DOI:10.19651/j.cnki.emt.2416218

面向复杂场景密集行人检测的 YOLOv8 改进模型*

胡伟超 皮建勇 胡 倩 黄 昆 王娟敏
(贵州大学计算机科学与技术学院公共大数据国家重点实验室 贵阳 550025)

摘 要:针对当前行人检测面临的环境复杂、目标尺寸多变和严重遮挡等挑战,导致现有检测技术在识别密集行人时容易发生误判和遗漏的问题,本文提出一种高效的面向复杂场景密集行人检测的 YOLOv8 改进模型。在骨干网络引入 DCNv2 设计 C2f_DCNetv2 替换 C2f 模块,提升骨干网络的特征提取能力;通过在架构中加入小目标检测头,增强模型对小尺寸目标的检测能力,提高对小目标的检测识别精度;基于四检测头改进 AFPN 设计出 AFPN-4H,优化特征层之间的信息融合,提高了模型对不同尺度目标的适应性和检测精度;最后,通过结合 Wise-IoU、Focaler-IoU 和 MPDIoU 得到 WFM-IoU,进一步提高了目标定位的准确性。实验结果表明,与原始的 YOLOv8n 模型相比,在 P、R、AP50 以及 AP50:95 等关键指标上分别提升 1.6%、4.0%、3.6%和 3.8%,也优于其他算法。验证了本文改进算法在复杂场景密集行人检测任务中具有较好的性能。

关键词:密集行人检测;YOLOv8;可变形卷积;多尺度特征融合;损失函数
中图分类号:TP 391.4;TN98 **文献标识码**:A **国家标准学科分类代码**:520.20

Improved YOLOv8 model for dense pedestrian detection in complex scenes

Hu Weichao Pi Jianyong Hu Qian Huang Kun Wang Juanmin
(State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang 550025, China)

Abstract: Aiming at the current challenges of pedestrian detection, such as complex environments, variable target sizes, and severe occlusions, which cause existing detection techniques to be prone to misjudgment and omission when recognising dense pedestrians, this paper proposes an efficient YOLOv8 improved model for dense pedestrian detection in complex scenes. DCNv2 is introduced into the backbone network, and C2f_DCNetv2 is designed to replace the C2f module, which improves the feature extraction capability of the backbone network; the detection capability of the model for small targets is improved by adding small-target detecting heads to the architecture, which improves the accuracy of small-target detection and recognition; based on the four detecting heads as well as the AFPN, the AFPN-4H is designed, which optimises the information fusion between the feature layers and improves the model′s adaptability and detection accuracy for targets of different scales; finally, through the combination of Wise-IoU, Focaler-IoU, and MPDIoU, the WFM-IoU is obtained, which further improves the target localisation accuracy. The experimental results show that compared with the original YOLOv8n model, it improves 1.6, 4.0, 3.6 and 3.8 percentage points in the key indexes of P, R, AP50 and AP50:95, respectively, which are also inferior to other algorithms. The improved algorithm in this paper has better performance in the dense pedestrian detection task in complex scenes.

Keywords: dense pedestrian detection; YOLOv8; deformable convolution; multi-scale feature fusion; loss function

0 引 言

近年来,随着智能安防、智能交通和自动驾驶等领域的快速发展,行人目标检测作为上述领域的共同基础得到了大量关注。在现实应用中,行人数据面临着环境复杂、目标尺寸多变和严重遮挡等挑战,导致现有检测技术在识别密集行人时容易发生误判和遗漏的问题,需要强大而精确的解决方案。

早在 2001 年,Viola 和 Jones 就提出了著名的 Viola-Jones (VJ) 探测器[1]。它结合了“整体图像”等多种重要技

收稿日期:2024-06-11
* 基金项目:贵州省科技支撑计划(黔科合支撑[2023]一般 430)项目资助

术，显著提高了检测效率和检测能力，首次实现了对固定物体的实时检测，有力地推动了目标检测领域的发展。2005 年 Dalal 和 Triggs[2] 提出了定向梯度直方图（histogram of gradient，HOG）特征描述符，设计了在密集均匀间隔的细胞网格上计算的 HOG 描述子，并采用重叠的局部对比度进行归一化以提高精度。2008 年，Felzenszwalb 等[3] 提出了可变形部件模型（deformable parts model，DPM）检测算法，可以将行人划分为不同的部分进行训练和学习，并在分类时将其视为不同部分检测的集合。然而，手动提取方法的提取步骤繁琐，计算成本高，实时性能不理想。

基于深度学习的目标检测的方法主要利用卷积神经网络（convolutional neural networks，CNN）[4] 来提取图像中的高层语义特征，并结合一些区域提名或者回归的技术来进行目标检测。主流的基于深度学习的目标检测算法根据检测步骤可分为两类：Two-stage 目标检测算法和 One-stage 目标检测算法[5]。

Two-stage 算法如 R-CNN[6]、Fast RCNN[7]、Faster RCNN[8] 和 Mask R-CNN[9] 等，这类算法使用独立的区域选取网络生成候选目标位置，进一步利用独立的目标检测网络进行分类和边界框精细化，得出最终结果。

One-stage 算法通过在特征图上设置一系列锚点，直接预测对象中心和对象边界框。代表算法有 YOLO（you only look once）[10-12] 系列以及 SSD[13] 目标检测算法。胡倩等[14] 提出 YOLOv5_Conv-SPD_DAFPN 模型，引入 Conv-SPD 网络模块缓解小目标或密集行人特征信息的丢失问题，提出了双层渐进金字塔网络（double asymptotic feature pyramid network，DAFPN）提升行人检测的准确性和精度，引入了 EfficiCIOU-Loss 定位损失函数，优化网络模型的定位性能，加快了模型的收敛速度。但使用 YOLOv5（https://github.com/ultralytics/yolov5）算法使得整体效果不佳并且计算量以及参数量较大。宁爽等[15] 提出帧间方向梯度直方图特征关联的行人检测方法，该方法通过扩展 YOLOv7[16] 模型并结合帧间方向梯度直方图特征对漏检的行人目标进行检测，有效提升了遮挡情况下的行人检测精度。但该方法仅针对遮挡行人，没有解决小目标行人漏检错检的问题。黄昆等[17] 提出 Crowd-YOLOv8，使用 nostride-Conv-SPD 模块，增强主干网络提取能力，引入小目标检测头和 CARAFE 上采样算子提高对小目标的检测效果。有效提升了算法对行人检测的精度，算法仍使用 PAN-FPN 结构进行特征融合容易导致信息丢失。综合比较，YOLOv8（https://github.com/ultralytics/yolov8）是目前 One-stage 目标检测算法中效率较高的，具有更好的应用前景。

