

文章编号 1004-924X(2023)10-1532-16

多分支无锚框网络密集行人检测算法

吕志轩, 魏霞*, 黄德启

(新疆大学 电气工程学院, 新疆维吾尔自治区 乌鲁木齐 830017)

摘要:针对街道等多人流量场景图像中人员密集、姿态变化多、人体遮挡严重造成的行人检测漏检问题,提出一种多分支无锚框网络(MBAN)行人检测方法。首先,在检测模型主干网络后加入多分支网络结构用以检测行人的多个关键区域局部特征;然后,设计了关键区域之间的距离损失函数引导分支网络对行人的局部检测位置进行差异化学习,接下来为了提高分支网络对行人局部特征空间信息的理解能力,在 Resnet50 网络尾部加入四个上采样块构成沙漏结构(Hourglass);最后,设计了一种局部特征选择网络自适应抑制多分支输出的非最优值,消除预测时的冗余特征框。实验结果表明 MBAN 方法对多人流量场景行人检测的 mAP 值、F1 值、Prec 和 Recall 分别达到 85.22%, 0.87, 80.07% 和 94.39%, 证明该方法对密集人群检测能力较强,与其他行人检测算法相比有较高的召回率。

关键词:无锚框网络;多分支网络;行人检测;局部特征;特征提取;特征选择

中图分类号:TP391.41; **文献标识码:**A **doi:**10.37188/OPE.20233110.1532

Dense pedestrian detection algorithm in multi-branch non-anchor frame network

LÜ Zhixuan, WEI Xia*, HUANG Deqi

(School of Electrical Engineering, XinJiang University, Urumqi 830049, China)

* Corresponding author, E-mail: 30462111@qq.com

Abstract: Considering the problem of missed pedestrian detection in dense pedestrian images, a multi-branch non-anchor frame network (MBAN) detection method is proposed to detect various posture changes and serious human occlusion in multi-person traffic scenes, such as streets. First, a multi-branch network structure is added after model backbone network detection to detect the local features of multiple key areas with pedestrians. Subsequently, the distance loss function between key areas is designed to guide the branch network to differentially learn the local detection position of pedestrians. Thereafter, four up-sampling blocks are added to the tail of the ResNet50 network to form an hourglass structure, thereby improving the branch network's ability to understand the spatial information of local features of pedestrians. Finally, a local feature selection network is designed to adaptively suppress the non-optimal values of the multi-branch output and eliminate the redundant feature box in prediction. In the experimental results, the mAP, F1, Prec, and Recall values of the MBAN method for pedestrian detection in multi-person scenes reached 85.22%, 0.87, 80.07%, and 94.39%, respectively. Therefore, this method is effective in detecting pedestrians in dense crowds and has higher recall rate compared with other pedestrian detection al-

收稿日期:2022-05-30;修订日期:2022-07-08.

基金项目:国家自然科学基金资助项目(No. 51468062)

gorithms.

Key words: non-anchor frame network; multi-branch network; pedestrian detection; local features; feature extraction; feature selection

1 引 言

随着我国社会稳步发展,社会人口数量也逐年增加,2021年我国人口数量14.11^[1]亿人,城镇人口数量达到9.02亿人,人口密度急剧增加,大量的人口对城市交通安全提出了挑战。近些年为了解决交通安全问题,无人驾驶汽车^[2]概念被提出,其中一个关键技术是对道路上的行人进行识别并做出预判来减小交通事故发生几率,因此让无人驾驶汽车能够精准并全面检测出街道上的行人非常重要。目前的行人检测算法对直立姿态下的稀疏人群有较好的检测效果,而街道上的行人数量多且姿态变化,相机拍摄下会出现遮挡情况,造成了很高的漏检率,离实用仍有距离。

使用深度学习神经网络对人体显著特征学习^[3-4]是目前解决行人检测问题的主要方法,该方法依靠深层次的卷积神经网络对大量图片计算来获得有用的特征信息,从而根据特征对人体进行高精度识别,基于深度学习的检测算法可分为有锚框模型算法与无锚框模型算法。

有锚框模型算法^[5]以锚框作为模型输出匹配基础,模型输出的每个特征点与多个尺度不同的锚框匹配,增加检测不同尺度物体的概率,该类算法可继续分成两阶段与单阶段两类,两阶段锚框模型算法 R-CNN(Region-based Convolution Neural Networks)由 GIRSHICK^[6]等在2014年提出,算法第一次将深度学习应用到目标检测领域;单阶段锚框算法 YOLO(You Only Look Once)由 REDMON^[7]等在2016年提出,该结构算法的检测速度基本达到实时性需求。有锚框模型算法出现早,检测能力较强,目前基于 R-CNN 与 YOLO 的改进模型算法仍被大多数行人检测模型使用,Roszyk^[8]等在锚框网络 YOLOv4 中加入多光谱融合机制,将行人 RGB 图像与热成像图输入网络后的结果融合,使行人检测精度达到70%,检测速度达到35 FPS;Gawande^[9]等在 R-CNN 网络中加入轮廓变换模块和定向滤波器组来优化行人在图像的旋转问题,然后使用仿射尺

度不变特征变化(Affine-Scale Invariant Feature Transform, ASIFT)来寻找图片中等比例行人特征点;李经宇^[10]等在 YOLOv3 网络中加入空间金字塔池化层(Spatial Pyramid Pooling)和特征金字塔网络(Feature Pyramid Networks)对多个尺度的行人特征进行融合,然后使用注意力模块 SE 模块来寻找车辆行人的重要特征,提高了行人车辆的检测精度;马立^[11]等将 YOLOv3 特征提取网络中的池化层换为卷积层,将卷积块换为深度可分离卷积块,大幅减小模型参数,并优化了交并比函数,对行人等小目标的检测准确率提高了9.8%。

由于有锚框模型算法中锚框超参数的数值设置依赖于先验知识难以泛化,并且锚框带来的巨大计算量使模型训练与预测时间过长,因此无锚框类型模型算法^[12-13]被提出与关注。

无锚框模型算法不再预设锚框,直接预测物体的边界框,以中心点或关键点作为模型输出的匹配基础,将模型输出与被检测物体的真实中心点或关键点匹配来学习显著特征,在保持检测精度与锚框模型算法无明显差异的情况下,使网络结构头部网络特征通道数量更少^[14]。2018年 HEI^[15]等提出的 CornerNet 与 2019 年 ZHOU^[16], DUAN^[17]等提出的 CenterNet 算法在 COCO 数据集上获得较好的检测效果,再一次证明了无锚框算法的优势,无锚框行人检测模型算法也在此基础上逐步发展。近几年 CAO^[18]等为了检测图像中的小目标行人,设计了多尺度无锚框网络(Multiscale Anchor-Free Region Proposal Network, MSAFNet),使用自适应特征融合(Adaptively Spatial Feature Fusion, ACFF)来自主选择主干网络较低层的小目标行人特征,并在网络尾部加入多尺度无锚框检测头(Multiscale Anchor-Free, MSAF)对不同尺度的特征进行解码;为了获得更多的行人特征信息,DING^[19]等设计了无锚框级联跨层融合网络(Cascaded Cross-Layer Fusion Network, CCFNet),在主干网络中使用 CCF 模块将不同层的特征合并获得更多的语义

信息,在检测头加入全局平滑图来获得全局特征信息;邹逸群^[20]等设计了无锚框尺度自适应行人检测算法(Anchor-Free Scale Adaptive, SFSA),在主干网络后加入FPN特征融合模块加强网络特征提取能力,在检测头加入SA注意力模块自适应给不同尺度的行人特征赋予权重,增强了网络对行人尺度变化的鲁棒性。

上述基于无锚框网络的行人检测算法虽然性能良好,但只利用了人体不同尺度下细节最显著也最易于分辨的全局特征,忽略了部分关键局部细节以及次易于分辨的区域,这造成了以上算法对拥挤、遮挡、姿态变化等情况下的行人检测效果不佳。对于现存问题,本文提出一种多分支无锚框网络(Multi Branch Anchor-free Network, MBAN)行人检测方法。

针对网络只关注全局特征的问题,设计了一种具有多条输出分支的主干网络,在主干网络中加入由多个局部特征提取模块组成的局部特征提取网络(Feature Extraction Net, FEN),用于在训练中寻找人体的显著特征和多个次显著特征、增大模型分辨行人的判断区域,并且为提高网络提取空间信息的能力,将原主干网络Resnet50替换为Resnet50-Hourglass。

设计了一种对多分支主干网络的多个输出之间进行相似度判断的局部特征距离函数(Local Feature Distance, LFD),该函数引导局部特征提取模块的权重学习方向向不同方向分离,避免聚集在同一特征区域学习到相同的局部特征。

