

·人工智能·

DOI:10.12454/j.jsuese.202401025



本刊网刊

基于全局特征聚焦与信息增强的遮挡行人检测算法

郑开魁¹, 吉康友¹, 李俊², 李琦铭^{2*}

(1. 福州大学 先进制造学院, 福建 晋江 362251; 2. 中国科学院海西研究院 泉州装备制造研究中心, 福建 泉州 362200)

摘要:随着行人检测技术在自动驾驶、机器人导航系统及智能监控等领域的广泛应用,对检测精度的要求日益提高。然而,在现实场景下行人遮挡问题突出,当严重遮挡时目标可视范围急剧减小,行人特征大量缺失,致使检测器难以有效区分目标与背景。针对此问题,创新性地提出全局特征聚焦与信息增强网络(GFFIE-Net),通过设计特征增强模块和分级融合机制,致力于解决行人遮挡检测难题。具体而言,针对主干网络输出的多个不同分辨率特征图,首先,通过卷积注意力模块(CBAM)动态调整特征图各通道与空间位置重要性,强化关键区域特征信息并抑制背景噪声,使网络聚焦目标区域;随后,顺序级联Mamba模块对聚焦去噪特征图进行全局信息的建模,提取被遮挡行人周围上下文信息,结合行人可见特征进行完整推理;最后,设计分级特征融合机制,高效融合高级语义信息与位置细节信息,以实现特征图间的多尺度信息交互。实验结果表明,在CityPersons数据集的严重遮挡子集上,漏检率二次指数达到43.7%,相比基准网络提升了4.4个百分点;在Caltech严重遮挡子集上,漏检率二次指数达到33.6%;在CrowdHuman数据集上,漏检率二次指数达到了43.2%,性能超过了一些主流算法。通过可视化分析,进一步证明GFFIE-Net具有高效处理行人遮挡的能力。

关键词:行人检测;Mamba;特征增强;CBAM

中图分类号:TP391

文献标志码:A

文章编号:2096-3246(2025)06-0104-15

行人检测是目标检测任务的重要分支之一^[1]。近年来,行人检测在自动驾驶、机器人导航系统以及智能监控等领域^[2-3]具有重要的实际应用价值,其核心目的在于精准识别静态图像中的全部行人目标。然而,行人常受其他行人或物体遮挡,致使难以提取完整特征信息,从而引发检测器漏检问题。因此,如何有效解决遮挡问题,成为进一步提升行人检测器性能的关键挑战。

随着深度学习在计算机视觉领域的快速发展,行人检测方法从传统的手工特征提取转向了基于深度学习的方法。早期的区域卷积神经网络(RCNN)系列方法^[4-6]将检测任务分为两个阶段,通过预设固定尺寸的锚框进行分类和回归,在当时取得了显著领先的检测精度,为行人检测技术的发展奠定了重要基础。然而,随着对行人检测研究的深入,发现锚框的超参数(如尺寸、纵横比和数量)对检测器性能影响极大,成为检测性能进一步提升的主要障碍。为了解决这一问

题,基于无锚框的行人检测方法应运而生,这类方法无须预设锚框,可以直接端到端地检测行人,避免了手工设计锚框的限制,因此逐渐受到研究者的青睐。例如,CornerNet^[7]和CenterNet^[8]通过重点关注目标的角点位置,使用角池化层增强特征以更准确地定位目标的左上角和右下角,还可以在角点对的基础上,加入目标中心关键点,组成三元组进行检测,既能捕捉目标的边界信息,又能兼顾目标的内部信息。中心与尺度预测(CSP)^[9]则将检测任务转化为中心点与相应尺度的预测任务,该方法简单且有效。本文提出的全局特征聚焦与信息增强网络(global feature focusing and information enhancement network, GFFIE-Net)同样采用基于中心点和尺度预测的方法,但在具体实现方面具有独特的设计和优势。为解决遮挡问题这一关键挑战,近年来,研究者们主要从后处理优化、基于特定模型改进以及基于身体部位特征处理3个方面深入

收稿日期:2024-12-09 修回日期:2025-02-20 网络出版日期:2025-03-24

基金项目:国家自然科学基金项目(52275178;62102394);福建省科技计划项目(2022L3094;2023N3010);泉州市科技计划项目(2024QZC001R)

作者简介:郑开魁(1986—),教授,博导。研究方向:摩擦学、增材制造、机器人技术等。E-mail:kuikui@fzu.edu.cn

* 通信作者:李琦铭,副研究员, E-mail:qimingli@fjirsm.ac.cn

探索解决方案。在后处理优化方面,研究者们通过设计特定损失函数和非极大值抑制(NMS)改进策略。例如,RepLoss^[10]等方法通过改进损失函数,对定位不准确的检测框进行惩罚,进一步提升了被遮挡行人的检测准确性。BIA-NMS^[11]和OTP-NMS^[12]等方法通过优化NMS策略,更灵活地抑制冗余检测结果,有效保留可能被遮挡的目标。然而,尽管研究者们持续对NMS进行优化,但由于其抑制样本框的特性,遮挡严重的正样本仍可能被误判为假阳性,从而使得遮挡问题难以有效解决。在基于特定模型改进方面,随着Transformer^[13-14]在计算机视觉领域的巨大成功,研究者们开始探索基于Transformer的目标检测算法(DETR)模型^[15]的方法。DETR采用一对一的标签匹配策略进行预测,有效规避了NMS的部分缺点。特别是Deformable DETR^[16],通过引入可变形注意力机制,自适应地调整输入特征,显著增强了处理遮挡的灵活性,并更好地捕获了目标间的联系。然而,Transformer引入的自注意力机制受到二次时间复杂度的限制,难以部署在高分辨率特征图上。此外,DETR需要手工设计Query数量及一些超参数,这限制了其在不同场景下的灵活性。特别是在面对不同人群密度的数据集时,如CityPersons^[17]和CrowdHuman^[18],在Query数量的设置上,后者必须设置为前者的两倍,否则性能就不如基于卷积神经网络(CNN)的基准网络。在基于身体部位特征处理方面,鉴于遮挡情况下行人身体部分可见的特点,研究者们进一步提出了基于身体部分模型的方法。该方法根据人体不同部位的特点,采用分而治之策略,对不同身体部位分别设计相应的检测器,以辅助行人整体检测。例如,Bi-Center^[19]和OAF-Net^[20]在后处理中设计多个中心预测分支,分别针对不同遮挡程度的行人进行检测。然而,上述3类方法存在一些共性问题,尚未得到有效解决。例如,未充分考虑在网络特征提取阶段面临的困难,致使在处理高分辨率特征图时,难以获取充足的全局信息,同时网络极易受到背景噪声的干扰。这些问题在一定程度上限制了现有方法对行人遮挡问题的处理能力。

本文针对上述共性问题进行了深入分析。一方面,由于网络注意力资源有限,而目标区域在整体图像中占比通常较小,背景像素常占据较大比例,同时图片常常因光照、天气和像素等噪声干扰较多,因此如何使网络聚焦目标区域、减少噪声干扰成为关键问题。卷积注意力模块(CBAM)^[21]通过动态调整特征图中每个通道和空间位置的重要性,实现了对特征信息的双重精炼,从而引导模型聚焦关键区域的特征信息,并有效抑制噪声干扰。黎瑞虹等^[22]将CBAM机制与自注意力机制相结合,应用于夜间行人目标检测任务,显著提升了网络的

特征提取能力与全局信息捕获能力,在精度指标上取得了显著进步。然而,与本文方法不同的是,其研究重点在于结合两种机制提升夜间检测精度,未针对行人遮挡场景下的特征缺失和复杂背景噪声干扰问题进行专门设计。Zang等^[23]提出的基于多重感受野和注意力机制的多光谱行人检测方法(MAPD),虽将多重感受野模块与CBAM机制相结合,构建了多重感受野注意力模型用于融合多模态特征,且其在多光谱目标检测任务中表现突出,但该方法在进行特征图融合前未关注噪声干扰问题。另一方面,CNN方法往往过于关注局部信息,却忽视了全局信息的重要性。而在行人检测任务中,需要提取被遮挡行人周围的上下文信息,特别是在高分辨率特征图上实现全局信息的有效提取。虽然自注意力机制的提出为网络提取全局特征信息提供了可能,但是其计算开销和参数量会随着特征图分辨率的提高而急剧上升,在某些对检测速度有一定要求的行人检测场景中应用受限。最近,为了克服这一难题,Mamba^[24]机制被提出,其通过独特的全局感受野和动态加权策略,突破了CNN的局部建模局限,兼具线性时间复杂度与强大的全局建模能力。进一步发展出的Vision Mamba^[25]模型通过构建Mamba模块在视觉数据处理上表现优秀。丁政泽等^[26]提出了一种跨模态图像融合网络,通过Mamba模块对高分辨率图像进行全局线性复杂度建模,在保证计算效率的同时能有效提取图像全局特征。石洋宇等^[27]利用Mamba模块长序列建模的特点,对图片进行正反双向2维选择性扫描,提取人体运动行为在不同层面的特征信息,构建了人体异常检测模型编码器。Mamba机制虽在部分视觉任务中凭借长距离建模特性表现突出,但在行人检测领域的应用极为有限。由于严重遮挡,行人目标的关键特征大量缺失,Mamba长距离建模过程中不可避免地会包含对噪声干扰特征的建模,在后续推理中易引发误导。这就需要在行人检测中先对特征图进行针对性优化,再利用Mamba模块挖掘被遮挡行人的隐藏特征和上下文信息,以提升遮挡检测性能。