Two-stage 和 One-stage 两种目标检测算法各有优缺点，前者在精度上往往表现出色，但是检测速度较慢[18]，后者虽然精度相对较低，但检测速度更快，能满足实时检测的要求。因此，One-stage 目标检测算法在实际应用中更受欢迎，应用更加广泛，本文的研究面向实际应用场景，所以选择 YOLOv8 算法进行实验。行人数据集具有背景环境复杂、目标小和被遮挡等干扰，对检测算法提出了重大挑战，需要强大而精确的解决方案。本文基于 YOLOv8 提出改进以解决上述问题，主要改进点如下：

1）基于第二代可变形卷积（deformable convNets v2，DCNv2）[19] 设计 C2f_DCNetv2 替换骨干网络的 C2f，增强了模型对目标形变的适应能力，提升模型的特征提取能力。

2）将小目标层加入到检测层中，提升了模型对小目标的感知能力，使模型能够更加精确的检测和定位小尺度目标。

3）利用自适应空间特征融合技术（adaptively spatial feature fusion，ASFF）[20] 添加小目标检测头改进渐近特征金字塔网络（asymptotic feature pyramid network，AFPN）[21] 设计出四检测头渐进金字塔网络（asymptotic feature pyramid network for four detection heads，AFPN-4H）。提高了模型对小目标的感知能力并减小非相邻特征层之间的语义间隙，充分融合了不同特征层的信息，提升检测头的检测精度。

4）结合 Focaler-IoU[22]、Wise-IoU[23] 和 MPDIoU[24] 设计全出 Wise-Focaler-MPDIoU（WFM-IoU），进一步提高了目标定位的准确性。

1　YOLOv8 模型

Ultralytics 公司在 2023 年 1 月发布的 YOLOv8 模型是 YOLO 系列中的最新进展，它在目标检测、图像分类、实例分割、关键点检测等多个计算机视觉任务中表现出色。YOLOv8 模型从小到大分为 v8n、v8s、v8m、v8l、v8x 五个版本，随着模型大小的增加，精度也相应提升，用户可以根据任务需求选择合适的网络模型。

YOLOv8 模型主要由 3 部分组成：骨干网络（Backbone）、颈部（Neck）和检测头（Head）。Backbone 采用 Darknet-53 框架，引入了 C2f 模块进行残差学习，增强了特征的表达能力。Neck 部分使用 C2f 模块替换了 C3 模块，采用 PAN-FPN 结构进行特征融合。Head 部分则发生了显著变化，由 YOLOv5 的 Anchor-Based 耦合头变为 Anchor-Free 的解耦头，使用 DFL（distribution focal loss，DFL）损失函数来提高检测精度。

这些改进使得 YOLOv8 在速度和精度之间实现了最佳平衡，特别是 YOLOv8n 版本，因其较小的参数量和计算量，已经成为嵌入式和低成本设备上的理想选择。这一模型在保持高速度检测的同时，也实现了高精度的目标检测，这对于行人检测等应用场景尤为重要，这些场景通常依赖于边缘摄像头和无人机等设备。考虑到这些设备的硬件限制，本文选取 YOLOv8n 模型，通过优化模型结构和损失函数，提供一个有效的解决方案，以在有限的资源下实现准确的行人检测。

YOLOv8 的不断更新和改进,为深度学习的目标检测算法的未来研究和应用奠定了坚实的基础。展示了 YOLOv8 系列模型的强大潜力和广泛适用性。

2　改进的 YOLOv8 模型

面对行人检测数据集中存在的人群密集导致的严重遮挡、环境复杂导致的光线干扰以及镜头纵深导致的目标尺度多样性等挑战,导致原始 YOLOv8n 存在漏检、错检、精度低的问题。针对上述问题,本文在 YOLOv8n 的基础上改进 YOLOv8 模型,改进点如下:

1)在骨干网络引入 DCNv2 设计 C2f_DCNetv2 模块改进骨干网络的 C2f 模块,使模型对目标形变和复杂场景具有更好的适应能力,改进骨干网络的特征信息提取能力;

2)针对行人检测数据集的目标尺度多样性导致小目标信息容易被忽略的问题,改进颈部网络,增加 160×160 的具有更丰富小目标特征的特征层。四个检测头辅助进行多尺度目标检测,有效提升模型对小目标行人的感知能力;

3)受 AFPN 的启发,基于 4 个检测头和 ASFF 原理设计 ASFF-4 模块,设计针对四检测头的渐进特征金字塔,称为 AFPN-4H,有效减少特征层融合时的信息丢失以及缩小非相邻特征层之间的语义差距,提升颈部网络的多尺度特征融合能力。

4)将 Focler-IoU、Wise-IoU 和 MPDIoU 结合起来,以 WFM-IoU 作为本模型的边界框回归损失函数,提升对目标的定位准确性。改进后的模型结构如图 1 所示。

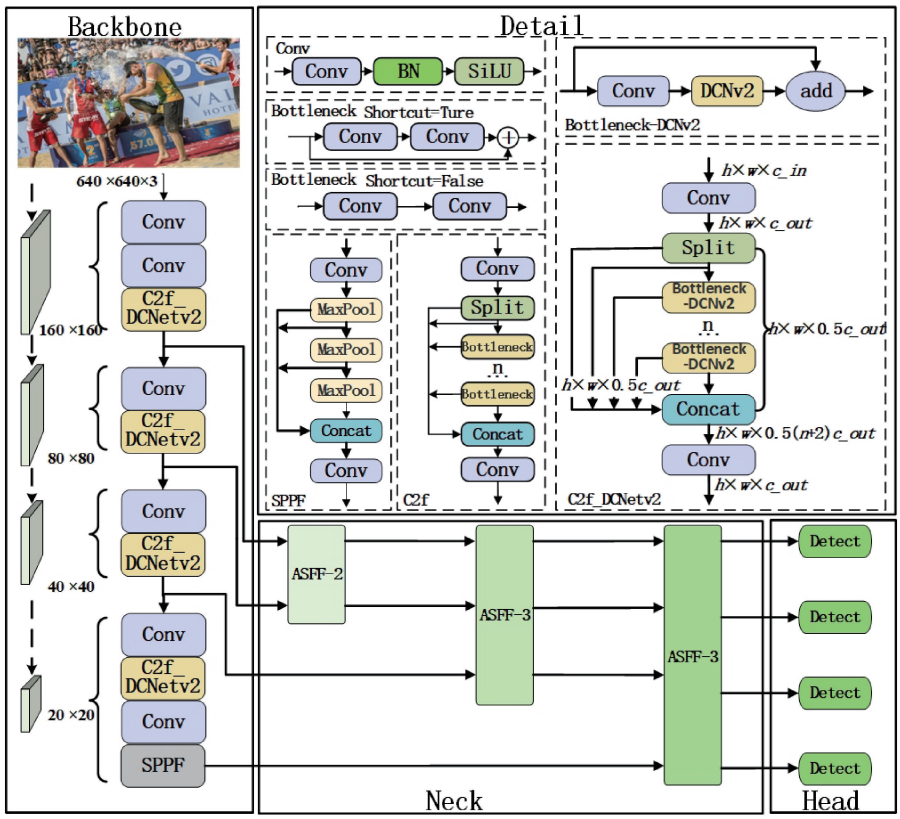


图 1　改进的 YOLOv8 网络结构

Fig. 1　Improved YOLOv8 network structure

2.1　C2f_DCNetv2 模块

行人检测任务场景复杂多变,目标密集且姿态各异,具有丰富的细节。面对上述问题,传统的卷积不能很好提取其特征信息。因此本文引入了 DCNv2 设计 C2f_DCNetv2,使模型对目标形变和复杂场景具有更好的适应能力。