在模型预测阶段,为避免多分支网络的各分支检测到同一物体产生冗余检测框,设计一种特征选择网络(Feature Select Net, FSN),为每个经过非极大值抑制后的分支输出分配优先级,按优先级排序选择输出优先级最大的分支特征,对其余分支输出进行非最优值抑制(Non-optimal suppression, Nos)。

为了验证算法性能,在WiderPerson数据集和CrowdHuman数据集上开展实验,文章进行实验的代码已开源,开源地址为<https://gitee.com/lv-zhixuan/pedestrian-detection>

2 多分支无锚框网络

本文提出的多分支无锚框网络(MBAN)模

型架构如图1所示,模型包括(a)主干网络、(b)多分支特征提取网络、(c)特征识别网络、(d)局部特征差异网络和(e)特征选择网络五个模块,(a)、(b)、(c)、(d)构成训练模型,(a)、(b)、(e)构成了预测模型。图1中, F 表示主干网络输出特征, F_g 表示全局特征, $F_i(i \in 1, 2)$ 表示第 i 个局部特征(本文设置最大局部特征数量为2,图1中为2个局部特征的示例), $D_i(i \in 1, 2)$ 表示第 i 个局部特征对应的特征值索引,Concat表示通道合并操作, L^* 表示各类损失函数。下面将对每个模块进行详细介绍。

2.1 主干网络

深度学习中越深层的网络越容易出现网络退化问题,为了解决这个问题,HE^[21]等人提出了基于差网络结构的Resnet50,该网络包含了4个残差模块,作用是防止网络层数过多导致模型出现梯度消失现象,使得Resnet50可以学习到物体更加丰富的语义信息。本文使用Resnet50作为MBAN模型主干网络,考虑到较深的网络为了提取出物体特征丢弃了与特征信息无关的空间像素,这会降低网络的空间理解能力,因此在Resnet50卷积块尾部加上上采样块构成Hourglass^[22]结构,多个上采样块通过扩大特征图的分辨率来减小单个特征像素对应真实图片的感受野,从而获取更多空间信息。改进得到的Resnet50-Hourglass网络如图1(a)所示,网络由1个卷积块(Conv1)、4个残差模块(Res1-4)和3个上采样块(Deconv1-3)构成,残差模块又由多个卷积块组成,卷积块中包括卷积层、BN层和ReLU激活函数。经过多次下和上采样,输入图像中行人的语义空间信息特征被提取出来,对应特征 $F \in R^{H \times W \times C}$,其中 C 代表通道数, H, W 代表高、宽。

2.2 多分支特征提取网络

多分支网络^[23]的提出改变了模型网络只能提取一个空间特征的情况,使得可以自主设置网络特征提取区域数量,MBAN多分支网络结构如图1(b)所示。多分支网络结构由多个特征提取(FEN)模块构成,每个模块用来独立提取互不相交区域中的行人特征^[24],FEN模块的详细结构见图2。以图2 FEN左分支为例,Resnet50-Hourglass的输出 F 作为FPN模块的Input输入高斯热力图^[25](Heatmap)卷积块,接着,使用 3×3 卷积

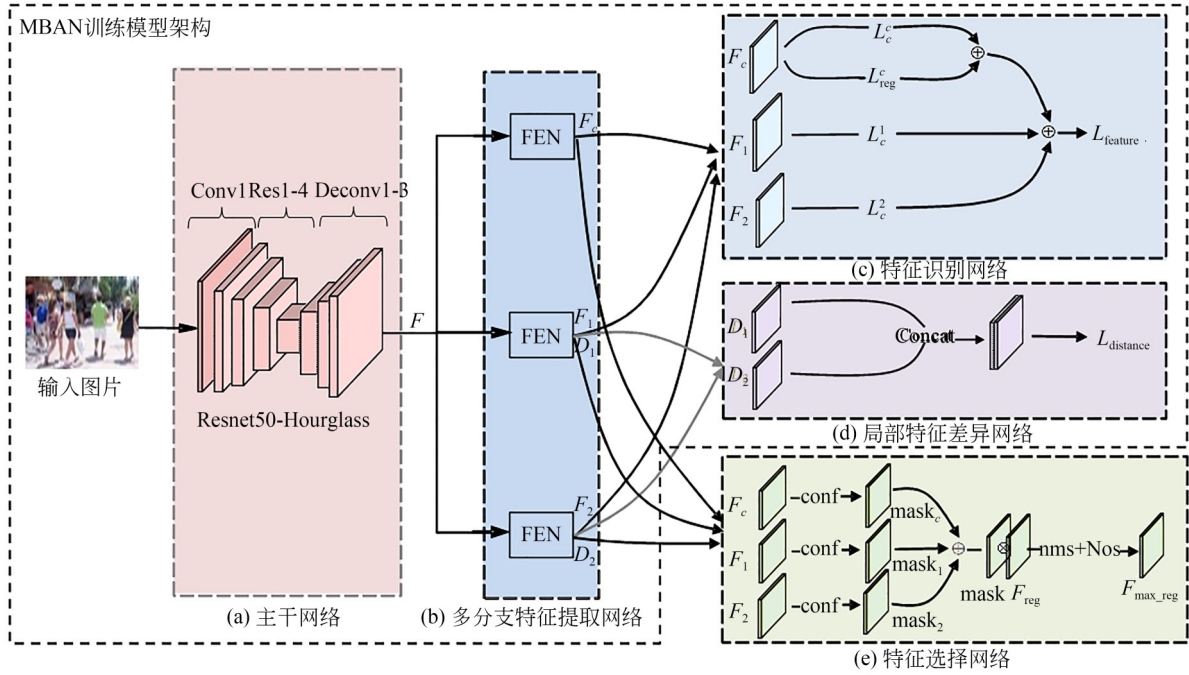


图1 MBAN模型架构

Fig. 1 MBAN model architecture

核的卷积层(Conv2d_1)卷积提取特征并用ReLU激活函数将特征非线性映射到同等大小空间中,最后,对经过BN层归一化后的特征输入进 1×1 卷积层调整特征通道数得到热力图特征 $F_{\text{Heatmap}} \in R^{H \times W \times 1}$ (检测类别数、通道数均为1)。同理可以得到高宽特征 $F_{HW} \in R^{H \times W \times 2}$ (检测框高度与检测框宽度各需要一个通道表示,通道数为2)与偏移量特征 $F_{\text{Reg. offsets}} \in R^{H \times W \times 2}$ (检测点对 x 轴的偏移量和 y 轴的偏移量各需要一个通道表示),将三个包含类别和位置信息的分支特征合并,得到FEN特征提取结果 $F_k \in R^{H \times W \times 5} \{k \in g, 1, 2\}$ 。

2.3 特征识别网络

基于一个全局特征与多个局部特征相结合的多分支思想,MBAN模型算法的网络结构中仅包含一条全局FEN分支与至少一条局部FEN分支,全局FEN分支用于检测行人全局最显著区域特征,局部FEN分支检测次要显著区域局部特征。因为检测位置的差异,全局FEN分支提取的特征包含类别信息与位置信息更全面,因此对全局FEN输出的热力图特征和位置特征都计算损失,而局部FEN分支提取的特征信息较少且局部特征的位置信息不能准确代表主体位置,因此仅计算局部FEN热力图特征的损失。

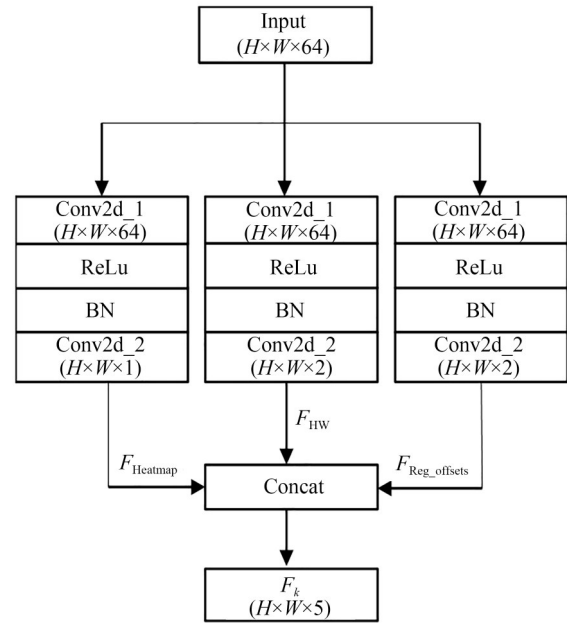


图2 特征提取网络

Fig. 2 Feature Extraction Net

识别检测网络结构如图1(c)所示,对于全局特征 $F_g \in R^{H \times W \times 5}$ ($F_{\text{Heatmap}} \in R^{H \times W \times 1}$, $F_{HW} \in R^{H \times W \times 2}$, $F_{\text{Reg. offsets}} \in R^{H \times W \times 2}$),按承载物体的信息类别不同将其分成上下分支分别计算损失值,上分支使用热力图损失(L_c^g)计算预测值与

