计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(9):302−308 [doi: 10.15888/j.cnki.csa.008108]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于嵌入式设备的 Anchor Free 行人检测①

张立国[1], 刘　博[1], 孙胜春[1], 张　勇[1], 金　梅[2]

[1](燕山大学 电气工程学院, 秦皇岛 066004)
[2](燕山大学 河北省测试计量技术与仪器重点实验室, 秦皇岛 066004)
通讯作者: 金　梅, E-mail: meijin297@126.com

摘　要: 通过嵌入式设备在边缘端进行行人检测能满足实时、安全与隐私保护等方面的基本需求. 由于原 CenterNet 检测网络模型 backbone 通常以 DLA、Hourglass 等复杂度较高的多层特征融合结构, 嵌入式设备的计算能力有限难以满足实时的要求, 因此基于 BiFPN 网络结构和加权特征融合方法, 通过对 backbone 中的不同特征层进行加权融合, 改进了原来的 backbone 方法, 在保证检测精度的同时提升了检测速度. 同时针对行人这一特定的检测类别, 通过修改训练期间 HeatMap 上高斯核分布, 增加对行人检测的适应性, 进一步减少了因行人之间相互遮挡而漏检造成的精度降低. 在 Jetson TX2 上的实验结果表明, 改进后的行人检测 *AP* 为 0.774, 同时单张图像的推理时间为 68 ms, 能够满足在嵌入式设备上的实时要求.

关键词: 嵌入式设备; CenterNet; 加权特征融合; 目标检测; 高斯核分布

引用格式: 张立国,刘博,孙胜春,张勇,金梅.基于嵌入式设备的 Anchor Free 行人检测.计算机系统应用,2021,30(9):302–308. http://www.c-s-a.org.cn/1003-3254/8108.html

Anchor Free Pedestrian Detection Based on Embedded Device

ZHANG Li-Guo[1], LIU Bo[1], SUN Sheng-Chun[1], ZHANG Yong[1], JIN Mei[2]

[1](School of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China)
[2](Key Laboratory of Measurement Technology and Instrument of Hebei Province, Yanshan University, Qinhuangdao 066004, China)

Abstract: Using embedded devices to detect pedestrians at the edge can meet the basic needs of real time, security and privacy protection. The original CenterNet backbone network model usually adopts Deep Layer Aggregation (DLA), Hourglass, etc. with high complexity for multi-level features fusion, which limits the computing power of embedded devices and thereby makes the real-time detection difficult. In view of this, BiFPN and weighted feature fusion are employed for the weighted fusion of feature layers in the backbone, by which the original backbone method is improved. This strategy enhances the detection speed while ensuring the detection accuracy. Further, the Gauss kernel distribution on the HeatMap during training was modified so that the adaptability to pedestrian detection can be increased. As a result, the accuracy reduction caused by missing detection due to pedestrian occlusion is lowered. The results of the experiment on Jetson TX2 show that the Average Precision (*AP*) of pedestrian detection with the improved method is 0.774, and the inference time of a single image is 68 ms, which can meet the requirements of embedded devices for real-time detection.

Key words: embedded device; CenterNet; weighted feature fusion; object detection; Gaussian kernel distribution

① 基金项目: 中央引导地方科技发展专项 (199477141G); 河北省引智项目
Foundation item: Special Fund of Central Government for Local Science and Technology Development (199477141G); Project for Intelligent Introduction of Hebei Province
收稿时间: 2020-12-20; 修改时间: 2021-01-18; 采用时间: 2021-02-03; csa 在线出版时间: 2021-09-02

行人检测是计算机视觉和数字图像处理的一个方向, 广泛用于安防、智能视频监控等领域, 将计算机视觉检测目标用在减少人力的使用的同时提高检测精度、提高灵活性具有重要意义. 目前已有的行人检测方法主要分为两大类, 一类是基于传统视觉处理的方法, 主要包括基于背景建模的算法、基于手工特征与机器学习的检测算法. 另一类主要是以神经网络为主的目标检测算法.

对以上算法中第一类算法的背景建模方法而言, 其主要是通过对背景进行建模, 然后将当前图像与背景模型进行比较, 确定前景, 如 ViBe 算法[1,2]、光流法 [3,4] 等, 该类方法通常受环境光照变化、背景的多模态性、运动物体的阴影等多方面因素的影响, 不具备较好的鲁棒性. 相比于背景建模算法, 基于手工特征与机器学习算法的方法主要通过特定的特征实现检测, 如 HOG+SVM[5,6]、HOG+DPM[7], 但该类方法很难处理遮挡问题, 人体姿势动作幅度过大或物体方向改变也不易检测.

