文章编号:1005-0523(2021)02-0115-07

一种基于多尺度特征融合的人头计数检测方法研究

张 泓,范自柱,石林瑞,符进武

(华东交通大学理学院,江西 南昌 330013)

摘要:通过摄像机等设备在特定场景下的人群计数在智能安防领域具有重大的研究意义。由于场景中人尺度变化大,背景杂乱,人群遮挡严重,传统的方法在这种场景中往往精度不高。对此提出了一种基于改进的Faster-RCNN人头检测模型,可以对场景中的人进行准确计数。该模型使用ResNet101作为特征提取网络,使用多尺度特征融合模块将提取的特征融合后分层进行检测,这样做的目的是为了检测不同尺度的人。此外,通过设计先验框的尺寸和使用Roi-Align代替Roi-Pooling层来进一步提高检测效果。实验表明,该方法在Brainwash和HollwoodHeads数据集上达到了最优的结果,精度分别达到了95.3%和89.1%。

关键词:卷积神经网络;深度学习;人头计数;多尺度特征融合

中图分类号:TP391.4 **文献标志码**:A

本文引用格式:张泓,范自柱,石林瑞,等. 一种基于多尺度特征融合的人头计数检测方法研究[J]. 华东交通大学学报,2021,38(2):115-121.

DOI:10.16749/j.cnki.jecjtu.20210416.007

A Head Detection Method Based on Multi-Scale Feature Fusion

Zhang Hong, Fan Zizhu, Shi Linrui, Fu Jinwu

(School of Science, East China Jiaotong University, Nanchang 330013, China)

Abstract: It is of great significance in the field of intelligent security to count the people in a specific scene with cameras or other devices. Due to the huge scale variation, messy background, and severe occlusion, the traditional method cannot get high precision accordingly. This paper proposed a head detection method based on an improved Faster-RCNN to accurately count the people. In this model, ResNet101, as a feature extraction network, uses a multi-scale feature fusion module to fuse the extracted features and perform hierarchical detection. The purpose is to detect people of different scales. In addition, by designing the size of an anchor and using Roi-Align instead of Roi-Pooling layer, the detection effect is further improved. Experiments show that the method achieves better results on the two Brainwash and HollwoodHeads datasets, and the accuracy reaches 95.3% and 89.1% respectively.

Key words: convolutional neural network; deep learning; crowd counting; multi-scale feature fusion

Citation format: ZHANG H, FAN Z Z, SHI L R, et al. A head detection method based on multi-scale feature fusion[J]. Journal of East China Jiaotong University, 2021, 38(2): 115-121.

收稿日期:2020-12-29

基金项目:国家自然科学基金项目(61991401,61673097,61702117);江西省自然科学基金重点项目(20192ACBL20010)

作者简介:张泓(1996—),男,硕士研究生,研究方向为图像处理和模式识别。E-mail:2571266306@qq.com。

通信作者:范自柱(1975—),男,教授,博士,研究方向为模式识别与机器学习。E-mail:zzfan3@163.com。

人群计数一直是计算机视觉领域的热门问题，有许多重要的应用，如公共安全管理，灾难管理，公共空间设计，情报收集及分析和嫌疑人搜索[1]，这些各式各样的应用促使研究人员去开发各种适用于不同环境的人群计数方法。目前的人群计数方法主要有两种：① 基于回归的方法；② 基于检测的方法。

在基于回归的方法中，研究人员[2-3]将人群整体当作一个对象来进行人群计数，Wang[2]提出了一个端到端的卷积神经网络模型来对密集场景中的人群进行计数，他们的模型直接输出密集人群图像中的人数。陆金刚[3]设计了一个基于多尺度多列卷积神经网络(multi-scale multi-column convolutional neural network, MSMCNN)，每一列网络使用不同大小的卷积核来提取不同尺度的特征以便感受到不同尺度的人头。但是这些方法都将人群整体作为研究对象，直接输出场景里的人数或与输入图像对应的密度图，无法精确的定位到场景中的每一个人。

在基于检测的方法中，研究人员[4-8]使用基于头部检测的方法来对场景中的人群进行计数。李欢[4]提出了一种基于SSD的行人头部检测方法，添加类别预测和位置预测两个旁支网络实现特征分离来改进SSD。张晓琪[5]提出了一种基于多特征协同的人头检测方法，相对于传统方法检测方法精度上更好。邢志祥[6]对经典的目标检测方法进行了实验，通过选取不同的特征提取网络和检测网络进行搭配，得到了一种最优的检测方法。Vora[7]设计了一个轻量级的人头检测网络，他们使用VGG16作为特征提取网络，然后将最后一层特征送入一个3*3大小的卷积核，最后分别使用两个1*1大小的卷积做分类和回归。Merad[8]提出了一种结合跟踪过程的头部检测方法，他们通过头部参考系统与摄像机参考系统之间的刚体变换来估计头部姿态。但是这些方法在高遮挡，尺度差异大的场景中精度很低。

