语义增强引导特征重建的遮挡行人检测

孙旭旦，吴　清，赵春艳，张满囤

(河北工业大学 人工智能与数据科学学院，天津 300401)

摘　要：行人被严重遮挡导致无法提取有效特征是行人检测中出现漏检的一个主要原因。为了解决该问题，提出一种语义增强引导特征重建的遮挡行人检测算法。首先，利用空间和通道之间的依赖性设计了语义特征增强模块，建立全局上下文信息用以增强遮挡行人特征。其次，为关注行人的可见区域，通过自适应特征重建模块生成语义分割图，自适应调整通道的有效权重，增强行人和背景的可判别性。最后，通过多层次级联语义特征增强和自适应特征重建两个模块得到多层次特征图，融合多特征用以最终的行人解析。实验结果表明，该方法在具有挑战性的行人检测基准 CityPersons 和 Caltech 上，对严重遮挡目标的漏检率分别实现了 47.28% 和 44.04%，在遮挡行人的检测上相较于其他方法具有较好的鲁棒性。

关键词：行人检测；　语义特征增强；　特征重建；　语义分割

中图分类号：TP391.4　　**文献标志码：**A　　**DOI：**10.3788/IRLA20210924

Semantic enhanced guide feature reconstruction for occluded pedestrian detection

Sun Xudan，Wu Qing，Zhao Chunyan，Zhang Mandun

(School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China)

Abstract: In pedestrian detection, the inability to extract effective features due to pedestrians being severely occluded is one of the main reasons for missing pedestrian detection. To solve this problem, a semantic enhanced guided feature reconstruction algorithm for occlusion pedestrian detection is proposed. Firstly, the semantic feature enhancement module is designed based on the dependency between space and channel, and the global context information is established to enhance the feature of occlusion of pedestrians. Secondly, in order to focus on the visible area of pedestrians, the adaptive feature reconstruction module is used to generate the semantic segmentation map, and adaptively adjust the effective weight of the channel, enhance the distinguishability of pedestrians and background. Finally, the multi-level feature map is obtained by multi-level cascade two modules of semantic feature enhancement and adaptive feature reconstruction, and the multiple features are fusion for the final pedestrian analysis detection. On the challenging pedestrian detection benchmark CityPersons and Caltech, experimental results show that the proposed method achieves the missed rate of 47.28% and 44.04%, respectively, which effectively robust compared with other methods in the detection of occluded pedestrian.

Key words: pedestrian detection; semantic feature enhancement; feature reconstruction; semantic segmentation

---

收稿日期:2021−11−30;　**修订日期:**2021−12−16

基金项目:河北省自然科学基金 (F2019202054, F2019202381)

作者简介:孙旭旦, 女, 硕士生, 主要从事图像处理与模式识别方面的研究。

导师简介:吴清, 女, 教授, 硕士生导师, 博士, 主要从事图像处理与模式识别方面的研究。

0 引 言

行人检测是智能交通系统的核心模块,在自动驾驶、视频监控、人机交互等领域有重要的应用[1−2]。与一般的目标检测[3−4]不同,行人检测需要解决遮挡问题,目前的行人检测算法虽能很好地处理未被遮挡或少量遮挡的行人场景,但在严重遮挡的行人检测上仍存在缺陷。自然环境下的行人目标不可避免会被遮挡,网络在提取特征时,这些被遮挡的目标往往会被认为是不感兴趣的背景区域,从而造成漏检。如何解决行人被遮挡带来的漏检问题是文中行人检测的研究重点。

目前,提出的许多有效解决方案均利用了卷积神经网络模型[5−6]。针对特定的遮挡模式,出现了训练局部检测器的方法,Noh 等人[7]提出学习并单独训练局部检测器,最后综合所有检测器的分数集成的结果来降低对行人检测的漏检率。Yang 等人[8]引入了并行分支,用具有部分感知的感兴趣池化去处理更大和更小的目标。优化损失函数调整预测框之间的距离,可以提高检测器的性能,Wang 等人[9]提出人群场景中行人之间的斥力损失,增加惩罚项,使预测框尽可能地接近真实框,远离其他目标真实框和预测框来提高行人的定位精度。Zhang 等人[10]通过设计一种新颖的损失函数 AggLoss 和遮挡感知的 Rol 池化,让模型去学习行人实例的不同部位来定位目标。为了处理严重的遮挡问题,出现了联合训练不同模式的方法,Zhou 等人[11]提出回归两个边界框,分别定位行人的全身和可见部分,以利用两个预测任务的互补性来处理遮挡。Zhang 等人[12]利用可见区域生成注意力机制,用作学习遮挡模式的外部监督。Luo 等人[13]采用双模全卷积网络检测行人。尽管这些方法在一些数据集上取得了出色的成就,但大部分的方法都是基于候选区域或基于 Anchor-base,需要预先设定锚框,在分配方式上很难达到最优,也会带来更多的 Anchor 超参数,降低检测器的性能。近年来提出的基于 Anchor-free 的方法省去了锚框的设定,它网络结构简单,可以充分利用关键点特征进行检测,提高了检测速度。由 Liu[14]提出的 CSP 检测算法,首次将 Anchor-free 应用到行人检测上,通过一些高级语义信息确定目标的中心点和尺度去预测行人检测框,摒弃了 NMS 操作,将图片输入到网络中,直接通过卷积将行人检测转化为中心点和尺度预测任务,这种检测方法简化了算法,减少了网络计算量和训练的时间,提高了检测速度和精度,在估计边界框的比例和纵横比方面更具灵活性。但该模型并未针对解决遮挡模式,同时在后处理阶段采用标准的非极大值抑制算法,使得检测效果不能达到最优。为解决这些问题,文中提出了一种用于鲁棒行人检测的语义增强引导的特征重建网络,可以充分利用特征进行检测,为遮挡目标的特征提取提供了新的途径。文中的主要贡献包括:

(1) 提出了语义特征增强模块,在空间和通道上进行全局上下文建模,建立特征与特征之间关联性,增强了遮挡行人的语义上下文信息。

(2) 提出了自适应的特征重建模块,利用像素级别的特征信息在动态过滤的模式下不断显式地调制输入特征,增加可见通道的有效权重,让网络去关注更多的行人信息,从而提高行人和背景的可分辨性。

(3) 引入了多层次级联方式,对模块进行不断的细化,让深层特征与浅层特征的有效融合以获得密集场景下更具鲁棒性的行人特征。在 Caltech[15]和 CityPersons[16]数据集上取得了较为优越的性能,与其他方法相比,能够更精准地检测出被遮挡的行人。

1 “面向遮挡行人”的语义增强引导特征重建网络架构

1.1 系统整体模型的构建

文中提出的基于语义增强引导特征重建的遮挡行人检测算法,由特征提取 (Backbone)、多层次级联网络 (Multi-level Cascade Network, MLCN) 和检测头 (Detection Head) 三部分组成。在此基础上,为了解决高重叠检测框带来的漏检问题,在后处理阶段采用 DIoU-NMS 算法,整个模型框架如图 1 所示。特征提取网络提取到多阶段特征后,进入多层次级联网络中,它有两个互补的模块,即语义特征增强模块 (Semantic Feature Enhancement Module, SFEM) 和自适应的特征重建模块 (Adaptivity Feature Reconstruction Module, AFRM)。在 SFEM 中通过改进的全局上下文块 (Improved Global Context Block, IGCB) 获取全局上下文信息,根据上下文信息去判断目标与背景之间的关联性以增强遮挡行人的特征;在 AFRM 中,将上一

阶段增强后的特征与骨干网络输出该级的特征进行通道间的融合形成语义分割掩码，编码得到的语义信息，自适应调制每个空间位置的特征表示，增强特征通道有效权重，让网络更加关注行人区域；自上而下级联这两个模块，使得提取的特征更加分离和清晰，尤其是对那些严重遮挡的行人目标。最后将每个阶段增强后的特征图融合到一起，进入检测头，分别预测行人的中心点、高度和偏移，最终生成行人边界框。

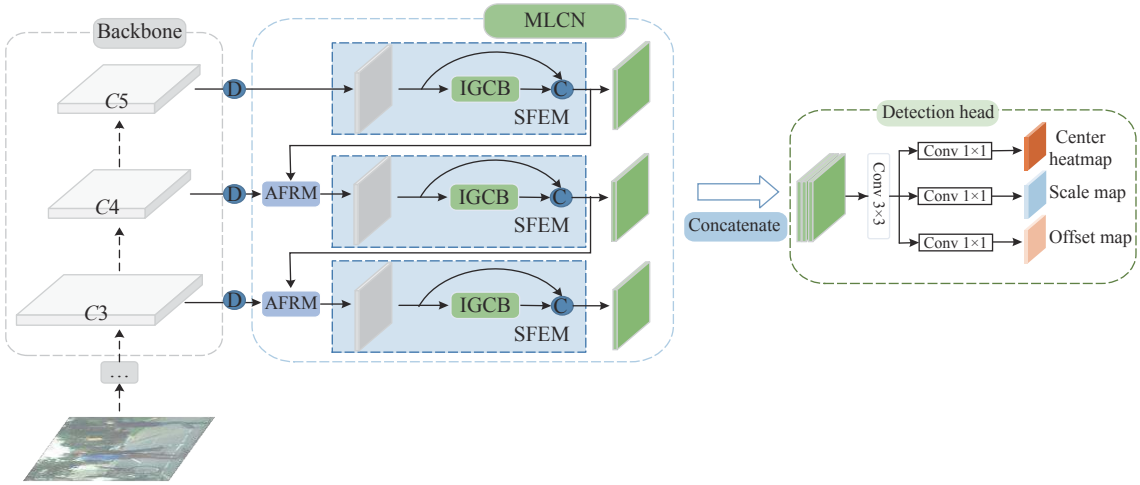


图 1 文中网络模型框架图

Fig.1 Framework diagram of proposed network model

1.2 特征提取网络

特征提取网络能够提取行人最具有判别力的特征，不同层的卷积特征代表不同的抽象级别，且具有不同的感受野域，可以为行人检测提供不同的线索。采用特征表达能力较强的 ResNet101 作为骨干网络，由于表征能力最强的特征存在于每个阶段的最深处，文中使用每个阶段的最后一个特征图作为每个分支输入的集合，同时也将这个输入集合以级联的方式进一步丰富。为了增大网络的感受野，增强网络的提取能力，利用空洞卷积将第五阶段的输出保持为输入图像尺寸的 1/16，最终得到的输出特征图为输入图像下采样的 2、4、8、16、16 倍。另外底层网络得到的特征图包含的信息较少，对检测结果影响不大，只采用阶段 3、4 和 5 的最后一个特征图 C3、*C*4和*C*5作为后续网络的输入。为了让不同尺度特征之间更好融合，将 *C*3、*C*4和*C*5通过反卷积进行上采样到输入图像的 1/4，将通道数降为 256。D 包括反卷积和归一化层操作，采用 3×3 的卷积核是为了消除卷积操作带来的冗余数据，使用归一化层来丰富浅层特征信息，可以在保证网络收敛的前提下有效降低网络对遮挡行人的漏检率，提升算法的收敛速度和鲁棒性。

