智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于改进FCOS的拥挤行人检测算法

齐鹏宇， 王洪元， 张继， 朱繁， 徐志晨

引用本文:
齐鹏宇, 王洪元, 张继, 等. 基于改进FCOS的拥挤行人检测算法[J]. 智能系统学报, 2021, 16(4): 811–818.
QI Pengyu, WANG Hongyuan, ZHANG Ji, et al. Crowded pedestrian detection algorithm based on improved FCOS[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(4): 811–818.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202010012

DOI: 10.11992/tis.202010012
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20210402.1043.004.html

基于改进 FCOS 的拥挤行人检测算法

齐鹏宇，王洪元，张继，朱繁，徐志晨
（常州大学 信息科学与工程学院，江苏 常州 213164）

摘　要：针对大规模拥挤场景视频中行人目标小、行人遮挡和行人交叠而导致的检测困难等问题，本文将逐像素预测目标检测框架——全卷积单阶段目标检测 FCOS（fully convolutional one-stage object detection）应用于行人检测，提出一种改进的主干网络用于提取行人特征，通过增加尺度回归实现目标行人的多尺度检测，同时减少其他特征层检测的目标数量，进而提升行人检测的能力。在拥挤行人场景数据集 CrowdHuman 和小目标行人数据集 Caltech 上的大量实验结果表明，和目前先进的方法相比，本文的方法对行人的检测精度有所提升，特别是对于小目标行人检测。与原始 FCOS 算法相比，在 CrowdHuman 上平均精度提升接近 15%，丢失率降低接近 33.0%；在 Caltech 上的平均精度提升 2%。在复杂拥挤场景下的实际应用也证明本文方法的有效性。

关键词：行人检测；多尺度检测；全卷积单阶段目标检测；拥挤行人场景；训练策略；小目标检测；尺度回归；逐像素预测
中图分类号：TP391.41　　**文献标志码**：A　　**文章编号**：1673−4785(2021)04−0811−08

中文引用格式：齐鹏宇，王洪元，张继，等. 基于改进 FCOS 的拥挤行人检测算法 [J]. 智能系统学报，2021，16(4)：811–818.
英文引用格式：QI Pengyu，WANG Hongyuan，ZHANG Ji，et al. Crowded pedestrian detection algorithm based on improved FCOS[J]. CAAI transactions on intelligent systems，2021，16(4)：811–818.

Crowded pedestrian detection algorithm based on improved FCOS

QI Pengyu，WANG Hongyuan，ZHANG Ji，ZHU Fan，XU Zhichen
(School of Information Science and Engineering，Changzhou University，Changzhou 213164，China)

Abstract: In view of the detection difficulty resulting from small pedestrian objects，pedestrian occlusion，and pedestrian overlap in large-scale crowded scene videos，this study applies a pixel-by-pixel prediction object detection framework，i.e.，fully convolutional one-stage object detection (FCOS)，for pedestrian detection. An improved backbone network is proposed to extract pedestrian features，achieve multi-scale detection of object pedestrians by increasing scale regression，reduce the number of objects detected by other feature layers，and thereby improve the ability of pedestrian detection. Several experiments have been performed on the crowded pedestrian scene dataset CrowdHuman and the small object pedestrian dataset Caltech. The results show that compared with current advanced methods，the proposed algorithm makes some improvements in the pedestrian detection accuracy，especially for small object pedestrian detection. Compared with the original FCOS framework，the average precision on CrowdHuman is increased by nearly 15% and the miss rate is decreased by nearly 33.0%. The average precision on Caltech is increased by 2%. Moreover，the actual use in complex，crowded scenarios proves the effectiveness of this algorithm.

Keywords: pedestrian detection; multi-scale detection; fully convolutional one-stage object detection; crowded pedestrian scene; training strategy; small object detection; scale regression; pixel by pixel prediction

行人检测属于计算机视觉领域一个重要的基础研究课题，对于行人重识别、自动驾驶、视频监控、机器人等领域有重要的意义[1-3]。而行人检测领域在实际场景下面临着行人交叠、遮挡等问题，此类问题依然困扰很多研究者，也是目前行人检测面临的巨大挑战。

在现有的目标检测算法[4]中，两阶段目标检测器（如 Faster R-CNN[5]、R-FCN[6]、Mask R-CNN[7]、RetinaNet[8]、Cascade R-CNN[9]）精度高但速度稍慢，单阶段目标检测器（如 YOLOv2[10]、

收稿日期：2020−10−14.　**网络出版日期**：2021−04−02.
基金项目：国家自然科学基金项目（61976028，61572085，61806026，61502058）；江苏省自然科学基金项目（BK20180956）.
通信作者：王洪元. E-mail: hywang@cczu.edu.cn.