真实值的损失。具体地,将热力图预测值 $F_{\text{Heatmap}} \in R^{H \times W \times 1}$ 和热力图真实值 $F_{\text{heat_true}} \in R^{H \times W \times 1}$ 输入 L_c^g 损失函数计算每一轮损失, L_c^g 损失计算公式如下:

$$L_c^g = - \sum \left(\frac{\log(p)(1-p)^2(y \cdot eq(1)) + \log(1-p)(p)^2(1-y)^4(y \cdot lt(1))}{\log(p)(1-p)^2(y \cdot eq(1)) + \log(1-p)(p)^2(1-y)^4(y \cdot lt(1))} \right), \quad (1)$$

其中: p 表示预测值, y 表示真实值, 公式(1)函数将等于1位置的数值设置为1, 其余位置数值设置为0, $lt(1)$ 函数将小于1位置的数值设置为1, $\sum$ 示对二维张量通道中的所有数值相加求和。在 L_c^g 损失函数中, 前半段用来计算二维张量中所有存在物体的特征点处的损失, 首先公式(1)将二维张量 F_{True} 中存在物体处的特征点值设置为1, 接着 $\log(p)$ 计算得到预测值对应位置处的损失, 损失值与预测值大小成反比, 最后 $(1-p)^2$ 使高预测值的简单样本特征点处的损失值更小, 从而模型在训练过程中更加关注损失值大的困难样本, 同理后半段计算不存在物体的特征点处的损失。在 L_c^g 损失函数中可以改变两个超参数, 分别是 $(1-p)^m$, $(p)^m$ 和 $(1-y)^n$ 右上角的指数值 m, n 。超参数 m 表示损失函数对简单样本的关注程度, m 值越大函数计算简单样本的得到的损失越小, 下次训练对该样本的关注程度越低; 考虑到过大的 m 同时会在一定程度上减小困难样本的损失值 ($0.9^2=0.81, 0.9^3=0.729$) 影响模型训练结果, 因此 m 值设置在2~3之间比较合适; 超参数 n 表示对二维张量 F_{True} 中物体中点周围半径范围内样本的关注程度, 较大的 n 使离中点越近的特征点损失值更小, 被关注程度低。因为对中心附近特征点给予过高关注会造成中点预测位置信息不准确导致预测框发生偏移, 因此实验将 n 设置为4, 降低周围特征点的关注度。

全局特征下分支使用L1函数对预测值高宽 $F_{HW} \in R^{H \times W \times 2}$ 、偏移量 $F_{\text{Reg_offsets}} \in R^{H \times W \times 2}$ 与对应真实值 $F_{HW_true} \in R^{H \times W \times 2}$ 、 $F_{\text{Reg_true}} \in R^{H \times W \times 2}$ 计算得到全局特征在位置上的损失, 公式为:

$$L_{\text{reg}}^g = \sum \left(\sum_{i=1}^2 (|y_i - p_i| * m) \right), \quad (2)$$

其中: y_i 表示真实值 ($y_1 = F_{HW_true}, y_2 = F_{\text{Reg_true}}$), p_i 表示预测值 ($p_1 = F_{HW}, p_2 = F_{\text{Reg_offsets}}$), 二维张量

$m \in R^{H \times W \times 1}$ 为图中所有行人位置索引, $\sum$ 表示对二维张量所有特征点相加求和。

局部特征损失函数 L_c^l 计算公式与全局特征上分支计算公式(1)相同, 至此, 识别检测网络计算多分支FEN模块的总损失函数 L_{feature} 如公式(3)所示:

$$L_{\text{feature}} = L_c^g + L_{\text{reg}}^g + L_c^l + L_c^2. \quad (3)$$

2.4 局部特征差异网络

训练过程中多分支FEN模块趋向于向物体最显著特征区域学习, 若不对每条分支进行区域划分则很容易学习到相同的网络权重, 使分支检测效果与全局FEN检测效果相同, 因此为了让各分支FEN特征提取方向逐渐向不同的次显著区域前进, 设计一种局部特征差异网络对每个分支FEN模块的权重进行调整, 扩大模块间权重差异, 减小不同FEN分支的特征提取区域重叠程度。局部特征差异网络的结构如图1(d)所示, 为保证经过训练后各分支负责的检测区域不重叠, 需要在训练过程中逐步增大分支间最大特征值特征点间的距离。以图3行人图像为例, 在训练前为两个局部分支设置检测范围, 上矩形为head区域, 下矩形为legs区域, head框上端、左端、右端位置与真实框对应位置相同, 下端位置设置在真实框中点位置的0.97倍处; 下矩形上端位置设置在真实框中点1.02倍处, 其余与真实框相同, 在每一轮训练后搜索legs与head区域内的所有数值, 数值最大点 $F_{1,2}$ 对应FEN分支检测到的最大特征值特征点, 接着对 $F_{1,2}$ 的位置索引 $D_{1,2}$ 使用欧式距离公式^[26] (Euclidean Distance) 得

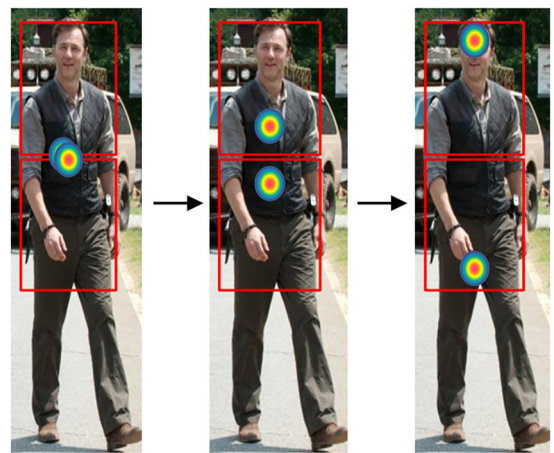


图3 FEN提取的局部特征点变化

Fig. 3 Local feature point change extracted by FEN

到分支间最大特征值特征点的距离,使用局部特征距离函数计算出的损失进行反向传播调整各分支 FEN 模块的权重,进而使检测特征点在每一轮训练后逐步从中间重叠部分向相反方向移动,到达上下限制框距离的最远处,即图 3 最后一幅图位置。

为使两个分支 FEN 检测到的特征点距离最大化,在模型训练中使用局部特征距离函数(Local Feature Distance, LFD)计算局部特征之间的距离损失 L_{distance} 。LFD 损失函数为公式(4),式中 F_i 表示经过分支 FEN 计算过后的特征分布二维张量, D_i 表示特征分布中特征值最大特征点的位置索引, index 为寻找位置索引函数。Sigmoid 函数将局部特征位置索引的距离映射到 $(0, 1)$ 之间, $-\log$ 函数对距离进行非线性变换,使特征点距离在变化域 $(0, 1]$ 内的输出值在 $(\infty, 0]$ 区间内递减得到 LFD 损失。

$$\begin{cases} D_i = \text{index}(\max(F_i)) \\ L_{\text{distance}} = -\log\left(\text{Sigmoid}\left(\sqrt{(D_1 - D_2)^2}\right)\right) \end{cases} \quad (4)$$

2.5 特征选择网络

在训练模型架构中,多分支特征提取网络提取出的多个局部特征可以提高网络对残缺或遮挡物体的检测能力,但同时网络在预测模型架构中会输出额外的局部检测框。如图 4 所示,右图是加入多分支特征提取网络预测架构的检测结果,与左图相比增加了两个局部特征检测框,为解决这些额外检测框带来的行人重复检测问题,

