计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2021.20200699
58(12):2724-2747,2021

基于深度学习的人群计数研究综述

余 鹰[1] 朱慧琳[1] 钱 进[1] 潘 诚[1] 苗夺谦[1,2]
[1](华东交通大学软件学院 南昌 330013)
[2](同济大学电子与信息工程学院 上海 201804)
(yuyingjx@163.com)

Survey on Deep Learning Based Crowd Counting

Yu Ying[1], Zhu Huilin[1], Qian Jin[1], Pan Cheng[1], and Miao Duoqian[1,2]
[1](*School of Software, East China Jiaotong University, Nanchang* 330013)
[2](*College of Electronics and Information Engineering, Tongji University, Shanghai* 201804)

Abstract Crowd counting, aiming to estimate the number, density or distribution of crowds in images or videos, belongs to the research category of object counting. It has been widely employed in crowd behavior analysis and public safety management to detect crowding or abnormal behavior in time to avoid accidents. In the past decades, although tremendous efforts have been made to enhance the performance of crowd counting algorithms, some long-standing challenges, such as cross-scene counting, perspective distortion and scale variation, remain unresolved. Along this line, an emerging research trend is to exploit the deep learning technologies for crowd counting. It has been proven to be an effective way to address the above issues. In this paper, crowd counting models based on deep learning are reviewed, analyzed, and discussed. Firstly, crowd counting models are introduced in details from the perspective of their principles, steps, and model variants, and the difference between the crowd counting models based on traditional methods and the crowd counting models based on deep learning are analyzed. Then the research status of crowd counting based on deep learning are expounded from four aspects: network structure, ground-truth generation, loss function and evaluation index. Meanwhile, the characteristics of various crowd counting data sets are compared and analyzed. Finally, some future directions of crowd counting are given.

Key words crowd counting; density map estimation; multi-scale; deep learning; convolutional neural network (CNN)

摘 要 人群计数旨在估计图像或视频中人群的数量、密度或分布,属于目标计数(object counting)领域的研究范畴,广泛应用于人群行为分析、公共安全管理之中,以便及时发现人群拥挤或异常行为,避免事故发生.鉴于人群计数系统强大的实用性,自21世纪以来,研究者对其方法及应用进行了大量广泛的研究.近年来,深度学习技术发展迅猛,很多工作发现深度学习技术可以有效地解决人群计数系统存在的一系列关键问题,例如跨场景计数、透视畸变、尺度变化等.因此,对基于深度学习的人群计数这一研究领域进行回顾、分析和展望.具体地,首先从概念、步骤、方法等维度详细介绍人群计数模型,分析基于传统方法和基于深度学习方法这2类人群计数模型的差异.然后,从计数网络结构、ground-truth生成、

收稿日期:2020-09-08;**修回日期**:2021-01-22
基金项目:国家自然科学基金项目(62163016, 62066014);江西省自然科学基金项目(20212ACB202001,20202BABL202018)
This work was supported by the National Natural Science Foundation of China (62163016, 62066014) and the Natural Science Foundation of Jiangxi Province (20212ACB202001, 20202BABL202018).

损失函数、评价指标这4个方面阐述基于深度学习的人群计数模型的研究现状.最后,比较分析了各种人群计数数据集的特点,并探讨和展望人群计数领域未来可能的研究方向.

关键词　人群计数;密度图估计;多尺度;深度学习;卷积神经网络

中图法分类号　TP391

人群计数是估计图像或视频中人群的数量、密度或分布[1],它是智能视频监控分析领域的关键问题和研究热点,也是后续行为分析[2-3]、拥塞分析[4]、异常检测[5-6]和事件检测[7]等高级视频处理任务的基础.随着城市化进程的快速推进,城市人口数量急剧增长,导致各种人员高度聚集的社会活动频繁发生,如果管控不当,极易发生拥挤踩踏事故.例如上海"12.31"外滩踩踏事故中,由于现场管理和应对措施不当,引发了人群拥挤和摔倒,最终造成了重大人员伤亡的严重后果[8-9].如果有精度良好的人群计数系统实时统计相关场所的人群数量、分布或密度等信息,及时发现人群拥挤和异常行为并进行预警,以便采取措施进行疏导,就可以避免悲剧的发生[10-11].性能良好的人群计数算法也可以迁移到其他目标计数领域,如显微图片中的细菌与细胞计数[12]、拥挤道路上的汽车计数[13]等,拓展人群计数算法的应用范围.因此,人群计数方法的研究有着重要的现实意义和应用价值.

