姜乃祺, 陈俊, 陈芳,等. 基于改进 YOLOv8n 的轻量型密集行人检测模型[J]. 智能计算机与应用,2025,15(11):137-141. DOI:10.20169/j.issn.2095-2163.251122

# 基于改进 YOLOv8n 的轻量型密集行人检测模型

姜乃祺, 陈 俊, 陈 芳, 孟伟强, 石浩铭
(福州大学 物理与信息工程学院, 福州 350108)

**摘 要**: 针对密集行人检测的难点和检测模型部署的资源限制,本文提出了一种基于改进 YOLOv8n 的轻量型密集行人检测模型,以 YOLOv8n 目标检测模型为基础,采用轻量型的 MobileViT 代替原主干网络,在图像特征提取时,实现局部和全局依赖关系的平衡;引入 BRA 注意力机制,在保证轻量化的同时,增强检测模型对长距离上下文依赖的捕获能力;加入 DyHead 动态检测头以增强检测头对行人多尺度目标和行人空间结构变化的敏感性,使检测模型具有更灵活的动态调整能力。在 CrowdHuman 行人数据集的实验结果表明,相较于 YOLOv8n,本文提出的改进模型 JI 指标提升 1.67,同时只有 3.98 M 的参数量和 6.32 GFLOPs 的计算量,能够满足工程应用的实际需求。

**关键词**: 密集行人检测; YOLOv8; MobileViT; BRA 注意力机制; DyHead
**中图分类号**: TP391.41 **文献标志码**: A **文章编号**: 2095-2163(2025)11-0137-05

# Research on lightweight dense pedestrian detection algorithm based on improved YOLOv8n

JIANG Naiqi, CHEN Jun, CHEN Fang, MENG Weiqiang, SHI Haoming
(College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China)

**Abstract**: In response to the difficulties of dense pedestrian detection and the resource limitations of detection model deployment, this paper proposes a lightweight dense pedestrian detection model based on YOLOv8n improvement. Based on the YOLOv8n object detection model, a lightweight MobileViT is used instead of the original backbone network to achieve a balance between local and global dependencies in image feature extraction; Introducing the BRA attention mechanism to enhance the detection model's ability to capture long-range contextual dependencies while ensuring lightweighting; Adding DyHead to enhance the sensitivity of the detection head to multi-scale pedestrian targets and changes in pedestrian spatial structure, and to enable the detection model to have more flexible dynamic adjustment capabilities. The experimental results on the CrowdHuman pedestrian dataset show that compared to YOLOv8n, the improved model proposed in this paper has a 1.67 improvement in JI index, with only 3.98 M of parameters and 6.32 GFLOPs of computation, which can meet the practical needs of engineering applications.

**Key words**: dense pedestrian detection; YOLOv8; MobileViT; BRA attention mechanism; DyHead

## 0 引 言

在如交通路口、娱乐场所和旅游景点等人员相对密集的区域,需要摄像头等监控设备对人群进行实时检测,以便在人员密度过高或出现安全事故时,及时疏散人群,做出合理的安全防患措施。人工监测需大量监控数据,人力成本高、检测效率低、漏检率高,而使用智能监控设备,利用行人目标检测技术,不仅可以极大的提高效率,同时可以保证全天时全天候地不间断工作,极大程度地提升了公共场所的安全系数。

密集行人检测是指在人员密集的场景下进行的行人检测,是行人检测任务的重要分支任务,存在诸多难点和挑战。由于行人的密集性,行人与行人之间的相互遮挡,容易造成检测模型的误检、漏检;工程部署的场景差异,会造成行人的多样性、多尺度问题和图像场景的复杂性问题,对检测模型的跨域性、泛化性和鲁棒性提出了较为苛刻的要求;检测模型

**作者简介**: 姜乃祺(1999—),男,硕士,主要研究方向:计算机视觉; 陈 芳(1998—),男,硕士,主要研究方向:电子与通信工程。
**通信作者**: 陈 俊(1978—),男,硕士,副教授,主要研究方向:物联网通信。Email:56851@qq.com。

**收稿日期**: 2024-02-25

的实际部署应当考虑硬件平台的算力,实现模型精度和轻量化的平衡。

现有的行人检测数据集,如 Caltech、CityPersons 和 KITTI 行人数据集,行人密度低,无法满足密集行人检测的任务需求,对此旷视公司发布 CrowdHuman 密集行人数据集,具有更多的行人数量、更高的行人密度和更复杂的行人遮挡关系[1]。为了解决高度行人带来的遮挡问题,Wang 等[2]提出 Repulsion Loss 聚集损失函数,通过吸引预测框和对应真实框的位置,同时排斥预测框和其它真实框的位置,以增强对密集行人的检测能力。根据 CrowdHuman 数据集提供的可见框标注,Huang 等[3]提出代表性区域非极大值抑制算法,利用较少遮挡的可见部分进行后处理,有效抑制了冗余的检测框,并减少了假正例。为了避免非极大值抑制错误地移除高度重叠的行人实例,Chu[4]对每个提议框,都预测一组可能高度重叠的实例。