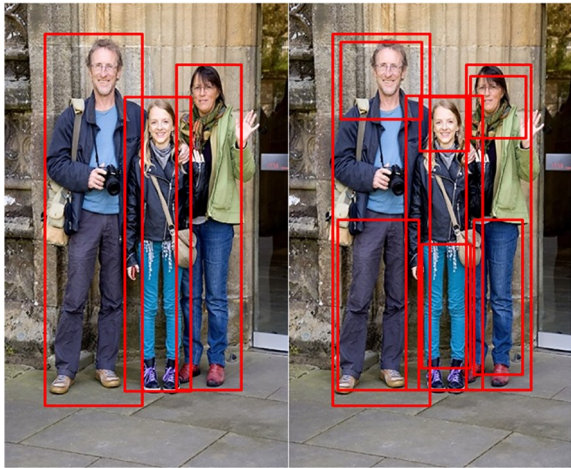


图 4 预测模型框架输出的结果

Fig. 4 Results of prediction framework output

使用特征选择网络 (Feature Select Net, FSN) 剔除重复检测到的冗余特征。

特征选择网络结构如图 1(e) 所示。首先选择 FEN 模块输出二维张量特征上达到置信度 (conf) 条件的所有特征点位置索引 $\text{mask}_k \{k \in g, 1, 2\}$, 接着根据索引数组筛选出对应位置特征 F_{reg} 的预测框信息 (box), 对所有 box 进行非极大值抑制^[27] (Non Maximum Suppression, NMS), 每一个分支对应的每个行人仅留下一个 conf 最高的 box, 最后对所有筛选出的 box 进行非最优值抑制 (Nos), 若一个行人存在多个分支的 box, 则根据设定的分支的优先级排序 (Rank) 依次选择优先级最大分支的 box 信息, Nos 的计算公式如式(5):

$$\begin{aligned} \text{Rank} &= \{F_g, F_i\} \\ B^{F_{(g,i)}} &= \{b_1^{F_{(g,i)}}, \dots, b_N^{F_{(g,i)}}\} \\ T^{F_{(g,i)}} &= \{t_1^{F_{(g,i)}}, \dots, t_N^{F_{(g,i)}}\} \\ b_i^F &= \begin{cases} 0, & \text{IoU}(b_i^{F_g}, t_i^{F_g}) \geq N_t \\ b_i^{F_g}, & \text{IoU}(b_i^{F_g}, t_i^{F_g}) < N_t, \text{IoU}(b_i^{F_i}, t_i^{F_i}) < N_t \\ b_i^{F_i}, & \text{Rank}(\text{IoU}(b_i^{F_i}, t_i^{F_i}) < N_t) \end{cases} \end{aligned} \quad (5)$$

其中: Rank 代表优先级排名, 设定 F_g 为第 1 优先级 $F_i \{i = 1, 2\}$ 的优先级排名依 i 值向后递增, $B^{F_{(g,i)}}$ 表示所有分支预测到的 box 信息, $T^{F_{(g,i)}}$ 表示每个行人的真实框信息, $b_i^{F_{(g,i)}}$ 表示一个预测 box, $t_i^{F_{(g,i)}}$ 表示一个行人的真实 box, N_t 代表阈值。Nos 公式中若 F_g 特征的 box 与真实框的交并比大于 N_t , 则进行非极大值抑制 (nms) 来剔除高重叠率 box, 若 F_g 与 F_i 的交并比均小于 N_t , 根据优先级排名仅留下第一优先级 F_g 特征的 box 删除其余特征, 如果多分支网络仅检测到多个局部分支特征 F_i 的 box, 则根据 Rank 选择优先级第一位的局部特征及其 box。使用 Nos 筛选得到图中每个行人最优位置检测框, 即最优位置二维张量 $F_{\text{max_reg}}$ 。

3 实验结果与分析

实验在 Win10 操作系统上进行训练和测试, 显卡使用 NVIDIA Tesla K80, 深度学习框架为 Pytorch1.7.1, 使用 GPU 进行运算, 使用 CUDA 并行架构来提高计算能力, 在训练过程中训练批

次大小(batch size)设置为 12,实验所需的具体运行环境如表 1 所示。

表 1 实验训练环境配置

Tab. 1 Experimental training environment configuration

名称	配置
CPU	Intel Xeon E5-2678 v3
显卡	NVIDIA Tesla K80
加速器	CUDA, CUDNN

3.1 数据集和评价指标

实验使用 WiderPerson^[28]数据集和 Crowd-Human^[29]数据集, WiderPerson 数据集共 13 382 张行人图片,有 9 000 张图片被标注约 40 万个标注框, CrowdHuman 数据集共有训练集 15 000 张图片,约有 470 万个标注,平均每张图片有 23 人,以上两个数据集中都包含大量拥挤、遮挡的行人,适合进行实验。为避免在同一个数据集上训练测试出现过拟合情况,实验使用 CrowdHuman 作为训练集,使用 WiderPerson 作为测试集。

在对算法模型进行测试时,需要通过模型的检测精度、检测速度和模型大小来判断检测性能,实验使用精确率(Precision)、平均精度均值(mean Average Precision)指标来评价模型的检测精度,使用帧率(Frames Per Second)指标来评价模型的检测速度,使用模型参数量(Parameter)评价模型的权重大小,使用计算量(GFLOPs)评价模型计算复杂度。

3.2 模型训练

使用 MBAN 算法模型在 CrowdHuman 数据集上进行了 120 个周期(epoch)的训练,前 85 轮周期冻结(freeze)模型中主干网络的参数,后 35 轮解冻(unfreeze)主干网络同时调整主干网络参数与后端网络参数,在训练中使用 Adam 优化器(Optimizer)调整参数,优化器中的变化量为模型当前网络参数(params)与下一次迭代使用的学习率(learning rate),params 数值在模型对每一批次训练产生的 loss 进行梯度下降(backward)后自动反向传播更新,learning rate 变化策略使用指数下降法,设置每个 epoch 学习率下降速率(learning decay)为 0.94,并在 unfreeze 后重新设定初值,变化情况如图 5 所示。

模型训练时的超参数设置使用了两种不同

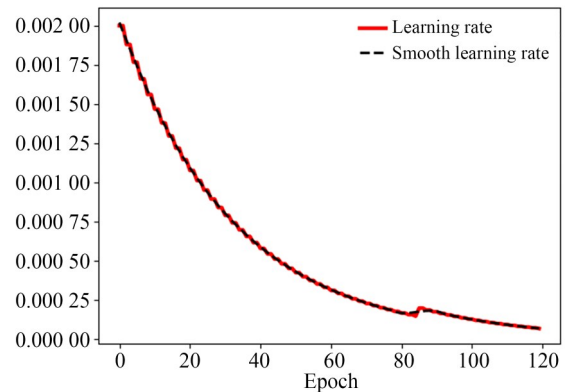


图 5 训练中学习率变化情况

Fig. 5 Learning rate changes during training

的策略。对于超参数 epoch 不设置固定数值,训练开始后模型会比较当前 epoch 对验证集的输出 loss 与上一轮 loss 的变化情况来判断模型是否逐步收敛,直至模型在一定 epoch 内 loss 不再下降,确定模型在当前 freeze 状态下已经达到收敛,之后模型 unfreeze 网络的所有权重再重新开始训练,当这次 loss 值稳定时模型训练结束,得到总训练周期、冻结训练周期、解冻训练周期数值;对于 learning rate 和 learning decay 超参数,参数值的改变只会影响模型的训练时长,不会影响模型最后收敛时的训练效果,实验中选择默认参数或根据经验设置。

每一轮 Optimizer 对网络参数进行改变后与目标数据进行对比,得到 MBAN 模型的损失值变化情况图 6,前 80 轮 epoch 中模型的损失值逐步减小趋于平稳,到 85 轮 epoch 后对主干模型解冻训练进一步使得损失值下降,在 120 轮周期处

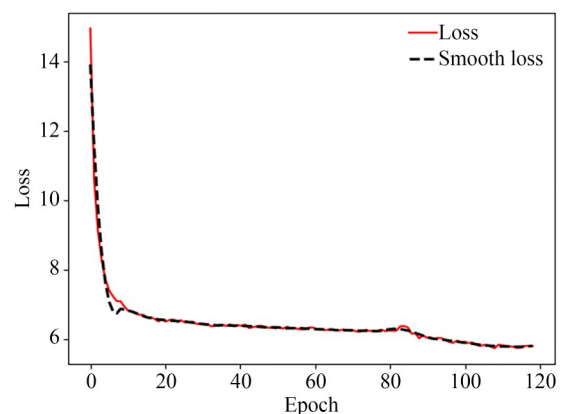


图 6 模型在训练过程中的损失变化情况

Fig. 6 Loss changes of the model during training

再次达到收敛,从整体趋势看出在模型训练中做的调整取得了让损失进一步下降的效果。

为了解训练过程中MBAN模型中各分支权重损失的变换情况,分开提取了 global, head, legs 的损失值,并绘制曲线图。

global 分支包含了两个损失为热力图损失 global_c_loss 和位置损失 global_reg_loss, 损失变化情况如图 7 所示。图中 global_c_loss 与 global_reg_loss 均逐渐降低并收敛,证明训练对 global 分支权重改变起到了作用。head 分支与 legs 分支只包含了热力图损失,损失变化如图 8 所示,图中 head_c_loss 与 legs_c_loss 损失也逐渐减小至收敛,说明权重向各自正确方向进行了变化,同时 legs 分支的损失相较 head 更大,说明对行人下半身提取特征与上半身相比更难找到一个较准确的泛化值。