可变形卷积的主要思想是在传统的卷积操作中引入可学习的偏移量,使得卷积核的位置和形状可以根据输入特征图的内容进行动态调整,从而适应不同物体的形状、大小和姿态等几何变换。其基本公式如下:

$$y(p)=\sum_{k=1}^{K} w_k \cdot x(p+p_k+\Delta p_k) \tag{1}$$

其中,$y(p)$ 是输出特征图的位置 p 处的值,w_k 是卷积核的权重,$x(p+p_k+\Delta p_k)$ 是输入特征图的位置 $p+p_k+\Delta p_k$ 处的值,p_k 是预定义的固定偏移量,Δp_k 是学习的动态偏移量。

DCNv2 是在 DCN 的基础上加入调制机制使得每个采样点不仅受到学习的偏移量的影响,还受到学习的特征幅

度的影响。

其基本公式如下：

$$y(p)=\sum_{k=1}^{K}w_k\cdot x(p+p_k+\Delta p_k)\cdot \Delta m_k \tag{2}$$

其中，Δm_k 是第 k 个位置的调制标量，其取值范围限定为[0,1]。

如图 2 所示，本文使用 DCNv2 替换 Bottleneck 中的第二个 3×3 的标准卷积，从而得到新的瓶颈模块 Bottleneck-DCNetv2。

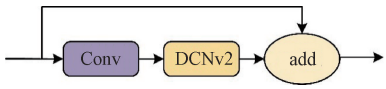


图 2　Bottleneck-DCNv2 结构图

Fig. 2　Structure of Bottleneck-DCNv2

再将 Bottleneck-DCNetv2 模块嵌入到 C2f 中得到 C2f_DCNetv2 模块，如图 3 所示。该设计使网络具备对目标形变更好适应能力并且能更准确地捕捉关键特征，提高对目标的识别和定位的准确性。

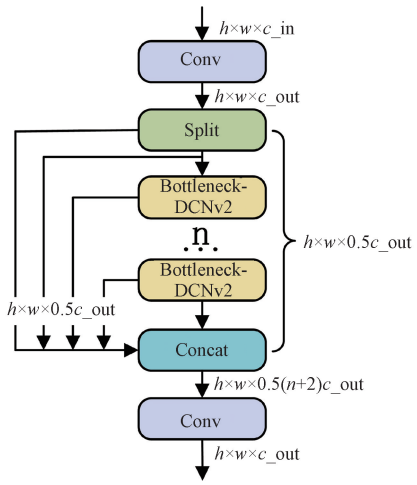


图 3　C2f_DCNetv2 模块结构图

Fig. 3　Structure of C2f_DCNetv2 module

2.2　小目标检测头

在行人检测任务中，由于人群密集和成像距离的差异，目标之间的严重遮挡以及尺寸差异导致了小目标特征信息在提取过程中的忽略，这通常会保留较大目标的关键信息，从而使小目标易于被漏检。

原始 YOLOv8 具有 3 种不同尺度(80×80、40×40 和 20×20)的特征检测层。然而，由于多倍数的下采样，小目标的特征信息仍难以保留。

为了增强对小目标的检测能力，本文提出在颈部网络中添加一个 160×160 尺度的特征图 P2，该特征图仅经过两倍的下采样(stride=2)，因此它保留了较高的分辨率和丰富的小目标特征信息。这个新的特征图增加了一个额外的金字塔层级，并引入了一个新的检测头，称为小目标检测头。这样的设计旨在提高模型对小目标的感知能力，从而减少漏检的情况。

2.3　AFPN-4H 模块

1) ASFF

ASFF 改善了单阶段目标检测器中特征金字塔的不一致性问题。在特征金字塔中，不同尺度的特征图之间存在冲突，这会干扰训练期间的梯度计算并降低特征金字塔的有效性。ASFF 允许网络自适应地学习如何在空间上过滤其他层级的无用信息，只保留有用信息进行融合，提高特征的尺度不变性。ASFF 分为两步：恒等缩放和自适应融合。

恒等缩放：不同层级的特征图通过恒等映射或通过空间变换(如上采样或下采样)被调整到相同的分辨率，确保在融合过程中，所有特征图都处于相同的空间尺度。

自适应融合：在这一过程中，模型通过网络训练自动学习确定每个特征图在融合时的权重。这些权重反映了模型根据当前任务需求和数据特性自适应调整每个特征图贡献的能力。学习到的权重随后用于加权合并不同层级的特征图，以产生最终的融合特征。

基于本文提出添加小目标检测层，在特征金字塔中需要融合 4 个不同尺度的特征图，提出了 ASFF_4。ASFF 的具体公式如下：

$$F_{ij}^{l}=\alpha_{ij}^{l}\cdot x_{ij}^{1\to l}+\beta_{ij}^{l}\cdot x_{ij}^{2\to l}+\gamma_{ij}^{l}\cdot x_{ij}^{3\to l}+\delta_{ij}^{l}\cdot x_{ij}^{4\to l} \tag{3}$$

其中，y_{ij}^{l} 表示输出特征图 y^{l} 中第 (i,j) 个位置上的向量；$x_{ij}^{n\to l}$ 表示将第 n 层级特征图调整为与第 l 层级特征图相同分辨率后得到的第 (i,j) 个位置上的向量；α_{ij}^{l}，β_{ij}^{l}，γ_{ij}^{l}，δ_{ij}^{l} 表示 4 个不同层级特征图对于第 l 层级特征图中第 (i,j) 个位置上的空间重要性权重，它们由网络自适应地学习，满足以下条件：

$$\alpha_{ij}^{l}+\beta_{ij}^{l}+\gamma_{ij}^{l}+\delta_{ij}^{l}=1,\alpha_{ij}^{l},\beta_{ij}^{l},\gamma_{ij}^{l},\delta_{ij}^{l}\in[0,1] \tag{4}$$