本文提出了一种基于改进的Faster-RCNN [9]的头部检测方法，该方法可以有效的感知多尺度信息，具有较高的准确性和鲁棒性。

1 基于多尺度特征融合的Faster-RCNN网络

事实上，在一张静态图片中，靠近摄像机的人头往往较大，远离相机的人头较小，使用单一的特征图进行检测往往不能取得很好的效果。为了解决这个问题，本文提出了一个新的人头检测方法，该方法能有效的利用多尺度特征来检测密集场景中的人群[10]。网络结构如图1所示，该的方法基于Faster-RCNN，为了进一步提高精度，设计了一个多尺度特征融合模块(multi-scale feature fusion, MSFF)，该模块可以融合不同层的特征，加强特征图之间的相关性。此外，通过先验框的设计和使用Roi-Align层代替了原来网络中的Roi-Pooling层，提高了在IoU(intersection over union)等于0.7时的检测效果。为了进一步提高检测效果使用大分辨率的图片进行训练，在保证图片的宽高比不变的情况下，将输入的图片放大到1 024*768，送入到特征提取网络中提取特征，然后将得到的特征送入多尺度特征融合模块进行融合，将融合后的特征图送入到RPN(region proposal network)网络中提取候选框，这些不同大小的候选框会通过Roi-Align映射成相同大小的特征图，最后将这些特征图送入到后续的网络中进行分类和回归。

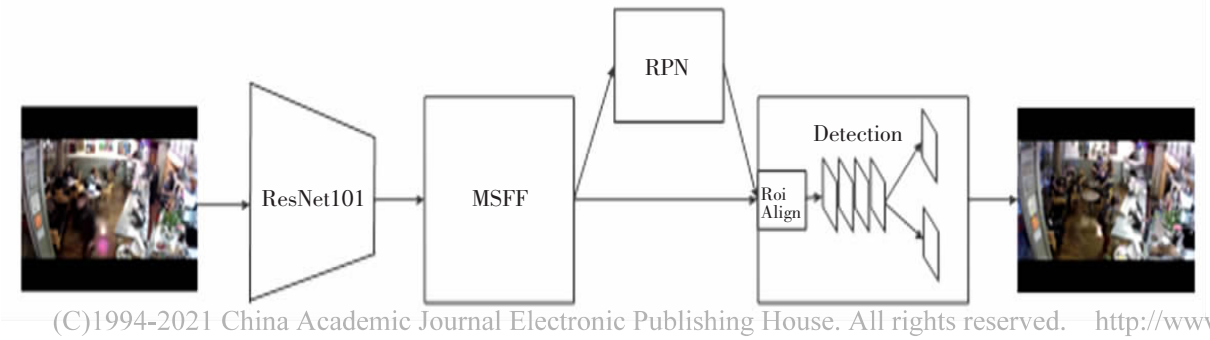

图1 基于多尺度特征融合的Faster-RCNN

Fig.1 Faster-RCNN based on multi-scale feature fusion

1.1　先验框的设计

通过对数据集中人头标注框的分析发现,人头的包围框可以近似看成一个正方形。去除了原 Faster-RCNN 网络中宽高比为 0.5 和 2 先验框,只保留了宽高比为 1 的先验框,即在每个特征图的每个像素点上只生成 1 种尺度的候选框。如图 2 所示,多尺度特征融合模块得到的 N2-N6 特征图分别对应的感受野(特征图上的像素点映射到原图所对应像素区域)的大小为{32^2, 64^2, 128^2, 256^2, 512^2}(32^2 表示 32*32 大小的像素区域)。