SSD[11]) 速度快但精度稍低。Zhi 等[12] 认为锚框 (anchor) 的纵横比和数量对检测性能影响较大，在需要预设候选框的检测算法中，这些 anchor 相关参数需要进行精准的调整。而在多数的两阶段算法中，由于 anchor 的纵横比不变，模型检测 anchor 变化较大的候选目标时会遇到麻烦，特别是对于小目标的物体。多数检测模型需要在不同的检测任务场景下重新定义不同的目标尺寸的 anchor，这是因为模型预定义的 anchor 对模型性能影响较大。在训练过程中，大多数的 anchor 被标记为负样本，而负样本的数量过多会加剧训练中正样本与负样本之间的不平衡。基于无预设候选框 (anchor-free) 的检测算法容易造成极大的正负样本之间不平衡，检测的精度也不如 anchor-base 算法。而近年来的全卷积网络 (fully convolutional network, FCN[13]) 在众多计算机视觉的密集预测任务中取得了好的效果，例如语义分割、深度估计[14]、关键点检测[15] 和人群计数[16] 等。由于预设候选框的使用，两阶段检测算法取得了好的效果，这也间接导致了检测任务中没有采用全卷积逐像素预测的算法框架。而 FCOS[12] 首次证明，基于 FCN 的检测算法的检测性能比基于预设候选框的检测算法更好。FCOS 结合 two-stage 和 one-stage 算法的一些特点逐像素检测目标，实现了在提高检测精度的同时，加快了检测速度。

由于拥挤场景下行人目标会出现交叠、遮挡和行人目标偏小等问题，本文提出新的特征提取网络提取更具判别性行人特征。对于 FCOS 检测算法，行人检测中行人尺度问题对模型性能的影响较大，针对该问题，本文改进多尺度预测用于检测小目标行人，有效地解决了行人目标偏小、拥挤等场景下行人检测精度不高的问题。

1　相关工作

1.1　FCOS 框架

FCOS 首先以逐像素预测的方式对目标进行检测，无需设置 anchor 的纵横比，然后利用多级预测来提高召回率并解决训练中重叠预测框导致的歧义，这种方法可以有效提高拥挤场景下行人检测精度，缓解行人拥挤而导致的检测困难的问题。实际上，诸如 Unitbox[17] 之类基于 DenseBox[18] 的 anchor-free 检测算法，难以处理重叠的预测框而导致召回率低的问题，该系列的检测算法不适合用于一般物体检测，FCOS 的出现打破这一局面。FCOS 表明，使用多级特征金字塔网络 (feature pyramid networks, FPN[19]) 预测可以提高召回率，提高检测精度。

FCOS 在训练中损失定义如下：

$$\text{Loss} = \frac{1}{N_{\text{pos}}}\sum_{x,y} L_{\text{cls}}(p_{x,y}, c^*_{x,y}) + \frac{1}{N_{\text{pos}}}\sum_{x,y} I_{\{c^*_{x,y}>0\}} L_{\text{reg}}(t_{x,y}, t^*_{x,y}) \tag{1}$$

式中：x、y 表示特征图上的某一位置；$p_{x,y}$ 表示预测分类分数；$c^*_{x,y}$ 表示真实分类标签；$t_{x,y}$ 表示回归预测目标位置；$t^*_{x,y}$ 表示真实目标位置，L_{cls} 是 Focal Loss 分类损失，L_{reg} 是 IOU Loss 回归损失，并且在预先的实验中发现，拥挤行人检测任务中，IOU Loss 效果要稍优于 GIOU Loss[20]。N_{pos} 表示正样本的个数，$I_{\{c^*_{x,y}>0\}}$ 表示激活函数，当 $c^*_{x,y}>0$ 时为 1，否则为 0。

此外，FCOS 还具有独特的中心度分支预测，可以抑制低质量框的比例。由于逐像素预测，很多像素点虽然处于真值框内，但是越接近真值框中心的像素点预测出高质量预测框的概率也越大，因此提出预测中心度损失函数，如式 (2) 所示：

$$\text{centerness}^* = \sqrt{\frac{\min(l^*, r^*)}{\max(l^*, r^*)} \times \frac{\min(t^*, b^*)}{\max(t^*, b^*)}} \tag{2}$$