1.3 行人检测器

行人检测器的任务是将提取的特征解析成行人检测框，文中采用基于中心点和尺度预测的方法，如图 1 所示的检测头模块。在融合了多层次级联网络输出的三阶段特征图后，附加检测头将其解析为检测结果。即先对融合的特征通过一个 3×3 卷积将特征通道数降为 256，再添加三个 1×1 卷积分别形成中心点热图、尺度特征图和偏移预测图。行人检测器的损失函数有三部分：中心点分类、尺度回归和偏移预测。

对于中心点分支，为了让训练过程更加收敛，在每个正样本中心位置使用了二维高斯掩码，如果这些掩码出现重叠，则将最大值作为该点的高斯值，表示为：

$$M_{ij}=\max\left(\mathrm{e}^{-\frac{(i-x_t)^2}{2\sigma^2 w_t}+\frac{(j-y_t)^2}{2\sigma^2 h_t}},0\right) \tag{1}$$

式中：t表示图片上第t个目标；(x_t,y_t,w_t,h_t)表示第t个目标中心点横纵坐标、宽度和高度；σ表示用于控制高斯掩模范围的高斯核。对中心点预测，将其转化为二分类问题，为解决正负样本失衡问题，参考 focal loss，其损失函数为：

$$L_c=-\frac{1}{N}\begin{cases}\sum_{i=1}^{\frac{W}{r}}\sum_{j=1}^{\frac{H}{r}}(1-p_{ij})^{\alpha}\log(p_{ij}), M_{ij}=1\\ \sum_{i=1}^{\frac{W}{r}}\sum_{j=1}^{\frac{H}{r}}(1-M_{ij})^{\beta}p_{ij}{}^{\alpha}\log(1-p_{ij}), \text{other}\end{cases} \tag{2}$$

式中：p_{ij}表示目标中心点预测的分数；M_{ij}表示真实的标签；α和β是超参数，设为α=2，β=4。

尺度预测分支使用卷积层生成尺度图，在固定纵横比 0.41 上预测高度比例，损失函数采用 $SmoothL_1$ 表示为：

$$L_s = \frac{1}{N}\sum_{t=1}^{N} SmoothL_1(\log h_t, p_t) \tag{3}$$

式中：$\log h_t$表示真实值；p_t表示预测值。

偏移预测分支使用卷积层生成中心点的偏移，在水平和垂直方向上调整中心点位置。偏移损失函数公式与公式 (3) 类似。

2 基于多层次级联网络的行人特征增强与重建

为了使提取的特征更加分离和清晰，提出由语义特征增强模块和自适应特征重建模块构成的多层次级联网络如图 1 中的 MLCN。它可以不断细化模型，提取高度抽象的特征，通过特征共享信息逐步改进每个阶段的特征达到更精确的结果，进一步为下一阶段的训练提供更具区别性的信息。同时，自上而下的级联能使每一级的特征不仅包含高层的语义信息，也包含底层的轮廓信息，这使得检测精度更高。

2.1 语义特征增强模块

网络在提取特征时，往往需要有对视觉情景的全局理解，需要根据上下文信息去判断目标与背景之间的关联性来推断定位目标。在存在严重遮挡的行人检测中，图像背景的全局视觉可以提供有用的语义上下文信息。对于行人检测，需要在图像中检测出行人，通常与行人目标共存的对象，例如：自行车、汽车等，可能会为检测提供有用的线索[17]，大量像素的关联才能行成目标。由于卷积神经网络 CNNs 的局部性，网络仅能通过堆叠多层卷积层进行远程依赖上下文建模，网络越深，优化难度越大。同时多个卷积层的堆叠只能达到局部检测的效果，不仅导致感受野受限，也会因计算量大带来难以优化的问题。Cao[18] 提出 GC-Block (Global Context Block) 去捕获长期的依赖关系，受上述方法的启发，为高效获取语义上下文信息，增强特征之间的相关性，引入了远程依赖建模思想，详细模型结构如图 2 (a) 所示。由于行人检测数据集过大，图片分辨率较高，为了增大感受野且不丢失特征分辨率，同时也为获得多尺度特征信息，对 GC-Block 进行改进，在空间建模阶段引入空洞卷积的思想如图 2 (b) 所示。对于输入特征F增加了空洞率为 2 和 3 的空洞卷积呈现出 5×5 和 7×7 大小的感受野，将其进行融合聚合更多的语义上下文信息，然后利用 1×1 的卷积核进行通道降维，通过 Softmax 进行非线性特征，得到空间注意力特征图F_s：

$$F_s = \phi\left(f_{conv}^{1\times1}\left(Concat\left(\left[f_{r=2}^{3\times3}(F), f_{r=3}^{3\times3}(F)\right]\right)\right)\right) \tag{4}$$

式中：F为输入特征；$\phi(\cdot)$为 Softmax 函数；F_s为空间注意力特征图。

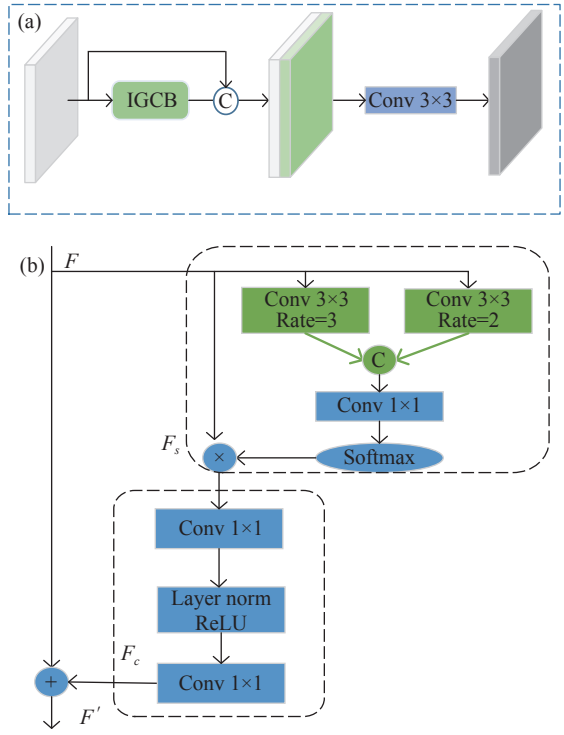


图 2 (a) 语义特征增强结构图；(b) 改进的全局上下文块

Fig.2 (a) Semantic feature enhancement structure diagram; (b) Improved global context bock