本文提出一种基于全局特征聚焦与信息增强的遮挡行人检测算法GFFIE-Net,通过CBAM与Mamba模块的顺序级联,首先聚焦网络注意力并减少噪声干扰,然后进行全局信息的建模,提取被遮挡行人周围的上下文信息;设计了一种分级特征融合机制,实现高级语义信息与位置细节信息的全面融合,提高对多尺度遮挡行人的检测能力;在3个公开数据集上进行实验验证,结果表明,本文方法在遮挡行人检测任务中取得显著的性能提升,展示了CBAM和Mamba模块在行人检测中的潜力。

1 GFFIE-Net 算法

1.1 GFFIE-Net 网络架构

为提高遮挡情况下的行人检测效果,本文从消除网络噪声干扰、聚焦注意力以及提取全局信息的角度考虑,提出了一种简单且有效的 GFFIE-Net 算法。

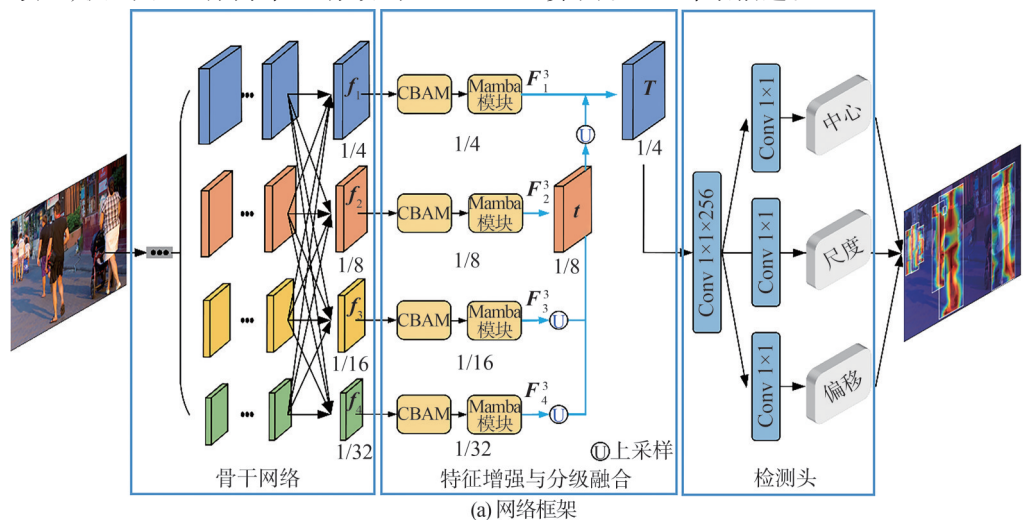


图 1 GFFIE-Net 网络框架和技术路线

Fig. 1 Overall architecture and technical flowchart of GFFIE-Net

1.2 特征提取

HRNet32^[28]最初在人类姿态估计领域崭露头角,其设计理念别具一格。HRNet32 采用并行处理多个分辨率分支的方式,实现不同分支间的信息共享,使特征图兼具丰富的语义信息和精准的位置信息。由于行人检测与人类姿态估计都关注人的动作和形态,许多研究都将 HRNet32 作为主干网络引入行人检测领域,并取得了显著成果,这充分证实了 HRNet32 在行人检测任务中的可行性和高效性。因此,本文选用 HRNet32 作为主干网络,最终输出 4 个不同分辨率、不同通道维度的特征图 f_1, f_2, f_3 和 f_4 , 其与输入图像的分辨率比例分别为 1/4、1/8、1/16 和 1/32, 通道维度分别为 32、64、128 和 256。其中:低维高分辨率特征图能够捕捉行人的细节信息,但难以获取全局信息;高维低分辨率特征图蕴含相对丰富的语义信息,但可能丢失部分细节且噪声较多。

1.3 特征增强与分级融合

在行人检测任务中,考虑到遮挡情况容易导致检测器误检或漏检,仅依靠 HRNet32 提取行人特征难以满足精度要求。因此,本文提出了特征增强与分级融合策略,通过对全局特征进行聚焦和信息增强,增强被遮挡行人的特征。

特征增强与分级融合旨在从主干网络输出的特征图挖掘多层次上下文语义信息,结合行人可见特征,精准推理被遮挡部分,辅助检测器精确检测遮挡

GFFIE-Net 的整体网络框架和技术路线如图 1 所示。在无锚检测框架^[9]基础上,对主干网络输出的 4 个不同分辨率、不同通道维度的特征图进行了特征增强操作。以下将从特征提取、特征增强与分级融合、行人检测头以及训练与测试这 4 个方面对 GFFIE-Net 进行详细描述。

行人。首先,嵌入 CBAM^[21],从空间注意力和通道注意力两个维度,自适应地重新校准特征图,以增强目标区域的显著性,并抑制背景噪声的干扰。接着,顺序级联 Mamba 模块^[25]对聚焦去噪特征图进行全局信息建模,以捕获 4 个尺度特征图的全局上下文信息。最后,本文设计了一种分级特征融合机制,该机制运用双线性插值算法对不同尺度的特征图进行适当的尺寸调整,保证其空间分辨率一致,并且采用通道拼接的融合方式,实现多尺度特征图的全面信息融合,从而进一步提升检测器对遮挡行人的检测能力。

1.3.1 卷积注意力模块(CBAM)

遮挡问题在行人检测任务中尤为突出,它不仅导致目标行人特征信息的缺失,还常常因各类噪声影响,进而干扰检测器。为了应对这一挑战,CBAM^[21]作为一种高效的特征增强手段被引入,如图 2 所示。图 2 中, H 为输入图像的高度, W 为输入图像的宽度, C 为通道维度(通道数)。CBAM 通过模拟人类视觉系统的工作机制,使模型能够自动聚焦于图像中的关键区域,同时忽略不重要的背景信息,从而增强特征表达的针对性和有效性。

CBAM 从通道注意力和空间注意力两个维度对输入特征图进行精细化处理。在通道注意力方面:首先,在空间维度执行平均和最大池化操作,提炼出全局描述向量,捕捉各通道在整个特征图上的全局重要性信息;随后,利用小型全连接神经网络学习并计算

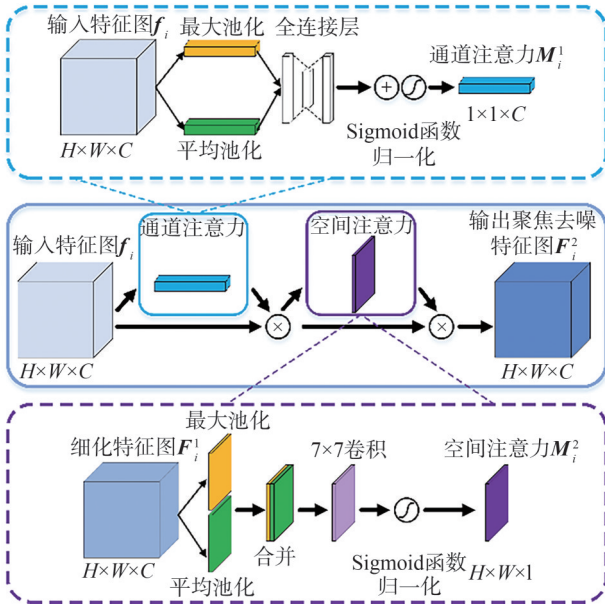


图2 卷积注意力模块

Fig. 2 CBAM

通道重要性权重,并通过 Sigmoid 函数进行归一化,输出表征各通道相对重要性的权重向量;最后,权重向量通过乘法作用于输入特征图,得到一次精炼特征图,实现特征响应重新加权。此过程不仅提升了模型对关键特征通道的敏感度,还助力网络动态聚焦信息丰富通道,增强模型表征能力与检测性能。在空间注意力方面,CBAM也在通道维度进行了平均和最大池化,重点捕捉目标区域的潜在信息,通过 7×7 卷积压缩空间信息,并使用 Sigmoid 函数进行归一化,生成单通道空间注意力图,最后将空间注意力图与一次精炼特征图相乘,再次加权特征,以获得二次精炼特征图。这种空间上的精细化处理提升了网络对关键空间位置的敏感度,使其关注目标行人区域,增强了网络的目标感知能力,公式如下:

$$F_i^1 = f_i M_i^1 + f_i \quad (1)$$

$$F_i^2 = F_i^1 M_i^2 + F_i^1 \quad (2)$$

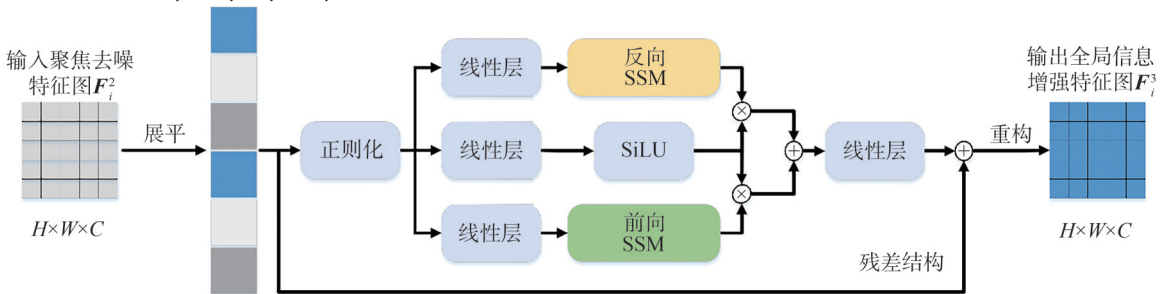


图3 Mamba模块

Fig. 3 Mamba module

在复杂遮挡情况下,由于行人部分特征缺失,通过传统方法难以获取足够的全局信息来进行准确判断。而 Mamba 模块利用其独特的结构,能够从多个尺

度式(1)~(2)中: f_i 为多个不同尺度的输入特征图, $i=1, 2, 3, 4$; F_i^1 为 f_i 经过通道注意力 M_i^1 加权后的细化特征图; F_i^2 为 F_i^1 经过空间注意力 M_i^2 再次加权后的聚焦去噪特征图。

在轻微遮挡场景中,空间注意力能够更准确地定位行人的边界信息,增强目标区域的特征表示,减少遮挡导致的特征模糊;在严重遮挡场景下,通道注意力可以突出未被遮挡部分的关键特征通道,为后续的推理提供有力支持,同时有效抑制背景噪声对检测的干扰。如图1所示,在 GFFIE-Net 中引入 CBAM,优化了特征图的注意力分布,缓解了噪声的干扰,增强了模型对遮挡行人的特征表示能力,有助于模型进一步理解图像上下文信息,进而实现对被遮挡行人的有效检测和识别。

1.3.2 Mamba 模块

在提取图像特征任务中,CNN 和 Transformer^[13]模型均存在各自的局限性:CNN 的感受野受限,难以捕捉全局上下文信息;Transformer 虽然借助自注意力机制具备提取全局信息的能力,但在面对高分辨率特征图时,却面临计算复杂度和内存开销方面的巨大挑战。为了突破这些限制,Vision Mamba 模型构建了 Mamba 模块^[25],如图3所示。在该模块中,特征图首先被展平成一维图像块向量,以保留其全局信息。随后,通过线性层进行进一步的特征提取与转换,这些线性层可以学习输入数据的非线性关系,并提取更高层次的特征。接着,前向状态空间模型(SSM)层利用其特性,前向处理图像块向量,在提取全局上下文信息的同时,捕捉特征向量之间的长距离依赖关系。同时,在反向 SSM 层中,图像块向量被反向传播,并通过 SSM 进行反向处理。这种双向处理方式不仅显著提升了模型对整体图像数据的理解和表示能力,而且在较低计算复杂度和较小内存开销的条件下,满足了提取高分辨率特征图全局信息的需求。

度的特征图中捕捉到长距离的特征依赖关系,结合可见特征进行完整推理。因此,在网络特征增强部分顺序级联 Mamba 模块对聚焦去噪特征图进行全局信息

建模,提取被遮挡行人上下文信息,增强 4 个尺度特征图对全局行人信息的关注。

1.3.3 分级特征融合机制

分级特征融合机制旨在对增强后的不同分辨率特征图进行深度融合,获取富含多尺度信息的行人特征图。特征融合部分如图 1 中蓝色箭头所示。 F_i^3 为特征图 F_i^2 经过 Mamba 模块后得到的全局信息增强特征图,为了使不同分辨率的特征图在空间维度上具有一致性,首先,采用双线性插值算法将分辨率较低的特征图 F_3^3 和 F_4^3 在空间分辨率上采样至与 F_2^3 相同维度(双线性插值算法是一种基于线性插值的图像缩放技术,通过对目标像素周围 2×2 邻域内的像素值进行两次插值计算,在保持一定图像平滑度的同时有效调整图像分辨率)。接着,将空间维度对齐后的 3 个特征图在通道维度进行拼接,合并不同尺度特征图的信息,生成初步多尺度融合特征图 t 。然后,为了将其与蕴含较多位置细节信息的高分辨率特征图 F_1^3 进行融合,再次运用双线性插值算法将 t 的空间分辨率上调至与 F_1^3 相同的空间维度。最后,将维度对齐后的 t 与 F_1^3 在通道维度完成拼接,最终得到蕴含丰富多尺度语义信息的特征图 T 。本文提出的融合方式能够更好地适应行人检测中不同尺度信息的需求,特别是在遮挡情况下,高分辨率特征图的细节信息和低分辨率特征图的语义信息能够相互补充,进一步提升对多尺度遮挡行人的检测能力,具体公式如下:

$$F_i^3 = \text{Mamba}(F_i^2) \quad (3)$$

$$t = \text{Concat}[F_2^3, U_{1/8}(F_3^3), U_{1/8}(F_4^3)] \quad (4)$$

$$T = \text{Concat}[F_1^3, U_{1/4}(t)] \quad (5)$$

式 (3) ~ (5) 中: $F_1^3 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 32}$, $F_2^3 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 64}$, $F_3^3 \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times 128}$, $F_4^3 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times 256}$, 特征图像素分别为原始输入图像像素的 $1/4$ 、 $1/8$ 、 $1/16$ 和 $1/32$; $U_{1/8}(F_3^3)$ 和 $U_{1/8}(F_4^3)$ 分别表示采用双线性插值算法将特征图 F_3^3 和 F_4^3 采样至 $1/8$ 尺度; $\text{Concat}[\cdot]$ 表示在通道维度进行拼接,输出初步多尺度融合特征图 $t \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 448}$; $U_{1/4}(t)$ 表示采用双线性插值算法将特征图 t 采样至 $1/4$ 尺度,输出多尺度语义特征图 $T \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 480}$ 。

这种分级融合机制具有显著优势。高分辨率特征图与低分辨率特征图之间的信息差异较大,直接拼接可能会导致特征图之间的信息冲突。先融合低分辨率特征图强化语义,再融合高分辨率特征图补充细节,逐步融合不同层次的特征,可以减少信息冗余,增强特征表示的层次性,助力网络理解上下文,

精准推理被遮挡行人特征。消融实验部分将进行合理性分析。

1.4 行人检测头

行人检测头用于将获得的特征图解析为检测结果。首先,采用一个 3×3 的卷积层,将特征图的通道维数减小到 256;接着,利用 3 个并行的 1×1 卷积层分别生成中心热图、尺度热图和偏移图;然后,根据中心热图和比例图中的相应比例,自动生成输入图像中行人的边界框;最后,通过偏移预测分支可以对行人的中心位置进行小幅调整,从而进一步提高检测性能。在行人检测头中,还需要为每个热图预测构造真实值(ground truth)。根据文献[9]中给出的输入图像的边界框注释,可以自动生成中心、尺度和偏移量的真实值。

1.4.1 中心点真实值

行人中心点所在的位置即为中心点真实值的正值。然而,在实际中很难确定行人准确的中心点用于训练。因此,为了减少围绕在正样本周围的大量负样本导致的模糊性,使用 2 维高斯掩模 $G(\cdot)$,公式如下:

$$M_{ij} = \max_k G(i, j, x_k, y_k, \sigma_{w_k}, \sigma_{h_k}), k = 1, 2, \dots, K \quad (6)$$

$$G(i, j, x_k, y_k, \sigma_{w_k}, \sigma_{h_k}) = e^{-\left(\frac{(i-x_k)^2}{2\sigma_{w_k}^2} + \frac{(j-y_k)^2}{2\sigma_{h_k}^2}\right)} \quad (7)$$