图 4 为 4 个不同层级特征图的自适应空间融合操作的示意图。

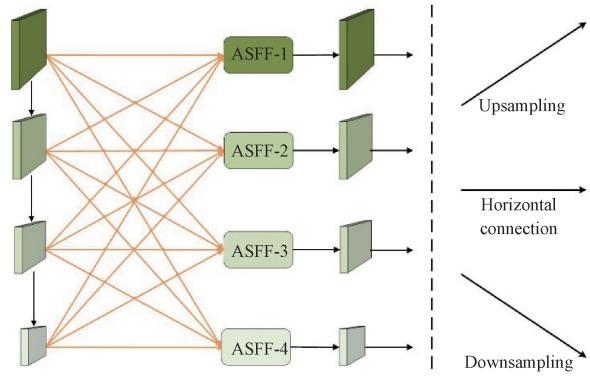


图 4　ASFF-4 网络结构

Fig. 4　ASFF-4 network structure

2) AFPN-4H

Yang 等提出 AFPN，利用 ASFF 抑制不同特征之间的

信息矛盾,保留更多有用的信息;还采用渐进式的特征层融合方法:首先融合两个相邻的低级特征,然后逐渐将高级特征纳入融合过程。这种方式避免了非相邻层次之间更大的语义鸿沟,提高了特征的表达能力。

AFPN 通过渐进式的特征融合减少了信息丢失并缩小了非相邻层间的语义差距,从而增强了颈部网络的特征融合能力。然而,AFPN 的原始设计仅考虑了 P3、P4、P5 三个特征层,未充分考虑小目标检测的需求。为了解决这一问题,本研究引入了 P2 层,创建了一个包含 4 个尺度特征图融合的渐进金字塔网络,称为四检测头渐进金字塔网络(asymptotic feature pyramid network for four detection heads,AFPN-4H)。

图 5 为 AFPN-4H 的网络结构图。在 AFPN-4H 的结构中,首先使用 ASFF 将 P2 和 P3 两个低级相邻特征图进行自适应空间融合,随后逐步将 P4 层特征纳入融合过程,并最终与 P5 层特征进行整合。这一渐进融合策略利用 ASFF 技术缓解了不同特征图之间的信息冲突,确保了融合过程中有效信息的保留,并减少了非相邻层间的语义差距。此外,通过将 P2 层纳入融合,模型对小目标的特征信息进行了更全面的整合,从而提升了对小目标的检测能力。

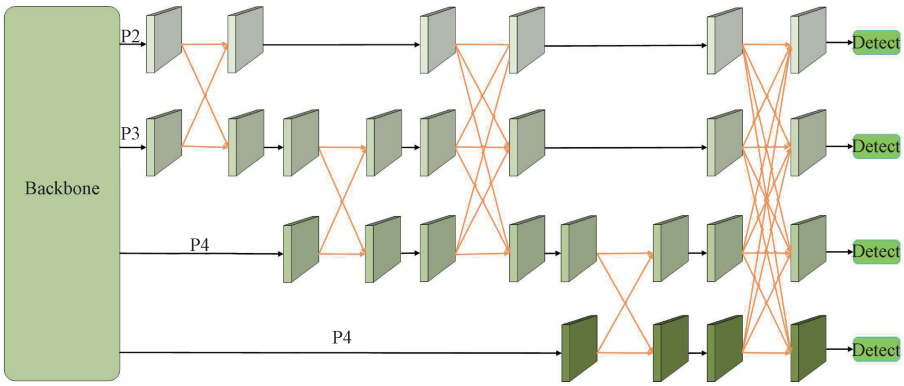


图 5 AFPN-4H 网络结构

Fig. 5 AFPN-4H network structure

2.4 改进的损失函数

损失函数用于衡量神经网络预测结果与实际期望结果之间的差异,其值越小表示预测结果越接近期望结果[25]。边界框回归则负责微调预测的边界框,使其更准确地贴合目标的真实边界框,这对提高目标检测的定位精度至关重要。交并比(intersection over union, IoU)是一种评估预测边界框与真实边界框之间重叠程度的指标。

IoU 公式如下:

$$IoU = \frac{|B^{gt} \cap B^{pred}|}{|B^{gt} \cup B^{pred}|} \tag{5}$$

其中,B^{gt} 和 B^{pred} 分别是真实边界框和预测边界框的面积。YOLOv8 采用 CIoU,考虑了边界框的中心点距离和长宽比,其公式为:

$$CIoU = IoU - \frac{d^2}{c^2} - \alpha v \tag{6}$$

其中,d 是预测框和真实框中心点的距离,c 是包含两个框的最小闭合区域的对角线长度,v 是长宽比的一致性项,α 是用于平衡 v 的参数。

Focaler-IoU 考虑了难易样本分布对回归结果的影响,通过关注不同的回归样本,可以在不同的检测任务中提高检测器的性能。为了在不同的回归样本中关注不同的检测任务,Focaler-IoU 使用线性间隔映射方法重构 IoU 损失,这有助于提高边缘回归。其公式如下:

$$\text{Focaler-IoU} = \begin{cases} 0, & IoU < d \\ \dfrac{IoU - d}{u - d}, & d \leqslant IoU \leqslant u \\ 1, & IoU > u \end{cases} \tag{7}$$

其中,u 和 d 为预设的阈值,调整可以使 Focaler-IoU 关注不同的样本。当 $IoU < d$ 时,表示无边界框重叠;当 $d \leqslant IoU \leqslant u$ 时,表示部分重叠;当 $IoU > u$ 时,表示完全重叠。其目的是在不同的 IoU 区间内调整损失函数的敏感度,以便更有效地处理不同程度的预测误差。其损失函数定义如下:

$$L_{\text{Focaler-IoU}} = 1 - \text{Focaler-IoU} \tag{8}$$

Wise-IoU 是一种基于 IoU 的边界框回归损失函数,它采用动态非单调聚焦机制来提高目标检测器的定位性能。WIoU 使用“离群度”来评估锚框的质量。这种方法可以智能地分配梯度增益,减少高质量锚框的竞争性,同时降低低质量示例产生的有害梯度。这样,WIoU 能够专注于普通质量的锚框,从而提高检测器的整体性能。

WIoU v1 使用两层注意力机制,其中第一层是距离注意力函数,第二层放大普通质量锚框的交集,并减少高质量锚框的交集。公式定义如下。WIoU v2 在 WIoU v1 中引入了单调聚焦系数 $L_{IoU}^{\gamma^*}$,降低了简单样本对损失值的影响,使模型能够专注于硬样本,提高分类性能,并使用批次中 L_{IoU} 的平均值动态归一化聚焦系数。公式定义如下:WIoU v3 使用基于离群度 β 的非单调聚焦系数,离群度 β

定义为批次内 L_{IoU} 与 L_{IoU} 平均值的比率。公式定义如下：

$$L_{WIoUv1}=R_{WIoU}\cdot L_{IoU} \tag{9}$$

$$R_{WIoU}=\exp\left(\frac{(x-x_{gt})^2+(y-y_{gt})^2}{(W_g^2+H_g^2)^*}\right) \tag{10}$$

$$L_{WIoU_{v2}}=\left(\frac{L_{IoU}^*}{L_{IoU}}\right)^{\gamma}\cdot L_{WIoU_{v1}} \tag{11}$$

$$L_{WIoU_{v3}}=r\cdot L_{WIoU_{v1}},\ r=\frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{12}$$