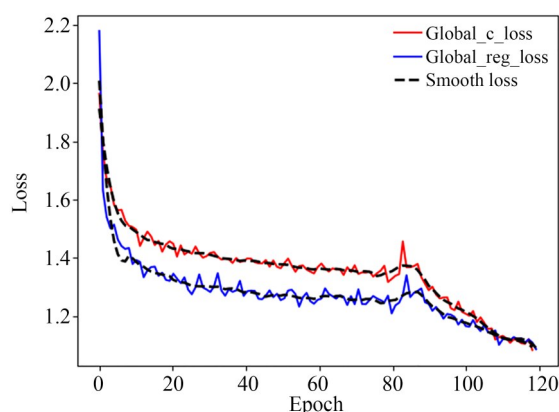


图 7 global 分支在训练过程中的损失变化情况
Fig. 7 Loss changes of the global during training

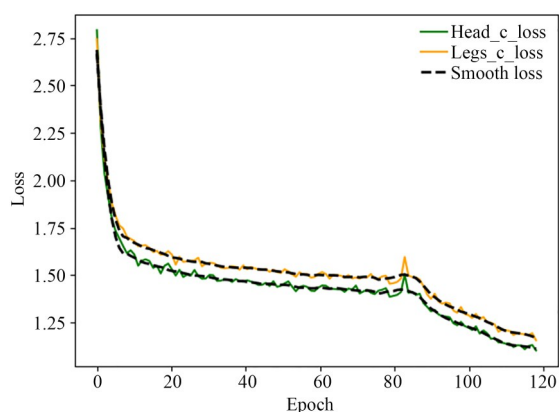


图 8 head 与 legs 分支训练过程中的损失变化情况
Fig. 8 Loss changes of the head and legs during training

3.3 消融实验

为了验证多分支无锚框网络模型各模块对整体检测效果的影响,在 WiderPerson 数据集上进行消融实验,实验以主干网络为 Resnet50 的无锚框模型作为基线模型(Baseline)。Baseline 网络结构如图 9 所示,由 Resnet50 和上采样卷积块构成,ResNet50 结构中除卷积块,还包含两个不同结构的残差块(Res_1, Res_2)来提高模型在较深层的特征提取能力,使行人图像在多个卷积块和残差块下采样后达到 2 048 深度,提取行人细节特征,随后再经过三个上采样卷积块对 Resnet50 特征提取结果上采样,得到预测需要的热力图特征(F_{Heatmap})、物体宽高特征(F_{HW})和中心点偏移量特征($F_{\text{Reg_offsets}}$)。与文献[16]中的网络结构相比,Baseline 在主干网络结构方面不同,使用了 Resnet50 替换了原主干网络,其余功能结构相同;与文献[17]中的网络结构相比,Baseline 也使用 Resnet50 主干网络替换了 Hourglass-104 网络结构,并将主干网络后的特征提取网络结构由分别生成中心点、角点两类关键

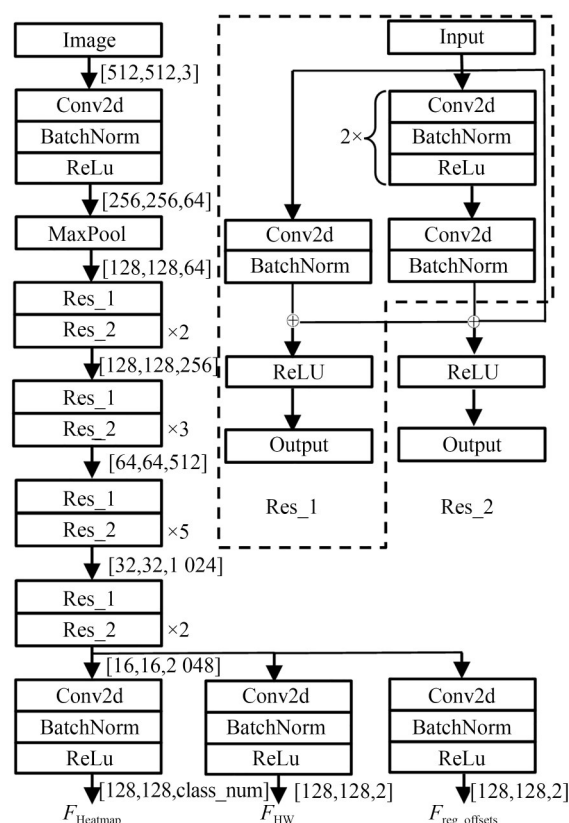


图 9 基线模型的网络结构

Fig. 9 Network structure of baseline model

点的分支结构变为仅生成一个中心关键点的分支结构。

消融实验基于基线模型逐步改进并测试性能,结果如表 2 所示。Baseline+hourglass+1FEN 和基线模型 Baseline 相比主干网络改进为 hourglass 了结构,模型平均检测精度(mAP)提高了 29.27%、F1 值提高 0.2、准确率提高 27.23%、召回率(Recall)提高 12.19%,表明 hourglass 结

构使模型对行人空间信息的提取能力得到了加强,因此检测到更多行人增强了模型检测性能。对比 Baseline+hourglass+1FEN(global)和 Baseline+hourglass+2FEN(global+head)性能,加入一个头部(head)FEN 分支后模型的 Prec 和 Recall 分别提高 6.38% 和 14.36%,证明该分支能有效增强网络对上半身特征的提取能力,学习到用局部特征判别行人。

表 2 消融实验结果

Tab. 2 Result of system scale calibration

Methods	mAP/%	F1	Prec/%	Recall/%	FPS	Params
Baseline	32.18	0.47	41.70	53.31	14.39	32.78
Baseline+hourglass+1FEN(global)	61.45	0.67	68.93	65.50	13.38	32.78
Baseline+hourglass+2FEN(global+head)	68.32	0.78	75.31	79.86	12.42	32.78
Baseline+hourglass+3FEN(global+head+legs)	75.04	0.83	78.89	86.55	11.19	32.78
Baseline+hourglass+3FEN+LFD	85.22	0.87	80.07	94.39	11.10	32.79

对比 Baseline+hourglass+2FEN(global+head)和 Baseline+hourglass+3FEN(global+head+legs)模型的性能,后者在 Prec 和 Recall 上分别提升 3.58% 和 6.69%,腿部(legs)分支 FEN 模块可以提取到下半身的局部特征,提升了对遮挡行人的检测能力,但由于腿部特征与头部和身体特征相比显著性较差,因此对这类特征学习提升的性能较小。Baseline+hourglass+2FEN+LFD 的检测性能在 mAP 值、F1 值、Prec、Recall 上均有提高,加入 LFD 模块监督各 FEN 向人体相反区域移动,学习不同身体部位间的互补特征,相比 Baseline+hourglass+2FEN 和 Baseline+hourglass+3FEN 固定位置学习,移动中的 FEN 能学习到更多的局部特征,提高模型的检测性能。

对比初始模型 Baseline 和所有改进模型,随着网络分支数量的增加模型参数量(Params)和检测速度(FPS)表现有些许降低,但牺牲一定空间和时间复杂度能较大提升检测能力,综合考虑 MBAN 仍是一个较好的检测方法。

3.4 可视化实验

为验证 MBAN 模型的每条分支均能正确检测到行人且检测到的部位不同,以及对模型的改进提升了检测效果,对各个阶段的模型输出可视化。

MBAN 中的 global, head 和 legs 分支对应着不同的检测区域,训练开始前确定各分支检测范围(range)以及高斯热力图中点位置(gspoint), global 分支的检测范围为整个真实框,其高斯热力图(mid_gspoint)为真实框中部;head 与 legs 以 global 作为基点来确定相对位置,head 检测框下端位置 head_down 设置范围为($head_up, mid$), legs 的上端位置 legs_up 为($mid, legs_down$)(mid 为真实框中间位置离 x 轴的距离),经过多次对参数的微调,head_down 设置为 $0.97*mid$, legs_up 为 $1.02*mid$ 有较好的效果,考虑到图片中行人的头部处于真实框中间偏上位置,head 的高斯热力图中点设置为 $0.85*mid_gspoint$,而在拥挤人群中由于行人的脚部与腿部经常被遮挡,过于下半身的特征很难被局部分支提取,因此将 legs 的高斯热力图设置在行人腰部略下位置 $1.09*mid_gspoint$,三个分支的可视化结果如表 3 所示,表中 F1 表示 head 分支检测到的行人上半部分特征, Fg 表示 global 分支检测行人中部全局特征, F2 为 legs 分支行人中间偏下部分特征。

对比三个分支的输出可视图,head 分支输出特征几乎覆盖了露出腰部以上位置的所有行人, global 分支主要识别行人特征最显著区域,因此输出仅覆盖了特征最显著的行人, legs 在训练时

表 3 MBAN 模型的各分支输出可视化图表

Tab. 3 Visualization of each branch output of MBAN

原图	
F1	
Fg	
F2	

设置高斯热力图中点设置在 global 中点偏下处,通过可视化图可以看出训练后 legs 分支检测到的特征也在 global 偏下位置,证明参数设置起到了作用。由于 WiderPerson 数据集的真实框不仅包括完整行人,还包括只有局部身体部分的行人,下半分支不仅能学习到人体下半部局部特征,还能学习到部分上半身偏下部分局部特征,legs 分支在可视图覆盖了更大的检测面积。