1.2　多尺度特征融合模块

在一个深度特征提取网络当中,深层的特征包含着丰富的语义信息但不利于小物体的检测,浅层的特征包含着丰富的位置信息但语义信息则较少。若只使用最后一层特征来进行检测,其它层的特征信息就会被忽略。Lin 等[11]设计了一种具有横向连接的自顶向下的特征金字塔网络结构(feature pyramid networks,FPN),该结构可以融合特征提取网络中不同尺度的特征来提升物体检测的准确率。我们参考了 Liu 等[12]在 FPN 结构上的改进,增加了一条自底向上路径,如图 2 所示,C1-C5 表示大小为 1 024*768 的输入图片经过 ResNet101 的第一到第五层卷积等得到的特征图,P5 由 C5 经过一个卷积核大小为 1*1 的卷积层得到,P4 由横向连接的 C4 和下采样的 C5 得到,P3-P2 以此类推,P2-P5 的通道数都为 512,特征图的大小分别为(256*192,128*96,64*48,32*24)。N2 由横向连接的 P2 得到,N3-N5 得到的方式与 P2-P3 类似,N6 由 N5 上采样得到。此外,为了降低上采样带来的混叠效应,在 N3-N5 之后分别加上了一个 3*3 的卷积层,但保持特征图的大小不变,N2-N6 的通道数都为 256,特征图的大小分别为(256*192,128*96,64*48,32*24)。这样做的目的是充分利用特征融合,使用五层不同大小的特征图分别进行预测,能够极大的提高网络对不同尺度人头的检测效果。

1.3　Roi-Align

此外,受 Mask-RCNN[13]的启发,将 Faster-RCNN 里的 Roi-Pooling 换成了 Roi-Align。为了使检测网络能够输入任意大小的图片,Roi-Pooling 被用来将感兴趣区域池化成固定大小的特征图,以便后续网络进行分类和回归的操作。但 Roi-Pooling 有一个局限性,Roi-Pooling 进行池化操作的时候会有两次