式(6)~(7)中: (i, j) 为预测的中心点位置; K 为图像中的目标数量; (x_k, y_k) 、 w_k 、 h_k 分别为第 k 个目标的中心点坐标、宽度和高度; σ_{w_k} 、 σ_{h_k} 分别为第 k 个目标对应的 2 维高斯掩膜在宽度和高度方向上的方差,其值与第 k 个目标的宽度 w_k 、高度 h_k 成比例,用于控制高斯掩膜的扩散范围; M_{ij} 为在预测的中心点位置 (i, j) 处,从图像中所有 K 个目标对应的高斯掩膜计算结果里选取的最大值,用于衡量该预测位置与各个目标真实中心的关联紧密程度,以辅助确定中心点真实值。

1.4.2 尺度真实值

尺度真实值通常可以定义为人体的高度或宽度。线标注在文献[17]中首次被提出,并通过统一的长宽比 0.41 自动生成行人检测框。根据线标注,先预测每个行人的高度然后生成具有预定长宽比(0.41)的包围框。尺度真实值被定义为:

$$S_{ij} = \lg h_k \quad (8)$$

式中, S_{ij} 为目标尺度真实值。为了减少模糊性, $\lg h_k$ 被赋予正样本半径 2 个像素范围内的所有负样本。

1.4.3 偏移量真实值

遵循 Liu 等^[9]的设定,在重新映射之前增加一个偏移分支以相应调整中心点的位置,用式(9)表示:

$$O_{ij} = \left(\left\lfloor \frac{x_k}{r} \right\rfloor - \left\lfloor \frac{x_k}{r} \right\rfloor, \left\lfloor \frac{y_k}{r} \right\rfloor - \left\lfloor \frac{y_k}{r} \right\rfloor \right) \quad (9)$$

式中, r 为下采样因子(设为4), $O_{ij}(\cdot)$ 表示输入图像的中心点坐标 (x_k, y_k) 被映射到输出图像的坐标 $\left(\left\lfloor \frac{x_k}{r} \right\rfloor, \left\lfloor \frac{y_k}{r} \right\rfloor \right)$ 。

2 训练与测试

2.1 训练

中心损失: 为了有效缓解正负样本极端不平衡问题, 本算法采用焦点损失(focal loss)将中心点预测任务的目标函数, 即中心损失 L_c 作为分类任务的损失函数来训练模型。其中, 中心点预测是模型对行人中心点位置的预测过程, 而 L_c 用于衡量该预测结果与真实中心点的差异, 如式(10)所示。

$$L_c = -\frac{1}{K} \sum_{i=1}^{W/r} \sum_{j=1}^{H/r} \alpha_{ij} (1 - p_{ij})^\gamma \lg p_{ij} \quad (10)$$

式中, p_{ij} 和 α_{ij} 分别由式(11)和(12)定义。

$$p_{ij} = \begin{cases} p_{ij}, & y_{ij} = 1; \\ 1 - p_{ij}, & \text{其他} \end{cases} \quad (11)$$

$$\alpha_{ij} = \begin{cases} 1, & y_{ij} = 1; \\ \left(1 - M_{ij}\right)^\beta, & \text{其他} \end{cases} \quad (12)$$

式(10)~(12)中: α_{ij} 为样本权重, 用于对不同样本进行加权, 进一步平衡正负样本; y_{ij} 为落在位置 (i, j) 的中心点数量; p_{ij} 为预测坐标点属于中心点的概率, $p_{ij} \in [0, 1]$; β 和 γ 为两个超参数, 采用文献[9]的设定方式将其设为 $\beta = 4$ 和 $\gamma = 4$ 。

尺度损失: 尺度预测是对行人目标高度的回归任务, 本算法使用 SmoothL1 损失来衡量尺度预测值与真实值的差异, 如式(13)所示。

$$L_s = -\frac{1}{K} \sum_{k=1}^K \text{SmoothL1}(s_k, \bar{s}_k) \quad (13)$$

式中, L_s 为尺度损失, s_k 和 $\bar{s}_k$ 分别为第 k 个目标尺度的预测值和真实值。

偏移量损失: 偏移量预测是对行人中心点坐标偏移的回归任务, 同样使用 SmoothL1 损失来约束预测结果与真实值的差异偏移量, 如式(14)所示。

$$L_o = -\frac{1}{K} \sum_{k=1}^K \text{SmoothL1}(o_k, \bar{o}_k) \quad (14)$$

式中, L_o 为偏移量损失, o_k 和 $\bar{o}_k$ 分别为第 k 个目标偏移量的预测值和真实值。

总损失 L 由中心损失、尺度损失以及偏移量损失构成, 如式(15)所示。

$$L = \lambda_c L_c + \lambda_s L_s + \lambda_o L_o \quad (15)$$

式中, λ_c 、 λ_s 和 λ_o 分别为 0.01、1 和 0.1。

2.2 测试

推断过程如下: 首先, 遵循 Liu 等^[9]的设定, 使用阈值为 0.01 的置信度分数筛选行人中心位置; 接着, 将筛选后的中心热图预测与相应的尺度预测相结合, 生成所预测行人的检测框; 然后, 使用预测的偏移图在重新映射到原始图像之前进一步调整行人的中心位置, 此时仍存在冗余的检测框; 最后, 通过 NMS 去除冗余的预测框, 得到最终的预测结果。

3 实验与分析

3.1 数据集与实验参数设置

3.1.1 数据集介绍

为了充分验证 GFFIE-Net 处理遮挡行人的鲁棒性, 在 3 个具有挑战性的行人数据集, 即 CityPersons^[17]、Caltech^[29] 和 CrowdHuman^[18] 上进行实验。其中: Citypersons 包含 5 050 张图像(2 975 张用于训练, 500 张用于验证, 1 575 张用于测试); Caltech 包含 42 782 张训练图像和 4 024 张测试图像。Citypersons 数据集比 Caltech 数据集具有更多样、更复杂的行人样本, 图像分辨率更高, 检测难度也更高。因此, 本文采用 Citypersons 数据集进行消融实验。

CityPersons 中的验证子集根据目标可见率不同被划分为 4 个遮挡子集, 以更公平地比较处理遮挡行人方法的有效性。具体来说, 对于像素值大于 50 的目标, 可见率在 [0.65, 1.00] 的被划分为合理遮挡子集(Reasonable), 可见率在 [0.65, 0.90] 的被划分为部分遮挡子集(Partial), 可见率在 [0.90, 1.00] 的被划分为少量遮挡子集(Bare), 而可见率在 [0.00, 0.65] 的被划分为严重遮挡子集(Heavy), 对遮挡子集的验证可作为判断检测器处理遮挡性能的重要指标。

Caltech 数据集给出了可视化区域的标注, 并根据被遮挡情况将行人分为 3 个子集: 对于像素值大于 20 的目标, 可见率在 [0.20, 1.00] 的被划分为全集(All); 对于像素值大于 50 的目标, 可见率在 [0.65, 1.00] 的被划分为部分遮挡子集(partial occlusion, PO), 可见率在 [0.20, 0.65] 的被划分为严重遮挡子集(heavy occlusion, HO)。

CrowdHuman 专门针对各类拥挤场景的行人, 从互联网上收集了 15 000、4 370 和 5 000 张图像, 分别用于训练、验证和测试子集。在训练和验证子集中总共有 470 000 个行人实例, 每幅图像大约有 24 个人。与其他的行人数据集相比, 其拥有更高的行人密度, 且行人之间也常常存在遮挡现象。该数据集丰富的拥挤场景及行人之间的遮挡情况, 能够有效验证本文方法在

处理各类拥挤场景中的行人时,对遮挡行人的定位准确性、特征提取与识别有效性。

3.1.2 评价指标

在行人检测中,为了更全面地评估检测器的性能,通常利用MR-I评价曲线。漏检率指所有真实值为正样本($P+N$)中被预测为负样本(N)的比例,记为 r 。 I 指待检测的 n 张图像中,平均每张图像中错误地将负样本分类为正样本(P)的比例。选取9个 $\lg r$ 在 $I \in [10^{-2}, 10^0]$ 的平均值并进行指数运算获得漏检率二次指数。漏检率二次指数越小,检测器的性能越高。在CityPersons^[17]和Caltech^[29]数据集文献中,以漏检率二次指数作为评价指标。而在CrowdHuman^[18]数据集文献中,除了漏检率二次指数外,同时使用召回率(Recall)和平均精度(AP)作为评价指标,召回率高说明算法能够尽可能多地检测出人群中的行人,而平均精度高则体现了检测结果的高精度。在实验过程中,采用与上述数据集文献相同的评价指标来评估本文方法的性能。

3.1.3 实验参数设置

本文所有实验基于的硬件设备为NVIDIA PCIe A100 GPU,编程语言及深度学习框架为Python3.8和Pytorch1.11.0。本算法使用HRNet32为主干网络,并使用其在ImageNet上的预训练权重进行训练以加快模型收敛,使用AdamW^[30]作为优化算法。另外,在数据加载阶段采用了图像翻转、随机缩放和裁剪作为数据增强策略。下面详细介绍在3个数据集上的参数设置:

CityPersons:输入图像大小为640×1 280像素。在训练过程中设置批大小 $b=16$ (每个GPU上设置4张图像,共使用4个GPU),设置初始学习率 $l=2 \times e^{-4}$,模型训练150 epoch,并且每训练50 epoch,学习率衰减为原来的1/2($l/2$)。

Caltech:输入图像大小为480×640像素。在训练过程中设置 $b=16$ (每个GPU上设置4张图像,共使用4个GPU),初始学习率 $l=2 \times e^{-4}$,模型训练150 epoch,并且每训练50 epoch,学习率衰减为原来的1/2($l/2$)。

CrowdHuman:输入图像大小为800×1 200像素。在训练过程中设置 $b=8$ (每个GPU上设置2张图像,共使用4个GPUs),初始学习率 $l=2 \times e^{-4}$,模型训练300 epoch,并且每训练100 epoch,学习率衰减为原来的1/2($l/2$)。

3.2 消融实验

通过在CityPersons数据集上的消融实验来验证GFFIE-Net的有效性。首先,研究不同全局信息提取方法对实验结果的影响;接着,分析不同模块对网络效果的影响;然后,探究特征增强在不同尺度特征图上对网络效果的影响,以及顺序级联结构和分级特征融合的合理性;最后,通过不同主干网络测试验证本文设计的增强模块的鲁棒性。

3.2.1 不同全局信息提取方法对实验结果的影响

表1为不同全局信息提取方法对网络性能的影响。除采用Mamba^[25]提取4个不同分辨率特征图的全局信息外,同时设计了基于Transformer自注意力机制^[13]的方法以及多层感知机(MLP)模块^[31]来进行对比实验。实验表明,采用Transformer、MLP和Mamba作为全局信息提取方法时,网络检测精度均显著提升。在Heavy子集上与基准网络相比,Transformer使 MR^{-2} 提升了3.9个百分点,MLP使其提升了1.2百分点,Mamba使其提升了4.0百分点。在3种全局特征提取方法中,Transformer和Mamba对网络性能的提升效果最为显著。然而,Transformer的计算开销(显存)和参数量远远高于Mamba。具体而言,Transformer的计算开销约是Mamba的3.6倍,参数量约是Mamba的1.7倍,这主要归因于其二次时间复杂度,在特征图分辨率增大时,计算开销和参数量呈指数级增长。考虑到实际应用场景对计算资源和实时性的要求,以及Mamba模块在与本算法其他模块配合时表现出的良好兼容性和协同效应,综合权衡两者性能后,确定采用Mamba模块作为全局特征提取方法。

表1 不同全局信息提取方法对网络性能的影响

Tab. 1 Impact of different global information extraction methods on network performance

方法	漏检率二次指数/%				显存/GB	参数量/M	单图推理时间/s
	Reasonable	Partial	Bare	Heavy			
CSP+HRNet32 ^[28] (基准网络)	10.4	9.3	6.8	48.1	30.6*	29.8*	0.14*
CSP+HRNet32+MLP ^[31]	10.8	9.2	7.8	46.9	43.2	34.7	0.19
CSP+HRNet32+Transformer ^[13]	10.3	9.0	6.7	44.2	135.2	54.2	0.42
CSP+HRNet32+Mamba ^[25]	9.8	8.8	6.8	44.1	37.5	31.3	0.17
GFFIE-Net+Mamba(本文方法)	9.5*	8.5*	6.5*	43.7*	37.9	31.5	0.17

注:*表示最佳结果。

3.2.2 核心模块对网络性能的影响

CBAM 模块、Mamba 模块和分级特征融合机制是 GFFIE-Net 的 3 个核心部分。为了验证 3 个核心部分对网络性能的影响,本文进行了消融实验,实验结果如表 2 所示。相比于 CSP+HRNet32(基准网络),CSP+HRNet32+CBAM 和 CSP+HRNet32+Mamba 在 Heavy 子集上的漏检率分别降低了 3.8 个百分点和 4.0 个百分点,两个模块顺序级联协同作用下的 CSP+HRNet32+

CBAM+Mamba 漏检率降低了 4.2 个百分点,这表明了 CBAM 和 Mamba 模块处理行人遮挡的有效性。对于单独增加分级特征融合机制的 CSP+HRNet32+Fuse 来说,虽然其在 Heavy 子集上的性能提升不多,但由于其高效特征融合的特性,在 Bare 子集上有显著提升。最后将 3 个核心部分共同作用得到的 GFFIE-Net(表 2 最后一行)在所有遮挡子集上相比于基准网络均有大幅度提升,这证明了本文模型处理行人遮挡的高效性。

表 2 核心部分对网络性能的影响

方法				漏检率二次指数/%			
CSP+HRNet32(基准网络)	CBAM	Mamba	Fuse	Reasonable	Partial	Bare	Heavy
√				10.4	9.3	6.8	48.1
√	√			10.3	9.6	6.7	44.3
√		√		9.8	8.8	6.8	44.1
√			√	10.6	8.6	6.3*	45.1
√	√	√		9.7	8.9	6.6	43.9
√	√	√	√	9.5*	8.5*	6.5	43.7*

注:√表示选用,*表示最佳结果。

3.2.3 不同尺度对网络性能的影响,顺序级联结构及分级特征融合的合理性

为了进一步探究本文特征增强方法对不同尺度特征图的增强效果,对主干网络输出的 4 个不同分辨率特征图进行了特征增强消融分析,10 组实验的设计及结果如表 3 所示。其中,特征图与输入图像的分辨率比例分别为 1/4、1/8、1/16 和 1/32。与未进行特征增强的 CSP+HRNet32(基准网络)对比,在前 7 组实验中(CBAM 顺序级联 Mamba),实验 1、2 和 5 是对 1/4 和 1/8 高分辨率特征图进行特征增强实验,在 Heavy 子集上,经特征增强后的网络性能提升较为明显,这主要归因于高分辨率特征图本身蕴含丰富的细节位置信息,在经过 Mamba 模块的全局建模后,能够充分学习被遮挡行人的上下文信息,并有效结合行人可见区域实现对整体行人的精准推理;实验 3、4 和 6 是对 1/16 和 1/32 低分辨率特征图进行特征增强实验,在 3 个低遮挡子集(Reasonable、Partial 和 Bare)上性能提升较突出,这是因为低分辨率特征图具有较多的通道数,承载着丰富的深层语义信息,而 CBAM 模块通过其独特的空间和通道维度重新加权机制,能够有效抑制存在较多干扰特征的通道,同时强化包含较多目标信息通道的权重,从而特别适合处理遮挡较少但受背景干扰(如强光、夜间、低分辨率等场景)的行人目标;实验 7 是同时对 4 个尺度特征图进行特征增强,其结合了 CBAM 处

理干扰背景下低遮挡行人和 Mamba 处理严重遮挡行人的优势,在 4 个遮挡子集的综合性能最好。实验 7、8 和 9 是对模块整合设计进行消融分析。实验 8 采用 Mamba 顺序级联 CBAM 的方式先全局建模后聚焦去噪,由于特征图中存在的噪声在早期未得到有效处理,严重干扰了 Mamba 模块的全局建模过程,最终检测性能不佳;实验 9 采用 CBAM 并联 Mamba 的方式,试图并行处理噪声与全局信息,然而在实际运行中,两者在并行处理过程中缺乏有效的协同机制,无法充分发挥各自的优势,导致无法实现有效的特征增强。综上所述,这 9 组实验的对比分析清晰地表明 CBAM 顺序级联 Mamba(实验 7)的设计,即先去噪后全局建模的方式,能够最有效地应对行人遮挡问题,显著提升模型的检测性能。此外,前 9 组实验采用了分级特征融合机制,先对 3 个高维低分辨率特征图进行初步融合,再与低维高分辨率特征图进行最终融合,为了充分证明分级融合的有效性,在第 10 组实验中直接对 4 个不同分辨率特征图进行了一次性融合。实验结果表明,分级融合机制在所有遮挡子集上均优于一次性拼接,特别是在 Heavy 子集上,性能提升了 1.1 个百分点。分级融合的设计不仅减少了信息冗余,还增强了特征表示的层次性,使模型能够更好地结合语义信息和细节信息,从而在处理被遮挡行人时表现出色。

表3 不同尺度对网络性能的影响,顺序级联结构和分级特征融合的合理性

Tab. 3 Impact of different scales on network performance, rationalization of sequential cascade structure and hierarchical feature fusion

