

双分流多尺度注意力与单层输出的 密集行人检测算法

金 静, 韩子阳, 翟凤文

(兰州交通大学 电子与信息工程学院, 兰州 730070)

摘 要: 针对密集行人检测因遮挡和小尺度目标导致的漏检与误检问题, 提出了一种基于双分流多尺度注意力与单层输出的密集行人检测算法 DSS R-CNN。首先, 设计了一种双分流多尺度注意力模块 (DBMAB), 将其嵌入 CloFormer 主干网络中以增强特征提取能力, 使模型更关注重要信息, 从而提升检测精度; 其次, 针对特征金字塔网络多层输出对 Sparse R-CNN 贡献有限的问题, 提出了一种基于特征调制模块 (FMB) 的单层输出颈 (SLON), 在保持检测精度的同时提高了推理速度; 最后, 采用 Focal-EIoU 损失函数进一步优化模型训练。在 CrowdHuman 数据集上的实验结果表明, 本文算法达到了 90.4% 的 AP、96.2% 的 Recall 和 44.1% 的 MR-2, 检测速度保持在 46 f/s, 优于现有密集行人场景下的检测算法。

关键词: 密集行人检测; Sparse R-CNN; CloFormer; 多尺度注意力; 单层输出颈

中图分类号: TP391 **文献标志码:** A **文章编号:** 1671-5497(2026)09-2476-13

DOI: 10.13229/j.cnki.jdxbgxb.20250284

Dense pedestrian detection algorithm with dual-branch multi-scale attention and single-layer output

JIN Jing, HAN Zi-yang, ZHAI Feng-wen

(School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China)

Abstract: Aiming at the problem of missed detection and false detection caused by occlusion and small-scale targets in dense pedestrian detection, the paper proposes a dense pedestrian detection algorithm DSS R-CNN (Dual-branch Single-layer Sparse R-CNN) based on dual-branch multi-scale attention and single-layer output. Firstly, a Dual-Branch Multi-scale Attention Block (DBMAB) is designed and embedded into the CloFormer backbone network to enhance the feature extraction ability, so that the model can pay more attention to important information, so as to improve the detection accuracy. Secondly, aiming at the problem that the multi-layer output of feature pyramid network has limited contribution to Sparse R-CNN, a Single-layer Output Neck (SLON) based on Feature Modulation Block (FMB) is proposed. It improves the inference speed while maintaining the detection accuracy. Finally, Focal-EIoU loss function is used to further optimize the model training. Experimental results on the CrowdHuman

收稿日期: 2025-04-03.

基金项目: 甘肃省高校教师创新基金项目 (2025B-060); 宁夏自然科学基金项目 (2023AAC03741); 甘肃省科技计划项目重点研发计划项目-工业类 (23YFGA0047).

作者简介: 金静 (1982-), 女, 副教授, 博士. 研究方向: 模式识别, 图像处理. E-mail: 28089092@qq.com

dataset show that the proposed algorithm achieves 90.4% AP, 96.2% Recall and 44.1% MR-2, and the detection speed remains at 46 f/s, which is better than the existing pedestrian detection algorithms in complex scenes.

Key words: dense pedestrian detection; Sparse R-CNN; CloFormer; multi-scale attention; single-layer output neck

0 引言

行人检测被广泛应用于自动驾驶、安全监控及人机交互等场景^[1]。随着深度学习技术的快速发展,行人检测算法已经取得了显著的进展,突破了传统算法所面临的一些限制。目前主流的基于深度学习的行人检测方法主要包括两阶段检测算法及单阶段检测算法,如 Faster R-CNN(Faster region-based convolutional neural network)^[2]、Cascade R-CNN^[3]、Mask R-CNN^[4]、Sparse R-CNN^[5]等属于两阶段检测算法,YOLO(You only look once)系列算法^[6-11]、SSD(Single shot multi box detector)^[12]系列算法、ViT^[13]以及 Swin-Transformer^[14]等属于后者。行人检测在密集人群场景中的应用面临诸多挑战,例如行人之间的互相遮挡、小尺度、低分辨率及复杂多变的背景环境等。其中,行人目标尺度多变的问题尤为明显,尤其是小目标包含的特征信息较少,难以进行特征提取,并且在训练阶段难以覆盖不同尺度的行人导致模型泛化能力下降,对较小的行人目标产生错误预测。而且,密集行人之间更易发生互相遮挡,容易导致检测精度的下降。

针对以上问题,Li等^[15]设计了一种新的遮挡感知检测头,包括3个独立的中心预测分支,并结合尺度和偏移量预测分支,以捕捉对被遮挡行人的硬训练样本的高层次理解;Wu等^[16]提出了一种融合多尺度特征融合模块和注意力机制的网络。为了增强语义特征的多尺度表示并灵活建模目标尺度变化,在主干网络中引入了隐式检测头(implicit detection heads, IDHs),帮助网络学习隐式知识,提高检测精度。柳长源等^[17]提出了一个基于多阶段提议稀疏区域卷积网络,在 Sparse R-CNN网络中设计了两个动态检测头,第一个负责不同尺度和空间的特征提取,第二个负责匹配可学习的提议特征,极大地提高了检测效率与特征的学习,可有效应对密集行人等复杂场景。Chen等^[18]提出了一种场景感知行人检测框架,用

于解决密集场景下的严重遮挡问题,通过显式建模场景上下文信息(如人群密度图与遮挡关系图),在检测头中引入 Scene-Aware Attention 模块,使网络能动态抑制背景干扰并增强被遮挡行人特征,从而在密集人群中实现更高的召回率与定位精度。Park等^[19]提出了一种融合注意力网络,联合利用稠密 SIFT 光流图像与 LiDAR 点云进行 3D 行人检测,为弥补单一模态的缺陷,设计了双流编码器分别提取光流运动特征与点云几何特征;随后通过跨模态 Fusion-Attention 模块实现两种特征的加权融合,使网络在光照变化或点云稀疏条件下仍保持鲁棒的 3D 行人检测性能。然而,现有方法仍存在一些不足:首先,遮挡感知检测头的设计虽然增强了对遮挡行人的理解,但其多分支结构可能导致计算复杂度较高,且在极端遮挡情况下性能仍有提升空间;其次,多尺度特征融合与注意力机制虽然提升了语义特征的表达能力,但可能过度依赖局部特征,忽略了全局上下文信息;此外,多阶段提议稀疏区域卷积网络虽然通过动态检测头提高了效率,但其性能高度依赖于提议生成的质量,而在复杂场景中提议生成可能不够准确;自适应提议生成机制虽然动态调整了提议分布,但其规则可能缺乏对复杂场景的充分适应性;最后,多尺度特征融合技术虽然增强了多尺度检测能力,但其残差结构和反卷积方法可能增加了模型复杂度,影响实时性。

针对 Sparse R-CNN 的不足,本文基于 Sparse R-CNN 模型提出一种基于双分流多尺度注意力与单层输出的密集行人检测算法 DSS R-CNN。首先设计了一个双分流多尺度注意模块 DBMAB,将其集成到 CloFormer 模型中作为本文模型的主干网络,同时对 CloFormer 模型中 Clo Block、ConvFFN 进行改进,这使改进后的模型更容易定位到图像中的感兴趣区域。然后,使用融合后的特征送入单层输出颈网络进行检测。观察到单尺度特征图上的可学习提议的大小可以覆盖多尺度特征图上的大多数锚衍生提议,使用