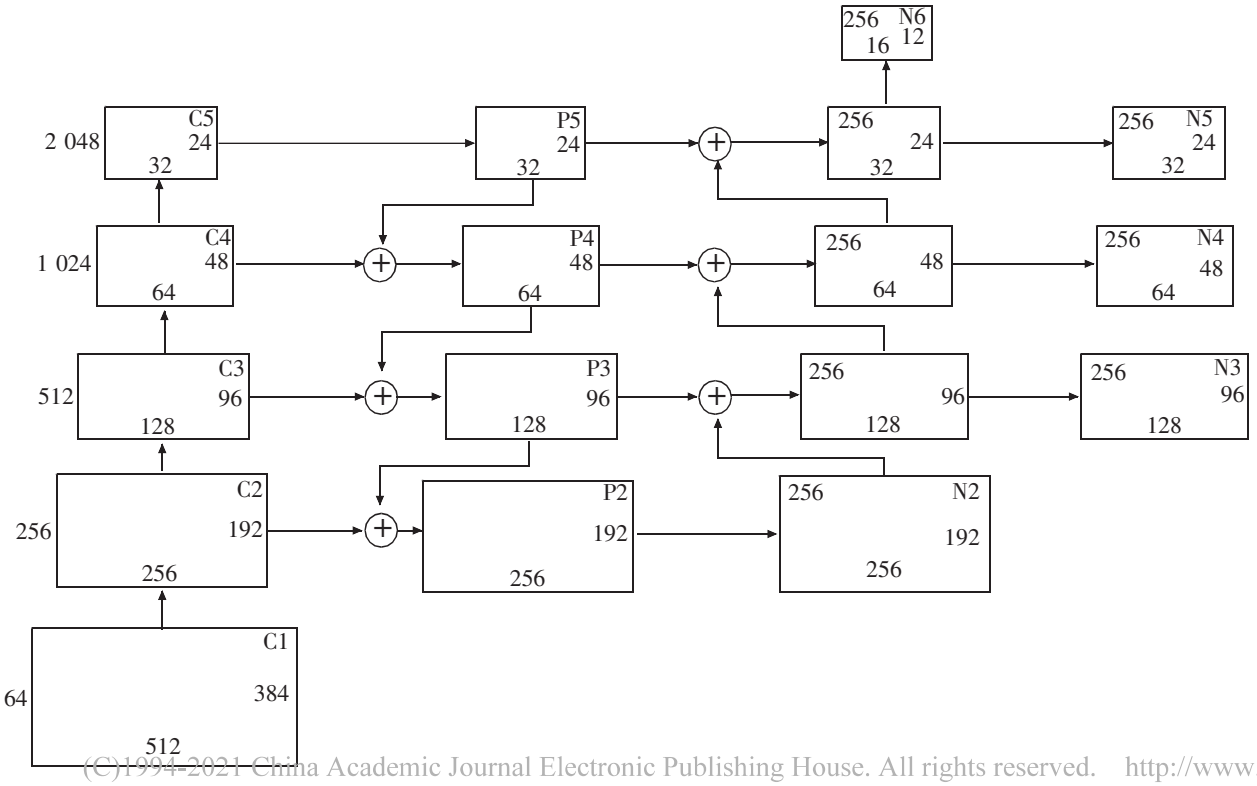

图 2　多尺度特征融合模块

Fig.2　Multi-scale feature fusion module

量化过程,即对浮点数进行向下取整。这样一来,经过 Roi-Pooling 得到的候选框和最开始候选框会有一些偏差,这在 IoU 较低的情况下影响不大,但 IoU 较高时精度会有很大的影响,特别是对于小物体。假设原图中有一个区域的大小为 655*655,经过 ResNet101 特征提取网络后在最后一层特征图的大小为 20.47*20.47,这个时候 Roi-Pooling 会进行向下取整操作,最后得到的特征图的大小为 20*20。将 20*20 大小的特征图映射成大小为 7*7 大小的特征图时,Roi-Pooling 平均将 20*20 的特征图划分成 7*7 的大小,也就是 49 个小区域,每个区域的大小为 2.86*2.86,这个时候 Roi-Pooling 又会进行第二次量化操作,小区域大小变成 2*2 的大小。Roi-Align 对 Roi-Pooling 进行了改进,Roi-Align 取消了量化操作,使用双线性插值法来计算最后的值,可以得到更精确的候选框。

2 实验

实验代码基于 Python3.5,使用的深度学习框架为 Pytorch(0.4 版本),显卡为 RTX 1080ti,内存为 16 G,运行环境 Ubuntu16.04。

2.1 实验数据集

本文分别在 HollwoodHeads[14],Brainwash[15]这两个公开的数据集上测试了该模型。如表 1 所示,Brainwash 数据集是一个密集人头检测数据集,拍摄的是在一个咖啡馆里出现的人群,然后对这群人进行标注而得到的数据集。该数据集包含 3 个部分:训练集由 10 769 张图像组成,共包含 81 975 个人头;验证集由 500 张图像组成,共包含 3 318 个人头;测试集由 500 张图像,共包含 5 007 个人头,平均每副图像包含 7.64 个人头。HollwoodHeads 数据集是在 21 部电影场景里收集的,由 224 740 张图像构成,共 369 846 个人头,平均每张图片包含 1.65个人头。

表 1 数据集介绍

Tab.1 Introduction to dataset

数据集	图片数/张	图片大小	总人数/人	平均人数/人
Brainwash	11 917	480*640	91 146	7.64
HollwoodHeads	224 740	–	369 846	1.65

2.2 超参数设置

训练过程中的初始学习率为 0.001,训练 16 万次,训练到 80 000 次时进行一次学习率的调整,调整大小为初始学习率的十分之一。使用随机梯度下降法(stochastic gradient descent,SGD)来优化模型,动量(momentum)的大小设置为 0.9。使用非极大抑制(non-maximum suppression,NMS)来去除重复比例较大的候选框。正负样本的判定与 Faster-RCNN 的设置一样,小于 0.3 的样本作为负样本,大于 0.7 的样本作为正样本,在此之间的样本忽略不计。使用 VOC[16]的评价标准来计算平均精度(AP),在测试时,对于输出的检测框如果与实际框的 IoU 比值大于预先设定的阈值(0.5 和 0.7),那么就认为这个检测框检测到了人头。

2.3 实验对比

对于 Brainwash 数据集,为了方便训练,去除了数据集中的没有人头的图片,构成的新数据集包含 10 461 张训练图片,484 张测试图片,493 张验证集图片。表 2 给出了该模型与之前一些经典方法的对比,Sermanet 等[17]提出了一种多尺度和滑动窗口的方法,他们称之为 Overfeat-AlexNet 模型。Stewart[15]用该方法进行了密集场景下的人头检测实验,并进一步提出了一个基于 LSTM 的人头计数方法,通过 GoogLeNet 来提取图片中的特征,然后将这些特征送入到一个 LSTM 中进行解码,输出一系列的人头检测框。通过使用不同的损失函数,他们提出了 3 种不同的检测模型,分别是:ReInspect,Lfix;ReInspect,Lfirstk;ReInspect,Lhungarian。除此之外,我们还与其他的一些经典的检测方法做了对比 SSD[18],YOLO9000[19],Tiny[20]和 HeadNet[21]。如表 2 所示,本文方法在 Brainwash 数据集上取得了最高的精度。图 3 展示了本文方法和 FCHD 模型在 Brainwash 数据集上的检测结果,从图 3(a)中可以看到 FCHD 模型对一些小人头很难进行有效的检测。图 3(b)展示了本文方法的检测结果,可以看到对一些小人头本文方法也能检测出来(其中黄色框是检测框,蓝色框是真实标注的人头)。值得一提的是,数据集中有一些没有标注的人头,如图 3(b)中红色箭头所指,这些数据或许会对精度的计算产生不利的影响。