为了验证多分支无锚框网络模型对密集行人的检测效果,将 Baseline, Baseline+2FEN, Baseline+3FEN+LFLoss 模型输出结果进行可视化分析。在 WiderPerson 检测数据集上选取多张存在人数密集、姿态变化多以及人体互相遮挡等拥挤道路上的常见问题的行人图片,使用三个模型对图片检测得到模型激活区域分布情况。如表 4 所示,MBAN 模型高激活区域与 Baseline 相比对人体的覆盖范围更广,由于 LFD 网络的原因,各分支特征提取网络激活区域并没有聚集覆盖在显著特征区域内,覆盖区域包括了人体头部、身体中部以及身体中部偏下位置,使得模型能够学习到更加全面的行人特征。MBAN 模型与 Baseline+2FEN 模型相比,在保持检测头部高

激活区域不变的情况下,增加了行人中部偏下部分的局部特征检测,因此对行人下半区域有了更高的关注度,下半区域的激活区域范围和激活值更大。

由可视化实验结果发现,MBAN 算法模型中的每条分支均正确检测到了行人的不同区域特征,且加入分支后的模型检测范围比 Baseline 模型检测范围变得更大,覆盖到更多被遮挡的行人,证明在无锚框网络中加入多分支模块可以对拥挤遮挡情况下的行人提取到更多的特征,从而提高检测时的召回率。

表 4 不同网络结构模型的输出可视化图表

Fig. 4 Output visual of different network structure models

原图	
Baseline	
Baseline+2FEN	
Baseline+3FEN+LFD (MBAN)	

3.5 与现有方法对比实验

为验证 MBAN 算法检测性能的优越性,实验选择目前行人检测领域的主流检测算法 CenterNet, ResNet50-FCOS^[30], YOLOv3^[31], VGG16-SSD^[32], YOLOv3-tiny^[33] 与本文算法 MBAN 进行比较,其中 VGG16-SSD, YOLOv3, YOLOv3-tiny 是基于锚框的类型检测方法, ResNet50-FCOS, CenterNet 与 MBAN 属于无锚框类型检测方法。陈一潇^[34]等使用 ResNet50-FCOS, YOLOv3, VGG16-SSD, YOLOv3-tiny 算法在 CrowdHuman 数据集上进行了训练与测试,

为保证检测结果一致性,实验中将 MBAN 算法模型在 CrowdHuman 数据集上训练与测试,训练样本与测试样本比例设置为 9:1,同时加入使用 CrowdHuman 数据集训练, WiderPerson 数据集测试的模型结果横纵向对比,验证模型泛化性能,各方法的检测性能结果如表 5 所示。

对比有锚框模型算法 VGG16-SSD、YOLOv3、YOLOv3-tiny 与无锚框模型算法 MBAN (CrowHuman) 的性能,有锚框算法检测精度普遍较高,平均检测精度(mAP)均达到了 65% 以上,其中 YOLOv3 检测精度最高达到 84.22%,但相应地模型参数量(Params)和计算量(GFLOPs)分别达到 61.5,154.9,计算复杂度在所有锚框算法中最大,推理时间(FPS)每秒 27 帧速度,而有锚框算法中计算复杂度最小模型 YOLOv3-tiny 使用 Darknet19 网络替代了 Darknet53,使得模型 GFLOPs 减小为 12.9,Params 为 8.0,FPS 达到每秒 100 帧,但同时 mAP 相比 MBAN 算法下降 16.99。CNN 主干网络虽然让 YOLOv3-tiny 在计算复杂度和推理时间上优于 MBAN (CrowHuman),但无法达到检测行人的精度要求。YOLOv3 的 mAP 值与实验中其余算法相比均处于领先,与经过所有改进后的 MBAN (CrowHuman)算法相比也仅低 1%,分析原因首先是因为 YOLOv3 的主干网络使用了 DarkNet53,该主干网络使用残差网络(Residual)加深网络深度,并用特有的卷积块(Conv2d+BN+LeakyReLU)来增强网络的特征提取能力,文章^[35]也证明 Darknet53 的 Top-5 检测性能与 ResNet152 相当,高于 ResNet50;其次 YOLOv3 在主干网络输出部位加入了 FPN 特征金字塔提取网络,每个特征输出层至少经过 5 次卷积(Conc2d)再上采样(upsampling)与下一层输出层结合,得到语义更加丰富的特征;最后在预测端,基于锚框类型的 YOLOv3 会在输出特征层的每一个特征点上生成 3 个锚框,更容易搜索到特征点处可能存在的行人。这些特点与技巧让 YOLOv3 在行人检测领域中有较好的检测精度,同时过深的卷积网络和大量锚框造成了高 Params 和高 GFLOPs。MBAN (CrowHuman)与 YOLOv3 相比在模型结构和大小方面更精简,Params 降低 46.7%,GFLOPs 降低 84.6%,主要

原因是因为 MBAN 在主干网络使用了参数量相对较小的 Resnet50-hourglass 代替 Draknet53,并使用多分支结构代替锚框结构,从而在减小模型复杂度的情况下提高模型检测精度。

对比无锚框模型算法 MBAN (CrowHuman)、ResNet50-FCOS 与 CenterNet, MBAN (CrowHuman)在 mAP、GFLOPs、Params 方面均优于 ResNet50-FCOS,与经典无锚框模型算法 CenterNet 相比,MBAN 仅在 GFLOPs 增加 1.7、Params 增加 7.1 的情况下将 mAP 提高了 22.06,对拥挤行人检测精度有了很大提升。MBAN 作为基于 CenterNet 的改进算法,保持了用关键点检测物体预测框中心点,回归预测框大小与类别属性的核心思想,在主干网络、特征提取网络、损失函数以及预测端方面进行了改进。主干网络方面,文章^[16]中 CenterNet 算法使用 ResNet18、DLA34、Hourglass104 三种主干网络进行实验,考虑到 ResNet18 网络的推理速度快、精度低,Hourglass104 的计算复杂度高、推理速度慢但检测精度高,为使 MBAN 能将两者优势结合,主干网络选择使用计算复杂度和推理速度适中的 ResNet50 替代,并按照 Hourglass 网络结构加入多个上采样模块改进得到 ResNet50-Hourglass。特征提取网络方面,CenterNet 在主干网络输出特征层尾部没有加入特征提取网络就输入预测端,而 MBAN 在主干网络尾部加入多分支特征提取网络(FEN)对输出特征层进行分支,因此 CenterNet 算法模型输出仅有一个全局特征,而 MBAN 算法模型输出为一个全局特征和多个局部特征。损失函数方面,CenterNet 使用 L1 损失函数计算位置损失,使用 Focal loss 损失函数计算热力图损失,MBAN 算法位置损失 L_{reg}^s 沿用了 L1,热力图损失 L_c^s 使用超参数测试改进后的 Focal loss 来适应不同分支的热力图变化情况,并增加了局部特征距离函数(LFD)计算局部特征提取网络输出的多个分支特征间距离损失。预测端方面,MBAN 相比 CenterNet 加入了特征选择网络(FSN)用以消除检测时的重复检测框和冗余检测框。基于 CenterNet 的以上改进使 MBAN 算法极大提高了对拥挤行人的检测精度,同时也增加了部分计算复杂度,使模型推理时间比 CenterNet 慢了约每秒 3 帧,但 FPS 仍大于 10 基本满

足实时检测要求。

对比 MBAN 在训练集 CrowHuman 和测试集 WiderPerson 上的性能,两者在检测精度方面差别不大,训练集的 mAP 值仅高于测试集 mAP 值 1.76% 左右,同时两种模型相比于其他行人检测算法检测精度均处于领先,表明 MBAN 对拥挤行人的检测能力较强。

在与现有行人检测算法对比实验中发现,MBAN 算法的检测精度超过了有锚框模型算法中领先的 YOLOv3 算法,并且计算复杂度与模型推理时间在无锚框算法中也较好,与 CenterNet 相比在检测精度方面有较大优势,在 3.6 小节中将对 MBAN 与 CenterNet 对行人图像的检测结果进行对比。

表 5 本文算法与其它算法的对比结果

Tab. 5 Comparison results between MBAN method and other methods

Methods	Backbone	Size	mAP/%	Params($\times 10^6$)	GFLOPs	FPS
VGG16-SSD	VGG16	512 $\times$ 512	69.16	60.3	22.9	35
YOLOv3	Darknet53	640 $\times$ 640	84.22	61.5	154.9	27
YOLOv3-tiny	Darknet19	640 $\times$ 640	68.23	8.0	12.9	100
ResNet50-FCOS	Resnet50	640 $\times$ 640	69.66	50.2	195.9	22
CenterNet	Hourglass104	512 $\times$ 512	63.16	25.7	22.1	14.
MBAN(CrowHuman)	Resnet50	512 $\times$ 512	86.98	32.8	23.8	11
MBAN(WiderPerson)	Resnet50	512 $\times$ 512	85.22	32.8	23.8	11