式中：l^*、r^*、t^*、b^*分别表示当前像素点到真值框边界的距离，这里使用开方来减缓中心损失的衰减。中心损失值在范围 [0,1]，因此使用二值交叉熵 (BCE) 损失进行训练，将中心度损失加到训练损失函数式 (1) 中。当回归中心在样本中心时，中心度损失会尽可能的接近 1，而当偏离时，中心度损失会降低。测试时，通过将预测框的中心损失与相应的分类分数相乘来计算最终分数，且该分数用于对检测到的预测框质量进行排序。因此，中心度可以降低远离目标中心的预测框的分数，再通过最终的非极大值抑制 (non-maximum suppression,NMS) 过程可以过滤掉这些低质量的预测框，从而显著提高行人检测性能。相比基于预设候选框的一类检测算法，FCOS 算法实现更好的检测性能。

1.2　原始 FCOS 特征提取网络

如图 1 所示，FCOS 算法的特征提取网络采用主干网络 (Backbone) 加上 FPN，Backbone 选用 ResNet[21] 提取特征，在 FPN 中，P_3、P_4、P_5 分别由 C_3、C_4、C_5 横向连接产生，P_6、P_7 由 P_5、P_6 通过步长为 2 的卷积产生。每层检测不同尺度大小的目标，P_i 层检测当前像素点处满足条件的目标，目标公式定义如下：

$$\max(l^*, r^*, t^*, b^*) \in [m_{i-1}, m_i] \tag{3}$$

式中：l^*、r^*、t^*、b^*分别表示当前像素点到真值框边界的距离；$[m_{i-1}, m_i]$ 表示 P_i 层回归目标范围，m_2、m_3、m_4、m_5、m_6 和 m_7 分别设置为 0、64、128、256、

512 和 ∞, 其中 ∞ 表示无穷大。这是一个非常有创造性的想法, 这样的设计使得 FCOS 检测算法是一个多尺度的 FPN 检测算法。

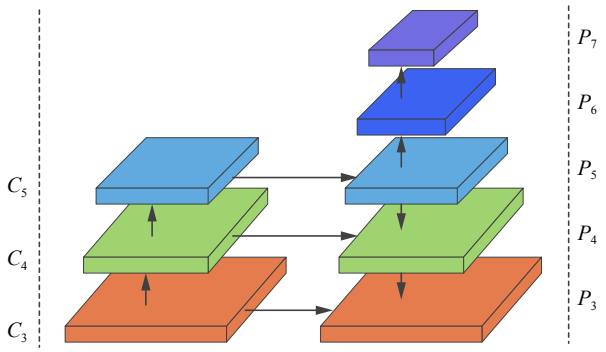


图 1　FCOS 特征提取网络

Fig. 1　FCOS feature extraction network

2　基于 FCOS 的行人检测

2.1　主干网络 VoVNet

深度学习中, 特征提取网络对于模型有着非常大的影响, 针对不同的数据集可以直接影响其检测性能。针对 ResNet 不足, 本文运用 VoVNet 作为行人特征的提取网络。

DenseNet[22] 在目标检测任务上展示出了较好的效果, 特别是基于 anchor-free 的目标检测模型, 这是因为相比于 ResNet, DenseNet 通过特征不断叠加达到好的效果, 其缺点是在后续特征叠加时, 通道数线性增加, 参数也越来越多, 模型花费时间增加, 影响模型速度。

VoVNet 认为在特征提取方面, 中间层的聚集强度与最终层的聚集强度之间存在负相关, 并且密集连接是冗余的, 即靠前层的特征表示能力越强, 靠后层的特征表示能力则会被弱化。VoVNet[23] 针对 DenseNet 做出改进, 提出一种新的模块, 即一次性聚合 (one-shot aggregation, OSA) 模块。OSA 模块将当前层的特征聚合至最后一层, 每一卷积层有两种连接方式, 一种方式是连接至下一层, 用于产生更大感受野的特征, 另一种方式是连接一次至最终输出的特征图上, 与 DenseNet 不同, 每一层的输出不会连接至后续的中间层, 这样的设计使得中间层的通道数保持不变。VoVNet 采用更加优化的特征连接方式, 通过增强特征的表示能力, 提高特征的提取能力, 进而提高模型的检测性能。

2.2　SE 模块

本文为了更好地契合复杂的行人特征, 在 VoVNet 上使用 SE 模块[24] 加强特征表示能力, 并且在特征图上使用 SE 模块进行权重分配, 使得深度特征更加多样化。