单层输出颈 SLON 代替 FPN, 加快推理速度时不牺牲检测性能, 同时在 SLON 的每一层加入特征调制模块 FMB, 以更好地融合不同层级的特征图。最后, 改进了边界框回归损失函数, 引入 Focal-EIoU 损失函数, 精确衡量预测框与真实框的重叠程度, 提升模型训练效果。通过上述改进, 本文方法在保持高效检测速度的同时, 进一步提升了模型在密集行人这一类复杂场景中对遮挡和小目标行人的检测精度, 解决了现有方法的不足。

1 Sparse R-CNN 模型介绍

Sparse R-CNN 是一种高效且创新的目标检测框架, 其核心架构由主干网络(Backbone)、特征金字塔网络(FPN)、动态实例交互头(Dynamic instance interactive head)以及两个特定于任务的预测层组成。与传统的目标检测方法不同, Sparse R-CNN 接收 3 种输入: 输入图像本身、一组稀疏的可学习提议框(Learnable proposal boxes)以及与之对应的提议特征(Proposal features)。其中, 提议框和提议特征是可学习的, 并且能够与网络中的其他参数一起进行端到端的优化, 从而在训练过程中动态调整以更好地适应目标检测任务。

在检测过程中, 主干网络首先对输入图像进行特征提取, 生成具有丰富语义信息的特征图。随后, 特征金字塔网络在此基础上构建多尺度的特征金字塔, 通过融合不同层级的特征图, 进一步增强对不同大小目标的特征表达能力。这些多尺度特征图被传递到动态实例交互头, 与稀疏的提议框和提议特征进行深度融合。动态实例交互头通过一系列迭代的特征交互和更新机制, 逐步优化提议框的特征表示, 使其能更准确地捕捉目标的语义和位置信息。最终, 经过动态交互优化的

对象特征被传递到两个特定于任务的预测层: 分类层负责预测目标的类别, 而回归层则对提议框的边界框进行精调, 最终实现目标的定位。

2 DSS R-CNN 模型介绍

本文提出的 DSS R-CNN 是基于 Sparse R-CNN 框架的改进模型, 旨在解决密集行人检测中的遮挡、小尺度目标检测以及复杂场景下的特征提取问题。针对 Sparse R-CNN 在密集场景中存在的局限性, DSS R-CNN 引入了双分流多尺度注意力机制(Dual-branch multi-scale attention block, DBMAB)、单层输出颈(Single-layer output neck, SLON)以及改进的边界框回归损失函数 Focal-EIoU。DSS R-CNN 算法框架如图 1 所示。以下详细介绍 DSS R-CNN 的核心改进模块及其设计思路。

2.1 基于 DBMAB 的 CloFormer 主干网络

Sparse R-CNN 的主干网络是 ResNet, 而 CloFormer^[20]相比 ResNet 具有更好的特征选择能力, 本文通过集成 DBMAB 注意力的 CloFormer 网络进一步提升模型整体性能。CloFormer 有 4 个阶段, 主要包括 Clo Block、ConvFFN、Convolution stem 和 DBMAB。输入图像首先经过 Convolution stem 来获得 Token, 然后经过一系列的 Clo Block 和 ConvFFN 提取特征。基于 DBMAB 的 CloFormer 主干网络整体架构图如图 2 所示。

2.1.1 改进的 CloFormer 模型

CloFormer 是针对图像任务优化的改进的 Transformer 模型, 适用于图像分类、目标检测和语义分割等领域。CloFormer 的核心创新在于其独特的 Clo block 设计。该模块在标准卷积操作中融入了注意力机制, 实现了跨位置特征的有效交互与深度融合, 提升了模型的特征提取能力。

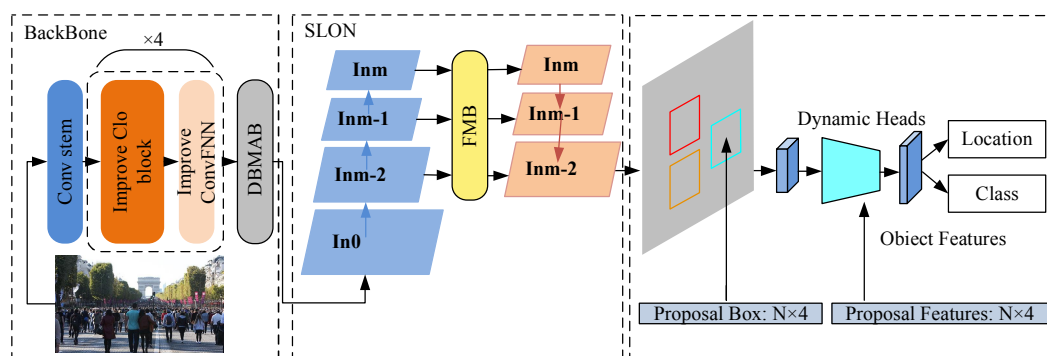


图 1 DSS R-CNN 算法框架

Fig. 1 Framework of DSS R-CNN algorithm

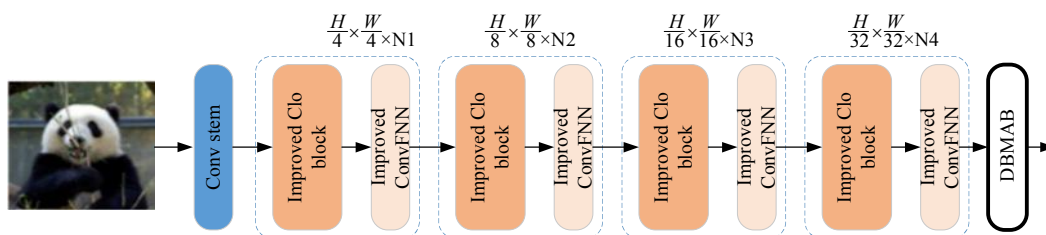


图2 基于DBMAB的CloFormer主干网络示意图

Fig. 2 Schematic diagram of the CloFormer backbone network based on DBMAB

为进一步增强网络性能,CloFormer还采用了优化的归一化策略和卷积前馈网络(Conv FFN)结构,这些改进有效提升了模型的表征能力和学习效率。

在CloFormer的设计中,深度可分离卷积^[21](DWConv)被广泛应用于Clo block和Conv FFN模块,以降低计算复杂度并提升效率。然而,DWConv在处理复杂特征时存在局限,如通道间交互能力较弱,可能影响全局信息捕捉。为此,本文采用SCConv(Split channel convolution)^[22]改进Clo block和Conv FFN模块,将DWConv替换为SSConv,从而加速计算并增强特征表示能力,如图3红色框所示,彩图参见电子版,下同。

2.1.2 双分流多尺度注意力模块DBMAB

为提升特征提取的效率和效果,本文提出双分流多尺度注意力块,该注意力机制融合了改进的多尺度大核注意力^[23](Multi-scale large kernel attention, MLKA)和门控空间注意力单元(Gated spatial attention Unit, GSAU)。

DBMAB的结构如图4所示。输入特征X首先经过层归一化(Layer normalization, LN)处理

后被送入MLKA模块,通过大核卷积来捕捉不同尺度的图像特征,整合全局和局部信息。MLKA生成的特征图随后进入GSAU模块,通过门控机制优化特征的空间分布,减少计算复杂度。经过GSAU处理后的特征图与输入特征进行逐元素相乘,并通过逐点卷积(Pointwise convolution)保持维度一致,最终输出处理后的特征图。具体公式如式(1)所示:

$$\begin{cases} N = \text{LN}(X) \\ X = X + \lambda_1 f_3(\text{MLKA}(f_1(N)) \otimes f_2(N)) \\ N = \text{LN}(X) \\ X = X + \lambda_2 f_6(\text{GSAU}(f_4(N), f_5(N))) \end{cases} \quad (1)$$