实验编号	协同方式	融合方式	分辨率比例				漏检率二次指数/%			
			1/4	1/8	1/16	1/32	Reasonable	Partial	Bare	Heavy
1	CBAM→Mamba	分级融合	√				10.4	9.3	7.0	44.5
2	CBAM→Mamba	分级融合		√			10.5	9.0	6.8	44.8
3	CBAM→Mamba	分级融合			√		9.9	8.8	6.8	45.9
4	CBAM→Mamba	分级融合				√	9.8	8.8	6.5	46.6
5	CBAM→Mamba	分级融合	√	√			10.2	8.9	6.8	43.9
6	CBAM→Mamba	分级融合			√	√	9.7	8.5	6.4*	45.4
7	CBAM→Mamba	分级融合	√	√	√	√	9.5*	8.5*	6.5	43.7*
8	Mamba→CBAM	分级融合	√	√	√	√	10.3	8.9	6.8	44.0
9	CBAM 并联 Mamba	分级融合	√	√	√	√	9.9	8.8	6.5	44.6
10	CBAM→Mamba	直接融合	√	√	√	√	10.2	9.1	6.9	44.8

注:→表示顺序级联,√表示选用,*表示最佳结果。

3.2.4 不同主干网络对性能的影响

为了验证增强模型在不同主干网络的输出特征图上进行特征增强的有效性和泛化性,进行了以下实验:通过将 GFFIE-Net 的主干网络由 HRNet32^[28] 替换为 ResNet50^[32]、VGG16^[33] 和 ConvMLP_L^[34] 来构建 GFFIE+ResNet50、GFFIE+VGG16 和 GFFIE+ConvMLP_L 模型。随后,将这些模型与仅使用 ResNet50、VGG16 和 ConvMLP_L 作为主干网络的初始模型 CSP 进行比较。表 4 为不同主干网络对性能的影响。由表 4 可见,GFFIE 在 HRNet32、ResNet50、VGG16 和 ConvMLP_L 为主干网络的情况下均表现出优异的检测性能。这不仅进一步证明了增强模型的有效性,也验证了其在不同主干网络的鲁棒性和泛用性。

表4 不同主干网络对性能的影响

Tab. 4 Impact of different backbone networks

方法	漏检率二次指数/%			
	Reasonable	Partial	Bare	Heavy
CSP+ResNet50 ^[32]	11.0	10.4	7.3	49.3
GFFIE+ResNet50 ^[32]	9.9	8.9	6.6	45.2
CSP+VGG16 ^[33]	11.6	11.9	7.1	51.1
GFFIE+VGG16 ^[33]	10.3	9.1	6.8	46.2
CSP+ConvMLP_L ^[34]	14.6	14.2	8.9	49.1
GFFIE+ConvMLP_L ^[34]	10.9	10.9	6.6	46.5
CSP+HRNet32 ^[31]	10.4	9.3	6.8	48.1
GFFIE+HRNet32 (本文方法)	9.5*	8.5*	6.5*	43.7*

注:*表示最佳结果。

3.3 对比实验

3.3.1 CityPersons 数据集对比实验

为了验证本文方法处理遮挡行人的有效性,将

GFFIE-Net 与当前的前沿算法在 CityPersons^[17] 数据集上进行了对比,实验结果如表 5 所示。

表5 CityPersons 数据集对比实验结果

Tab. 5 Comparative experimental results on the CityPersons dataset

方法	漏检率二次指数/%				单图 推理时间/s
	Reasonable	Partial	Bare	Heavy	
TLL ^[35]	15.5	17.2	10.0	53.6	
LBST ^[36]	12.6	18.6		48.7	
Cascade R-CNN ^[6]	11.2				0.73
Bi-box ^[37]	15.1	15.3	11.0	60.0	0.27
RepLoss ^[10]	13.2	16.8	7.6	56.9	
PR ² NMS ^[38]	11.1			53.3	0.22
OR-CNN ^[39]	12.8	15.3	6.7	55.7	
AP ² M ^[40]	10.4	9.7	6.2	48.6	
EGCL ^[41]	10.9	11.6	7.4	46.4	0.25
CSP+HRNet32 ^[9] (基准网络)	10.4	9.3	6.8	48.1	0.14*
PRNet++ ^[42]	10.7	9.9	6.9	51.2	0.20
BIA-NMS ^[11]	10.1	10.2	6.1*	47.3	0.48
MHAS ^[43]	10.7				
SMPD ^[44]	9.9	9.0	6.5	45.6	0.17
CaSe+ OTP-NMS ^[12]	9.5*			47.8	
CrowdDet+ OTP-NMS ^[12]	10.4			48.6	
VA-CSP ^[45]	9.6	9.1	6.6	48.1	
DAFC+ ^[46]	11.8			44.8	
GFFIE-Net (本文方法)	9.5*	8.5*	6.5	43.7*	0.17

注:*表示最佳结果。

首先,为了确保与CSP+HRNet32(基准网络)的公平比较,在相同的配置下对基准网络进行了复现,本文方法在4个子集上均实现了不同程度的性能增益,其中,在Heavy子集上更是实现了4.4个百分点的性能增益。其次,与SMPD^[44]、MHAS^[43]、EGCL^[41]、DAFC+^[46]等单阶段遮挡处理算法对比,其中,DAEC+和EGCL都与特征增强相关,前者通过特征校准重点关注可见区域,后者采用对比学习以指导特征学习,本文方法在4个子集上均表现更优,特别是在Heavy子集上漏检率二次指数达到43.7%。即使与最近提出的针对行人严重遮挡问题的改进NMS策略模型BIA-NMS^[11]、CaSe+OTP-NMS^[12]和CrowdDet+OTP-NMS^[12]相比,本文方法在Heavy子集上的性能也实现了超越。在模型推理速度方面,本文方法的检测速度为0.17 s/图,比两阶段方法Cascade R-CNN^[6]和BIA-NMS^[11]有大幅度提升,即使与上述单阶段方法SMPD^[44]、MHSA^[43]、EGCL^[41]相比,本文方法在模型整体推理速度上也实现了超越。综上所述,本文方法能在较快的推理速度下有效地处理静态图片中不同遮挡程度的行人目标。

3.3.2 Caltech数据集对比实验

表6为Caltech数据集对比实验结果,呈现了本文方法和近年来的主流方法在Caltech^[29]数据集上的性能对比。

表6 Caltech数据集对比实验结果
Tab. 6 Comparative experimental results on the Caltech dataset

方法	漏检率二次指数/%		
	All	PO	HO
FRCN ^[4]	62.6	8.7	53.1
Bi-box ^[37]	56.8	6.4	38.7
RepLoss ^[10]	59.0	5.0	47.9
MRCN ^[5]	59.8	7.4	48.7
CSP+HRNet32(基准网络) ^[9]	57.2	5.2	45.3
AP ² M ^[40]	55.9	3.3	42.2
SMPD ^[44]	55.2	4.2	44.8
MHAS ^[43]		4.3	
VLPD ^[47]	52.4 [*]	2.3 [*]	37.7
F-CSP ^[48]	60.5	6.0	51.5
VA-CSP ^[45]		3.2	34.2
GFFIE-Net(本文方法)	52.9	3.1	33.6 [*]

注:*表示最佳结果。

由表6可见,本文方法相比于CSP+HRNet32(基准网络)在3个子集上有大幅度性能提升。与最新遮挡处理方法AP²M^[40]、SMPD^[44]和MHAS^[43]相比,也

有着很强的竞争力。在PO子集和All子集上,本文方法的性能略低于VLPD^[47],这是由于VLPD针对检测器对类人物体的漏检问题提出了基于语义自监督的视觉语言模型,对未被遮挡或轻度被遮挡行人的特征提取和处理更为精细,但本文方法相比VLPD在处理HO子集时上漏检率二次指数提高了4.1个百分点,表现最好。与同样使用中心和尺度预测策略的方法F-CSP^[48]和VA-CSP^[45]相比,本文方法也表现更好。这些结果充分证明了本文方法在处理行人遮挡问题上的有效性。

3.3.3 CrowdHuman数据集对比实验

表7为CrowdHuman数据集对比实验结果,呈现了GFFIE-Net与最新研究方法在CrowdHuman数据集上的性能对比。

表7 CrowdHuman数据集对比实验结果
Tab. 7 Comparative experimental results on the Crowd-Human dataset

方法	漏检率二次指数/%	召回率/%	平均精度/%
RFB-Net ^[49]	65.2	94.1	78.3
GossipNet ^[50]	49.4		80.4
RelationNet ^[51]	48.2		81.6
RepLoss ^[10]	45.7	88.4	85.6
FRCN+A+Sp ^[52]	45.1		
PR ² NMS ^[38]	43.4	93.3	89.3
FCOS+AEVB ^[53]	47.7		
Deformable DETR ^[16]	54.0	92.5	86.7
POTO ^[54]	52.2	96.3	88.5
POTO+3DMF+Aux ^[54]	48.9	96.5	89.1
OAF-Net ^[20]	45.0	96.5	89.8
Bi-Center ^[19]	44.6	96.5	89.9
CFRLA-Net ^[55]	44.8	96.7	89.9
OPLA ^[56]	51.6		88.9
DLA ^[57]	44.3	96.1	90.1
DAFC+ ^[46]			89.8
GFFIE-Net(本文方法)	43.2 [*]	96.7 [*]	90.7 [*]