SE 模块首先依照空间维度来进行特征压缩, 将每个二维的特征通道变成一个实数, 输出一个二维空间, 它的维度与特征通道数相等, 即二维空间表示对应特征通道上的分布结果。之后生成一个具有权重的二维空间, 表示特征通道间的相关性。最后将对应的特征图乘上权重特征, 实现一个特征的权重分配, 突出重要的特征, 完成在通道维度上对原始特征通道上重要性的重标定。

SE 模块类似于注意力机制, 本文将其使用在 VoVNet 上, 如图 2 所示, 在每层特征下采样时, 将特征进行 SE 权重分配。根据 VoVNet 的特征连接方式添加 SE 模块权重机制, 本文方法可以提供更加多元化的特征, 使得行人特征更好地表达, 提高行人检测的精度。并且 SE 模块可以在几乎不增加模型时间复杂度的情况下提升模型的检测性能。

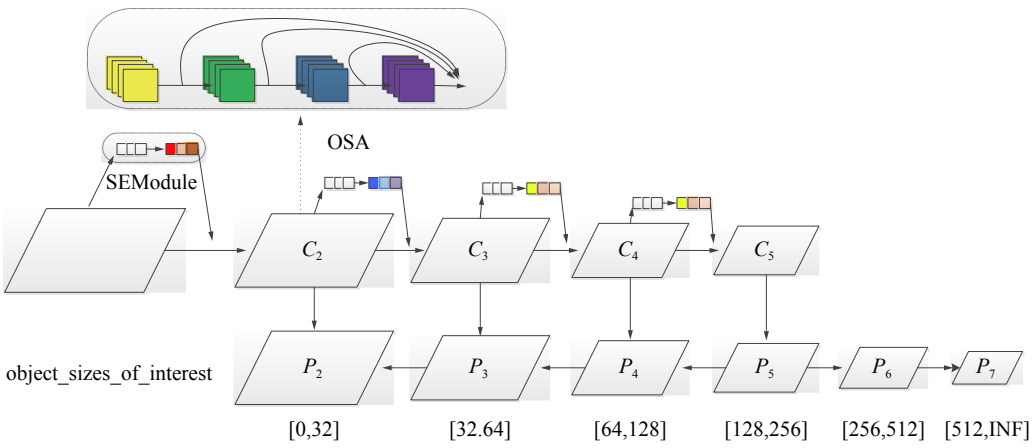


图 2　修改后框架

Fig. 2　Update framework

2.3　多尺度检测

原始模型 FPN 采用 5 层不同尺度回归目标，这 5 层尺度回归的目标大小分别为 [0,64]、[64,128]、[128,256]、[256,512] 和 [512, ∞]，分别对应 FPN 中的 P_3、P_4、P_5、P_6 和 P_7。针对行人目标的特点，本文发现，不论是在常用的行人数据集中，还是在真实检测场景中，行人检测的难点在于拥挤行人和小目标行人的检测。对于 FCOS 模型，每层每个像素点都会回归固定尺度大小范围内的目标。相对地，如果目标行人拥挤在某个尺度范围内，将会使得检测层的任务过重，导致检测效果降低，此问题也是影响模型性能效果的原因之一，在多目标检测场景中会导致 FCOS 模型的检测性能稍有降低，同时也说明，当检测任务复杂，检测目标数量较多时，本文提出的多尺度检测会使 FCOS 检测性能提高。

如图 2 所示，减小 P_3 层的回归尺度，设置 P_3 层回归尺度为 [32,64]，减少 P_3 层的检测任务量；增加 P_2 层，P_2 层由 C_2 层横向连接和 P_3 层向下连接组成，P_2 层回归尺度为 [0,32] 的目标，这样的网络设计既能减少 P_3 层的回归目标数，也能更好地利用特征检测小目标行人，提高行人检测精度。在最终的 FPN 上，本文的方法在 FPN 上拥有 6 层特征图以检测 6 个不同尺度范围的目标。

总体网络框架如图 3 所示，相较于未改进 FCOS 算法，预测特征图由 5 个增加到 6 个，而后对特征图上每个点进行逐像素预测，每个点均需预测目标回归框、目标类别、目标中心度，以上 3 种预测结果对应图 3 中 3 个预测分支，假设当前特征图大小为 $W\times H$，则有 $W\times H$ 像素点需要进行预测。