为了捕获通道之间的依赖性，进行了通道注意力转换。为了实现轻量级特征，采用 1×1 卷积将通道数降为原通道的$1/r(r=4)$；然后添加层归一化和 ReLU 操作可以显著增强目标检测，降低优化难度；再用 1×1 的卷积恢复到原来的通道数，这种编码-解码的特征转换方式能够学习通道之间的关联性，突出目标区域通道信息，更好的学习遮挡行人特征。利用空间与通道的建模来提升网络的识别能力，可以在不显著增加计算量的情况下提升网络的表达能力。

为了提取特征的多样性，将 IGCB 模块输出的具有较大感受野的特征图F'与初始特征图F以通道叠加

的操作方式融合，直接优化从输入中选择的有用信息，这带来了显著的性能；然后采用 3×3 的卷积处理特征融合带来的混叠效应，最终得到语义增强后的特征图。

2.2 自适应的特征重建模块

网络经过语义特征增强模块学习了全局上下文信息，能够根据上下文信息推测出行人部分，但当出现严重遮挡时，提取的特征仍然包含着较多的背景信息。为了让网络关注更多的行人信息，设计了一个自适应的特征重建模块，采用像素级别的语义分割图，通过输入特征模式的复杂性计算关注点位置，确定哪一个通道的关注点更多，自适应的给其增大权重，加强行人可见部分并调整整体特征以抑制不可见区域，让被遮挡的行人实例可以接收到对可见身体部位的强响应，使网络更加关注行人区域，从而增加检测的可判别性。

自适应的特征重建模块如图 3 所示，它接收两种输入，来自上一级特征增强后的特征F_p和骨干网络输出该级的特征图F，输出为重建后的多通道特征F_r。F_p中包含丰富的语义上下文信息，对于遮挡目标有更高的响应，由于骨干网络在卷积的过程中会丢失大量的细节，受 FPN 启发，将深层特征与浅层特征进行交互生成语义分割图F_m：

$$F_m = \sigma(\delta(Concat[F_p, F])) \tag{5}$$

式中：F_m表示行人位置区域的概率图；$\delta(\cdot)$表示 3×3 卷积和 ReLU 激活；$\sigma(\cdot)$表示 Sigmoid 函数。

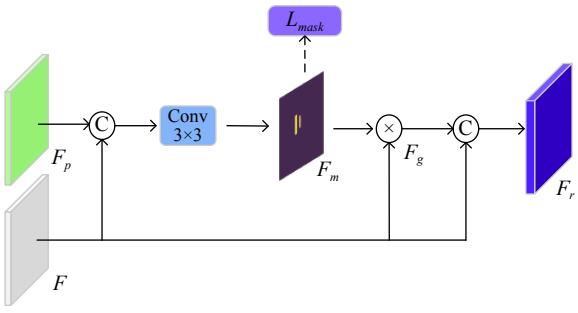


图 3 自适应特征重建结构图

Fig.3 Structure diagram of adaptive feature reconstruction

对于任意一个位置$\alpha \in F_m$，如果α上包含目标，则$m_\alpha \to 1$；如果α为背景，则$m_\alpha \to 0$。F_m用于调制输入特征的每个位置的特征表示，根据输入特征模式的复杂性来选择最佳匹配的参数大小。对于骨干特征图F上任意一点β位置，对应像素值为t_β，F将会跟F_m上对应的β位置的像素值m_β相乘。当$m_\beta \to 1$时，β位置的像素值会根据m_β值重新进行自适应的权重分配。相反，$m_\beta \to 0$时，β位置的像素值也会自适应的进行权重的重新分配。重新分配后的权重表示为：

$$m_\beta = \begin{cases} \to 1, \beta\text{为目标} \\ \to 0, \beta\text{为背景} \end{cases} \tag{6}$$

$$F_{g\beta} = t_\beta \cdot m_\beta \tag{7}$$

式中：m_β为F_m上β位置的像素值；t_β为F上β位置的像素值；$F_{g\beta}$为特征重建后的特征图F_g上β点像素值。

因为$F \in [H \times W \times C]$为多通道特征图，其上的$C$个特征通道的位置$(i, j)$上的特征值都将由$F_m(i, j, 1)$进行调制；为了给后续 SFEM 提供更详细的信息，还将原始特征F与激活的特征图F_g相结合，得到重建后的特征图$F_r \in [H \times W \times C]$：

$$F_r = \delta(Concat([F, F_g])) \tag{8}$$

式中：$\delta(\cdot)$表示 3×3 卷积和 ReLU 激活函数。

目前在包含行人检测在内的许多计算机视觉任务中，精细像素级别的分割注释是匮乏的。在 Caltech 数据集中，68% 的行人身高不到 80 pixel，随着网络不断的卷积，这相当于 conv 5 处的 3×5 pixel，特别是对于小目标和严重遮挡的目标来说，使用粗略的边界框标注和精细逐像素的标注几乎一致。图 4 对比了精细分割图和粗略分割图，可以看出这种粗略的分割在较大分辨率下与精细分割效果不相上下。由于笔者的分割任务只是用来辅助监督，不需要分割出行人的精确形状，同时也为了节省成本，选择基于目标边界框标注的像素分割方式。这个标记过程创建了一个粗略的分割，如果像素位于可视边界框标注内，将该像素点设置为前景，像素值设为 1，否则，背景设为 0。语义分割图采用交叉熵损失函数：

$$L_m = BCELoss(G_t, S_t) \tag{9}$$

式中：G_t代表粗略标注的真实值；S_t代表语义分割 mask 预测值。

总的损失函数表示为：

(10)