表 2　在 Brainwash 上的对比实验
Tab.2　The comparison experiment on Brainwash

数据集	拥挤程度	方法	AP(IoU=0.5)	AP(IoU =0.7)
Brainwash	咖啡厅密集场景,平均一张图片 7.64 个人头	Over Feat – Alex Net[17]	62.0	–
		Re Inspect,Lhungarian[15]	78.0	–
		YOLO9000[19]	62.5	19.3
		SSD[18]	56.8	15.2
		Tiny[20]	89.3	46.0
		Head Net[21]	91.3	51.4
		本文	95.3	70.7

(a) FCHD[7]模型的检测结果

(b) 本文方法的检测结果

图 3　模型检测效果对比
Fig.3　Model detection result comparison

HollwoodHeads 数据集是一个电影场景里收集的数据集,训练集,验证集,测试集分别包含 216 694,6 719,1 297 张图片。我们和一些基线方法进行了对比,如 SSD [18],YOLO9000 [19],Tiny[20],HeadNet [21]的人头检测方法,Merad[8]提出的两种人头检测模型 Context–Local 和 Context–Local+Global+Pairwise 方法。表 3 展示了本文方法在 IoU=0.5 和 0.7 的情况下都优于以前的方法。图 5 展示了一些本文方法的检测结果,对于电影场景中正常的人头都能有一个很好的检测,但是对于一些影子数据,该方法会有漏检测的结果。

表3 在HollwoodHeads上的对比实验
Tab.3 The comparison experiment on Hollwood Heads

数据集	拥挤程度	方法	AP(IoU =0.5)	AP(IoU =0.7)
HollwoodHeads	电影场景,平均每张图像包含1.65个人头	Context – Local+ Global + Pairwise[8]	72.7	–
		YOLO9000[19]	73.2	51.3
		SSD[18]	81.1	59.5
		HeadNet[21]	83.0	52.6
		Tiny[20]	81.3	49.7
		本文	89.1	68.5

图4 HollwoodHeads数据集检测示例
Fig.4 Detection result on HollwoodHeads

2.4 进一步的实验

为了验证多尺度融合模块和Roi-Align的有效性,在Brainwash数据集和HollwoodHeads数据集上进行了一些实验。IoU的大小表示检测框和真实框之间的重合度,IoU越大则表明检测框和真实框之间越接近。一般IoU大于等于0.5就可以认为是一个正确的检测,但在某些场景中对检测框有更高的要求,本文在IoU=0.5和0.7的这两种情况下进行了实验。表4展示了在Brainwash数据集上的实验结果,可以看到本文方法比原始的Faster-RCNN网络提高了15.7%,将RoI_Pooling替换为Roi_Align在IoU=0.5时候精度相差不大,而在IoU=0.7的情况下可以带来6%的提升。这说明Roi_Align使得模型输出得框更接近真实得框。

表4 在Brainwash数据集上的实验结果
Tab.4 Experimental results on Brainwash

方法	AP(IoU=0.5)	AP(IoU=0.7)
Faster-RCNN[9]	79.8	–
Faster-RCNN+RoI_Pooling+MSFF	95.2	64.7
Faster-RCNN+Roi_Align+MSFF(本文)	95.3	70.7

表5展示了在HollwoodHeads数据集上的实验结果,Faster-RCNN(C4)表示用特征提取网络第4层特征(C4)进行的实验。在HollwoodHeads数据集上本文方法比原始的Faster-RCNN网络提高了8.28%,将RoI_Pooling替换为Roi_Align在IoU=0.7的情况下可以在此基础上再带来1.4%的提升。

表5 在HollwoodHeads数据集上的实验结果
Tab.5 Experimental results on HollwoodHeads

方法	AP(IoU=0.5)	AP(IoU=0.7)
Faster-RCNN(C4)[9]	74.01	51.6
Faster-RCNN+RoI_Pooling+MSFF	89.5	67.1
Faster-RCNN+Roi_Align+MSFF(本文)	89.1	68.5