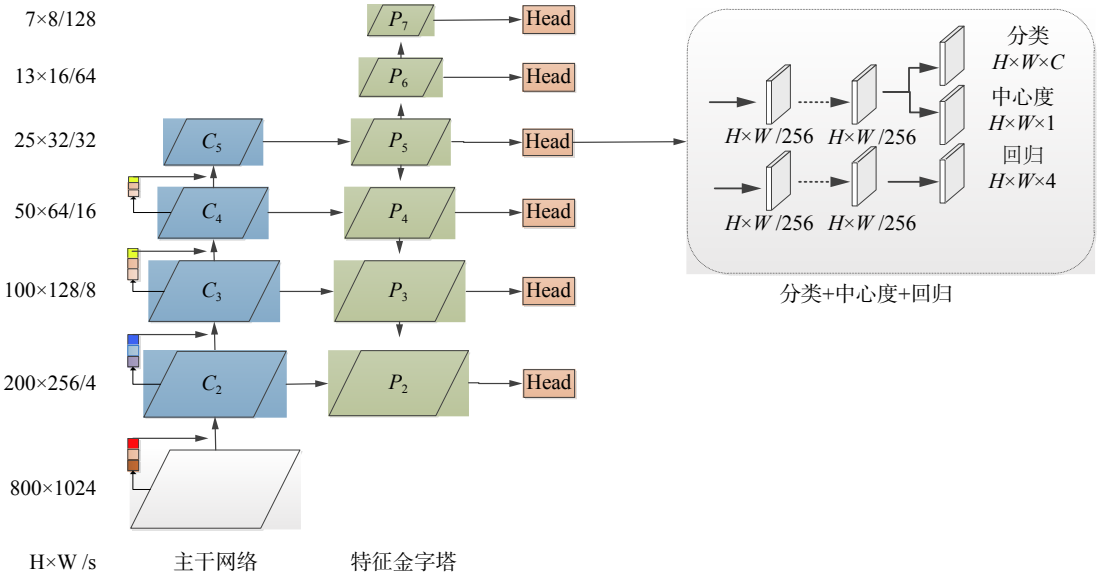


图 3　总体框架

Fig. 3　Final framework

3　数据集和评估

本文实验主要使用 CrowdHuman[25] 和 Caltech 行人数据集。行人数量多、场景拥挤是行人检测中一个巨大的挑战，针对这一问题，旷视发布 CrowdHuman 数据集，用于验证检测算法在密集人群行人检测任务中的性能。CrowdHuman 数据集中 15 000、4 370 和 5 000 个图片，分别用于训练、验证和测试。针对 CrowdHuman 数据集，本文只使用全身区域标注用于训练和评估，由于还未公布测试集，参考相关文献 [25-26] 后，实验结果在验证集上进行测试。Caltech 行人数据集时长约为 10 h 城市道路环境拍摄视频，数据集中随机分配训练集、测试集、验证集，其对应比例为 0.75∶0.2∶0.05，3 个集相互独立，测试集图片约为 24 438 张。

本文采用 MR^{-2}(miss rate) 和 AP 的评估准则，MR^{-2} 表示在 9 个 FPPI(false positive per image) 值下 (在值域 [0.01,1.0] 以对数空间均匀间隔) 的平均丢失率值，FPPI 定义如下：

$$\mathrm{FPPI}=\frac{\mathrm{FP}}{N} \tag{4}$$

式中：N 表示图片的数量；FP 表示未击中任意一个真值框的预测框数量。MR^{-2} 是目前衡量行人检测一个非常重要的指标，也是本文主要采用的评价指标。其数值越低说明行人检测模型性能越好。

AP 表示平均精度，PR(Precision-Recall) 曲线所围成的面积即为 AP 值大小，AP 值越大检测精度越高，其中 AP、Recall、Precision 计算公式如下：

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{6}$$

$$\text{AP} = \int_0^1 P(R)\,\mathrm{d}R \tag{7}$$

式中: TP 是检测出正样本的概率; FN 是正样本检测出错误样本的概率; FP 是负样本检测出正样本的概率。

4　实验

本文实验环境为 Ubuntu18.04、Cuda10 和 Cudnn7.6, 使用 4 块 2080Ti 的 GPU, 每个 GPU 有 11G 内存, 由于 FCOS 算法要求较高, 存在内存不够的问题, 实验通过线性策略[27]调整了 batch_size 大小和 IMS_PER_BATCH 的数量。其余参数沿用 FCOS 在 COCO 数据集上基础参数配置, 算法基于 detectron 框架。