式中：λ_s为权重因子，分别设置为 0.01、1 和 0.1；λ_c、λ_s和 λ_o 分别为中心点、尺度和偏移的损失函数；L_m为语义分割掩码损失函数。

(a)

(b)

(c)

图 4 原图与不同分割图对比。(a) 原图; (b) 精细分割图; (c) 粗略分割图

Fig.4 Comparison of original image and different segmentation images. (a) Original image; (b) Fine segmentation map; (c) Rough segmentation map

3 实验与结果分析

3.1 实验环境及评价指标

实验选择的 CityPersons 和 Caltech 行人检测数据集都来源于车载摄像机, 与现实生活中图像的实际遮挡频率一致, 含有大量的遮挡行人图片, 适用于检验文中方法所要解决问题的有效性。CityPersons 数据集包括训练集、验证集和测试集, 训练集包括 2 975 张城市街道的行人图片, 分辨率大小为 1 024×2 048。验证集有 500 张图片, 由于测试集没有真实标注, 采用验证集作为测试数据。Caltech 行人数据集是目前规模较大的行人数据集, 视频约 10 h 左右, 分辨率为 640×480, 30 frame/s, 标注了约 250 000 frame (约 137 min), 350 000 个矩阵框, 2 300 个行人, 另外还对矩阵框之间的时间对应关系及遮挡的情况进行标注。数据集为 set00~set10, 其中 set00~set05 为训练集, set06~set10 为测试集。

实验是在 Pytorch1.2.0 框架下进行的。网络训练是两块 Nvidia GeForce RTX 3090 GPU, 测试使用一块 GPU。实验的骨干网络为预训练的 ResNet50, 利用 Adam 算法作为随机梯度优化器优化训练模型, 共训练 150 epoch。对于 Caltech 数据集, 训练初始学习率设置为 0.000 1, 训练批次大小为 32, 训练和测试图片尺寸输入为 336×448。对于 CityPersons 数据集, 训练初始学习率设置为 0.000 2, 训练批次大小为 8, 训练和测试图片尺寸输入分别为 1 280×640 和 2 048×1 024。测试结果采用阈值为 0.5 的 DIoU-NMS 算法滤除冗余预测结果。

评估遵循加州理工学院评估标准, 即每张图像假阳性 (FPPI) 的对数平均漏检率 (log-average Miss Rate, MR^{-2}), 范围为 $[10^{-2}, 10^{2}]$。笔者评估了具有不同行人高度范围 (High, PXs) 和不同能见度水平 (Vis) 的五个子集, 如表 1 所示。

表 1 数据集子集划分标准

Tab.1 Standards for dividing data set subsets

Type	Bare	Reasonable	Heavy	Small	All
High	h>50	h>50	h>50	50<h<75	h>20
Vis	v>90%	v>65%	0<v<65%	v>65%	v>20%

3.2 消融实验

对 CityPersons 数据集进行消融研究来验证提出方法的有效性, 使用对数平均漏检率 MR^{-2} 作为评价指标。

特征融合方式有通道拼接 (Cat)、逐像素相加 (Sum) 和逐像素相乘 (Multiply) 三种, 为了验证提出的 AFRM 采用什么融合方式进行特征重建有更好的效果, 表 2 对比了各种融合方式得到的漏检率。实验利用 CSP[14] 作为验证基准, 采用 Cat 融合方式增加了特征通道, 但每个特征下的信息并没有加强, 导致网络不能提取到更多有用的信息, 对漏检结果影响不大。采用 Sum 的融合方式虽然没有增加特征通道, 但很多背景信息都被加入到重建的特征上, 导致网络不能很好地区分前景与背景, 在 Reasonable 子集上表现出更差的

结果。而采用 Multiply 融合方式取得了最好的结果，这种逐像素相乘操作，利用空间注意图去激活原始特征图的方式，不仅没有增加特征通道数，反而加强了通道上关注点的权重，利用可见部分的关注点调整整体特征以抑制不可见区域，能让网络对行人和背景更加有分辨性，更有利于检测器去定位目标。

表 2 AFRM 的融合方式对比

Tab.2 Comparison of AFRM fusion methods

Baseline	Cat	Sum	Multiply	Reasonable	Heavy
√				11.00%	49.30%
√	√			10.60%	48.74%
√		√		11.20%	49.10%
√			√	10.26%	48.21%