在另一大类基于神经网络的算法中, 主要是以特征网络提取特征然后组合头部网络回归定位具体位置的方法定位检测目标为主, 近年来衍生出多种系列的检测算法, 如 YOLO 系列[8−11]、RCNN 系列[12−14]、Anchor Free 系列[15−17], 在实际的嵌入式设备应用上, 主要是以 YOLO 系列的阉割版和 Anchor Free 系列为主, 相对而言, YOLO 系列的阉割版虽然能取得较高的模型推理速度, 但是当出现部分遮挡, 行人部分超出视野范围等情况, 精度会严重降低. 而既有的 Anchor Free 方法虽然整体结构较为简单适用于嵌入式设备的部署, 但是如 CenterNet[16]、FCOS[17] 等特征提取网络结构特征提取层和参数量较多会严重导致推理速度变慢, 所以本文通过优化特征网络结构进行特征提取, 从而保证头部网络输入特征的有效性, 此外针对行人间的相互遮挡情形, 提出针对行人的高斯核分布改进方式, 保证了模型的检测精度.

1 相关理论基础

1.1 CenterNet 算法原理

CenterNet 是基于中心点的检测方法, 使用图像作为输入, 然后经过骨干网络提取特征, 最后在头部网络经过 3 个分支, 一个分支预测中心点的位置 (HeatMap), 一个分支预测因下采样过程带来中心点位置误差的修正量 (offset), 最后一个分支预测检测框的大小 (scale), 其抽象结构见图 1. 一般而言, 输入图像可用$I \in R^{W\times H\times 3}$表示, 其中$W$表示图像的宽度, H 表示图像的高度, 3 为图像的通道数, 预测中心点的分支最后得到预测结果$\hat{Y} \in [0,1]^{\frac{W}{R}\times\frac{H}{R}\times C}$, 其中 R 代表下采样率, C 代表类别数, 表示与检测物体的中心点相关, $\hat{Y}=0$则表示与背景相关, 预测修正量的分支会得到$\frac{W}{R}\times\frac{H}{R}\times 2$大小的预测值来表示每个中心点的修正值, 同样的预测修正量的分支会得到$\frac{W}{R}\times\frac{H}{R}\times 2$大小的预测值来表示每个中心点的修正值, 预测检测框大小的分支会得到$\frac{W}{R}\times\frac{H}{R}\times 2$的预测值来表示检测物体的宽和高. 而骨干网络一般采用 Hourglass[18]、DLA[19] 等多层特征融合的模型.

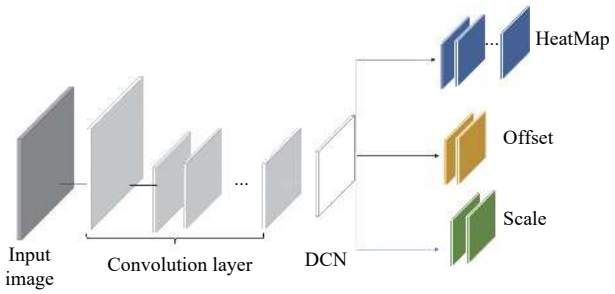


图 1 CenterNet 的网络结构图

在训练 CenterNet 过程中, 通常按以下方式设置 ground truth 和损失函数, 记检测物体在原图上的中心点为p, 计算图像经下采样后低分辨率的同一位置为$\tilde{p}=\left\lfloor\frac{p}{R}\right\rfloor$, 对于 HeatMap 分支设置 ground truth 为高斯核的分布形式:

$$Y_{xyc}=\exp\left(-\frac{(x-\tilde{p}_x)^2+(y-\tilde{p}_y)^2}{2\sigma^2}\right) \tag{1}$$

其中, σ_p是标准差, Y_{xyc}依据标注的中心点生成的高斯分布, $\tilde{p}_x$、$\tilde{p}_y$表示$\tilde{p}$在x、y方向上的分量, 此分支训练过程通过 focal loss[20] 定义损失函数:

$$L_k=\frac{-1}{N}\sum_{xyc}\begin{cases}\left(1-\hat{Y}_{xyc}\right)^{\alpha}\ln\left(\hat{Y}_{xyc}\right), Y_{xyc}=1\\ \left(1-Y_{xyc}\right)^{\beta}\left(\hat{Y}_{xyc}\right)^{\alpha}\ln\left(1-\hat{Y}_{xyc}\right), \text{其他}\end{cases} \tag{2}$$