4.1　CrowdHuman 数据集实验结果

如表 1 消融实验所示, 其中 6stage 表示多尺度检测方法, SE 表示 SE 模块。在 FCOS 上采用 VoVNet 作为 Backbone 起到了极大的提升作用, 相较于主干网络为 ResNet, AP_{50} 提升 11.2%。在 FPN 中多添加一个尺度的回归层, 对于行人检测的效果有极大的提升, 这是因为密集的行人检测受尺度变化影响较大。相较于原始 FCOS 方法, 本文方法在指标 AP_{50} 上提升了 15.0%。针对于不同主干网络, SE 模块在指标 AP_{50} 上有 0.2%~0.3% 的提升, 说明 SE 模块能增强行人特征提取能力。模型由 5 个尺度增加到 6 个尺度, 指标 AP_{50} 提升 3.5%, 并且对于模型检测小目标行人有着极大的提升, 可以看到指标 AP_S 提升 8.5%, 实验结果也印证多尺度改进能有效地提升模型检测小目标行人的性能。

表 1　CrowdHuman 数据集 AP
Table 1　AP on CrowdHuman

方法	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN[5]	36.7	68.3	35.2	23.4	37.2	40.4
FCOS+ResNet50	40.1	70.0	40.3	16.3	39.1	53.6
FCOS+VoVNet39	53.6	81.2	58.7	25.5	52.3	66.9
FCOS+ VoVNet39+SE	53.6	81.4	58.8	25.2	52.4	67.0
FCOS+ VoVNet39+6stage	57.7	84.7	64.0	34.0	55.0	70.1
FCOS+ VoVNet39+6stage+SE	**58.3**	**85.1**	**64.7**	**34.5**	**55.8**	**70.5**

CrowdHuman[25] 数据集中采用指标 MR^{-2}, 本文采用相同指标并对比了 CrowdHuman[25] 中部分实验, 表 2 可以看到, 在 CrowdHuman 数据集上, 通过消融实验表明: 采用 VoVNet 相较于采用 ResNet, 指标 MR^{-2} 降低 26.91%。拥有 SE 模块的检测模型相较于没有 SE 模块的检测模型, 指标 MR^{-2} 降低 0.9%。改进多尺度回归后的检测模型相较于未改进的检测模型, 指标 MR^{-2} 降低 6%。本文提出的方法相较于原始方法, 指标 MR^{-2} 降低了 33.62%。实验结果证明, 本文的方法在拥挤场景下的行人检测效果提升较为明显。

表 2　CrowdHuman 数据集 MR^{-2}
Table 2　MR^{-2} on CrowdHuman

方法	MR^{-2}	AP_{50}
RetinaNet[8]	63.33	80.83
FPN[25]	50.42	84.95
RFB Net[26]	65.22	78.33
FCOS+ ResNet50	83.62	70.0
FCOS+ VoVNet39	56.71	81.2
FCOS+ VoVNet39+SE	56.09	81.4
FCOS+ VoVNet39+6stage	50.90	84.7
FCOS+ VoVNet39+6stage+SE	**50.02**	**85.1**

如表 3 所示, 针对 CrowdHuman 数据集, NMS 的 IOU 阈值设定也是不同的, 原始 FCOS 算法在 COCO 数据集上 IOU 阈值设置为 0.7, 而针对拥挤行人场景, 本文发现 IOU 阈值设置为 0.5 时, 模型整体性能较好。图 4(a) 表示 PR 曲线图, 图 4(b) 表示 MR-FPPI 曲线, 可以清晰地看到本文方法总体上提升较大。在采用了 VoVNet 后, 对模型性能有了极大的提升, 说明 VoVNet 更加适合于 FCOS 在拥挤场景下提取行人特征。多尺度检测方法在拥挤场景下的行人检测也是有效的, 提升效果明显。

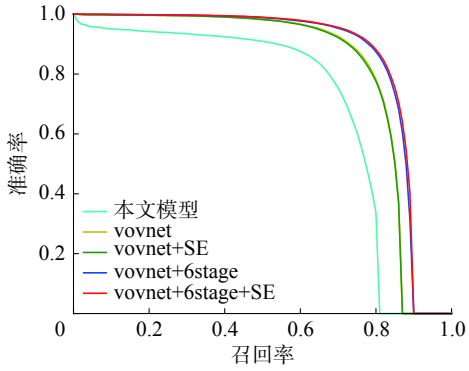
表 3　CrowdHuman 数据集 IOU 阈值
Table 3　IOU threshold on CrowdHuma

IOU	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
0.3	55.2	81.0	61.6	33.8	53.6	65.7
0.4	57.2	83.9	63.9	34.2	55.0	68.8
0.5	58.3	**85.1**	64.7	**34.5**	**55.8**	70.5
0.6	**58.4**	84.7	65.4	34.4	55.7	71.0
0.7	58.1	83.7	**66.0**	34.0	55.6	71.1
0.8	57.5	81.5	65.7	33.0	54.9	**71.2**
0.9	54.8	76.2	62.9	29.8	52.1	69.8