注:*表示最佳结果。

由表7可见,本文方法在3个指标上均为最佳。首先,Deformable DETR^[16]主要借助可变形注意力机制来调整输入特征,以处理遮挡问题,然而在实际应用中,其在复杂遮挡场景下的信息捕捉和特征整合能力仍存在一定局限性,与其相比,本文方法在3个指标上分别提升了10.8个百分点、4.2个百分点和4.0个百分点。其次,与最新提出的DLA^[57]、OPLA^[56]和DAFC+^[46]相比,本文方法检测性能更佳。最后,与先进的遮挡处理方法OAF-Net^[20]、Bi-Center^[19]、CFRLA-Net^[55]、

DAFC+^[46]相比,本文方法在 CrowdHuman 数据集上实现了显著的性能提升。这主要得益于 CBAM 模块对特征图的聚焦去噪以及 Mamba 模块对全局信息的高效建模,使模型能够更准确地识别和定位遮挡行人。实验结果进一步证明了本文方法可以有效地在实用场景下处理遮挡行人目标。

4 实验结果可视化

图 4 为检测框与中心热图可视化结果,从 3 个数据集中提取了 7 种具有代表性的实用场景图片进行分析。由图 4 可见,第 1 列为输入图像,直观呈现了原始场景中的行人分布与遮挡状况;第 2 列为基准网络

(CSP+HRNet32)的检测框与中心热图可视化结果,红色框为输出检测框,蓝色虚线框标识了漏检的遮挡行人目标,可以明显看出在多种场景中,由于遮挡行人的特征严重缺失,基准网络出现了不同程度的漏检现象,如在交通密集场景中,当多个行人相互遮挡时,基准网络的检测框出现了大量遗漏,中心热图对被遮挡行人的响应也十分微弱;相比之下,第 3 列的本文方法对多尺度特征图进行了聚焦去噪,提取了全局上下文信息推理被遮挡行人,对被遮挡行人的中心响应更为显著,检测框定位更为准确。实验结果可视化进一步验证了本文方法在处理不同场景行人遮挡问题上的有效性。



图 4 检测框与中心热图可视化结果

Fig. 4 Detection box and center heat map visualization result

5 结 论

本文针对行人遮挡问题,通过设计 CBAM 顺序级联 Mamba 模块结构以及分级特征融合机制,提出了一种基于全局特征聚焦与信息增强的遮挡行人检测算法 GFFIE-Net。具体而言,在处理行人遮挡场景时,

CBAM 独特的双重精炼结构能精准地分析特征图各通道和空间位置的重要性,有效增强了被遮挡行人的微弱特征,同时强力抑制复杂背景噪声的干扰,从而引导网络将注意力聚焦于关键区域;顺序级联的 Mamba 模块则针对聚焦去噪后的特征图进行高效全局信息建模,提取被遮挡行人上下文信息,结合可见

特征推理完整行人;分级特征融合机制实现了多尺度特征图的信息交互,提升了算法对多尺度遮挡行人的检测能力。在3个数据集上的实验结果表明,GFFIE-Net相比现有方法取得了显著的性能提升。在CityPersons的Heavy子集上,漏检率二次指数达到43.7%,较基准网络提升4.4个百分点;在Caltech的HO子集上,漏检率二次指数达到33.6%;在CrowdHuman数据集上,漏检率二次指数达到43.2%,召回率达到96.7%,平均精度达到90.7%,与其他先进算法相比均处于领先地位,充分证明了本文方法处理遮挡行人检测任务上的高效性。

基于当前研究成果及实际应用需求,未来研究可考虑利用有限的标注数据进行半监督或自监督学习,减少对大规模标注数据的依赖,同时也可提高模型的泛化性,从而进一步提升算法在不同场景下的适用性和准确性。

参考文献:

- [1] Guo Yongcun, Yang Tun, Wang Shuang. Multi-object real-time detection of mine electric locomotive based on improved YOLOv4-tiny[J]. *Advanced Engineering Sciences*, 2023,55(5):232–241.[郭永存,杨豚,王爽.基于改进YOLOv4-tiny的矿井电机车多目标实时检测[J]. *工程科学与技术*, 2023,55(5):232–241.]
- [2] Chen Wei, Zhu Yuxuan, Tian Zijian, et al. Occlusion and multi-scale pedestrian detection A review[J]. *Array*, 2023, 19:100318.
- [3] Zhang Guanghua, Li Congfa, Li Gangying, et al. Small target detection algorithm for UAV aerial images based on improved YOLOv7-tiny[J]. *Advanced Engineering Sciences*, 2025,57(3):235–246.[张光华,李聪发,李钢硬,等.基于改进YOLOv7-tiny的无人机航拍图像小目标检测算法[J]. *工程科学与技术*, 2025,57(3):235–246.]
- [4] Ren Shaoqing, He Kaiming, 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.
- [5] He Kaiming, Gkioxari G, Dollár P, et al. Mask R-CNN[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020,42(2):386–397.
- [6] Cai Zhaowei, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection[C]//*Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 2018:6154–6162.
- [7] Law H, Deng Jia. CornerNet: Detecting objects as paired keypoints[C]//*Computer Vision–ECCV 2018*. Cham: Springer, 2018:765–781.
- [8] Duan Kaiwen, Bai Song, Xie Lingxi, et al. CenterNet: Key-point triplets for object detection[C]//*Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Piscataway: IEEE, 2019:6568–6577.
- [9] Liu Wei, Liao Shengcai, Ren Weiqiang, et al. High-level semantic feature detection: A new perspective for pedestrian detection[C]//*Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach: IEEE, 2019:5182–5191.
- [10] Wang Xinlong, Xiao Tete, Jiang Yuning, et al. Repulsion loss: Detecting pedestrians in a crowd[C]//*Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 2018:7774–7783.
- [11] Abdelmutalab A, Wang Chunyan. Pedestrian detection using MB-CSP model and boosted identity aware non-maximum suppression[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2022,23(12):24454–24463.
- [12] Tang Yi, Liu Min, Li Baopu, et al. OTP-NMS: Toward optimal threshold prediction of NMS for crowded pedestrian detection[J]. *IEEE Transactions on Image Processing*, 2023, 32:3176–3187.
- [13] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. (2021–06–03)[2024–10–20]. <https://doi.org/10.48550/arXiv.2010.11929>.
- [14] Meng Lingbing, Yuan Mengya, Shi Xuehan, et al. RGB-D salient object detection with three-branch multi-level transformer feature interaction[J]. *Advanced Engineering Sciences*, 2023,55(6):245–256.[孟令兵,袁梦雅,时雪涵,等.3分支多层次Transformer特征交互的RGB-D显著性目标检测[J]. *工程科学与技术*, 2023,55(6):245–256.]
- [15] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//*Computer Vision–ECCV 2020*. Cham: Springer, 2020:213–229.
- [16] Zhu Xizhou, Su Weijie, Lu Lewei, et al. Deformable DETR: Deformable transformers for end-to-end object detection [EB/OL]. (2021–03–18)[2024–10–20]. <https://doi.org/10.48550/arXiv.2010.04159>.
- [17] Zhang Shanshan, Benenson R, Schiele B. CityPersons: A diverse dataset for pedestrian detection[C]//*Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway: IEEE, 2017:4457–4465.
- [18] Shao Shuai, Zhao Zijian, Li Boxun, et al. CrowdHuman: a benchmark for detecting human in a crowd[EB/OL]. (2018–04–30)[2024–10–20]. <https://doi.org/10.48550/arXiv.1805.00123>.
- [19] Li Qiming, Bi Yuquan, Cai Rongsheng, et al. Occluded pedestrian detection through Bi-Center prediction in anchor-free network[J]. *Neurocomputing*, 2022,507:199–207.
- [20] Li Qiming, Su Yijing, Gao Yin, 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.
- [21] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module[C]// *Computer Vision—ECCV 2018*. Cham: Springer, 2018: 3–19.
- [22] Li Ruihong, Fu Zhitao, Zhang Shaochen, et al. Nighttime object detection in infrared and visible images based on multi-attention mechanism[J]. *Infrared Technology*, 2024, 46(12): 1371–1379. [黎瑞虹, 付志涛, 张韶琛, 等. 基于多注意力机制的红外与可见光图像夜间目标检测[J]. *红外技术*, 2024, 46(12): 1371–1379.]
- [23] Zang Ying, Cao Runlong, Li Hui, et al. MAPD: Multi-receptive field and attention mechanism for multispectral pedestrian detection[J]. *The Visual Computer*, 2024, 40(4): 2819–2831.
- [24] Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces[EB/OL]. (2023–12–01)[2024–10–20]. <https://doi.org/10.48550/arXiv.2312.00752>.
- [25] Zhu Lianghui, Liao Bencheng, Zhang Qian, et al. Vision Mamba: efficient visual representation learning with bidirectional state space model[EB/OL]. (2024–11–14)[2024–10–20]. <https://arxiv.org/abs/2401.09417v3>.
- [26] Ding Zhengze, Nie Rencan, Li Jintao, et al. MTFuse: An infrared and visible image fusion network based on mamba and transformer[J]. *Computer Science*, 2025, 52(8): 188–194. [丁政泽, 聂仁灿, 李锦涛, 等. MTFuse: 基于 Mamba 和 Transformer 的红外与可见光图像融合网络[J]. *计算机科学*, 2025, 52(8): 188–194.]
- [27] Shi Yangyu, Xie Chengjie, Zheng Diwen, et al. Multi-scale anomaly behavior detection based on Mamba–CNN[J/OL]. *Journal of Beijing University of Aeronautics and Astronautics*, [2025–01–08]. [石洋宇, 谢承杰, 郑棣文, 等. 基于 Mamba–CNN 的多尺度异常行为检测方法[J/OL]. *北京航空航天大学学报*, [2025–01–08].]
- [28] Sun Ke, Xiao Bin, Liu Dong, et al. Deep high-resolution representation learning for human pose estimation[C]// *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway: IEEE, 2019: 5686–5696.
- [29] Dollar P, Wojek C, Schiele B, et al. Pedestrian detection: An evaluation of the state of the art[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 34(4): 743–761.
- [30] Loshchilov I, Hutter F. Decoupled weight decay regularization[EB/OL]. (2017–11–14)[2024–10–20]. <https://doi.org/10.48550/arXiv.1711.05101>.
- [31] Quan Yu, Zhang Dong, Zhang Liyan, et al. Centralized feature pyramid for object detection[J]. *IEEE Transactions on Image Processing*, 2023, 32: 4341–4354.
- [32] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]// *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway: IEEE, 2016: 770–778.
- [33] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014–09–04)[2024–10–20]. <https://doi.org/10.48550/arXiv.1409.1556>.
- [34] Li Jiachen, Hassani A, Walton S, et al. ConvMLP: Hierarchical convolutional MLPs for vision[C]// *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Piscataway: IEEE, 2023: 6307–6316.
- [35] Song Tao, Sun Leiyu, Xie Di, et al. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation[M]// *Computer Vision—ECCV 2018*. Cham: Springer International Publishing, 2018: 554–569.
- [36] Cao Jiale, Pang Yanwei, Han Jungong, et al. Taking a look at small-scale pedestrians and occluded pedestrians[J]. *IEEE Transactions on Image Processing*, 2020, 29: 3143–3152.
- [37] Zhou Chunlun, Yuan Junsong. Bi-box regression for pedestrian detection and occlusion estimation[C]// *Computer Vision—ECCV 2018*. Cham: Springer International Publishing, 2018: 138–154.
- [38] Huang Xin, Ge Zheng, Jie Zequn, et al. NMS by representative region: Towards crowded pedestrian detection by proposal pairing[C]// *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway: IEEE, 2020: 10747–10756.
- [39] Zhang Shifeng, Wen Longyin, Bian Xiao, et al. Occlusion-aware R–CNN: Detecting pedestrians in a crowd[C]// *Computer Vision—ECCV 2018*. Cham: Springer International Publishing, 2018: 657–674.
- [40] Liu Mengyin, Zhu Chao, Wang Jun, et al. Adaptive pattern-parameter matching for robust pedestrian detection[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 35(3): 2154–2162.
- [41] Lin Zebin, Pei Wenjie, Chen Fanglin, et al. Pedestrian detection by exemplar-guided contrastive learning[J]. *IEEE Transactions on Image Processing*, 2022, 32: 2003–2016.
- [42] Song Xiaolin, Chen Binghui, Li Pengyu, et al. PRNet++: Learning towards generalized occluded pedestrian detection via progressive refinement network[J]. *Neurocomputing*, 2022, 482: 98–115.
- [43] Yuan Jing, Stathaki T, Ren Guangyu. Mean height aided post-processing for pedestrian detection[EB/OL]. (2024–08–24)[2025–01–08]. <https://doi.org/10.48550/arXiv.2408.13646>.
- [44] Jiang Hangzhi, Liao Shengcai, Li Jinpeng, et al. Urban scene