本文针对密集行人检测的难点和检测模型轻量化部署的问题,以 YOLOv8n 目标检测模型为基础,使用轻量型 MobileViT 网络替代原模型的主干网络(Backbone),实现图像特征的高效提取,维持局部和全局依赖关系的平衡;在主干网络的尾部添加 BRA(Bi-level Routing Attention)注意力机制,在不大量增加网络参数的前提下,增强主干网络的特征提取能力,使检测模型获得对长距离上下文依赖的捕获能力;在目标检测模型的检测头(Head)部分添加 DyHead(Dynamic Head)动态检测头,使模型获得动态调整的能力,以此构成效果良好的轻量型密集行人检测模型。

## 1 YOLOv8 目标检测模型

YOLOv8 目标检测模型是由 Ultralytics 公司在 YOLOv5 版本的基础上,提出的升级版本,是 YOLO(You Only Look)系列的继承和延伸,融入了更多目标检测和深度学习图像处理技术的新方法,在该系列中达到了 SOAT(State Of the Art)的效果,网络结构如图 1 所示。

YOLO 系列模型的设计思想不同于 Two-Stage 模型或大规模的模型,YOLO 作为一个 One-Stage 模型,其精度和模型的检测速度都是模型设计的重要考量,该系列模型设计的初衷在于实时性,因此,YOLOv8 提供了 5 种不同尺寸的模型,通过调节网络深度缩放因子、网络宽度缩放因子和网络通道缩放因子实现模型的放缩,以此实现不同尺寸模型的设计,可以满足不同场景、不同设备的任务需求。

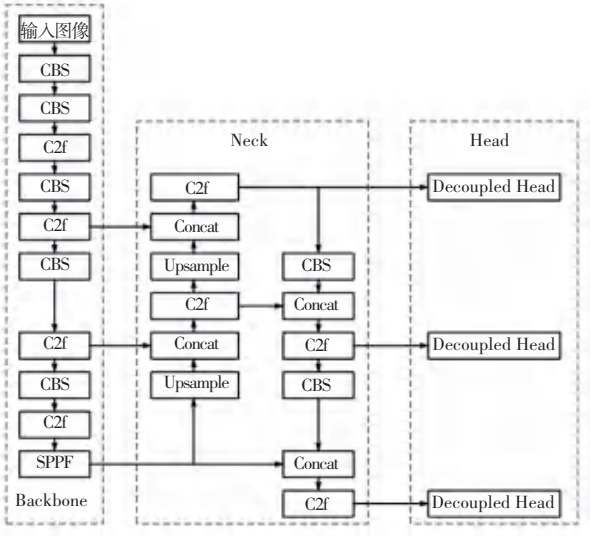


**图 1 YOLOv8 网络结构**

**Fig. 1 YOLOv8 network structure**

## 2 基于改进 YOLOv8n 的轻量型改进的密集行人检测模型

### 2.1 MobileViT 网络

CNN(Convolutional Neural Network)因其特有的归纳偏置,被广泛应用于图像处理的各个领域,其卷积核的设计使网络模型更关注于图像局部的纹理特征。虽然 CNN 可以通过堆叠卷积层扩大感受野,但获得的感受野区域对输出结果的影响服从高斯分布,影响权重以感受野中心向外扩散并迅速衰减,真实感受野与理论感受野存在差异[5]。

ViT(Vision Transformer)通过 QKV(Query Key Value)建立全局依赖关系,弥补了 CNN 全局归纳建模能力的不足,对于遮挡、干扰、域迁移等问题的处理有较好的表现,但通常需要大量的训练数据、图像增强和 L2 正则化的支撑,当模型大小被约束,ViT 的表现往往不如轻量型的 CNN。

MobileViT 网络结合两者优势,互补劣势,实现了局部和全局依赖关系的平衡[6]。MobileViT 网络在网络浅层采用轻量型 MobileNetV2 卷积块以提取局部细节信息,在网络深层采用轻量型 MobileViT 卷积块以提取高级语义信息,并获得捕获全局感受野的能力。这种设计方式,既获得了 CNN 对局部纹理信息的提取能力,又获得了 ViT 灵活、动态的感受野,同时保持了整个网络模型的轻量化。

