHANG F F, LEONG L V, YEN K S, et al. An enhanced lightweight model for small-scale pedestrian detection based on YOLOv8s [J]. Digit Signal Process, 2025, 156: 104866.

[11] TANG Y, LIU M, LI B P, et al. NAS-PED: Neural architecture search for pedestrian detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2025, 47(3): 1800-1817.

[12] CAO J, PANG Y W, XIE J, et al. From handcrafted to deep features for pedestrian detection: A survey [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44: 4913-4934.

[13] CHOLLET F. Xception: Deep learning with depthwise separable convolutions [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1251-1258.

[14] ZHANG X D, ZENG H, GUO SH, et al. Efficient long-range attention network for image super-resolution [J]. European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 649-667.

[15] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions [J]. ArXiv preprint arXiv: 1511.07122, 2015.

[16] TAN M, LE Q. Efficientnetv2: Smaller models and faster training [C]. International Conference on Machine Learning. PMLR, 2021: 10096-10106.

[17] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks [C]. International Conference on Machine Learning. PMLR, 2019: 6105-6114.

[18] HU J, SHEN L, SUN G. Squeeze-and-excitation networks. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2019, 42: 2011-2023.

[19] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2117-2125.

[20] CHEN G, WANG H T, CHEN K, et al. A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(2): 936-953.

[21] DING X, ZHANG X, MA N, et al. Repvgg: Making vgg-style convnets great again [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13733-13742.

[22] CHEN Y, XU L, LIU K, et al. Event extraction via dynamic multi-pooling convolutional neural networks[C]. 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015: 167-176.

[23] SHAO SH, ZHAO Z J, LI B X, et al. Crowdhuman: A benchmark for detecting human in a crowd [J]. ArXiv preprint arXiv:1805.00123, 2018.

[24] ZHANG SH F, XIE Y L, WAN J, et al. WiderPerson: A diverse dataset for dense pedestrian detection in the wild [J]. IEEE Transactions on Multimedia. 2020, 22(2): 380-393.

[25] CAI X H, LAI Q X, WANG Y W, et al. Poly kernel inception network for remote sensing detection [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 27706-27716.

[26] LAU K W, PO L, REHMAN Y A U. Large separable kernel attention: Rethinking the large kernel attention design in CNN [J]. Expert Systems with Applications. 2023, 236: 121352.

[27] SI Y ZH, XU H Y, ZHU X ZH, et al. SCSA: Exploring the synergistic effects between spatial and channel attention [J]. Neurocomputing, 2025, 634: 129866.

[28] ZHAO Y, LYU W, XU S, et al. Detrs beat yolos on real-time object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 16965-16974.

[29] WANG C Y, YEH I H, MARK LIAO H Y. Yolov9: Learning what you want to learn using programmable gradient information [C]. European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 1-21.

作者简介

黄思禄,硕士研究生,主要研究方向为深度学习、计算机视觉和目标检测。

E-mail:2024212121@stu.ppsuc.edu.cn

钟寒(通信作者),博士,副教授,主要研究方向为数据分析、计算机视觉和图像处理。

E-mail: zhonghan@ppsuc.edu.cn