其中，W_g 和 H_g 是最小封闭框的尺寸上标 * 表示一个分离操作，将 W_g 和 H_g 从计算图中分离出来，用于防止 R_{WIoU} 产生阻碍收敛的梯度。$\frac{L_{IoU}^*}{L_{IoU}}$ 是锚框的离群度，而 γ 是一个超参数。当 $\beta=\delta$ 时，$r=1$。

MPDIoU 是基于最小点距离的一种新型边界框相似度比较指标，考虑相关因素，如重叠或非重叠区域、中心点距离以及宽度和高度的偏差，同时简化了计算过程：

$$MPDIoU=IoU-\frac{\rho^2(P_{1_{pred}},P_{1_{gt}})}{w^2+h^2}-\frac{\rho^2(P_{2_{pred}},P_{2_{gt}})}{w^2+h^2} \tag{13}$$

其中，$\rho^2(P_{1_{pred}},P_{1_{gt}})$ 和 $\rho^2(P_{2_{pred}},P_{2_{gt}})$ 分别是预测边界框的左上角和右下角与真实边界框对应角点之间的欧氏距离。w 和 h 分别是边界框的宽度和高度。

本文结合上述 3 种边界框回归方法，得到全新的 WFM-IoU，它不仅考虑难易样本分布对回归的影响，还利用动态非单调聚焦机制来提高目标检测器的定位性能，再通过最小化预测边界框与真实边界框的距离实现检测器的综合定位准确性。由于 WIoU 有 3 个版本，WFM-IoU 也存在对应的 3 个版本。

$$L_{WIoU-MPDIoU}=R_{WIoU}\cdot L_{MPDIoU} \tag{14}$$

$$L_{WFM-IoU}=L_{WIoU-MPDIoU}+IoU-\text{Focaler}-\text{IoU} \tag{15}$$

3　实验设计与结果分析

3.1　数据集与实验环境

为了验证新算法的效果，本文采用了 CrowdHuman 数据集[26]进行了一系列实验。该数据集共计 24 370 张图像，分为 15 000 张训练图像、4 370 张验证图像和 5 000 张测试图像。训练和验证图像中，行人实例总数接近 470 000 个，平均每幅图像包含约 23 个行人实例。这些图像涵盖了各种密集人群场景和复杂的环境。每个行人实例都标注了头部、可见身体部分和全身 3 种类型的标签，但本研究仅使用了全身标签。

实验操作系统为 ubuntu20.04，GPU 显卡使用 RTX 3090，运行 CUDA 版本为 CUDA 11.3。基于深度学习框架 Pytorch1.11，由 PyCharm Python3.8 编译。具体实验参数如表 1 所示。

表 1　训练参数设置

Table 1　Training parameter setting

参数	值	参数	值
epochs	100	optimizer	SGD
batch_size	16	close_mosaic	10
image_size	640	lr0	0.01
worker	12	weight_decay	0.000 5

3.2　评价指标

本实验采用准确率 P(Precision)、召回率 R(Recall)和平均精度(mean average precision，mAP)作为模型性能的评价指标。其公式分别为：

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

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

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

其中，TP 为模型检测的目标为正确目标数量；FP 为模型错误检测的目标数量；FN 为模型误检及漏检的数量；平均精度 mAP 是所有类别 AP 的均值，但在本实验中仅有行人一个类别，所以 mAP 即为 AP。AP50 表示 IoU 取值为 0.5。AP50:95 表示 IoU 的值从 0.5 到 0.95，步长为 0.05，然后计算这些 IoU 下的平均 AP。

3.3　特征金字塔网络的对比实验

为了验证本文提出的 AFPN-4H 模块的性能，将该改进与其他优秀的特征金字塔网络进行对比实验，实验结果如表 2 所示。

表 2　特征金字塔网络对比实验

Table 2　Comparative experiment on feature pyramid networks

YOLOv8n+model	Parameters/M	GFLOPS/G	P/%	R/%	AP50/%	AP50:95/%
YOLOv8	3 157 184	8.9	84.2	70.2	81.4	50.2
小目标检测头（P2）	3 354 128	17.4	84.6	71.8	83.0	52.0
BiFPN-P2[27]	2 105 436	7.8	84.2	70.1	81.1	50.0
RepGFPN[28]	3 436 816	9.1	84.6	69.7	81.0	50
HSFPN[29]	2 046 480	7.6	84.3	68.9	80.5	49.0
AFPN-4H	3 402 169	16.8	85.0	72.3	83.4	52.9

对表 2 数据分析可得,AFPN-4H 在各个指标上的提升较为明显,在 P、R、AP50 和 AP50:95 上分别提升 0.8%、2.1%、2.0%和 2.7%。相对于添加小目标层的 YOLOv8,AFPN-4H 的参数量和计算量都有明显的减少,证明 AFPN-4H 是轻量的模块,减少计算负担。BiFPN-P2 和 HSFPN 的参数量和计算量显著减少,但是在各评价指标都存在明显的下降。RepGFPN 效果较差,在各个指标均有小幅下降。综合分析得出本文提出的 AFPN-4H 具有更好的多尺度特征融合能力,减少了信息丢失,保留了更多关键信息,并且不增加计算负担。

3.4 损失函数对比实验

为了验证本文改进的损失函数 WFM-IoU 的效果,与 SIoU[30]、GIoU[31]、DIoU[32] 等经典损失函数进行对比实验,与 MPDIoU 和 Focaler-IoU 以及 WFM-IoU 的 v1～v3 做消融实验,直观看出改进损失函数的卓越性能。所有实验均在统一的实验环境下进行。实验结果如表 3 所示,各损失函数均未增加模型的参数量和计算量,因此不作为参考指标。

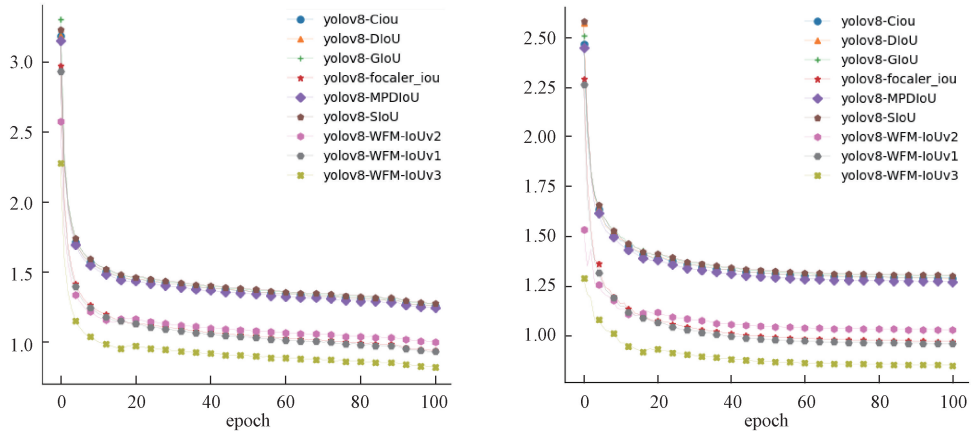
根据表 3 数据分析,相对于 CIoU,其他损失函数在精确度(P)上均有提升,其中 Focaler-IoU 实现最大增长幅度。在 R、AP50、AP50:95 等关键指标上,SIoU、GIoU、DIoU、MPDIoU、Focaler-IoU 等损失函数无明显提升,甚至部分损失函数展现出性能下降。相较之下,本文提出的 WFM-IoU v1～v3,在 P、R 和 AP50 上均有提升,特别是 WFM-IoU v3 在这些指标上的增长最为显著,分别提升了 0.6%、0.3%和 0.4%。WFM-IoU v2 在 AP50:95 上表现较差。