式中: $\text{LN}(\cdot)$ 和 λ 分别、层归一化和可学习的缩放因子; $\otimes$ 和 $f_i(\cdot)$ 分别表示保持维度的逐元素乘法和第 i 个逐点卷积。

(1)多尺度大核注意MLKA

多尺度大核注意力机制通过自适应权重分配,使网络聚焦于关键特征区域,同时抑制非重要信息的干扰。现有的密集行人检测模型通常采用通道注意力和自注意力,但这些方法无法同时捕

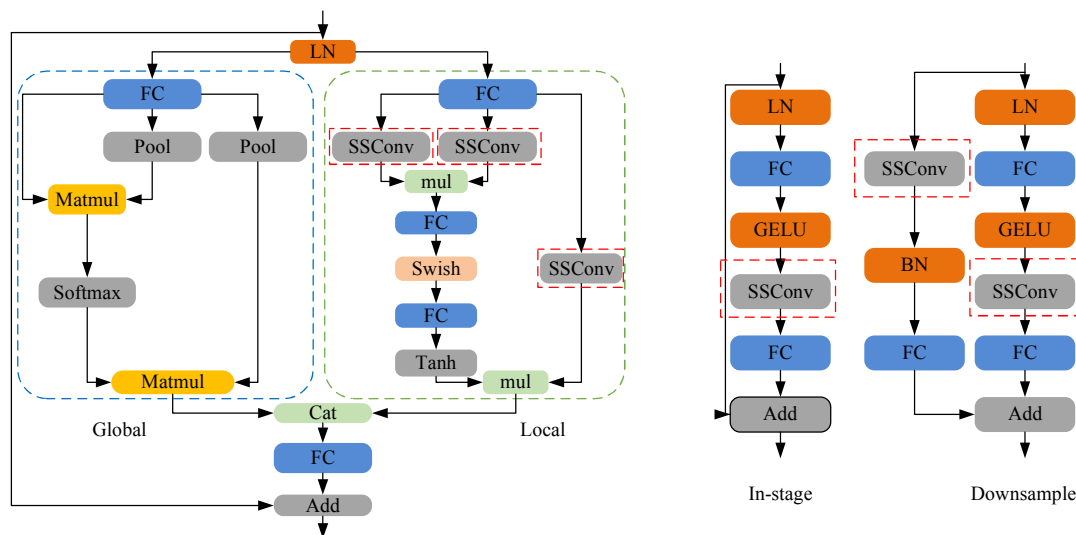


图3 改进的Clo block和Conv FFN模块示意图

Fig. 3 Schematic diagram of the improved Clo block and Conv FFN module

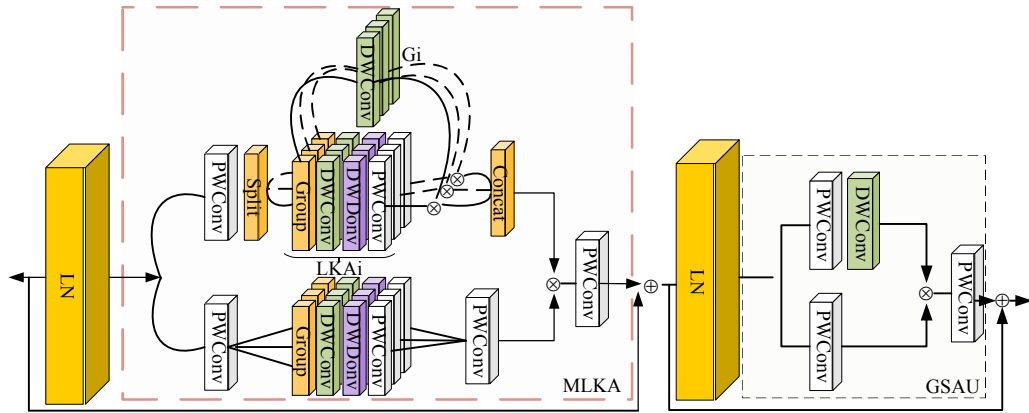


图4 DBMAB模块架构图

Fig. 4 DBMAB module architecture diagram

捉局部信息和长距离相关性,并且通常只考虑固定位置的注意力特征图。因此,本文改进了MLKA,将大核分解和多尺度学习相结合来解决上述问题。同时,为了进一步提升MLKA的特征提取能力,在只有点卷积(PWConv)的支路中也引入了LKA,并与上路分支进行特征融合。这一改进不仅增强了局部特征的表达能力,还进一步提升了模型对多尺度信息的捕捉能力,从而在密集行人检测任务中表现出更强的鲁棒性和精度。MLKA由以下3个主要部分组成:

大核注意力(Large kernel attention, LKA): 给定输入特征映射 $X \in \mathbb{R}^{C \times H \times W}$, LKA将 $K \times K$ 卷积分解为深度卷积 $f_{\text{DW}}(\bullet)$ 、深度扩张卷积 $f_{\text{DWD}}(\bullet)$ 和点向卷积 $f_{\text{PW}}(\bullet)$,如式(2)所示:

$$\text{LKA}(X) = f_{\text{PW}}(f_{\text{DWD}}(f_{\text{DW}}(X))) \quad (2)$$

多尺度机制(Multi-scale mechanism): 为了学习具有全尺度信息的注意力图,使用分组多尺度机制来增强LKA。假设输入特征映射 $X \in \mathbb{R}^{C \times H \times W}$,该模块首先将其分割为 n 个块 $X_1, X_2, \dots, X_n$ 。对于第 i 组特征 X_i ,使用分解的LKA生成均匀尺度的注意力图 LKA_i 。通过这种方式,模型可以在不同尺度上生成注意力图,从而捕捉到不同尺度的特征信息。

门控聚合(Gated aggregation): 尽管较大的LKA捕获了更宽的像素响应,但其生成的注意力图中可能出现块效应。为了避免块效应并学习更多局部信息,利用空间门控通过以下方式将 $\text{LKA}_i(\bullet)$ 动态地适应为 $\text{MLKA}_i(\bullet)$,如式(3)所示:

$$\text{MLKA}_i(X_i) = G_i(X_i) \otimes \text{LKA}_i(X_i) \quad (3)$$

式中: $G_i(\bullet)$ 为由 $a_i \times a_i$ 深度卷积生成的第 i 个门控; $\text{LKA}_i(\bullet)$ 为分解后的LKA。

该方法消除了块效应,使具有较大感受野的 MLKA_i 更多地对长距离依赖性作出反应,而较小的 MLKA_i 则倾向于保持局部纹理。

(2) 门控空间注意单元GSAU

在Transformer块中,前馈网络(Feed-forward network, FFN)是增强特征表示的重要组成部分。然而,传统的具有中间层的MLP对于Sparse R-CNN网络来说过于复杂,特别是对于大尺度图像输入。因此,将空间注意力和门控线性单元相结合,构建了GSAU模块,以实现自适应门控机制并减少参数和计算。该设计不仅实现了自适应特征选择,还显著降低了模型复杂度和计算开销。