其中$\hat{Y}_{xyc}$表示 HeatMap 分支相对应的Y_{xyc}的预测值, α、β是 focal loss 的参数, N代表中心点 (检测到的目标物体) 数量, 一般设置$\alpha=2$, $\beta=4$. 在 offset 分支, 为了恢复因下采样造成的误差, ground truth 设置为

$\left(\frac{p}{R}-\tilde{p}\right)$, 使用 $L1$ loss 进行回归:

$$L_{\text{off}}=\frac{1}{N}\sum_{p}\left|\hat{O}_{\tilde{p}}-\left(\frac{p}{R}-\tilde{p}\right)\right| \tag{3}$$

同样的在 scale 分支, 假设$\left(x_1^{(k)},y_1^{(k)},x_2^{(k)},y_2^{(k)}\right)$是目标 k的左上角的位置和右下角的位置, 其类别设定为c_k. 中心点设为$p_k=\left(\frac{x_1^{(k)}+x_2^{(k)}}{2},\frac{y_1^{(k)}+y_2^{(k)}}{2}\right)$, 设置关于大小的 ground truth $s_k=\left(x_2^{(k)}-x_1^{(k)},y_2^{(k)}-y_1^{(k)}\right)$用 $L1$ loss 去预测在这个中心点的尺寸:

$$L_{\text{size}}=\frac{1}{N}\sum_{k=1}^{N}\left|\hat{S}_{p_k}-s_k\right| \tag{4}$$

最后对以上 3 个分支的损失进行平衡:

$$L_{\text{det}}=L_k+\lambda_{\text{size}}L_{\text{size}}+\lambda_{\text{off}}L_{\text{off}} \tag{5}$$

式中, 一般在实验中设置$\lambda_{\text{size}}=0.1$, $\lambda_{\text{off}}=1$.

1.2 BiFPN 结构

在 Backbone 的研究中, 模型有效性是一个重要的概念, 在目标检测网络中都需要一个特征提取层提取深度特征, 一般的做法是对高层的语义特征和低层的细节特征进行融合, 即 FPN[21] 结构, 这样用在检测过程中可以提高位置检测和分类的精度, 但同时也会极大的增加参数量, 致使检测的速度降低, 所以越来越多的 FPN 结构尝试在尽可能少的增加参数量的同时能保证一定的检测精度, 以满足在嵌入式设备上的实时性要求. 其中 EffcientDet[22] 提出加权融合的 BiFPN 结构就可以有效的对下采样或上采样后的不同分辨率的特征图进行有效的融合, 其结构如图 2 所示. 图中假设 P_3、P_4、P_5、P_6、P_7为经过在初始输入图像上逐级下采样后得到的不同分辨率的特征图, 然后通过跳跃连接使双向网络层结构 (top-down 和 bottom-up) 提取的特征进行特征融合. 此外在融合的过程中, 为了区分不同特征层对最后输出特征的不同贡献和提高特征提取结构的有效性, 可以通过对不同层级的特征进行加权实现:

$$O=\sum_i\frac{w_i}{\varepsilon+\Sigma_j w_j}\cdot I_i \tag{6}$$

其中, I_i 表示融合过程中的所有被融合前的特征图, w_i为其对应的权值, 是一个可训练的参数, $\Sigma_j w_j$表示所有的权值之和, O表示融合后结果特征图的输出. 以 P_6和P_7融合的过程为例, 设输入的特征层为P_6^{in}和P_7^{in}, 第 6 层中间的特征设为P_6^{td}, 输出特征设为P_6^{out}, 同理第 6 层的输出特征为P_5^{out}, 计算第 6 层的输出如下:

$$P_6^{td}=Conv\left(\frac{w_1\cdot P_6^{\text{in}}+w_2\cdot Resize(P_7^{\text{in}})}{w_1+w_2+\varepsilon}\right) \tag{7}$$

$$P_6^{\text{out}}=Conv\left(\frac{w_{1'}\cdot P_6^{\text{in}}+w_{2'}\cdot P_6^{td}+w_{3'}Resize(P_5^{\text{out}})}{w_{1'}+w_{2'}+w_{3'}}\right) \tag{8}$$

经过以上计算就能更加有效的提取到低层细节特征和高层语义特征的混合特征, 用于头部网络的检测和分类任务.