3.6 模型检测结果

本节使用 CenterNet 与 MBAN 算法对拥挤环境下的行人图像进行检测,在数据集中选取了四张具有代表性的图片,图片中的行人均具有拥挤、部分遮挡、姿态变化等现实拥挤街道上会出现的情况。使用 CenterNet 算法检测结果为图 10,第一张图片行人遮挡与姿态变换程度相对较小,模型检测到了大部分行人,而后三张图片行人姿态变换较多导致模型检测能力显著下降,仅能检测到靠前或特征明显的行人,另外 CenterNet 也在大量数据集训练中学习到了部分行人局部特征,但仅能够检测到一些特征明显的行人头



图 10 CenterNet 模型检测结果

Fig. 10 Prediction results of CenterNet model

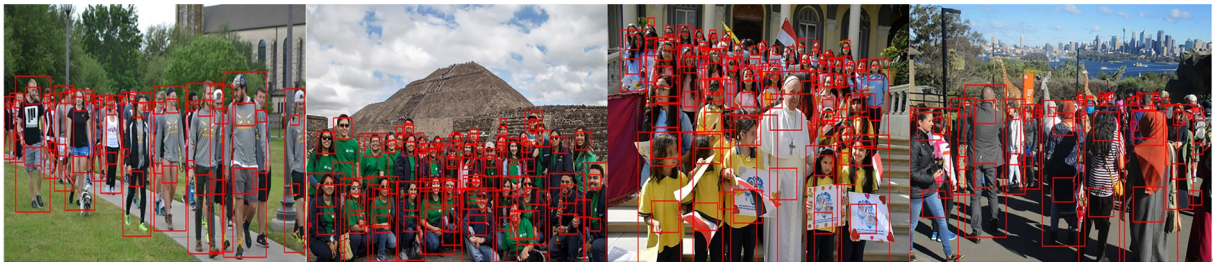


图 11 MBAN 模型检测结果

Fig. 11 Prediction results of MBAN model

部。MBAN算法对行人检测结果为图 11,由图 11 结果可以看到 MBAN 对检测到的每个无遮挡行人输出了三个检测框,大框对应 global 分支检测到的 box,小框对应 head 与 legs 分支检测 box,对于图中受到遮挡的行人,也会输出局部分支 head 和 legs 检测到的 box。MBAN 检测结果与图 8 的 CenterNet 检测结果相比检测到了更多行人,可以再次证明 MBAN 检测性能高于 CenterNet。

MBAN 使用两个局部分支检测遮挡行人在提高了召回率的同时也出现了大量冗余框,如图 11 所示,如果行人已经被 global 分支的 box 检测

到,则局部分支检测到的 box 成为冗余框,如果行人上半身被 head 分支 box 检测到,legs 分支检测到的 box 则为冗余。为解决检测框冗余问题,实验在模型预测框架中加入 FSN 网络,加入 FSN 网络后的检测结果为图 12,与图 11 相比加入 FSN 的 MBAN 算法检测无遮挡行人时仅保留了 global 分支的 box,在检测遮挡行人时优先保留 head 分支的 box,head 局部特征无法检测到的情况下最后保留下 legs 分支的 box。由图 12 检测结果可以看到加入 FSN 网络的 MBAN 算法正确按分支优先级排序筛选检测框,剔除了多分支带来的冗余检测框。



图 12 在预测框架中加入 FSN 的 MBAN 模型检测结果

Fig. 12 Prediction results of MBAN model with FSN

4 讨 论

MBAN 算法模型使用多分支网络对行人特征检测,在提高了拥挤状态下行人召回率和精度的同时,也出现了对遮挡严重的小目标行人特征定位偏差问题,如图 12 第三幅图中最上排检测框的位置相较于脸部区域偏上,这是在训练中加入局部特征距离损失函数(LFD)产生的问题。训练中使用 LFD 损失判断局部分支间的距离,局部分支距离较近时损失越大,由此训练两个局部检测分支检测区域向上下两个区域移动,为了保证两个分支距离不会太远或太近,设置上下限超参数限制移动范围;在实验中发现若检测对象为目标较大的无遮挡行人,那么设置较大的上下限超参数才能使上下分支检测区域到达脸部和腿部位置,但使用该参数下的模型对小目标遮挡行人检测则会使局部检测框发生上下两边的距离偏差,原因是 LFD 导致局部分支的检测范围达到或超出了小目标的全局检测框边线位置,同理

若将上下限参数设置较小来满足小目标的检测范围,会出现无遮挡行人的检测范围过近,上下分支检测区域与全局检测区域合并到一起的情况,因此为了同时满足模型对无遮挡行人和小目标行人的检测效果,经过反复实验选取了适中的上下限参数,但对于遮挡严重的极小目标,LFD 函数仍会使局部检测框出现偏差,需要分类讨论。

虽然 MBAN 算法模型对密集环境下的行人有较好的检测效果,但模型仍存在一些局限性,首先是对 LFD 函数上下限参数的设置问题,经过人为实验经验设置的超参数不能同时满足模型对小目标和大目标的检测效果,因此在后期实验打算设计一种方法对行人的真实框大小分类,在训练时按类别不同使用相对应的上下限参数计算 LFD 损失,较大的目标给予更大的上下限参数,小目标给予较小的参数,解决统一参数造成距离偏差造成的检测误差;其次,MBAN 模型的参数量为 32.8×10^6 ,相较于轻量级模型仍有

一定距离,在之后的实验中为将模型部署在移动设备当中,要对模型进行剪枝(Pruning),在保证模型精度不发生较大下降的基础上,逐步减少主干网络卷积层的数量或使用参数量较小的深度可分离卷积(Depthwise Separable convolution)进行替换,最后,实验后序使用知识蒸馏方法(Knowledge Distillation)进一步减小模型的参数量和计算量,尝试建立一个新的轻量级模型作为学生模型,拟合 MBAN 模型输出结果,得到与 MBAN 模型相似的检测效果。

5 结 论

在拥挤街道场景下,无人驾驶汽车对街道人流检测时可能因为行人互相遮挡或多姿态变化造成漏检问题,为此,本文提出多分支无锚框网络(MBAN)模型算法。MBAN 算法使用局部特征差异网络(LFD)训练局部特征提取网络(FEN)中每条分支的权重,使各分支检测区域差异化,并在原主干网络 ResNet 尾部加入 Hourglass 结构提高模型对空间信息的敏感度,最后使用特征选择网络(FSN)按优先级思想选择合适

的检测框分支,剔除冗余结果。

对 MBAN 算法模型进行了消融实验、可视化实验和对比实验。消融实验中,经过改进的 MBAN 模型平均检测精度达到 85.22%,模型召回率达到 94.39%,比基线模型提高 41.08%,证明改进模型加强了遮挡与多姿态行人的特征提取能力,从而提高召回率;对检测结果可视化显示特征提取分支的检测区域存在差异,改进后模型检测区域大小增加,证明局部特征距离函数(LFD)能够有效指导分支网络向不同方向区域学习,增加模型对检测行人的敏感度;对比实验中 MBAN 算法在不同数据集下的 mAP 值分别为 85.22 和 86.98,领先于其他行人检测算法,证明了算法的优异性。

以上实验结论证明本文算法拥有较好的检测性能,有能力检测出拥挤环境下受到遮挡的多姿态行人,可以应用于对拥挤行人的检测任务中。在未来的工作中计划将讨论中的工作进一步完成。并使用 Transformer^[36-37]等优秀算法中较为先进的理念与技巧加入到本文算法中进一步提升检测能力,配置到移动端进行测试。

参考文献:

- [1] 侯佳伟. 从七次全国人口普查看我国人口发展新特点及新趋势[J]. 学术论坛, 2021, 44(5): 1-14.
HOU J W. Looking at the new characteristics and trends of China's population development from the seven national censuses [J]. *Academic Forum*, 2021, 44(5): 1-14. (in Chinese)
- [2] LUO C X, YANG X D, YUILLE A. Exploring Simple 3D Multi-Object Tracking for Autonomous Driving [C]. 2021 *IEEE/CVF International Conference on Computer Vision (ICCV)*. 10-17, 2021, *Montreal, QC, Canada*. IEEE, 2022: 10468-10477.
- [3] LIU C, TANG X, MA J J, *et al.* Remote Sensing Images Feature Learning Based on Multi-Branch Networks[C]. *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*. September 26-October 2, 2020, *Waikoloa, HI, USA*. IEEE, 2021: 2057-2060.
- [4] KE W, ZHANG T L, HUANG Z Y, *et al.* Multiple Anchor Learning for Visual Object Detection [C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13-19, 2020, *Seattle, WA, USA*. IEEE, 2020: 10203-10212.
- [5] LI S, YANG L X, HUANG J Q, *et al.* Dynamic Anchor Feature Selection for Single-Shot Object Detection[C]. 2019 *IEEE/CVF International Conference on Computer Vision (ICCV)*. October 27 - November 2, 2019, *Seoul, Korea (South)*. IEEE, 2020: 6608-6617.
- [6] GIRSHICK R, DONAHUE J, DARRELL T, *et al.* Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation [C]. 2014 *IEEE Conference on Computer Vision and Pattern Recognition*. 23-28, 2014, *Columbus, OH, USA*. IEEE, 2014: 580-587.
- [7] REDMON J, DIVVALA S, GIRSHICK R, *et al.* You Only Look Once: Unified, Real-Time Object Detection[C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 27-30,