为了更有效地捕获空间信息,采用单层深度卷积来加权特征图。给定密集变换的 X 和 Y , GSAU的过程如式(4)所示:

$$\text{GSAU}(X, Y) = f_{\text{DW}}(X) \otimes Y \quad (4)$$

式中: $f_{\text{DW}}(\bullet)$ 和 $\otimes$ 分别表示深度卷积和元素乘法。

通过引入空间门控机制, GSAU去除了非线性层来降低计算复杂度并捕获局部空间信息。

MLKA通过大核卷积和多尺度机制,有效整合了全局和局部信息,解决了传统方法在处理遮挡和尺寸变化时的局限性。GSAU则通过门控机制优化了空间信息的整合,减少了计算复杂度,同时增强了特征的表达能力。DBMAB模块的这些特性使其能够更准确地捕捉行人特征,提高检测精度,尤其在行人密集、尺寸多样和背景复杂的场景中表现突出。

2.2 基于FMB的单层输出颈-SLON

特征金字塔网络(FPN)在目标检测中具备两大功能:多尺度特征融合和多尺度检测。多尺

度特征融合能够从不同尺度的特征图中编码出丰富的上下文信息。它将特征图划分为多个尺度进行检测,其优点是可以为目标分配一个大小相似的候选框。本文研究了 Sparse R-CNN 中这两个功能的贡献,发现仅保留多尺度特征融合对平均精度的影响可以忽略不计,但仅保留多尺度检测平均精度将会大幅度下降。这说明多尺度的检测所起作用较小。其原因在于:传统检测器中,锚点生成的候选框(Proposals)受限于固定尺度和单层特征图,因此需要依赖多层特征图来覆盖不同尺度的目标。而 Sparse R-CNN 的可学习提议框通过自适应调整大小,能够直接匹配单层特征图中的多尺度目标,从而大幅降低了对多尺度特征图的依赖性。因此,设计了图 5(b)中的单层输出颈 SLON,过程如式(5)所示:

$$\begin{cases} F_i = \text{Conv}(In_i) + F_{i+1} \uparrow, i = m-1, m-2 \\ \text{Out} = \text{Conv}(F_{m-2}) \end{cases} \quad (5)$$

式中: m 表示 FPN 的顶层; $m-2$ 表示 FPN 的第 3 高层。

SLON 只有一个输出层,并且需要一个 RoI 头来生成候选框,从而在不降低性能的情况下减少了其他输出层的计算。

对于单层输出颈 SLON,在其主干网络输出的特征后引用了 FMB^[24]模块,这可以使网络更好地对特征进行提取。FMB 由 SMFA 和 PCFN 两个模块组成,如图 6 所示。以下将详细介绍 FMB 模块。

2.2.1 自调节特征融合模块 SMFA

非局部特征能捕捉图像中较大范围或远距离像素之间的相互作用,反映图像的全局上下文信

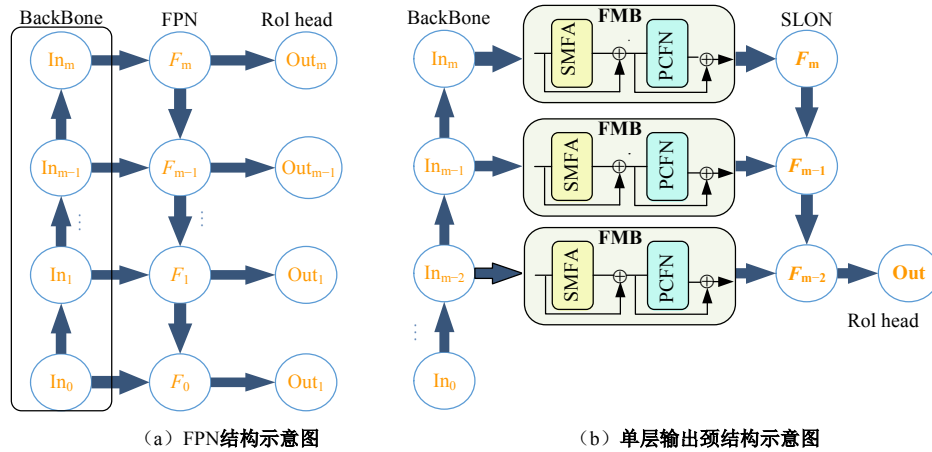


图 5 不同颈部结构的图解

Fig. 5 Illustration of the different neck structures

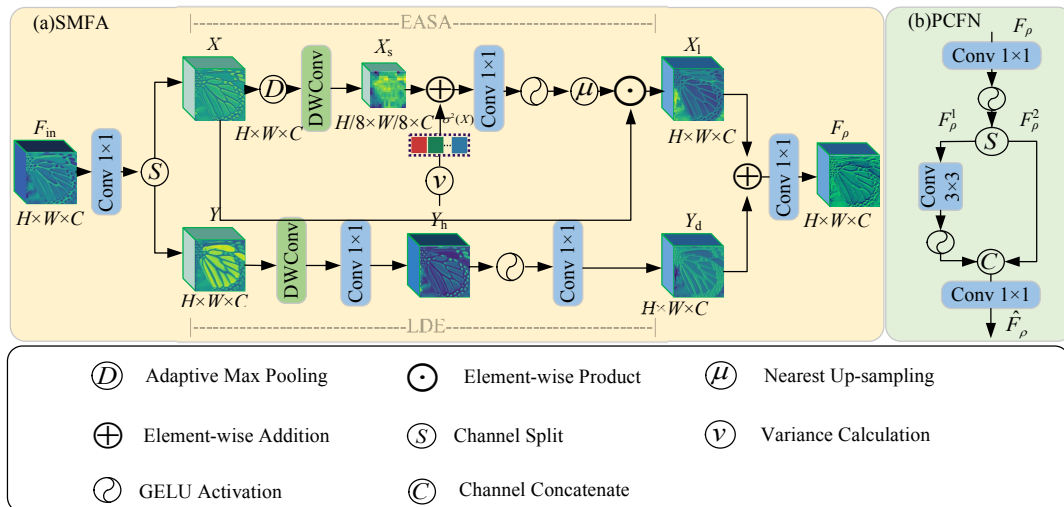


图 6 FMB 模块

Fig. 6 FMB module

息和长距离依赖性,对于中小目标的检测具有显著优势。SMFA 模块通过轻量级的设计,协同建模局部和非局部特征,具体包括两个分支:EASA (Efficient approximation of self-attention)分支和LDE(Local detail estimation)分支。

给定输入特征 $F_{in} \in \mathbb{R}^{H \times W \times C}$,其中 $H \times W$ 表示空间大小, C 是通道数,首先通过 1×1 卷积扩展通道数,并将其分为两部分,作为 EASA 和 LDE 分支的输入,如式(6)所示:

$$\{X, Y\} = S(\text{Conv}_{1 \times 1}(\|F_{in}\|_2)) \quad (6)$$

式中: $\|\cdot\|_2$ 是 L2 归一化; $\text{Conv}_{1 \times 1}(\cdot)$ 表示 1×1 卷积层; $S(\cdot)$ 表示通道分割操作。

其次,EASA 和 LED 分支分别处理特征 X 和 Y ,生成非局部特征 X_l 和局部特征 Y_d 。最后,通过元素加法将 X_l 和 Y_d 融合,并通过 1×1 的卷积生成 SMFA 模块的输出。该过程如式(7)所示:

$$F_\rho = \text{Conv}_{1 \times 1}(X_l + Y_d) \quad (7)$$

式中: $F_\rho \in \mathbb{R}^{H \times W \times C}$ 是输出特征。

2.2.2 基于部分卷积的前馈网络 PCFN

常规前馈网络(FFN)对每个像素位置进行相同的操作,这缺乏空间维度上的信息交换。因此改进了 FFN,提出了一种有效的基于部分卷积的前馈网络(PCFN),以进一步细化从 SMFA 中获得的代表性特征。