表 3 损失函数对比、消融实验

Table 3 Comparison of loss functions and ablation experiments

YOLOv8+损失函数		P/%	R/%	AP50/%	AP50:95/%
CIoU		84.2	69.7	81.1	49.8
SIoU		84.3	69.7	81.0	49.9
GIoU		84.3	69.7	80.9	49.9
DIoU		84.7	69.3	81.0	49.9
MPDIoU		84.4	69.7	81.0	49.9
Focaler-IoU		85.1	69.7	81.1	49.9
WFM-IoU	v1	**84.6**	**69.8**	**81.2**	**49.9**
	v2	**84.5**	**69.8**	**81.2**	49.7
	v3	**84.8**	**70.0**	**81.5**	**49.9**

图 6 展示了各损失函数的 box_loss 的变化趋势,图 6(a)和(b)分别为训练验证阶段的 box_loss 变化趋势。从图中可以明显看出,Focaler-IoU 和 WFM-IoU v1～v3 的损失值明显低于其他损失函数,证明 Focaler-IoU 和 WFM-IoU v1～v3 都有效地减小了预测边界框与真实边界框之间的差异,表明其在模型定位准确性上具有明显优势,尤其是 WFM-IoU v3 的效果最为突出。这一结果表明,WFM-IoU 损失函数能够有效提升目标检测器的定位性能。综合分析得出,本文提出的 WFM-IoU 的整体性能优于其他损失函数,其中 WFM-IoU v3 更适合本实验的数据集,因此本文选择 WFM-IoU v3 作为损失函数。



(a) train/box_loss变化趋势 (a) train/box_loss trends

(b) val/box_loss变化趋势 (b) val/box_loss trends

图 6 box_loss 变化趋势

Fig. 6 box_loss trends

3.5 模型对比实验

为了验证本文改进检测算法的性能,将该模型与经典的二阶段检测算法以及其他优秀的 YOLO 系列及其改进算法在 CrowdHuman 数据集上进行对比实验。所有实验均在统一的实验环境下进行。实验结果如表 4 所示。

分析表 4 数据可知,本文改进算法在 AP50 和 AP50:95 上优于大部分其他算法,在检测精度相当的情况下,本文改进模型具有参数量小和计算量小的优势。Fast-RCNN 算法在 AP50 上高出本文改进算法 0.8%,但是在 AP50:95 这个更加严格的指标上低 4.0%,并且 Fast-RCNN 是双阶

表 4　模型对比实验

Table 4　Model comparison experiments

模型	AP50/%	AP50:95/%	Parameters/M	GFLOPS/G
YOLOv8n(baseline)	81.1	49.8	3.0	8.9
Fast-RCNN	85.5	49.8	41.5	52
RetinaNet[33]	81.9	49.8	45.7	50.0
YOLOv5s	81.8	48.4	7.2	16.5
YOLOv5m	83.8	52.9	21.2	49.0
YOLOv3-tiny	77.0	49.3	11.6	19.1
文献[14]	83.2	51.0	11.4	30.2
Crowd-YOLOv8[17]	84.5	53.4	3.6	21.9
YOLOv8-MobileVit[34]	80.1	51.9	1.4	6.6
YOLOv8n+ours	**84.7**	**53.6**	**3.6**	**16.3**

段目标检测算法，Parameters 和 GFLOPS 远远大于本文提出算法。综合分析可以得出本文提出算法具有更好的性能，既能提供高效的检测性能，也不增加过大的计算负担。

3.6　消融实验

为了进一步直观地了解提出的各模块对 YOLOv8 网络结构的增益效果，在 CrowdHuman 数据集上进行了多组的消融实验。消融实验的结果如表 5 所示。

表 5　消融实验

Table 5　Ablation experiments

序号	YOLOv8n	C2f_DCNetv2	P2	AFPN-4H	WFM-IoUv3	Parameter/M	GFLOPS/G	P/%	R/%	AP50/%	AP50:95/%
1	√					3157 184	8.9	84.2	69.7	81.1	49.8
2	√				√	3157 184	8.9	84.8	70.0	81.5	49.9
3	√	√				3 204 512	9.1	85.2	71.4	82.2	50.5
4	√		√			3 354 128	17.4	84.6	71.8	83.0	52.0
5	√			√		3 402 169	16.8	85.0	72.3	83.4	52.9
6	√	√		√		3 558 599	16.3	85.4	73.9	84.4	53.6
7(本文)	√	√	√	√	√	3 558 599	16.3	**85.8**(↑**1.6**)	**74.2**(↑**4.0**)	**84.7**(↑**3.6**)	**53.6**(↑**3.8**)

从表 5 中可以看出，本文提出的各种改进方法都能够对 YOLOv8n 算法的性能产生积极的影响。在各个评价指标上，本文提出的算法相比于原始 YOLOv8n 算法都有提升。改进后的算法与原始 YOLOv8n 相比，对行人的检测精度(P)、召回率(R)、AP50 和 AP50:95 分别提高了 1.6%、4.0%、3.6%和 3.8%，这说明本文提出的改进算法能够有效地提高 YOLOv8n 算法在行人检测任务上的鲁棒性和准确性。

与 YOLOv8n 相比，骨干网络使用 C2f_DCNetv2 的模型在 P、R、AP50 和 AP50:95 分别提升了 1.0%、1.7%、1.1%和 0.7%。证明了 C2f_DCNetv2 对目标形变有更好适应能力并且能更准确地捕捉关键特征，提高目标的定位准确性，能够适应不同场景和不同密集程度下的目标检测，使骨干网络特征提取能力得到显著的提高。