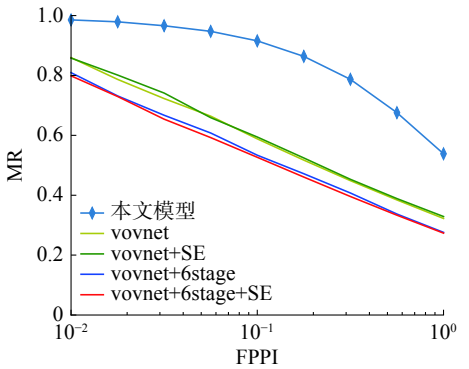
4.2　Caltech 数据集结果

如表 4 所示, 在车载摄像头的行人数据集

Caltech 上本文提出的方法也有一定提升，相较于原始 YOLOv2 方法，AP 实现了 2% 的提升。在 Caltech 数据集上的提升，说明本文模型的鲁棒性较好。



(a) PR 曲线图



(b) MR-FPPI 曲线图

图 4　CrowdHuman 曲线图

Fig. 4　CrowdHuman curves

表 4　Caltech 行人数据集

Table 4　Caltech pedestrian datasets

方法	AP
YOLOv2[10]	88.32
FCOS+ResNet50	89.36
FCOS+VoVNet39	90.20
FCOS+VoVNet39+6stage	90.35
本文方法	**90.39**

4.3　实际场景检测结果

本文的模型使用 CrowdHuman 训练集进行训练，在实际场景下的检测也有不错的效果，本文挑选出实际场景下一张室内行人和一张室外行人进行检测。因为本文算法无需设置 anchor 的尺寸和纵横比，所以在实际场景中的行人检测鲁棒性较好。如图 5 所示，图 5(a)、(c) 表示原始 FCOS 方法在拥挤行人中的效果，图 5(b)、图 5(d)表示本文方法的最终效果，可以看到，原始 FCOS可以较好地检测出图片中的行人，漏检率较低，但是仍存在伪正例，相比于图 5(b)，可以看到图 5(a) 右上角小目标行人未检测出来，远处的行人检测效果也不如图 5(b) 的检测效果,而相比于图 5(d)，可以看到图 5(b) 右边出现置信度为 0.64 的错误预测框。本文提出的方法可以较好地检测行人，减少 FP 出现的情况，在实际拥挤场景下能较好地检测目标行人。但当行人目标交叠时，或者对于有遮挡的行人，检测的效果大部分仅能检测出可视的部分，无法将全身区域标注出来，导致与真值框交并比的值较低，被视为负类。这也是目前本文方法面临的主要问题之一。

(a) 原始 FCOS 方法检测效果图 1

(b) 改进 FCOS 方法检测效果图 1

(c) 原始 FCOS 方法检测效果图 2

(d)改进 FCOS 方法检测效果图 2

图 5　实际场景检测效果

Fig. 5　Actual scene detection effect

5　结束语

针对行人目标检测中行人拥挤、目标偏小等问题，本文提出一种基于 FCOS 框架的行人检测算法。通过融入新的 Backbone 并且在 FPN 中添加一层 P_2 层，实现行人目标的多尺度检测。通过融入 SE 模块进行特征的权重分配，更好地提取行人特征，提高行人检测精度。本模型方法无需设置 anchor 纵横比等参数，参数设置少。相较于目前先进方法，可以达到有较强竞争力的检测效果。在实验中也发现，本文提出的方法受行人深度特征影响较大，如何在拥挤遮挡等实际场景下进行更高精度行人检测是我们进一步要研究的内容。

参考文献：

[1] NI Tongguang, DING Zongyuan, CHEN Fuhua, et al. Relative distance metric leaning based on clustering centralization and projection vectors learning for person re-identification[J]. IEEE access, 2018, 6: 11405–11411.

[2] WANG Hongyuan, DING Zongyuan, ZHANG Ji, et al.

Person reidentification by semisupervised dictionary rectification learning with retraining module[J]. Journal of electronic imaging, 2018, 27(4): 043043.

[3] 戴臣超, 王洪元, 倪彤光, 等. 基于深度卷积生成对抗网络和拓展近邻重排序的行人重识别 [J]. 计算机研究与发展, 2019, 56(8): 1632–1641.
DAI Chenchao, WANG Hongyuan, NI Tongguang, et al. Person re-identification based on deep convolutional generative adversarial network and expanded neighbor reranking[J]. Journal of computer research and development, 2019, 56(8): 1632–1641.