如图 6(b)所示,PCFN 模块首先通过 1×1 卷积配合 GELU 激活函数在扩展的隐藏特征空间中进行跨通道特征交互。随后,该模块将隐藏特征沿通道维度分割为两个子块 $\{F_\rho^1, F_\rho^2\}$,其中子块 F_ρ^1 经过 3×3 卷积和 GELU 激活函数处理以编码局部上下文信息。最后,将处理后的子块 F_ρ^1 与子块 F_ρ^2 进行通道拼接,并通过 1×1 卷积实现特征融合,同时将通道维度缩减至原始输入尺寸。该过程如式(8)所示。

$$\begin{cases} \{F_\rho^1, F_\rho^2\} = S(\phi(\text{Conv}_{1 \times 1}(\|F_\rho\|_2))) \\ \hat{F}_\rho = \text{Conv}_{1 \times 1}(C([\phi(\text{Conv}_{3 \times 3}(F_\rho^1)), F_\rho^2])) \end{cases} \quad (8)$$

式中: $F_\rho^1 \in \mathbb{R}^{H \times W \times C/2}$, $F_\rho^2 \in \mathbb{R}^{H \times W \times 3C/2}$; $C(\cdot)$ 表示级联操作。

2.3 损失函数 Focal-EIoU

Sparse R-CNN 的损失函数采用 GIoU(Generalized intersection over union loss),计算公式如式(9)所示:

$$\text{GIoU} = \text{IoU} - \frac{|C/(A \cup B)|}{|C|} \quad (9)$$

式中: C 为预测框和真实框的最小外接矩形; A 和 B 分别为真实框和预测框。

GIoU 是一种用于评估和优化目标检测中边界框(Bounding Box)回归的指标。它不仅考虑预测框与真实框的重叠区域,还引入了一个惩罚项,用于衡量两个框的最小外接矩形(Bounding rectangle)的面积差异。然而,它并未显式考虑两个框之间的相对位置关系(如中心点距离)。损失函数 Focal-EIoU^[25]通过引入中心点距离和宽高比的惩罚项,进一步优化了边界框回归。Focal-EIoU 的参数示意图如图 7 所示。其中, A 为真实框, B 为预测框, C 为 A 和 B 的最小封闭框; $\rho^2(A, B)$ 为 A 、 B 中心点的距离, A' 为 A 点中心点, B' 为 B 点中心点; c 为最小封闭框 C 的对角线长度, w, h 为 A 、 B 、 C 3 个框的宽和高。Focal-EIoU 的计算公式如公式(10)所示:

$$\begin{aligned} L_{\text{EIoU}} &= L_{\text{IoU}} + L_{\text{dis}} + L_{\text{asp}} = \\ 1 - \text{IoU} &+ \frac{\rho^2(A, B)}{c^2} + \frac{\rho^2(w^A, w^B)}{(w^C)^2} + \frac{\rho^2(h^A, h^B)}{(h^C)^2} \\ L_{f(x)} &= \begin{cases} -\frac{\alpha x^2 [2\ln(\beta x) - 1]}{4}, & 0 < x \leq 1; \frac{1}{e} \leq \beta \leq 1 \\ -\alpha \ln(\beta)x + D, & x > 1; \frac{1}{e} \leq \beta \leq 1 \end{cases} \end{aligned} \quad (10)$$

式中: β 用于控制曲线的弧度; e 为自然常数; x 代表真实值与预测值的差值; D 为一个常数; $\rho^2(w^A, w^B)$ 和 $\rho^2(h^A, h^B)$ 分别为宽的差的平方和高的差的平方。

Focal-EIoU 的公式如下所示:

$$L_{\text{Focal-EIoU}} = \text{IoU}^\gamma L_{\text{EIoU}} \quad (11)$$

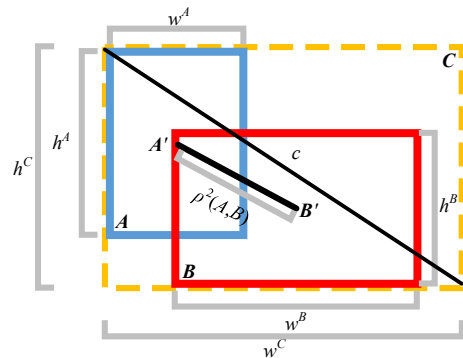


图 7 Focal-EIoU 的参数示意图

Fig. 7 Schematic diagram of the parameters of Focal-EIoU

式中: γ 为控制异常值抑制程度的参数。

根据式(11),IoU越高的目标损失就越大,而IoU低的目标的损失会降低,从而对高质量目标进行加权,提高回归精度。

3 实验及分析

3.1 实验数据集与评估指标

本文算法在配备3张RTX 3080Ti显卡的服务器上实现,该服务器采用Ubuntu操作系统及Pytorch框架。实验选用专为密集行人检测设计的CrowdHuman数据集^[26],旨在应对行人密集场景下的检测挑战。该数据集包含大量行人图像,涵盖街道、广场等复杂场景,总共有24 370张图片,分为15 000张训练集、5 000张测试集和4 370张验证集。训练集和验证集中共有约470 000个实例,平均每张图片包含23个人,并存在各种遮挡情况。每个行人实例均用头部边界框、可见区域边界框和全身边界框进行标注。

行人检测任务中主要采用平均精度(AP),对数平均漏检率(MR-2)和召回率(Recall)对算法性能进行评估。

平均精度(AP):是行人检测任务最常用的评估指标,其为P-R曲线下的面积,平均精度越高意味着模型能够更准确地检测出目标物体,公式如式(12)所示。

$$AP = \int_0^1 P(R) dR \quad (12)$$

式中: R 为召回率,指实际为正类别的样本中模型成功预测为正类别的样本所占的比例; $P(R)$ 指的是精准率(Precision)关于召回率(Recall)的函数。

漏检率(MR):表示没能检测到目标的比例,如式(13)所示:

$$MR = 1 - \frac{TP}{GT} \quad (13)$$

式中:GT表示真实行人目标总数;TP表示预测正确且实际为正类。

对数平均漏检率(MR-2)如式(14)所示:

$$MR^{-2} = e^{\left(\frac{1}{9} \sum_{i=1}^9 \ln(MR_i)\right)} \quad (14)$$

其表示在指定误检率的情况下行人检测器的漏检率,MR-2越低,表示模型漏检率越低,检测性能越好。

召回率(Recall):反映的是模型检测到的正样本数量与实际正样本数量的比例,召回率越

高,模型对正样本的检测能力越强,如式(15)所示:

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

式中:FP表示预测错误且实际为正类;FN表示预测错误且实际为负类。

3.2 对比实验分析

3.2.1 不同颈部效果对比

将单层输出颈与FPN、以及其他改进版本的BiFPN^[27]进行了对比实验,具体实验结果如表1所示,FPN作为默认的颈部结构,在Sparse R-CNN中达到了87.1%的平均检测精度(AP),召回率(Recall)为95.6%,计算量(FLOPs)为112G,帧率为43 f/s。BiFPN通过改进特征融合机制,进一步提升了检测精度,AP达到了87.4%,召回率提升至95.8%,但计算量增加至129G,帧率略有下降至40 f/s。相比之下,本文提出的SLON在保持高检测精度(AP为87.6%)和高召回率(95.8%)的同时,显著降低了计算量(FLOPs为92G),并且帧率提升至46 f/s。此外,SLON在误检率(MR-2)方面也表现优异,仅为45.9%,优于FPN和BiFPN。