MobileViT 卷积块由 Local Representation 和 Global Representation 两部分构成,依次对局部信息和全局信息进行建模。Local Representation 仍然利用

CNN 的偏置归纳特性，采用传统卷积实现，先通过卷积核大小为 $N$ 的卷积块，再通过卷积核大小为 1 的卷积块改变特征图的维度。Global Representation 则充分考虑 CNN 和 ViT 两者的计算机制，并在 ViT 的基础上加以改进。

标准卷积根据其滑动窗口的计算机制，可以等效为 unfold、特征提取、folding 3 个阶段的操作，Global Representation 在 ViT 的基础上，将 ViT 对 patch 的展平和线性映射，替换为 unfold 操作，然后经过 ViT 的 Transformer 后，做 fold 操作，将 Transformer 生成的 Token 恢复为 patch。

Global Representation 的实现方式有别于 ViT 的自注意力机制或轻量化的稀疏注意力机制，既不会丢失图像像素，又不会丢失图像像素的空间顺序，可以保证其感受野仍为特征图大小。

MobileViT 的设计方式既获得了 CNN 对局部信息的提取能力，又获得了 ViT 灵活、动态的感受野，同时保持了整个网络模型的轻量化，减轻了训练过程中对大批量训练数据、图像增强和正则化的依赖，是 CNN 和 ViT 两者的优质结合，也是两者融合使用的重要探索。

## 2.2　BRA 注意力机制

注意力机制已广泛应用于图像处理、自然语言处理、语音识别等多个领域，仿照人类与世界的交互模式，通过可训练的网络权重使模型聚焦在图像中重要的区域，这种机制能灵活地嵌入到深度学习的网络之中。通过注意力机制可以在不大量增加网络参数和计算量的同时，带来网络检测性能的提升；而纯粹堆叠卷积层、加深和加宽网络，虽然能丰富网络模型的非线性特征表达能力，带来一定的性能提升，但与其付出的算力代价不成正比。

为了解决自注意力机制计算量大和内存负担的问题，本文引入 BRA 注意力机制（Bi-level Routing Attention）用于增强网络模型对长距离上下文依赖的捕获能力。通过稀疏性的结构设计，在保证不大量增加计算量的同时，实现具有内容感知的更灵活的计算分配[7]。BRA 注意力机制结构示意如图 2 所示，在粗略的区域级别过滤掉相关性不强的键值对，对剩余的区域即路由区域，做对应关联区域的相关操作，以此实现模型的轻量化设计。

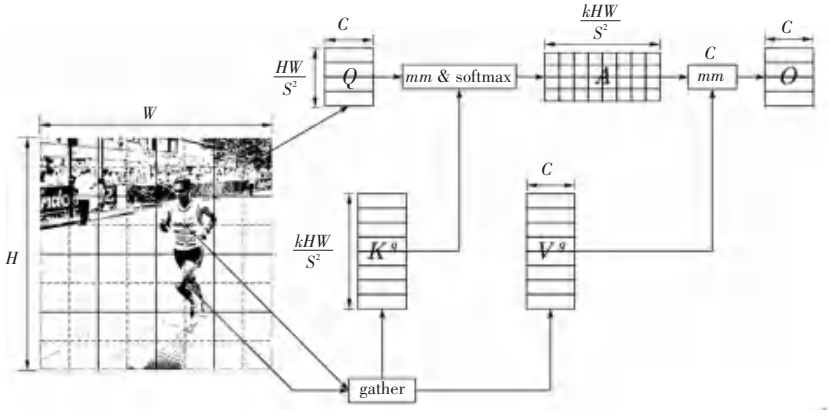


**图 2　BRA 注意力机制结构示意图**

**Fig. 2　Schematic diagram of BRA attention mechanism structure**

## 2.3　DyHead 动态检测头

为了增强 YOLO 架构中检测头的检测能力，以应对复杂的行人检测场景，本文引入了 DyHead 动态检测头，在不大量增加检测头计算量的同时，根据行人检测的检测要点和检测难点，做出针对性优化[8]。

行人检测头的作用是将行人检测模型提取并将特征融合的图像特征转化为行人检测所需的输出结果。行人检测头应当对检测目标的尺寸具有敏感性，以应对图像中大小不一的行人目标，满足行人多尺度和小目标行人的任务要求；行人检测头应当具有行人目标的空间结构敏感性，以应对图像中行人目标的形态差异、角度差异和位置差异；根据检测头输出结果的不同，任务的不同需要，如检测的行人目标为边界框、中心点或关键点等其他任务需求，检测头应当根据任务的不同设定进行动态调整。