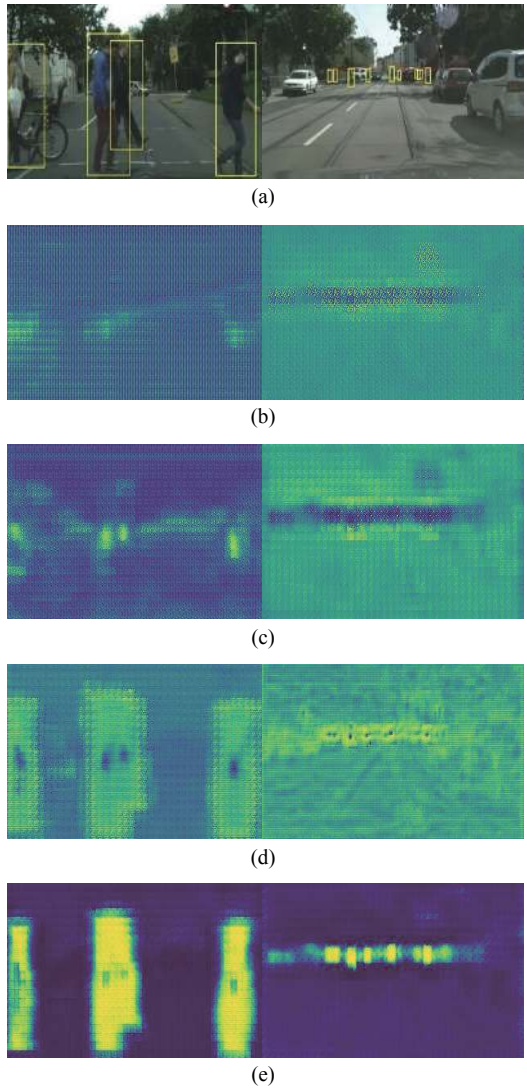
为了验证文中提出 SFEM 和 AFRM 模块的有效性，表 3 比较了各模块添加前后的对比结果。可以看出，基线上单独添加 SFEM 或 AFRM，在 Reasonable 和 Heavy 子集上漏检率都有所下降，表明提出的 SFEM 或 AFRM 模块是有效的。SFEM 能够获取全局上下文信息，学到行人与背景之间的关系加强行人特征，使网络提取到更多有用的信息，但并不是所有的上下文信息都有利于检测，可能会带来更多的冗余信息，致使网络提取到无用的特征，因此，添加互补的 AFRM 能够加大背景与前景的权重差距，让网络在提取特征时更关注行人区域。级联 SFEM 和 AFRM 两个模块为网络特征提取提供了深层和浅层特征模式的交互，迫使负样本远离正样本，让检测器在密集场景下减少误检，在 Heavy 遮挡子集上漏检率降低了 1.98%。

表 3 各模块消融对比实验

Tab.3 Comparison experiment of ablation of each module

Baseline	SFEM	AFRM	Reasonable	Heavy
√			11.00%	49.30%
√	√		10.82%	48.77%
√		√	10.26%	48.21%
√	√	√	9.85%	47.32%

为更直观地观察所提算法的处理效果，图 5 列出添加各模块前后 Conv5_3 的特征可视化对比图。可以看出图 5 (b) 在基准下，由于行人与背景相似且行人被严重遮挡等原因，提取的行人特征不明显。图 5 (c) 只添加 SFEM，网络通过上下文信息，建立了特征与特征之间的关联性，增强了特征的表征能力，相比基准网络提取到更多的特征。但当行人与背景相似时，更多的行人信息被误认为背景，造成特征的缺失。图 5 (d) 只添加 AFRM，可以基于可见部分推断被遮挡的部分，使行人特征更加突出，从而将背景与前景分开。图 5 (e) 为文中提出的级联 SFEM 和 AFRM 两个模块后的整体处理效果，对遮挡目标和小目标可提



(a) (b) (c) (d) (e)

图 5 添加各模块前后 Conv5_3 层的特征可视化对比图。(a) 原图; (b) 基准; (c) 添加语义增强; (d) 添加特征重建; (e) 文中方法

Fig.5 Comparison of feature visualization of Conv5_3 layer before and after adding each module. (a) Original image; (b) Baseline; (c) SFEM; (d) AFRM; (e) Proposed method

取出更有区分性的增强特征，并很好地区分了行人与背景。

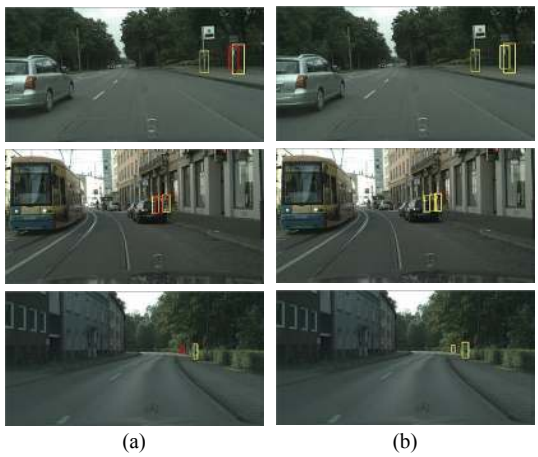
后处理是行人检测中重要的一部分，能够对预测的结果去除重复的检测框，特别是针对行人之间的遮挡，检测框之间的重叠面积比较大，标准的非最大值抑制算法会过滤掉很多有用的信息。采用交并比阈值较低会导致高度重叠的行人漏检，阈值较高又会显著地增加假阳性样本，带来大量的误检。因此，文中引入 DIoU(Distance-IoU)[19] 的 NMS 进行后处理。为了证明采用 DIoU-NMS 的有效性，在 heavy 子集下给出不同的 IoU 阈值进行实验结果对比。从表 4 可以看出，在阈值为 0.5 时，检测结果最好。对相互遮挡的行人，通过 DIoU-NMS 可以捕获两个行人之间重叠程度的相对位置，保留其他行人框，减少漏检；对单个的行人，会删除多余的检测框，减少误检。

表 4 不同 NMS 类型在不同阈值下的对比结果

Tab.4 Comparison results of different NMS types at different thresholds

Type	IoU=0.5	IoU=0.6	IoU=0.7
Baseline (NMS)	49.30	49.91	53.47
Proposed method (NMS)	47.32	47.85	51.00
Proposed method (DIOU-NMS)	47.28	47.85	50.75

图 6 对比了文中方法与 CSP 的部分检测结果。红色框表示 CSP 与文中算法相比漏检的行人目标，可以看出，CSP 在严重遮挡时，行人与背景相似上检测效果较差，该模型对那些特征不明显的目标并不敏感，从而造成漏检问题。而文中提出的方法，在遮挡和背景物相似场景下，都有很好的检测能力。



(a) (b)