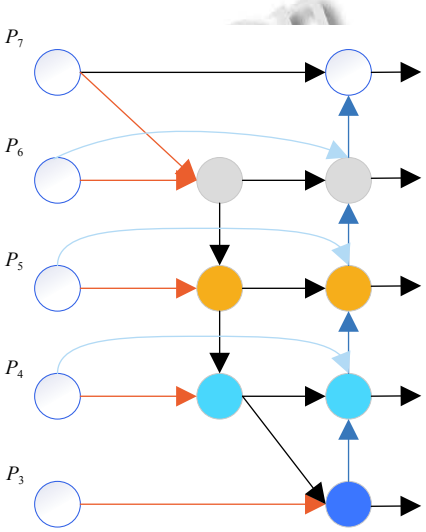


图 2 BiFPN 结构示意图

2 行人检测算法设计

2.1 行人检测模型

原 CenterNet 使用的 backbone 是多层特征融合的 DLA34 和 Hourglass101, 这类模型参数量大, 前向传播速度较慢, 不适合使用在嵌入式这类计算能力有限的设备上, 所以根据 BiFPN 结构提出一种新的特征提取结构, 其参数量在嵌入式设备上可以满足实时性的同时, 保证了精度不会出现大幅降低, 改进后的网络结构如图 3 所示, 从网络结构的图中可以看出输入图像首先经过一个 Conv1(卷积)→Bn1(批标准化)→ReLU(激活层)→maxpool(最大池化) 的结构得到一个 64 维的特征图, 然后使用 ResBlock(残差块) 进一步提取特征, 分别将残差块下采样输出的特征相对应的按式 (7), 式 (8) 进行分辨率调整, 并且按图中 BiFPN 结构进行融合, 经过 BiFPN 结构之后得到对应输入的不同层融合之后的特征, 将这些特征经过 Conv (卷积), DeConv[23](可变形卷积) 送入到头部分支, 最后再在不同的分支分别进行

卷积, 得到各个头部检测分支的对应结果, 综合 3 个分支的结果即可得到最终的检测结果.

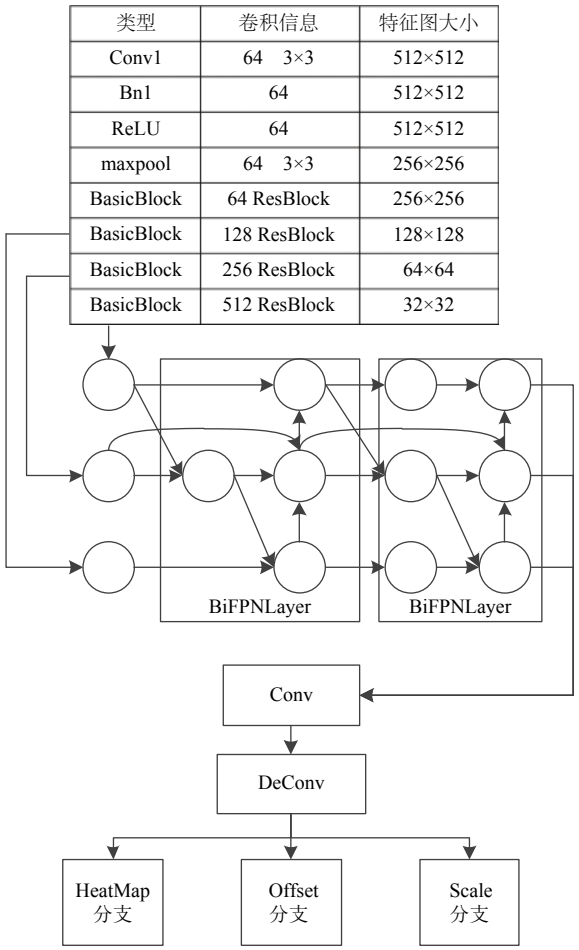


图 3 基于 CenterNet 改进的网络结构

2.2 检测网络训练及损失

本文所提方法训练时的输入量和原 CenterNet 网络的输入量相同, 不过因行人这一检测类别的特殊性进行微调, 并改进其相应的损失函数.