实验结果充分证明了SLON在密集场景下的行人检测任务中的优势。SLON不仅能够有效减少计算复杂度,提升推理速度,还能在保持高检测精度的同时降低误检率。因此,SLON在实时行人检测系统中具有广泛的应用前景,尤其是在计算资源有限且对检测精度要求较高的场景中。

表1 不同颈部应用于Sparse R-CNN的效果对比

Table 1 Comparison of effects of different necks applied to Sparse R-CNN

Neck	AP /%	MR-2 /%	Recall /%	FLOPs /G	帧率/ (f·s ⁻¹)
FPN(default)	87.1	46.9	95.6	112	43
BiFPN	87.4	46.2	95.8	129	40
SLON	87.6	45.9	95.8	92	46

3.2.2 不同损失函数效果对比

为了验证使用不同交并比(IoU)作为定位损失函数对行人检测准确率的影响,以Sparse RCNN模型为例,分别使用以下几种定位损失函数与原模型默认的GIoU loss进行对比,实验结果如表2所示。

由实验结果可知,使用Focal-EIoU通过综合评估预测框与真实框的几何关系,有效提升了模

表 2 不同 IoU 方式的定位损失函数对实验结果的影响

Table 2 Effects of localization loss functions with different IoU modalities on the experimental results

算法	AP /%	MR-2 /%	Recall /%	帧率/ (f·s ⁻¹)
Sparse RCNN+IoU	86.8	47.5	95.4	43
Sparse RCNN+GIoU loss	87.1	46.9	95.6	42
Sparse RCNN+DIoU loss	87.2	47.3	95.6	42
Sparse RCNN+CIoU loss(default)	87.1	47.0	95.6	43
Sparse RCNN+SIoU loss	87.3	46.5	95.6	42
Sparse RCNN+Alpha-IoU loss ^[28]	87.4	46.3	95.7	43
Sparse RCNN+Focal-EIoU	87.6	46.1	95.8	43

型对不同尺度目标的适应能力,从而进一步提高模型检测的精度。具体而言,Focal-EIoU 不仅继承了 IoU、GIoU、DIoU 和 CIoU 的优点,还通过引入 Focal Loss 的思想,增强了模型对难样本的学习能力。这种改进使得模型在密集行人检测任务中表现尤为突出,能够更准确地处理目标重叠、遮挡等复杂场景。此外,Focal-EIoU 在保持高检测精度(AP 为 87.6%)的同时,误检率(MR-2)降低至 46.1%,召回率(Recall)提升至 95.8%,且推理速度(帧率)保持在 43 f/s,展现了其在精度和效率之间的良好平衡。

3.2.3 不同算法实验对比

将 DSS R-CNN 与多种主流目标检测算法进行了对比实验,包括 Faster R-CNN 和 Cascade R-CNN 等二阶段检测算法,以及 Sparse R-CNN、YOLOX^[29]、CenterNet^[30]、YOLOv8^[31] 和 Swin-Transformer、文献[16]、YOLOv11^[32] 等一阶段检测算法,实验结果如表 3 所示。结果表明,DSS R-CNN 在 AP、Recall 和 MR-2 等关键指标上均优于大多数对比算法,展现了其在复杂行人场景下的良好的检测能力。DSS R-CNN 的 AP 值达到了 90.4%,MR-2 值为 44.1%,Recall 值为 96.2%。与基准算法 Sparse R-CNN 相比,DSS R-CNN 在 AP、MR-2 和 Recall 上分别提升了

3.3%、2.8% 和 0.6%。此外,DSS R-CNN 在帧率上也保持了较高的推理速度(46 f/s),能满足实时检测的需求。YOLOv11 在通用目标检测任务中综合性能处于 SOTA 水平,尤其在速度与精度的平衡、动态结构设计及多场景适配性上表现突出,与其性能比较能较为充分地验证本文方法的先进性。

表 3 不同目标检测算法对比

Table 3 Comparison of different object detection algorithms

算法	Image-size	AP /%	MR-2 /%	Recall /%	帧率/ (f·s ⁻¹)
Faster R-CNN	640×640	72.4	51.2	88.5	29
Cascade R-CNN	640×640	74.6	44.2	87.2	36
YOLOX	640×640	86.5	46.3	95.9	79
CenterNet	640×640	87.0	45.7	96.1	48
YOLOv8	640×640	90.2	44.5	96.2	76
Sparse R-CNN(Baseline)	640×640	87.1	46.9	95.6	43
Swin-Transformer	640×640	90.1	44.8	95.9	51
文献[16]	640×640	87.4	45.4	95.6	44
YOLOv11	640×640	91.1	44.3	97.2	62
DSS R-CNN(Ours)	640×640	90.4	44.1	96.2	46

为深入探究本文算法对不同尺度目标的检测效果,采用 AP_l、AP_m、AP_s 等评价指标来讨论模型性能。其中,AP_l 代表目标面积大于 96×96 像素的大目标平均精度;AP_m 代表目标面积介于 32×32 与 96×96 像素之间的中等目标平均精度;AP_s 则代表目标面积小于 32×32 像素的小目标平均精度;AP50 和 AP75 代表将交并比(IoU)阈值设定为 0.5 和 0.75 时计算得到的平均精度值。

从表 4 能出,本文算法在大、中、小目标方面的平均检测精度都比基准算法要高。尤其在小尺度行人中的平均检测精度提升最多,提升了约 3.3%,这表明该算法在提取目标尺度信息方面表现较好,而且能够把全局信息和局部信息相结合,从而有效应对密集行人场景中的遮挡问题、尺度变化问题以及目标重叠问题,提升检测的鲁棒性

表 4 Crowdhuman 数据集上的比较实验

Table 4 Comparative experiments on the Crowdhuman dataset

算法	AP	AP50	AP75	AP _l	AP _m	AP _s
RetinaNet ^[33]	30.16	65.87	22.88	33.15	30.85	16.96
Faster R-CNN	38.78	73.12	35.53	40.79	38.71	29.05
Cascade R-CNN	42.64	77.63	42.69	41.56	41.97	31.13
Sparse R-CNN	45.37	80.19	45.07	48.43	43.85	32.86
DSS R-CNN(Ours)	47.46	81.38	47.26	49.62	44.93	36.21

和准确性。

图8展示了所提模型与表4中列出的其他先进模型在 AP_s (小目标平均精度)、 AP_m (中目标平均精度)和 AP_l (大目标平均精度)3个指标上的P-R(精确率-召回率)曲线对比。其中,图8(a)展示了各模型在 AP_s 上的P-R曲线,图8(b)展示了 AP_m 的P-R曲线,图8(c)则展示了 AP_l 的P-R曲线。P-R曲线是目标检测领域中用于评估模型性能的核心指标之一,其横轴表示召回率(Recall),纵轴表示精确率(Precision)。通常,P-R曲线覆盖的面积(Area under curve, AUC)越大,表明模

型在目标检测任务中的综合性能越优,能在高召回率的同时保持较高的精确率。

从图8中可以明显观察到,DSS R-CNN模型在 AP_s 、 AP_m 和 AP_l 三个指标上的P-R曲线覆盖面积均优于其他对比方法,尤其是在 AP_s 指标上,其P-R曲线覆盖面积显著高于其他模型,优势最为明显。这一结果充分证明了该模型在密集多尺度目标检测任务中的有效性和鲁棒性。