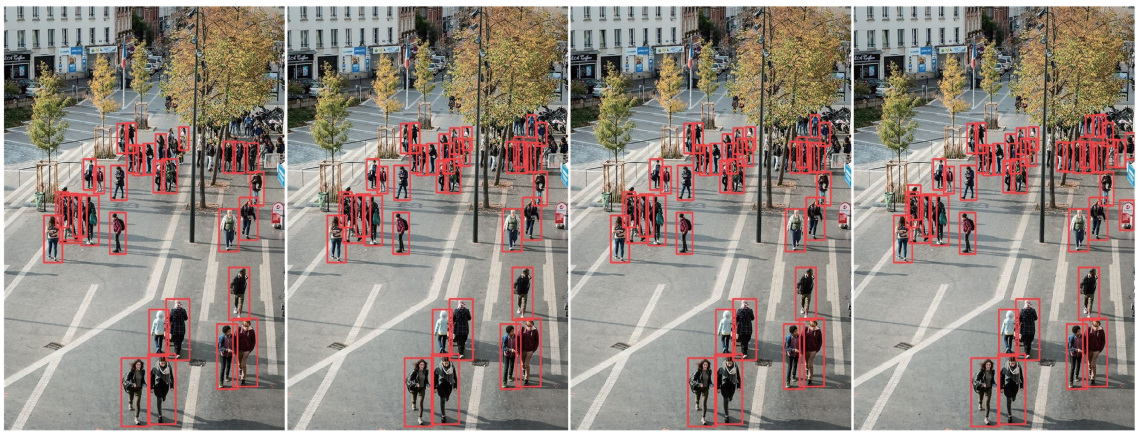
加入小目标检测头后模型的 P、R、AP50 和 AP 50:95 的值分别提升了 0.4%、2.1%、1.9%和 2.2%，证明了小目标检测头有效地捕捉到了更多小目标的特征信息，使模型对小目标的感知能力得到了提升。

颈部网络使用 AFPN-4H 后，与 YOLOv8n 相比，其 P、R、AP50 和 AP50:95 分别提升了 0.8%、2.6%、3.3%和 3.1%；相对于 P2，分别提升了 0.4%、0.5%、0.4%和 0.9%，这证明 AFPN-4H 能充分融合不同特征层的特征信息，减少特征层在融合时的信息丢失和退化，有效提升了模型对各个尺度目标的检测精度。损失函数的加入如 3.4 节所述一致，选用 WFM-IoUv3。在改进算法的基础上在 P、R 和 AP50 等指标上分别提升了 0.4%、0.3%和 0.3%。最关键的是 WFM-IoUv3 能够减小预测边界框与真实边界框的位置差距，使模型对行人的定位更加精准无误。

3.7　检测结果可视化

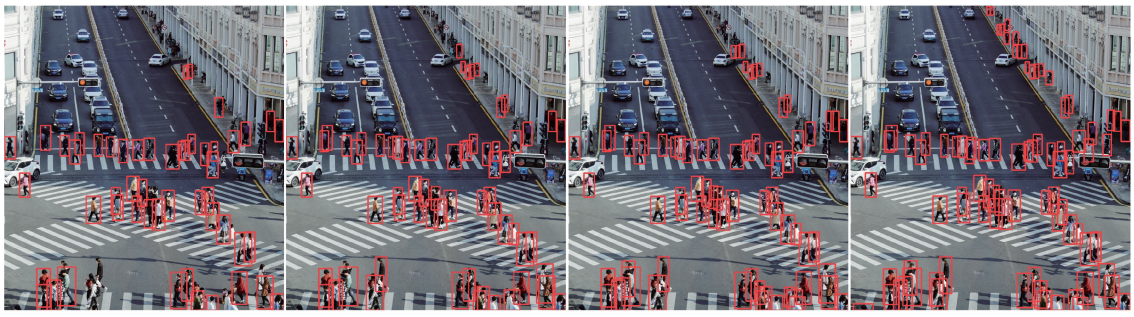
图 7 展示了在相同实验环境下，YOLOv8n 原始模型

与文献[14]、文献[17]所提算法以及本文提出的改进模型在密集、多尺度、高遮挡环境中行人的检测效果。其中,图 7(a)~(c)分别展示了场景一、场景二和场景三的 4 种模型的行人检测效果,从左至右分别是:YOLOv8n 模型、文献[14]所提模型、文献[17]所提模型以及本文改进的模型。从图中可以观察到,在处理密集和多尺度目标的检测场景任务中,本文改进的模型能够更准确地识别出更多目标,特别是小尺寸目标。



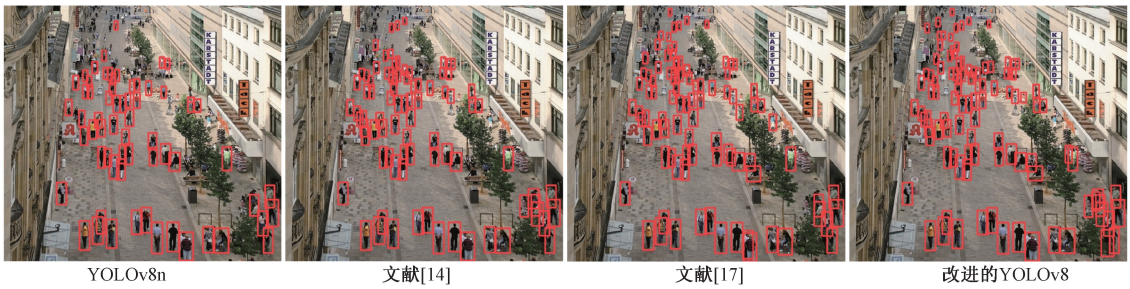
YOLOv8n　　文献[14]　　文献[17]　　改进的YOLOv8

(a) 场景一的各模型检测效果
(a) The detection effect of each model in scene one



YOLOv8n　　文献[14]　　文献[17]　　改进的YOLOv8

(b) 场景二的各模型检测效果
(b) The detection effect of each model in scene two



YOLOv8n　　文献[14]　　文献[17]　　改进的YOLOv8

(c) 场景三的各模型检测效果
(c) The detection effect of each model in scene three

图 7　检测效果对比

Fig. 7　Comparison of detection effects

在场景一中,存在人与人、人与物之间的遮挡,原始的 YOLOv8n 模型漏检了许多被遮挡的行人,文献[14]和文献[17]模型的检测效果好于 YOLOv8n,但仍然存在漏检。相比之下,本文改进模型识别出了更多被遮挡的目标以及小目标行人;在场景二中,由于道路两侧行人尺寸较小,原始的 YOLOv8 模型漏检了许多远处的小目标行人和骑电动车的人甚至较大的遮挡目标也被漏检,文献[14]和文献[17]模型检测出了大大部分大目标,但对小目标存在漏检。而改进模型则成功检测出了所有的行人目标;在场景三中,存在行人密集、目标尺寸变化大和遮挡等情况,原始的 YOLOv8n 模型、文献[14]和文献[17]的模型都漏检了部分遮挡目标和图片远端的小目标行人。与此同时,改进模型则检测出了更多行人。