针对行人之间容易出现遮挡的情况, 如图 4 所示, 通过改进训练过程中 HeatMap 的高斯核分布形式, 来提高准确度, 即将 (1) 式修改为:

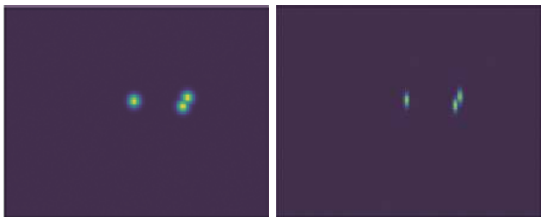
$$Y_{xyc}=\exp\left(-\frac{(x-\hat{p}_x)^2}{2\sigma_x^2}-\frac{(y-\tilde{p}_y)^2}{2\left(\frac{g_w}{g_h}\right)^2\sigma_x^2}\right) \tag{9}$$

其中, σ_x为原方差σ_p, g_w和g_h为标注框 ground truth 对应的宽和高, HeatMap 的变化如图 5 所示, 图 5(a) 中采用的是式 (1) 中的分布形式, 如果以这种形式表达行人中心点的分布, 在预测过程中当行人的距离较近时, 很容易在预测过程中导致两个响应峰值距离较近, 导致最后使用最大池化或 soft-NMS[24] 等过滤手段时将其过滤, 即响应更为强烈的预测中心点将另一个中心点“吞并”, 而如果按式 (5) 的形式绘制 HeatMap, 则会在响应图 5(b) 上产生一条明显的界限, 避免因行人相互遮挡或距离较近产生漏检. 相应的在 HeatMap 分支设置损失函数:

$$L_k=\frac{-1}{N}\sum_{xy}\begin{cases}\left(1-\hat{Y}_{xy}\right)^{\alpha}\ln\left(\hat{Y}_{xy}\right), Y_{xy}=1\\ \left(1-Y_y\right)^{\beta}\left(\hat{Y}_{xy}\right)^{\alpha}\ln\left(1-\hat{Y}_{xy}\right), \text{其他}\end{cases} \tag{10}$$

式 (10) 与 (2) 的不同之处在于, 首先因为应用的是针对于行人这一个单类别的检测, 所以 HeatMap 的结构没有类别对应的维度, 仅在x、y方向是有效的, 其次针对于 HeatMap 上$Y_{xy}\neq 1$的情况, 由于改进后高斯核在两个方向上的分布形式不同, 所以仅在y方向上对 focal loss 损失进行衰减, 这有利于检测行人时生成更加符合其长宽比的检测框, 提高模型的精度.

图 4 行人之间的相互遮挡



(a) 改进前的高斯核分布 (b) 改进后的高斯核分布

图 5 改进前后的高斯核分布形式

3 实验结果及分析

3.1 实验数据、环境及评价指标

实验过程中使用 CityPerson[25] 数据集首先进行

30 个 epoch 的预训练, 然后使用 CrowdHuman 行人数据集进行 130 个 epoch 训练和评测. CrowdHuman 数据集是密度较高的行人检测数据集, 平均每张图片有 22.64 个行人检测框, 在训练过程中使用 15 000 张训练集图像和 4370 张验证集图像进行训练, 使用 5000 张测试集图像进行评测.

实验训练过程中所用硬件环境 Inter Core i7 9400, GPU2080Ti, 操作系统为 Ubuntu 16.04, 训练深度学习框架为 MXNet 1.5.0, 最终应用的嵌入式平台为 Jetson TX2, 在模型移植过程中采用 TensorRT 加速, 训练和推理过程图片采用 512 的大小作为输入.

在评价指标上主要采用平均精度 (Average Precision, *AP*) 作为主要的评价依据, 其计算过程如式 (11):

$$AP = \int_0^1 P(R)\mathrm{d}R \tag{11}$$

其中, P 表示精确度, R 表示召回率. 精确度是指正确检测到的物体在所有目标中所占的比例, 而召回率是指正确检测到的物体在所有检测到的目标中所占的比例, 式 (12) 表示精确度的计算, 式 (13) 表示召回率的计算, 两式中 TP (IoU 不小于阈值) 为正确检测出目标, FP (IoU 小于阈值) 为错误检测目标, FN 为没有被检测出目标. 本文 IoU 阈值设置为 0.3, 当 IoU ≥0.3 时, 则认为检测正确, 否则为错误.

$$Precision = \frac{TP}{TP+FP} \tag{12}$$

$$Recall = \frac{TP}{TP+FN} \tag{13}$$

参数量和检测速度的评测采用固定大小图片 512×512 作为输入, 分别计算模型前向传播过程的权重参数的总量及传播时间来实现.