- based semantical modulation for pedestrian detection[J]. *Neurocomputing*,2022,474:1–12.
- [45] Lin Xincheng, Tang Yang, Zhao Chaoqiang, et al. Visible attention mechanism-based anchor-free model for pedestrian detection[J]. *Control Engineering of China*, 2024, 31(3):535–544.[林鑫辰,唐漾,赵超强,等.基于可视注意力机制的非锚点行人检测模型[J]. *控制工程*,2024,31(3):535–544.]
- [46] Tang Shuyuan, Zhou Yiqing, Li Jintao, et al. Dual attention pedestrian detector for occlusion scenario based on feature calibration[J]. *Journal of Xidian University*, 2024, 51(6):25–39.[汤书苑,周一青,李锦涛,等.基于特征校准的双注意力遮挡行人检测器[J]. *西安电子科技大学学报*,2024,51(6):25–39.]
- [47] Liu Mengyin, Jiang Jie, Zhu Chao, et al. VLPD: Context-aware pedestrian detection *via* vision-language semantic self-supervision[C]//*Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway:IEEE,2023:6662–6671.
- [48] Zhang Tao, Cao Yahui, Zhang Le, et al. Efficient feature fusion network based on center and scale prediction for pedestrian detection[J]. *The Visual Computer*, 2023, 39(9):3865–3872.
- [49] Liu Songtao, Huang Di, Wang Yunhong. Receptive field block net for accurate and fast object detection[M]//*Computer Vision—ECCV 2018*. Cham: Springer International Publishing, 2018:404–419.
- [50] Hosang J, Benenson R, Schiele B. Learning non-maximum suppression[C]//*Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway:IEEE,2017:6469–6477.
- [51] Hu Han, Gu Jiayuan, Zhang Zheng, et al. Relation networks for object detection[C]//*Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway:IEEE,2018:3588–3597.
- [52] Zhang Shanshan, Chen Di, Yang Jian, et al. Guided attention in CNNs for occluded pedestrian detection and re-identification[J]. *International Journal of Computer Vision*, 2021, 129(6):1875–1892.
- [53] Zhang Yuang, He Huanyu, Li Jianguo, et al. Variational pedestrian detection[C]//*Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway:IEEE,2021:11617–11626.
- [54] Wang Jianfeng, Song Lin, Li Zeming, et al. End-to-end object detection with fully convolutional network[C]//*Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, Piscataway:IEEE,2021:15844–15853.
- [55] Li Jun, Bi Yuquan, Wang Sumei, et al. CFRLA-net: A context-aware feature representation learning anchor-free network for pedestrian detection[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, 33(9):4948–4961.
- [56] He Haoyang, Li Zhishan, Tian Guanzhong, et al. Towards accurate dense pedestrian detection *via* occlusion-prediction aware label assignment and hierarchical-NMS[J]. *Pattern Recognition Letters*, 2023, 174:78–84.
- [57] Zhang Yi, Luo Chen. A dynamic label assignment strategy for one-stage detectors[J]. *Neurocomputing*, 2024, 577:127383.

Global Feature Focusing and Information Enhancement Network for Occluded Pedestrian Detection

ZHENG Kaikui¹, JI Kangyou¹, LI Jun², LI Qiming^{2*}

(1.School of Advanced Manufacturing, Fuzhou University, Jinjiang 362251, China;

2.Quanzhou Institute of Equipment Manufacturing, Haixi Institute, Chinese Academy of Science, Quanzhou 362200, China)

Abstract:

Objective Pedestrian detection is a crucial task in computer vision, particularly in applications such as autonomous driving, robot navigation, and intelligent surveillance. However, pedestrian occlusion in real-world scenarios remains a significant challenge. Occlusion causes a sharp reduction in the visible range of targets and a substantial loss of pedestrian features, making it difficult for detectors to effectively distinguish between targets and pedestrians. Existing methods, including post-processing optimization, specific model-based improvements, and body-part feature-based approaches, have limitations such as inaccurate handling of heavily occluded positive samples, high computational complexity, and susceptibility to background noise. Therefore, developing a more effective method to address pedestrian occlusion detection is essential to enhance the performance of pedestrian detectors.

Methods The proposed global feature focusing and information enhancement network (GFFIE-Net) employed HRNet-W32 as the backbone network to generate multi-scale feature maps with different resolutions (1/4, 1/8, 1/16, and 1/32 of the input image). These feature maps captured both high-level semantic information and low-level spatial details, which were essential for detecting pedestrians in complex scenes. The convolutional block attention module (CBAM) was embedded after the feature maps to enhance the feature representation and reduce background noise interference. CBAM adjusted the importance of each channel and spatial location in the feature maps through operations such as global average