图 6 部分检测结果对比。(a) CSP 检测结果; (b) 文中方法检测结果

Fig.6 Comparison of some test results. (a) CSP test results; (b) Proposed method test results

3.3 实验结果分析

文中方法在 CityPersons 和 Caltech 数据集上与当前同类研究的最新模型进行比较，采用评价指标为 MR^{-2}，数据集子集划分标准按照表 1。

表 5 对比了在 CityPersons 数据集上进行验证的各个子集的 MR^{-2} 值，选取不同遮挡处理的方法包括 CSP[14] 、R2NMS[20] 、MSAF[21]、PRNet[22]、PEN[23]、CSANet[24]、APD[25]、Couple[26]、IDC[27] 等，并在最后一列给出各方法的处理时间。从表 5 中可以看出，RepLoss、OR-CNN 和 PRNet 是针对行人模式进行损失函数优化的，因此在少部分遮挡的行人 Bare 子集上表现出较好的效果；MSAF 是一个多级融合的多尺度的检测算法，能够减少负样本的产生，所以针对遮挡范围较小的 Reasonable 子集行人表现出较好的结果。与这些方法相比，文中提出的方法主要解决严重遮挡的行人目标，由于网络能根据上下文信息去判断行人与背景之间的关系，通过自适应的参数匹配模式能力，对行人给予更大的关注点，让网络更加关注行人区域；同时，将深层与浅层增强后的特征进行融合能够提高网络的特征表示能力，提取更为抽象的信息，大大提升了对遮挡目标的检测效果，因此在 heavy 子集上实现了 47.28% 的 MR^{-2}，表现出最好的性能，在 All 子集

表 5 CityPersons 数据集对比

Tab.5 Comparison of CityPersons data sets

Method	Reasonable	Bare	Heavy	Small	All	Time
RepLoss	13.20	7.60	56.90	42.60	44.45	-
OR-CNN	12.80	**6.70**	55.70	42.30	42.32	-
PRNet	10.80	**6.80**	53.30	-	-	-
PEN	10.40	7.00	47.40	-	-	0.36
MSAF	**9.50**	7.10	48.40	15.50	-	-
IDC	10.70	-	50.60	14.70	41.40	-
R2NMS	11.10	-	53.30	-	-	-
CSANet	12.00	7.30	51.30	-	-	0.32
APD	10.60	7.10	49.80	15.70	-	0.16
Couple	12.30	-	49.81	38.31	40.39	-
CSP	11.00	7.03	49.30	16.00	-	0.33
Proposed	**9.85**	**6.82**	**47.28**	**13.93**	**36.65**	0.36

上也超过了最优的方法实现 36.65% 的 MR^{-2}；针对小目标，在 Small 子集上也表现出很好的性能。相比其他算法，文中算法的时间效率虽未达最优，但实现了更低的漏检率。

表 6 对比了在 Caltech 数据集上的检测结果，将文中方法与三个遮挡子集下的最先进方法进行比较：PAMS_FCN[8]、RepLoss[9]、Bi-box[11]、ATT-part[12]、CSP[14]、SSNet[28]、AR-Ped[29] 等，可以看出文中方法在不同的子集上呈现出优越的性能。在 Heavy 子集上实现了 44.04% 的 MR^{-2} 最优性能，由于文中针对严重遮挡的检测，在遮挡范围较大的子集上性能优越。同时在 Reasonable 和 All 子集分别实现了 4.32% 和 56.34% 的 MR^{-2}，仅次于第一好的性能。这些结果表明文中提出的基于 Anchor-free 的检测算法没有复杂的先验框设置，直接通过目标中心点、尺度和偏移来进行预测，在一定程度上能够解决检测框重合带来的影响，让检测结果的精度更高。

表 6 Caltech 数据集对比

Tab.6 Comparison of caltech data set

Method	Backbone	Reasonable	Heavy	All
ATT-part	VGG-16	10.30	45.18	-
RepLoss	ResNet-50	5.00	47.90	59.00
OR-CNN	ResNet-50	4.10	45.00	-
SSNet	ResNet-50	6.30	-	-
PAMS_FCN	ResNet-50	4.50	47.40	**53.70**
Bi-Box	VGG-16	7.61	44.40	-
AR-Ped	ResNet-50	4.36	48.80	-
CSP	ResNet-50	4.54	45.80	56.94
Proposed	ResNet-50	**4.32**	**44.04**	**56.34**

另外，为更好地观察文中算法对拥挤场景下严重遮挡目标的检测误差，图 7 给出了一个存在模型失效的可视化检测结果示例，虚线检测框是文中算法漏检的行人目标。在行人密集场景中，当遮挡比例大于 90% 以上时，行人目标呈现出的视觉线索信息过少，造成检测器可利用的信息太少，以至于无法准确检测出存在大面积重叠的所有行人目标。

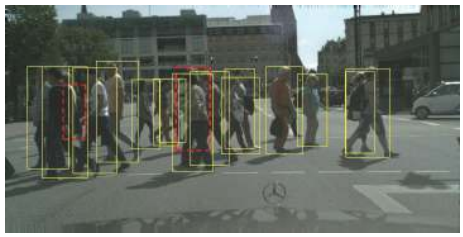


图 7 模型失效场景

Fig.7 Model failure scene

4 结 论

文中提出了一种具有鲁棒性的语义增强引导特征重建的遮挡行人检测算法，通过语义特征增强模块建立远程依赖获取更多的语义信息来增强特征；在特征重建模块中生成语义分割图，自适应地调制全身特征，增加关注点通道权重，以突出可见区域，抑制背景区域；通过级联分割和重建，在抑制干扰特征的同时保留有用或重要的特征，为行人检测提供更有区别的信息，以解决严重遮挡的行人检测问题；在后处理阶段采用 DIoU-NMS 算法为行人之间的遮挡保留了更多有用的信息。这种联合机制得到的特征信息可综合考虑更复杂的因素，在 CityPersons 和 Caltech 数据集上对严重遮挡行人具有较好的检测效果。由于行人检测对实时性的要求较高，未来将着手优化网络模型，以满足在实际应用中对实时性的要求。