3.2 实验设置和结果

为了更好的对改进后的模型及输入进行评估, 分别设置了不同的对比实验, 首先针对模型结构的适应性进行评估, 对其 Backbone 分别使用 DLA34、ResNet34、MobileNet_v2[26] 及本文改进后的结构进行对比, 训练过程的损失曲线和精度曲线分别如图 6 和图 7 所示, 几个 Backbone 的参数量如表 1 所示, 由于 MobileNet_v2 的参数量较少, 所以很快训练权重就完成拟合, 但最后的损失也略大, 所以精度较低, 相比于 ResNet34, DLA34 融合了更多高级语义信息和低层细节信息, 所以 DLA34 在整个训练过程中以比 ResNet34 更少的参数量达到更优的效果. 最后从本文算法的损失曲线和精度曲线来看, 虽然相比于 DLA34、和 ResNet34 的精度略低, 但从表 1 可以看出其参数量相较于其他两个 Backbone 分别少了 34.4% 和 56.2%, 充分说明了本文所用方法提取特征的有效性.

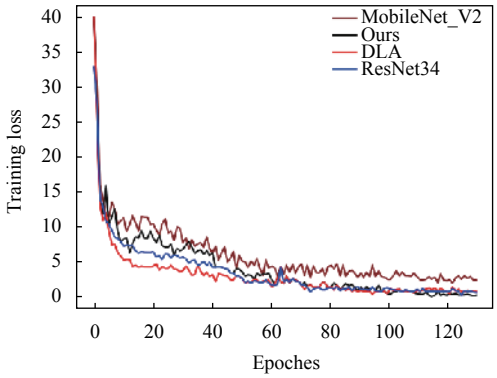


图 6 训练损失变化曲线

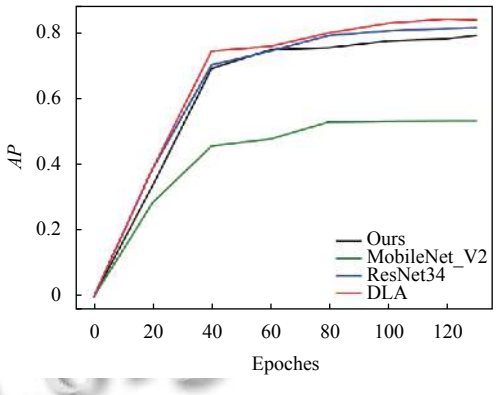


图 7 训练精度变化曲线

表 1 不同 Backbone 的参数量

Backbone	DLA34	ResNet34	MobileNet_v2	本文
Params	15 742 104	21 797 672	3504 960	**10 326 540**

为了体现对行人检测框训练时 HeatMap 的改进及相应的损失函数的修改带来的效果增益, 首先用未改进 HeatMap 输入的方法进行检测, 从中挑选出 128 处因遮挡导致漏检的结果, 如图 8(a) 所示, 然后使用改进后的训练方式及损失函数进行重新训练, 得到的结果如图 8(b), 改进前后在精度和因遮挡造成的评测如表 2 所示.

结合图 8 和表 2 可以看出, 经过对 HeatMap 训练的高斯核分布改进后, 不仅能够提高检测精度和减少因遮挡造成的漏检, 而且因为新的高斯核分布形式与

行人这一类别更加匹配, 所以也会提高新人检测的置信度.

(a) 改进前检测效果 (b) 改进后检测效果

图 8 高斯核改进前后检测效果

表 2 改进 HeatMap 上高斯核的分布方式对行人遮挡效果的提升效果

Period	*AP*	Occultation detection
改进前	0.773	0/128
改进后	0.786	34/128

最后为了准确的评估模型在 传播速度 (FPS) 来综合比较经 TensorRT 加速后在 Jetson TX2 上的表现, 其中在移植时, 权重参数全部量化为 float8, 结果如表 3 所示.

表 3 Jetson TX2 移植后的效果

Backbone	DLA34	ResNet34	MobileNet_v2	本文
Inference time (ms)	189	275	23	68
AP	0.820	0.801	0.514	**0.774**

从表 3 可以看出, 相比于其他几种 Backbone 在 Jetson TX2 上的表现, 本文所提方法精度仅略微降低, 但 68 ms 的推理速度足以保证模型在嵌入式平台 Jetson TX2 上的实时性.