3.3 消融实验分析

为了验证本文提出的改进算法中各个组成部分的有效性,以Sparse R-CNN模型为基准算法,在CrowdHuman数据集的基础上进行了大量的消融实验,结果如表5所示。第一行为没有添加任何模块的基准算法,二、三、四行分别为在基准算法中仅使用了改进的CloFormer主干、单层输入颈和Focal-EIoU损失函数,最后一行为添加了所有方法的DSS R-CNN算法。

由实验结果可见,本文算法在CrowdHuman数据集上的AP达到了90.4%,检测速度为46 f/s。改进的CloFormer主干可以获得1.1%的AP、1.6%的MR-2和0.2%的Recall提升,表明了改进的CloFormer主干可以更好地对特征进行提取,促进性能的提升。单层输出颈可以获得0.5%的AP、1.0%的MR-2和0.2%的Recall提升,表明了单层输出颈可以充分融合来自于主干网络输入的不同层级的特征。Focal-EIoU损失函数可以获得0.5%的AP、0.8%的MR-2和0.2%的Recall提升,说明Focal-EIoU损失函数减少了检测目标中假阳性的数量,由消融实验结果验证了各模块的有效性。

表5 本文数据集上的消融实验结果

Table 5 Results of ablation experiments on the proposed dataset

Model	Improved scheme			AP /%	MR-2 /%	Recall /%	帧率/ (f·s ⁻¹)
	DBMAB- CloFormer	SLON	Focal- EIou				
A	×	×	×	87.1	46.9	95.6	43
B	✓	×	×	88.2	45.3	95.8	40
C	×	✓	×	87.6	45.9	95.8	46
D	×	×	✓	87.6	46.1	95.8	43
E	✓	×	✓	88.9	44.8	96.0	41
F	✓	✓	×	89.6	44.5	96.1	42
H	✓	✓	✓	90.4	44.1	96.2	46

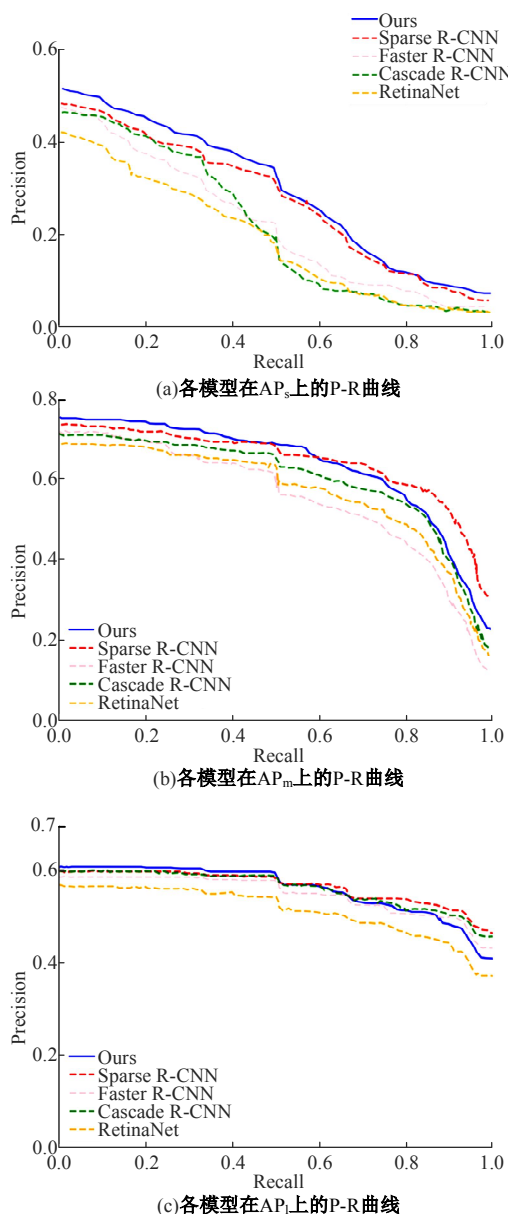


图8 五种目标检测方法的 AP_s 、 AP_m 和 AP_l 的P-R曲线
Fig. 8 P-R curves of AP_s , AP_m , and AP_l for the five object detection methods

3.4 不同算法可视化实验结果

图9展示了RetinaNet、Faster R-CNN、Sparse R-CNN以及本文提出的模型在目标检测任务中的可视化结果对比。为了更直观地展示各模型的检测性能,图中使用绿色边界框标注模型能够正确检测到的目标,而红色边界框则表示模型未能检测到的目标(即漏检目标)。

如图9所示,在CrowdHuman数据集上,本文所提出的DSS R-CNN算法的检测结果,相比基

准算法优势明显,也超越了RetinaNet和Faster R-CNN这些较新的以及经典的检测算法。在处理混合尺度目标时,DSS R-CNN的检测性能更为出色。原因在于本文设计的双分流多尺度注意力模块(DBMAB)借助并行注意力机制,有效地捕捉到图像的局部细节特征与全局特征,让模型对多尺度信息的敏感度得以提升,漏检现象也大大减少。这一可视化结果进一步验证了本文模型在实际应用中的优越性和鲁棒性。