- 2016, Las Vegas, NV, USA. IEEE, 2016: 779-788.
- [8] ROSZYK K, NOWICKI M R, SKRZYPCZYŃSKI P. Adopting the YOLOv4 architecture for low-latency multispectral pedestrian detection in autonomous driving[J]. *Sensors*, 2022, 22(3): 1082.
- [9] GAWANDE U, HAJARI K, GOLHAR Y. SIRA: Scale illumination rotation affine invariant mask R-CNN for pedestrian detection[J]. *Applied Intelligence*, 2022, 52(9): 10398-10416.
- [10] 李经宇, 杨静, 孔斌, 等. 基于注意力机制的多尺度车辆行人检测算法[J]. *光学精密工程*, 2021, 29(6): 1448-1458.
- LI J Y, YANG J, KONG B, *et al.* Multi-scale vehicle and pedestrian detection algorithm based on attention mechanism [J]. *Opt. Precision Eng.*, 2021, 29(6): 1448-1458. (in Chinese)
- [11] 马立, 巩笑天, 欧阳航空. Tiny YOLOV3 目标检测改进[J]. *光学精密工程*, 2020, 28(4): 988-995.
- MA L, GONG X T, OUYANG H K. Improvement of Tiny YOLOV3 target detection[J]. *Opt. Precision Eng.*, 2020, 28(4): 988-995. (in Chinese)
- [12] JIN Z C, LIU B, CHU Q, *et al.* SAFNet: A Semi-Anchor-Free Network with Enhanced Feature Pyramid for Object Detection [C]. *IEEE Transactions on Image Processing*, 7, 2020, IEEE, 2020: 9445-9457.
- [13] ZHU C C, HE Y H, SAVVIDES M. Feature Selective Anchor-Free Module for Single-Shot Object Detection [C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15-20, 2019, Long Beach, CA, USA. IEEE, 2020: 840-849.
- [14] ZHANG S F, CHI C, YAO Y Q, *et al.* Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection [C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13-19, 2020, Seattle, WA, USA. IEEE, 2020: 9756-9765.
- [15] LAW H, DENG J. CornerNet: detecting objects as paired keypoints [J]. *International Journal of Computer Vision*, 2020, 128(3): 642-656.
- [16] ZHOU X, WANG D, KRÄHENBÜHL P. Objects as points [J]. *arXiv preprint arXiv:1904.07850*, 2019.
- [17] DUAN K W, BAI S, XIE L X, *et al.* Centernet: keypoint triplets for object detection [C]. 2019 *IEEE/CVF International Conference on Computer Vision (ICCV)*. October 27 - November 2, 2019, Seoul, Korea (South). IEEE, 2020: 6568-6577.
- [18] CAO Z W, YANG H H, XU W J, *et al.* Multi-scale anchor-free region proposal network for pedestrian detection [J]. *Wireless Communications and Mobile Computing*, 2021, 2021: 1-12.
- [19] DING Z F, GU Z C, SUN Y P, *et al.* Cascaded cross-layer fusion network for pedestrian detection [J]. *Mathematics*, 2022, 10(1): 139.
- [20] 邹逸群, 肖志红, 唐夏菲, 等. Anchor-free 的尺度自适应行人检测算法[J]. *控制与决策*, 2021, 36(2): 295-302.
- ZOU Y Q, XIAO Z H, TANG X F, *et al.* Anchor-free scale adaptive pedestrian detection algorithm [J]. *Control and Decision*, 2021, 36(2): 295-302. (in Chinese)
- [21] HE K M, ZHANG X Y, REN S Q, *et al.* Deep Residual Learning for Image Recognition [C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 27-30, 2016, Las Vegas, NV, USA. IEEE, 2016: 770-778.
- [22] CHU S W, SONG Y, ZOUO J J, *et al.* Human Pose Estimation Using Deep Convolutional Densenet Hourglass Network with Intermediate Points Voting [C]. 2019 *IEEE International Conference on Image Processing (ICIP)*. 22-25, 2019, Taipei, China. IEEE, 2019: 594-598.
- [23] WENG X, YAN Y, DONG G S, *et al.* Deep multi-branch aggregation network for real-time semantic segmentation in street scenes [C]. *IEEE Transactions on Intelligent Transportation Systems*. 16, 2022, IEEE, 2022: 17224-17240.
- [24] CHEN Z F, QIN X, YANG C, *et al.* Composite localization for human pose estimation [J]. *arXiv preprint arXiv:2105.07245*, 2021.
- [25] DU C J, YU H, YU L. A scale-sensitive heatmap representation for multi-person pose estimation [J]. *IET Image Processing*, 2022, 16(4): 1194-1207.
- [26] WU H, CAO Y, WEI H P, *et al.* Face recognition based on haar like and euclidean distance [J]. *Journal of Physics: Conference Series*, 2021, 1813(1): 012036.
- [27] CHOI S B, LEE S S, PARK J, *et al.* Standard

- greedy non maximum suppression optimization for efficient and high speed inference[C]. 2021 *IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*. 1-3, 2021, Gangwon, Korea, Republic of. IEEE, 2021: 1-4.
- [28] ZHANG S F, XIE Y L, WAN J, *et al.* WiderPerson: a diverse dataset for dense pedestrian detection in the wild[J]. *IEEE Transactions on Multimedia*, 2020, 22(2): 380-393.
- [29] SHAO S, ZHAO Z, LI B, *et al.* Crowdhuman: A benchmark for detecting human in a crowd[J]. *arXiv preprint arXiv:1805.00123*, 2018.
- [30] TIAN Z, SHEN C H, CHEN H, *et al.* FCOS: A Simple and Strong Anchor-Free Object Detector [C]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. October 19, 2020, IEEE, 2020: 1922-1933.
- [31] 王宸, 张秀峰, 刘超, 等. 改进 YOLOv3 的轮毂焊缝缺陷检测[J]. *光学精密工程*, 2021, 29(8): 1942-1954.
- WANG CH, ZHANG X F, LIU CH, *et al.* Detection method of wheel hub weld defects based on the improved YOLOv3[J]. *Opt. Precision Eng.*, 2021, 29(8): 1942-1954. (in Chinese)
- [32] KANG H J. Real-time Object Detection on 640x480 Image with VGG16 SSD[C]. 2019 *International Conference on Field-Programmable Technology (ICFPT)*. 9-13, 2019, Tianjin, China. IEEE, 2020: 419-422.
- [33] WANG Y M, JIA K B, LIU P Y. Impolite pedestrian detection by using enhanced YOLOv3-tiny [J]. *Journal on Artificial Intelligence*, 2020, 2(3): 113-124.
- [34] 陈一潇, 阿里甫·库尔班, 林文龙, 等. 面向拥挤行人检测的 CA-YOLOv5[J]. *计算机工程与应用*, 2022, 58(9): 238-245.
- CHEN Y X, ALIFU K, LIN W L, *et al.* CA-YOLOv5 for crowded pedestrian detection [J]. *Computer Engineering and Applications*, 2022, 58(9): 238-245. (in Chinese)
- [35] REDMON J, FARHADI A. YOLO9000: Better, Faster, Stronger[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 21-26, 2017, Honolulu, HI, USA. IEEE, 2017: 6517-6525.
- [36] CAO J, PANG Y, ANWER R M, *et al.* PSTR: End-To-End One-Step Person Search with Transformers[C]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022: 9458-9467.
- [37] YU R, DU D, LALONDE R, *et al.* Cascade Transformers for End-to-End Person Search[C]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022: 7267-7276.

作者简介:



吕志轩(1998—),男,江苏徐州人,硕士研究生,2020年于南京邮电大学获得学士学位,主要从事深度学习图像处理、目标检测方面的研究。E-mail: lvzhixuan368@163.com

通讯作者:



魏霞(1977—),女,新疆呼图壁人,研究生导师,副教授,主要从事智能数据采集、网络安全、智能运维方面的研究。E-mail:30462111@qq.com