参考文献：

[1] Xu Xinkai, Ma Yan, Qian Xu, et al. Scale-aware EfficientDet: Real-time pedestrian detection algorithm for automated driving [J]. ***Journal of Image and Graphics***, 2021, 26(1): 93-100. (in Chinese)

[2] Sun J B, Ji J. Memory-augmented deep autoencoder model for pedestrian abnormal behavior detection in video surveillance [J]. ***Infrared and Laser Engineering***, 2022, 51(6): 20210680. (in Chinese)

[3] Zhang Ruiyan, Jiang Xiujie, An Junshe, et al. Design of global-contextual detection model for optical remote sensing targets [J]. ***Chinese Optics***, 2020, 13(6): 1302-1313. (in Chinese)

[4] Wang Jianlin, Fu Xuesong, Huang Zhanchao, et al. Multitype cooperative targets detection using improved YOLOv2 convolutional neural network [J]. ***Optics and Precision Engineering***, 2020, 28(1): 251-260. (in Chinese)

[5] Wang Chunzhe, An Junshe, Jiang Xiujie, et al. Region proposal optimization algorithm based on convolutional neural networks [J]. ***Chinese Optics***, 2019, 12(6): 1348-1361. (in Chinese)

[6] Xue Shan, Zhang Zhen, Lv Qiongying, et al. Image recognition method of anti UAV system based on convolutional neural network [J]. ***Infrared and Laser Engineering***, 2020, 49(7): 20200154. (in Chinese)

[7] Noh J, Lee S, Kim B, et al. Improving occlusion and hard negative handling for single-stage pedestrian detectors[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 966-974.

[8] Yang P, Zhang G, Wang L, et al. A part-aware multi-scale fully convolutional network for pedestrian detection [J]. ***IEEE Transactions on Intelligent Transportation Systems***, 2020, 22(2): 1125-1137.

[9] Wang X, Xiao T, Jiang Y, et al. Repulsion loss: Detecting pedestrians in a crowd[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7774-7783.

[10] Zhang S, Wen L, Bian X, et al. Occlusion-aware R-CNN: Detecting pedestrians in a crowd[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 637-653.

[11] Zhou C, Yuan J. Bi-box regression for pedestrian detection and occlusion estimation[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 135-151.

[12] Zhang S, Yang J, Schiele B. Occluded pedestrian detection through guided attention in cnns[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6995-7003.

[13] Luo Haibo, He Miao, Hui Bin, et al. Pedestrian detection algorithm based on dual-model fused fully convolutional networks (Invited) [J]. ***Infrared and Laser Engineering***, 2018, 47(2): 0203001. (in Chinese)

[14] Liu W, Liao S, Ren W, et al. High-level semantic feature detection: A new perspective for pedestrian detection[C]// Proceedings of the Conference on Computer Vision and Pattern Recognition, 2019: 5187-5196.

[15] 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***, 2011, 34(4): 743-761.

[16] Zhang S, Benenson R, Schiele B. Citypersons: A diverse dataset for pedestrian detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3213-3221.

[17] Li Jingyu, Yang Jing, Kong Bin, et al. Multi-scale vehicle and pedestrian detection algorithm based on attention mechanism [J]. ***Optics and Precision Engineering***, 2021, 29(6): 1448-1458. (in Chinese)

[18] Cao Y, Xu J, Lin S, et al. Gcnet: Non-local networks meet squeeze-excitation networks and beyond[C]//Proceedings of the International Conference on Computer Vision Workshops, 2019: 1971-1980.

[19] Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12993-13000.

[20] Huang X, Ge Z, Jie Z, et al. NMS by representative region: Towards crowded pedestrian detection by proposal pairing[C]// Proceedings of the Conference on Computer Vision and Pattern Recognition, 2020: 10750-10759.

[21] Cao Z, Yang H, Xu W, et al. Multiscale anchor-free region proposal network for pedestrian detection [J]. ***Wireless Communications and Mobile Computing***, 2021, 2021: 5590895.

[22] Song X, Zhao K, Chu W S, et al. Progressive refinement network for occluded pedestrian detection[C]//Computer Vision–ECCV 2020: 16 th European Conference, 2020: 32-48.

[23] Jiao Y, Yao H, Xu C. PEN: Pose-embedding network for pedestrian detection [J]. ***IEEE Transactions on Circuits and Systems for Video Technology***, 2020, 31(3): 1150-1162.

[24] Zhang Y, Yi P, Zhou D, et al. CSANet: Channel and spatial mixed attention CNN for pedestrian detection [J]. ***IEEE Access***, 2020, 8: 76243-76252.

[25] Zhang J, Lin L, Zhu J, et al. Attribute-aware pedestrian detection in a crowd [J]. ***IEEE Transactions on Multimedia***, 2020, 23: 3085-3097.

[26] Liu T, Luo W, Ma L, et al. Coupled network for robust pedestrian detection with gated multi-layer feature extraction and deformable occlusion handling [J]. ***IEEE Transactions on Image Processing***, 2020, 30: 754-766.

[27] Ding M, Zhang S, Yang J. Improving pedestrian detection from a long-tailed domain perspective[C]//Proceedings of the 29th ACM International Conference on Multimedia, 2021: 2918-2926.

[28] Yang X, Liu Q. Scale-sensitive feature reassembly network for pedestrian detection [J]. ***Sensors***, 2021, 21(12): 4189-4208.

[29] Brazil G, Liu X. Pedestrian detection with autoregressive network phases[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 7231-7240.