综上所述,本文提出的改进模型显著提升了

YOLOv8n 在复杂环境中对密集多尺度行人的检测能力，有效地解决了原始模型在遮挡情况下的漏检问题。

4　结　　论

本文提出一种高效的面向复杂场景密集行人检测的 YOLOv8 改进模型。该模型通过引入 DCNv2 设计 C2f_DCNetv2，有效提升了骨干网络的特征提取能力，增强了网络对于复杂图像内容的理解；通过在架构中加入 P2 层作为新的小目标检测头，增强模型对小尺寸目标的检测能力提高了对小目标的检测识别精度；基于四检测头改进 AFPN-4H，使其能够有效融合四层特征层，从而提升了颈部网络的多尺度特征融合能力。不仅优化了特征层之间的信息流动，还提高了模型对不同尺度目标的适应性和检测精度；通过结合 Wise-IoU、Focaler-IoU 和 MPDIoU 得到的 WFM-IoU，综合提升了模型的检测能力，尤其是在目标定位准确性方面取得了显著进步。通过实验对比，该模型表现优秀，与原始的 YOLOv8n 相比，在 P、R、AP50 以及 AP50:95 等关键指标上分别提升 1.6、4.0、3.6 和 3.8 个百分点，具有很强的应用前景。

该模型主要是针对正常天气情况下的行人检测，没有对大雾、雨雪等恶劣天气条件下进行实时性的测试，未来的研究可以尝试完善模型功能实现恶劣天气情况下的高效检测以及轻量化模型，以便在边缘设备上运行高速、高精度、低功耗的行人检测模型。

参考文献

[1] HARIHARAN B, ARBELAEZ P, GIRSHICK R, et al. Simultaneous detection and segmentation[C]. Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13. Springer International Publishing, 2014: 297-312.

[2] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C]. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005,1: 886-893.

[3] FELZENSZWALB P, MCALLISTER D, RAMANAN D. A discriminatively trained, multiscale, deformable part model[C]. 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008: 1-8.

[4] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning ap-plied to document recognition[J]. IEEE, 1998,86(11):2278-2324.

[5] 王永生，姬嗣愚. 基于深度学习的目标检测算法综述[J]. 计算机与数字工程，2023,51(6):1231-1237.
WANG Y SH, JI S Y. Review of target detection algorithms based on deep learning[J]. Computer & Digital Engineering, 2023,51(6):1231-1237.

[6] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580-587.

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

[8] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks [C]. Advances in Neural Information Processing Systems 28,2015.

[9] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]. IEEE International Conference on Computer Vision, 2017:2961-2969.

[10] 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.

[11] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7263-7271.

[12] REDMON J, FARHADI A. YOLOv3: An incremental improvement[J]. ArXiv preprint arXiv: 1804.02767,2018.

[13] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot Multibox detector[C]. In Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I 14 2016. Springer International Publishing, 2014: 21-37.

[14] 胡倩，皮建勇，胡伟超，等. 基于改进 YOLOv5 的密集行人检测算法研究[J/OL]. 计算机工程，1-12[2024-05-13]. https://doi. org/10.19678/j. issn. 1000-3428.0068753.
HU Q, PI J Y, HU W CH, et al. Research on dense pedestrian detection algorithm based on lmproved YOLOv5[J/OL]. Computer Engineering: 1-12[2024-05-13]. https://doi. org/10.19678/j. issn. 1000-3428.0068753.

[15] 宁爽，宋辉. 帧间方向梯度直方图特征关联的行人检测方法[J/OL]. 电子测量与仪器学报，1-7[2024-07-22]. http://kns. cnki. net/kcms/detail/11.2488. TN.20240523.1757.018.html.
NING SH, SONG H. Pedestrian detection method based on inter-frame directional gradient histogram feature correlation [J/OL]. Journal of Electronic Measurement and Instrumentation, 1-7[2024-07-22].

http://kns.cnki.net/kcms/detail/11.2488.TN.20240523.1757.018.html.

[16] WANG C Y, BOCHKOVSKIY A, LIAO H Y. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023:7464-7475.

[17] 黄昆,齐肇建,王娟敏,等.基于改进 YOLOv8 的密集行人检测模型[J/OL].计算机工程,1-11[2024-10-22]. https://doi.org/10.19678/j.issn.1000-3428.0069026.
HUANG K, QI ZH J, WANG J M, et al. A pedestrian aggregation detection model based on animproved YOLOv8[J/OL]. Computer Engineering: 1-11[2024-10-22]. https://doi.org/10.19678/j.issn.1000-3428.0069026.

[18] 张福豹,吴婷,赵春峰,等.基于弱光增强与 YOLO 算法的锯链缺陷检测方法[J].电子测量技术,2024,47(6):100-108.
ZHANG F B, WU T, ZHAO CH F, et al. Saw chain defect detection system based on low-light Enhancement and YOLO algorithm[J]. Electronic Measurement Technology, 2024, 47(6): 100-108.

[19] ZHU X, HU H, LIN S, et al. Deformable ConvNets v2: More deformable, better results[C]. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 9308-9316.

[20] LIU S, HUANG D, WANG Y. Learning spatial fusion for single-shot object detection[J]. ArXiv preprint arXiv:1911.09516, 2019.

[21] YANG G, LEI J, ZHU Z, et al. AFPN: Asymptotic feature pyramid network for object detection[C]. 2023 IEEE International Conference on Systems, Man, and Cybernetics(SMC). IEEE, 2023: 2184-2189.

[22] ZHANG H, ZHANG S. Focaler-IoU: More focused intersection over union loss[J]. ArXiv preprint arXiv: 2401.10525, 2024.

[23] TONG Z, CHEN Y, XU Z, et al. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism[J]. ArXiv preprint arXiv: 2301.10051, 2023.

[24] SILIANG M, YONG X. Mpdiou: A loss for efficient and accurate bounding box regression[J]. ArXiv preprint arXiv:2307.07662, 2023.

[25] 王亚鹏,韩文花.改进 YOLOv5 算法下的无人驾驶道路行人识别研究[J].国外电子测量技术,2024,43(6):170-178.
WANG Y P, HAN W H. Pedestrian recognition research on unmanned roads with improved YOLOv5 algorithm[J]. Foreign Electronic Measurement Technology, 2024, 43(6): 170-178.

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

[27] TAN M, PANG R, LE Q V. Efficientdet: Scalable and efficient object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10781-10790.

[28] XU X, JIANG Y, CHEN W, et al. Damo-yolo: A report on real-time object detection design[J]. ArXiv preprint arXiv:2211.15444, 2022.

[29] CHEN Y, ZHANG C, CHEN B, et al. Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases[J]. Computers in Biology and Medicine, 2024, 170: 107917, DOI: 10.1016/j.compbiomed.2024.107917.

[30] GEVORGYAN Z. SIoU loss: More powerful learning for bounding box regression[J]. ArXiv preprint arXiv:2205.12740, 2022.

[31] REZATOFIGHI H, TSOI N, GWAK J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 658-666.

[32] ZHENG Z, WANG P, LIU W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]. AAAI Conference on Artificial Intelligence, 2020, 34(7): 12993-13000.

[33] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]. IEEE International Conference on Computer Vision, 2017: 2980-2988.

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

作者简介

胡伟超,硕士研究生,主要研究方向为计算机视觉、目标检测。

E-mail:2789336014@qq.com

皮建勇(通信作者),副教授,博士,主要研究方向为类脑智能、分布式计算。

E-mail:pijianyong@139.com