4 结论与展望

本文主要是针对原 CenterNet 检测网络在嵌入式设备上检测速度较慢, 提出了一种满足实时要求又不大幅降低检测精度的网络模型. 然后针对于行人这一检测类别通过改进头部网络 HeatMap 分支的高斯核分布进一步降低因遮挡带来漏检的方法. 实验结果表明, 本文所提方法在嵌入式设备上与其他方法相比具有一定的优势, 在保证检测精度的同时, 通过有效的检测模型极大的减少了参数量并提高了检测速度, 同时在行人检测的相互遮挡问题上进行了研究. 如何进一步在现有部署框架上尽可能少的使精度下降, 是下一阶段的研究方向.

参考文献

1 张磊, 傅志中, 周岳平. 基于 HSV 颜色空间和 Vibe 算法的运动目标检测. 计算机工程与应用, 2014, (4): 181–185.

2 胡小冉, 孙涵. 一种新的基于 ViBe 的运动目标检测方法. 计算机科学, 2014, 41(2): 149–152. [doi: 10.3969/j.issn.1002-137X.2014.02.033]

3 Zhang GF, Chanson H. Application of local optical flow methods to high-velocity free-surface flows: Validation and application to stepped chutes. Experimental Thermal and Fluid Science, 2018, 90: 186–199. [doi: 10.1016/j.expthermflusci.2017.09.010]

4 潘光远. 光流场算法及其在视频目标检测中的应用研究 [硕士学位论文]. 上海: 上海交通大学, 2008.

5 Han F, Shan Y, Cekander R, *et al*. A two-stage approach to people and vehicle detection with HOG-based SVM. Performance Metrics for Intelligent Systems 2006 Workshop. Piscataway, NJ, USA. 2006. 133–140.

6 李文书, 韩洋, 阮梦慧, 等. 改进的基于增强型 HOG 的行人检测算法. 计算机系统应用, 2020, 29(10): 199–204. [doi: 10.15888/j.cnki.csa.007587]

7 Girshick R, Iandola F, Darrell T, *et al*. Deformable part models are convolutional neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 437–446.

8 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 779–788.

9 Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017: 6517–6525.

10 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv: 1804.02767, 2018.

11 Wang XL, Shrivastava A, Gupta A. A-fast-RCNN: Hard positive generation via adversary for object detection. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 3039–3048.

12 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

13 Elhabian SY, El-Sayed KM, Ahmed SH. Moving object detection in spatial domain using background removal

techniques-state-of-art. Recent Patents on Computer Science, 2008, 1(1): 32–54. [doi: 10.2174/2213275910801010032]

14 He KM, Gkioxari G, Dollár P, *et al*. Mask R-CNN. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy. 2017. 2980–2988.

15 Law H, Deng J. CornerNet: Detecting objects as paired keypoints. Proceedings of the 15th European Conference on Computer Vision. Munich, Germany. 2018. 765–781.

16 Duan KW, Bai S, Xie LX, *et al*. CenterNet: Keypoint triplets for object detection. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea. 2019. 6568–6577.

17 Tian Z, Shen CH, Chen H, *et al*. FCOS: Fully convolutional one-stage object detection. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea. 2019. 9626–9635.

18 Newell A, Yang KY, Deng J. Stacked hourglass networks for human pose estimation. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands. 2016. 483–499.

19 Yu F, Wang DQ, Shelhamer E, *et al*. Deep layer aggregation. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 2403–2412.

20 Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice. 2017. 2999–3007.

21 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 936–944.

22 Tan MX, Pang RM, Le QV. Efficientdet: Scalable and efficient object detection. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA. 2020. 10778–10787.

23 Zhu XZ, Hu H, Lin S, *et al*. Deformable convnets v2: More deformable, better results. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA. 2019. 9300–9308.

24 Bodla N, Singh B, Chellappa R, *et al*. Soft-NMS—improving object detection with one line of code. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy. 2017: 5562–5570.

25 Zhang SS, Benenson R, Schiele B. Citypersons: A diverse dataset for pedestrian detection. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 4457–4465.

26 Sandler M, Howard A, Zhu ML, *et al*. Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 4510–4520.