针对以上问题，DyHead 动态检测头依次添加 $\pi_L$ 尺度感知注意力（Scale-aware Attention）、$\pi_S$ 空间感知注意力（Spatial-aware Attention）和 $\pi_C$ 任务感知注意力（Task-aware Attention），DyHead 动态检测头结构如图 3 所示，三者级联以增强检测头在密集行人检测任务中的动态调整能力。尺度感知注意力采用空间注意力机制，以获得目标尺度变化的敏

感性;空间感知注意力采用可变形卷积,以获得灵活多变的感受野;任务感知注意力采用动态激活函数,以获得灵活多变的非线性表达能力,根据任务的不同来动态调整模型。

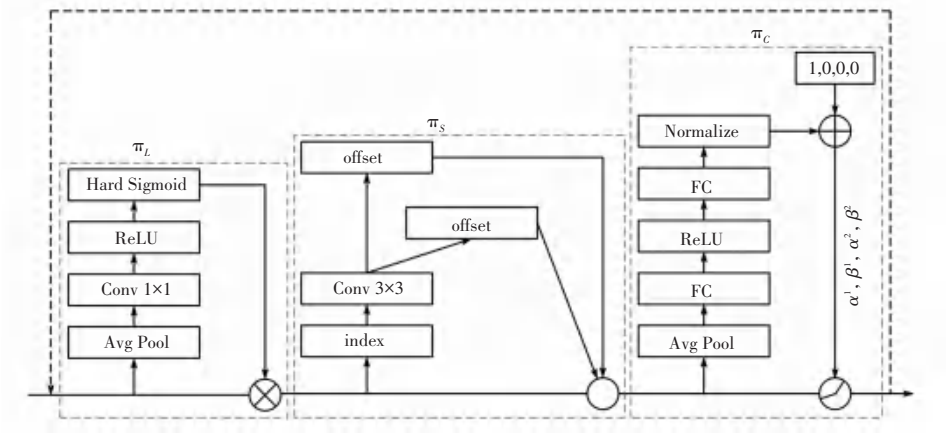


**图 3 DyHead 动态检测头结构图**

**Fig. 3 Diagram of DyHead structure**

## 3 实验结果与分析

### 3.1 密集行人检测数据集

CrowdHuman 数据集是旷视科技发布的用于行人检测的数据集,其包含了各种行人的实际生活场景,是一个评价行人检测模型检测性能的常用标准数据集。相较于其他行人检测数据集,CrowdHuman 数据集的行人数量更多,单张图片的行人密度更大,行人与行人之间的遮挡关系更复杂,因此也称为密集行人检测数据集,其训练数据集共计 15 000 张图像,339 565 个行人目标,验证数据集共计 4 370 张图像,99 481 个行人目标。

### 3.2 实验环境及训练参数

本文的实验环境配置:用于网络训练和推理的操作系统为 Ubuntu 20.04,CPU 为 Intel(R) Xeon(R) Silver 4214R @2.4 GHz;显卡为 4 张 RTX3080Ti(12 GB),深度学习框架为 Pytorch 1.10.0;Cuda 版本为 CUDA 11.3。

本文模型的训练参数设置:输入图像分辨率为 640×640,批处理量(Batch Size)为 8;采用 SGD(Stochastic Gradient Descent)作为模型的优化器;共训练 300 轮次,在最后 10 轮次关闭 Mosaic 数据增强;初始学习率为 0.01;动量为 0.937;权重衰减为 5e-4。

### 3.3 模型评价指标

本文实验使用 JI(Jaccard Index)、AP(Average Precision)和 MR(Miss Rate)来衡量密集行人检测模型的检测精度,公式如下:

$$AP = \int_0^1 P(R)\,dR \tag{1}$$

$$MR = \frac{FN}{TP + FN} \tag{2}$$

$$JI = \frac{|\ IoUMatch(D,G)\ |}{|\ D\ | + |\ G\ | - |\ IoUMatch(D,G)\ |} \tag{3}$$

其中,$D$(Detection Box)为行人检测模型的预测框,$|\ D\ |$ 为预测框的数量;$G$(Grounding Box)为数据集的真实标注框;$|\ G\ |$ 是标注框的数量;$|\ MatchIoU(D,G)\ |$ 为预测框和真实框匹配的数量,当预测框和真实框的交并比大于阈值,即认为两者匹配成功。