此外,我们还在不同分辨率下的训练图片进行了实验,实验结果如表6所示,在保持图片宽高比不变的情况下1 024*768表示将图片放大到1 024*768的大小,640*480相同,IoU的取值为0.5,由表6可知在相同条件下,输入图片分辨率越大精度越高。

表 6　在不同分辨率下的实验结果
Tab.6　Experimental results at different resolutions

数据集	AP(1 024*768)	AP(640*480)
Brainwash	95.3	93.6
HollwoodHeads	89.1	85.5

3　结论

本文提出了一种端到端的人头检测模型,该模型能够很好的检测到密集场景的人头。

1) 使用 ResNet101 作为特征提取网络提取特征,然后将提取到的特征送入一个多尺度特征融合模块(MSFF)进行特征融合,该模块将融合后的特征分别送入后续的网络进行检测。

2) 通过先验框的设计和使用 Roi-Align 代替了 Roi-Pooling 进一步的提高了精度。由于融合了不同层的特征,该模型能够有效的检测到不同尺度的人头。实验表明,该方法在 Brainwash 和 HollwoodHeads 数据集上达到了最优的结果。

参考文献:

[1] SINDAGI V A,PATEL V M. A survey of recent advances in CNN-based single image crowd counting and density estimation[J]. Pattern Recognition Letters,2017:3-16.

[2] WANG C,ZHANG H,YANG L,et al. Deep people counting in extremely dense crowds[C]//Proceedings of the 23rd ACM International Conference on Multimedia,2015.

[3] 陆金刚,张莉. 基于多尺度多列卷积神经网络的密集人群计数模型[J]. 计算机应用,2019,39(12):3445-3449.

[4] 李欢,陈先桥,施辉,等. 基于 SSD 的行人头部检测方法[J]. 计算机工程与设计,2020,41(3):827-832.

[5] 张晓琪,宋钢. 基于多特征协同的人头检测新方法[J]. 西南师范大学学报(自然科学版),2018,43(7):46-52.

[6] 邢志祥,顾凰琳,魏振刚,等. 基于卷积神经网络的行人人头检测方法对比研究[J]. 安全与环境工程,2019,26(1):77-82.

[7] VORA A,CHILAKA V. FCHD:Fast and accurate head detection in crowded scenes[C]//Conference on Image Prolessing,2019.

[8] MERAD D,AZIZ K E,THOME N,et al. Fast people counting using head detection from skeleton graph[C]//Advanced Video and Signal Based Surveillance,2010.

[9] REN,HE K,GIRSHICK R,et al. Faster R-CNN:Towards real-time object detection with region proposal networks[C]//Neural Information Processing Systems,2015.

[10] 郑晓芳,黄鹿鸣,傅军栋. 基于元胞自动机的建筑火灾预测与疏散系统[J]. 华东交通大学学报,2020,37,172(2):126-132.

[11] LIN T,DOLLAR P,GIRSHICK R,et al. Feature pyramid networks for object detection[C]//Computer Vision and Pattern Recognition,2017.

[12] LIU S,QI L,QIN H,et al. Path aggregation network for instance segmentation[C]//Computer Vision and Pattern Recognition,2018.

[13] HE K,GKIOXARI G,DOLLAR P,et al. Mask R-CNN[C]//International Conference on Computer Vision,2017.

[14] VU T,OSOKIN A,LAPTEV I,et al. Context-aware CNNs for person head detection[C]//International Conference on Computer Vision,2015.

[15] STEWART R,ANDRILUKA M,NG A Y,et al. End-to-end people detection in crowded scenes[C]//Computer Vision and Pattern Recognition,2016.

[16] EVERINGHAM M,ESLAMI S M,VAN G L,et al. The pascal visual object classes challenge:A retrospective[J]. International Journal of Computer Vision,2015,111(1):98-136.

[17] SERMANET P,EIGEN D,ZHANG X,et al. OverFeat:Integrated recognition,localization and detection using convolutional networks[C]//Computer Vision and Pattern Recognition,2013.

[18] LIU W,ANGUELOV D,ERHAN D,et al. SSD:Single shot multibox detector[C]//European Conference on Computer Vision,2016.

[19] REDMON J,FARHADI A. YOLO9000:Better,faster,stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017.

[20] HU P,RAMANAN D. Finding tiny faces[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017.

[21] LI W,LI H,WU Q,et al. Head Net:an end-to-end adaptive relational network for head detection[J]. IEEE Transactions on Circuits and Systems for Video Technology,2019,30(2):482-494.