[4] JIAO Licheng, ZHANG Fan, LIU Fang, et al. A survey of deep learning-based object detection[J]. IEEE access, 2019, 7: 128837–128868.

[5] 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.

[6] DAI Jifeng, LI Yi, HE Kaiming, et al. R-FCN: object detection via region-based fully convolutional networks[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 379−387.

[7] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2961−2969.

[8] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2980−2988.

[9] CAI Zhaowei, VASCONCELOS N. Cascade R-CNN: delving into high quality object detection[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6154−6162.

[10] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 7263−7271.

[11] LIU Wei, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 21−37.

[12] TIAN Zhi, SHEN Chunhua, CHEN Hao, et al. FCOS: fully convolutional one-stage object detection[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 9627−9636.

[13] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 4438−4446.

[14] LIU Fayao, SHEN Chunhua, LIN Guosheng, et al. Learning depth from single monocular images using deep convolutional neural fields[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 38(10): 2024–2039.

[15] CHEN Yu, SHEN Chunhua, WEI Xiushen, et al. Adversarial PoseNet: a structure-aware convolutional network for human pose estimation[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy, 2017: 1212−1221.

[16] BOOMINATHAN L, KRUTHIVENTI S S S, BABU R V. CrowdNet: a deep convolutional network for dense crowd counting[C]//Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam, The Netherlands, 2016: 640−644.

[17] YU Jiahui, JIANG Yuning, WANG Zhangyang, et al. UnitBox: an advanced object detection network[C]//Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam, The Netherlands, 2016: 516−520.

[18] HUANG Lichao, YANG Yi, DENG Yafeng, et al. DenseBox: unifying landmark localization with end to end object detection[EB/OL]. (2015−09-19)[2021−05-07] https://arxiv.org/abs/1509.04874.

[19] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2117−2125.

[20] REZATOFIGHI H, TSOI N, GWAK J Y, et al. Generalized intersection over union: a metric and a loss for bounding box regression[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 658−666.

[21] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770−778.

[22] HUANG Gao, LIU Zhuang, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 4700−4708.

[23] LEE Y, HWANG J W, LEE S, et al. An energy and GPU-computation efficient backbone network for real-time ob-

ject detection[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA, 2019: 752−760.

[24] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7132−7141.

[25] SHAO Shuai, ZHAO Zijian, LI Boxun, et al. CrowdHuman: a benchmark for detecting human in a crowd[EB/OL]. (2018−04-30)[2021−05-07] https://arxiv.org/pdf/1805.00123.pdf.

[26] LIU Songtao, HUANG Di, WANG Yunhong. Adaptive NMS: refining pedestrian detection in a crowd[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 6459−6468.

[27] GOYAL P, DOLLÁR P, GIRSHICK R, et al. Accurate, large minibatch SGD: training ImageNet in 1 hour [EB/OL]. (2018−04-30)[2021−05-07] https://arxiv.org/pdf/1706.02677.pdf.

作者简介：

齐鹏宇，硕士研究生，主要研究方向为计算机视觉和行人检测。

王洪元，教授，博士，主要研究方向为人工智能和模式识别。承担国家自然科学基金项目、省市科技研究基金项目等多项课题研究，发表学术论文百余篇。

张继，讲师，主要研究方向为计算机视觉和行人检测。

2021 中国“AI+”创新创业大赛——智能信息创新与应用大赛

人工智能技术已经深度融入信息生产和传播的各个环节，智能化也成为媒体未来的趋势和发展方向，智能信息发展需要更多技术和应用创新。由中国人工智能学会主办，新浪新闻承办的 2021 中国“AI+”创新创业大赛——智能信息创新与应用大赛诚挚邀请研究人员、产业从业人员、高校学生以及爱好者参赛，助力智能信息发展。本次大赛将采用线上初赛和答辩方式进行，最终取得名次的队伍将进入 2021 中国“AI+”创新创业大赛全国总决赛。

赛程安排：

报名截止日期：2021 年 8 月 20 日

初赛作品提交截止日期：2021 年 9 月 5 日

答辩名单公布日期：2021 年 9 月 10 日

答辩和颁奖时间：2021 年 9 月中旬

2021 中国“AI+”创新创业大赛全国总决赛：2021 年 10 月

竞赛秘书处联系方式：

报名网站：http://2021aichina.caai.cn/

联系邮箱：ai_media@vip.sina.com