根据检测模型的部署问题,使用模型参数 Params 和模型计算量(FLOPs)衡量密集行人检测模型的轻量化程度。综合评估密集行人检测模型的精度和轻量化程度,才能更好地反映检测模型在低算力平台的检测表现。

### 3.4 模型消融实验

为了验证本文提出的基于改进 YOLOv8n 的轻量型密集行人检测模型对密集行人检测任务的有效性,对模型进行消融实验,实验结果见表 1。

实验结果表明,使用 MobileViT 网络代替原 YOLOv8n 的 Backbone,不仅带来检测性能的提升,漏检率大幅降低,同时轻量化的网络结构降低了密集行人检测模型整体的参数量和计算量;在 MobileViT 的尾部添加 BRA 注意力机制,模型的参数量和计算量略微增多,且模型检测性能获得了一定的提升;在 YOLOv8n 的检测头部分添加动态检测头后,模型的检测效果进一步提升,JI 指标达到了 77.70%。说明该改进模型很好的兼顾了轻量化和检测效果。

表 1　消融实验结果

Table 1　Results of ablation experiment

| YOLOv8n | MobileViT | BRA | DyHead | JI/% | AP/% | MR/% | Params/M | FLOPs/G |
|---|---|---|---|---|---|---|---|---|
| √ | × | × | × | 76.03 | 85.84 | 58.14 | 3.01 | 8.12 |
| √ | √ | × | × | 77.22 | 86.84 | 48.38 | 2.72 | 6.05 |
| √ | √ | √ | × | 77.37 | 87.03 | 48.58 | 2.76 | 6.15 |
| √ | √ | √ | √ | 77.70 | 87.13 | 47.97 | 3.98 | 6.32 |

### 3.5　模型对比实验

为了进一步验证基于改进 YOLOv8n 的轻量型密集行人检测模型和常规的目标检测模型进行对比实验，实验结果见表 2，模型实际的检测效果如图 4 所示。

表 2　对比实验结果

Table 2　Results of comparative experiment

| 模型 | JI/% | AP/% | MR/% | Params/M | FLOPs/G |
|---|---|---|---|---|---|
| YOLOv5n | 70.26 | 81.67 | 55.10 | 1.76 | 4.17 |
| YOLOv5s | 75.07 | 85.59 | 48.26 | 7.02 | 15.83 |
| YOLOv6n | 76.79 | 85.58 | 59.24 | 4.30 | 11.00 |
| YOLOv7tiny | 77.65 | 86.97 | 48.37 | 6.01 | 13.09 |
| CrowdDet | 82.30 | 90.70 | 41.40 | 41.35 | 208.80 |
| 本文模型 | 77.70 | 87.13 | 47.97 | 3.98 | 6.32 |

图 4　检测效果图

Fig. 4　Detection rendering

实验结果表明，本文提出的优化模型，相较于 One-Stage 的 YOLO 系列的模型，在模型的轻量化程度和检测精度上，都有较好的表现；和 Two-Stage 的 CrowdDet 相比，虽然检测精度仍有差距，但本文模型的参数量和计算量更适合在低算力平台部署，满足工程部署的实际要求。

## 4　结束语

本文提出了一种基于 YOLOv8n 改进的轻量型密集行人检测模型，针对密集行人检测的难点和检测模型部署的有限资源，以 YOLOv8n 目标检测模型为基础，采用 MobileViT 代替原 Backbone 部分，并在 Backbone 的尾部添加 BRA 注意力机制，同时在 Head 部分添加 DyHead，在保证轻量化的同时，使密集行人检测模型具有跨域性、泛化性和鲁棒性。

## 参考文献


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

[2] WANG X, XIAO T, JIANG Y, et al. Repulsion loss: Detecting pedestrians in a crowd[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7774-7783.

[3] HUANG X, GE Z, JIE Z, et al. Nms by representative region: Towards crowded pedestrian detection by proposal pairing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10750-10759.

[4] CHU X, ZHENG A, ZHANG X, et al. Detection in crowded scenes: One proposal, multiple predictions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 12214-12223.

[5] NASEER M M, RANASINGHE K, KHAN S H, et al. Intriguing properties of vision transformers[C]//Proceedings of Advances in Neural Information Processing Systems. NIPS, 2021: 23296-23308.

[6] MEHTA S, RASTEGARI M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer[J]. arXiv preprint arXiv, 2110.02178, 2021.

[7] ZHU L, WANG X, KE Z, et al. BiFormer: Vision transformer with bi-level routing attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 10323-10333.

[8] DAI X, CHEN Y, XIAO B, et al. Dynamic head: Unifying object detection heads with attentions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 7373-7382.