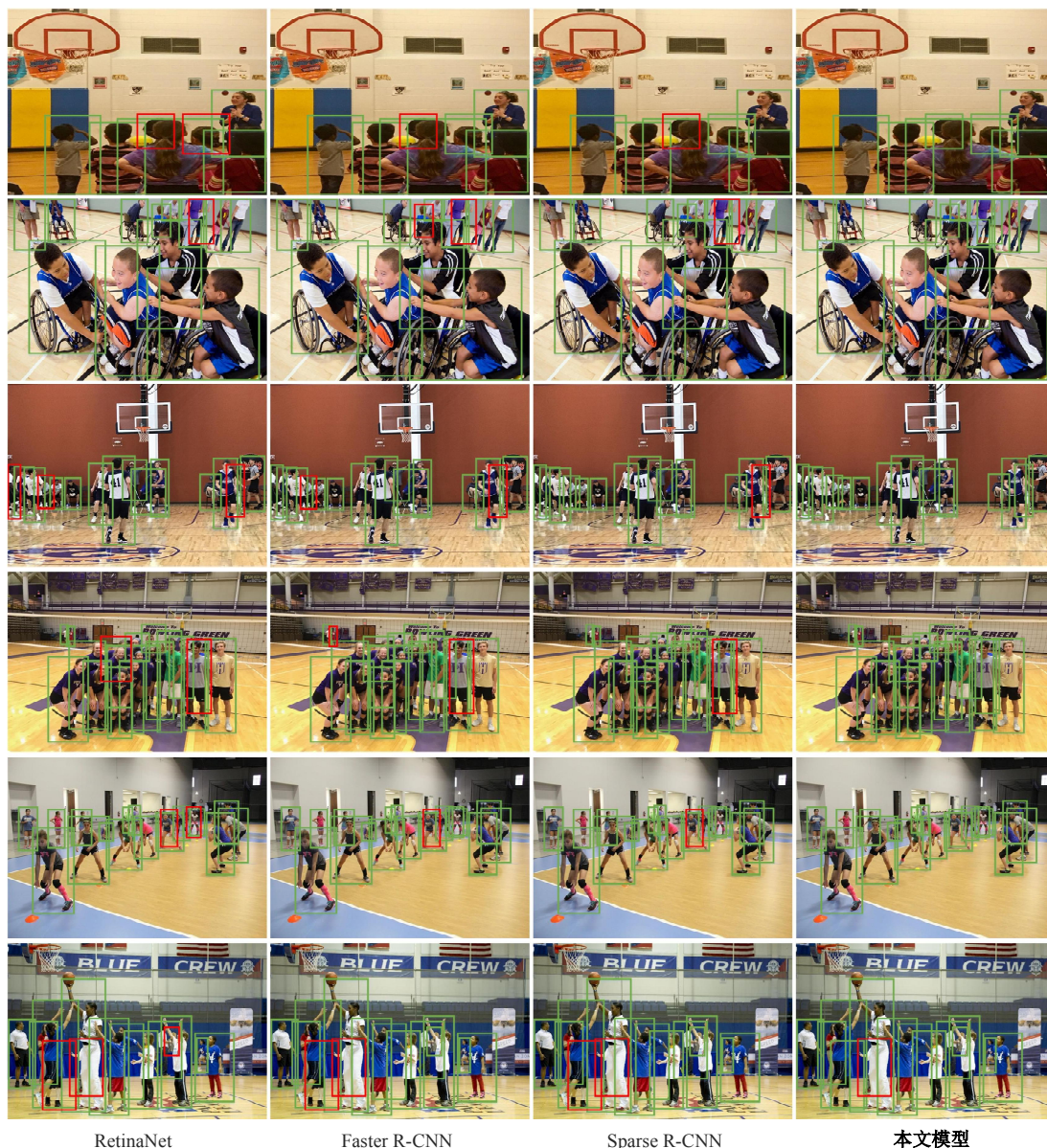


图9 CrowdHuman数据集上的检测结果
Fig. 9 Results on CrowdHuman dataset

4 结束语

针对密集行人场景中行人检测出现遮挡、小尺度导致漏检、误检的问题,基于Sparse R-CNN

骨干网络提出了一种融合双分流多尺度注意力与单层输出的密集行人检测算法。首先,提出了一个双分流多尺度注意模块DBMAB,并将该注意

力块嵌入CloFormer模型中来对主干进行修正;其次,设计了基于特征调制模块FMB的单层输出颈SLON;最后,改进损失函数为Focal-EIoU。在CrowdHuman数据集上进行实验,DSS R-CNN算法的AP值为90.4%、MR-2值为44.1%、Recall值为96.2%,同基准算法Sparse R-CNN相比,分别提高了3.3%、2.8%和0.6%,同时检测速度提高了3 f/s。结果表明,本文模型在这两类具有挑战性的场景中具有更好的行人检测效果。

参考文献:

- [1] 齐向明,刘晓暄,王子键. 关键特征感知并行细粒度特征提取的密集行人检测[J]. 计算机工程与应用, 2026, 62(1): 297-306.
Qi Xiang-ming, Liu Xiao-xuan, Wang Zi-jian. Dense pedestrian detection based on key feature perception and parallel fine-grained feature extraction[J]. Computer Engineering and Applications, 2026, 62(1): 297-306.
- [2] Ren S 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] Cai Z W, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection[C] //IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 6154-6162.
- [4] He K, Gkioxari G, Dollár P, et al. Mask R-CNN [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 2017: 2961-2969.
- [5] Sun P Z, Zhang R, F Jiang Y, et al. Sparse r-CNN: end-to-end object detection with learnable proposals [C] //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Nashville, USA, 2021, 14449-14458.
- [6] 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, Las Vegas, USA, 2016: 779-788.
- [7] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C] //IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6517-6525.
- [8] Redmon J, Farhadi A. YOLOv3: an incremental improvement [EB/OL]. [2018-04-08]. <https://arxiv.org/abs/1804.02767>.
- [9] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: optimal speed and accuracy of object detection [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, 11: 829-838.
- [10] Zhu X K, Lyu S C, Wang X, et al. TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios [C] //2021 IEEE/CVF International Conference on Computer Vision Workshops, Montreal, Canada, 2021:2778-2788.
- [11] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Vancouver, Canada, 2023: 7464-7475.
- [12] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 2016, 9905: 21-37.
- [13] Alexey D. An image is worth 16x16 words: Transformers for image recognition at scale[J]. Arxiv Preprint arxiv: 2010.11929, 2020.
- [14] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows [C] //Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021:9992-10002.
- [15] Li Q M, Su Y J, Gao Y, et al. OAF-Net: an occlusion-aware anchor-free network for pedestrian detection in a crowd[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(11): 21291-21300.
- [16] Wu J T, Dai G J, Zhou W H, et al. Multi-scale feature fusion with attention mechanism for crowded road object detection[J]. Journal of Real-Time Image Processing, 2024, 21(2): 29.
- [17] 柳长源, 张玉亮, 毕晓君. 基于多阶段提议稀疏区域卷积网络的城市交通目标检测[J]. 电子学报, 2023, 51(1): 26-31.
Liu Chang-yuan, Zhang Yu-liang, Bi Xiao-jun. Ur-

- ban traffic object detection based on multi-stage proposal sparse regional convolutional neural network[J]. Chinese Journal of Electronics, 2023, 51(1): 26-31.
- [18] Chen T, Chen J H, Gao T, et al. A novel scene-aware pedestrian detection in dense scenes[C]//2024 27th International Conference on Computer Supported Cooperative Work in Desig, Tianjin, China, 2024: 1316-1321.
- [19] Park S K, Chung J H, Pae D S, et al. Fusion-attention network using dense scale-invariant feature transform flow image and point cloud for 3D pedestrian detection[J]. Multimed Tools and Applications, 2025, 84(14): 12813-12833.
- [20] Fan Q, Huang H B, Guan J Y, et al. Rethinking local perception in lightweight vision transformer[J]. arXiv, 2023. arXiv:2303.17803.
- [21] Chollet F. Xception: deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, USA, 2017:1800-1807.
- [22] Li J F, Wen Y, He L H. Seconv: spatial and channel reconstruction convolution for feature redundancy[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 6153-6162.
- [23] Wang Y, Li Y S, Wang G, et al. Multi-scale attention network for single image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024: 5950-5960.
- [24] Zheng M J, Sun L, Dong J X, et al. SMFANet: A lightweight self-modulation feature aggregation network for efficient image super-resolution[C]//European Conference on Computer Vision, Milan, Italy, 2024:359-375.
- [25] Zhang Y F, Ren W, Zhang Z, et al. Focal and efficient IOU loss for accurate bounding box regression[J]. Neurocomputing, 2022, 506: 146-157.
- [26] Shao S, Zhao Z J, Li B X, et al. Crowdhuman: a benchmark for detecting human in a crowd[J]. arXiv, 2018. arXiv:1805.00123.
- [27] Tan M X, Pang R M, Le Q V. EfficientDet: scalable and efficient object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020:10778-10787.
- [28] He J B, Erfani S, Ma X J, et al. Alpha-IoU: a family of power intersection over union losses for bounding box regression[C]//2021 Conference on Neural Information Processing Systems, San Diego, USA, 2021: 20230-20242.
- [29] Ge Z, Liu S T, Wang F, et al. YOLOX: exceeding YOLO series in 2021[J]. arXiv, 2021. arXiv: 2107.08430.
- [30] Zhou X, Wang D, Krähenbühl P. Objects as points[J]. arXiv, 2019. arXiv:1904.07850.
- [31] Varghese R, Sambath M. YOLOv8: a novel object detection algorithm with enhanced performance and robustness[C]//2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems, Chennai, India, 2024: 1-6.
- [32] Khanam R, Hussain M. Yolov11: An overview of the key architectural enhancements[J]. arXiv, 2024. arXiv:2410.17725.
- [33] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017:2999-3007.