随着人工智能、计算机视觉等技术的不断发展,人群计数受到了国内外众多学者的广泛关注和研究.早期人群计数主要使用传统的计算机视觉方法提取行人特征[14],然后通过目标检测[15-19]或回归[20-21]的方式获取图像[22-25]或视频[26-28]中人群的数量.传统方法具有一定局限性,无法从图像中提取更抽象的有助于完成人群计数任务的语义特征,使得面对背景复杂、人群密集、遮挡严重的场景时,计数精度无法满足实际需求.近年来,深度学习技术发展迅猛,在许多计算机视觉任务中得到成功应用[29],促使研究人员开始探索基于卷积神经网络(convolutional neural network, CNN)[30]的人群计数办法.相比于传统方法,基于CNN的人群计数方法在处理场景适应性、尺度多样性等问题时表现更优.而且由于特征是自学习的,不需要人工选取,可以显著提升计数效果,因此已经成为当前人群计数领域的研究热点.使用CNN的人群计数方法主要分为直接回归计数法和密度图估计法2类.直接回归法只需向CNN送入人群图片,就可以直接输出人群数量,适用于人群稀疏场景.在密度图法中,CNN输出的是人群密度图,再以数学积分求和的方式计算出人数.这类方法性能的好坏一定程度上依赖于密度图的质量.为了提升密度图质量,会引入新的损失函数[31]来提高密度图的清晰度和准确度.无论采用哪种方法,都需要先进行特征提取.为了提升特征的鲁棒性,常使用多尺度预测、上下文感知、空洞卷积、可形变卷积等方法改进特征提取过程,以增强特征的判别能力.

得益于深度学习模型强大的特征提取能力,基于深度学习的人群计数方法的研究已经取得了很多优秀的成果.根据计数对象,可以将这些方法归纳为基于图像和基于视频的2类;根据网络模型结构,可将它们划分为单分支结构、多分支结构和特殊结构3类;根据度量规则,可将它们划分为基于欧氏距离损失、基于SSIM损失和基于对抗损失等多类.

本文重点讨论基于深度学习的静态图像人群计数方法,主要贡献可以归纳为3个方面:

1) 从不同层面,对人群计数领域的研究现状进行系统全面的总结和深入的探讨,包括计数网络结构、损失函数、性能评价指标等.这种全面梳理可以帮助研究人员快速了解基于深度学习的人群计数算法的研究现状和关键技术.

2) 基于数据比较了不同模型的计数效果,分析了计数模型性能优劣的原因,为未来研究人员设计更加优化的计数模型提供借鉴.

3) 归纳总结了在模型设计、损失函数定义、ground-truth生成等方面存在的问题,为未来该领域的研究指明了方向.

1　人群计数网络

1.1　单分支结构计数网络

早期使用CNN的人群计数网络均为只包含一条数据通路的单分支网络结构.Wang等人[32]最先将CNN引入人群计数领域,提出了一种适用于密集人群场景的端到端CNN回归模型.该模型对AlexNet网络[33]进行改进,将最后的全连接层替换

为单神经元层,直接预测人群数量.由于没有预测人群密度图,所以无法统计场景中的人员分布情况.此外,虽然该模型通过 CNN 自动学习了有效的计数特征,但是由于 AlexNet 的宽度较窄,深度也较浅,导致特征鲁棒性不够强,在人群密集场景下的计数效果较差,并且在跨场景计数时,效果不甚理想,缺乏足够的泛化性.

为了解决跨场景问题,Zhang 等人[24]提出了一种基于 AlexNet 的跨场景计数模型 Crowd CNN,首次尝试输出人群密度图,其总体结构如图 1[24]所示:

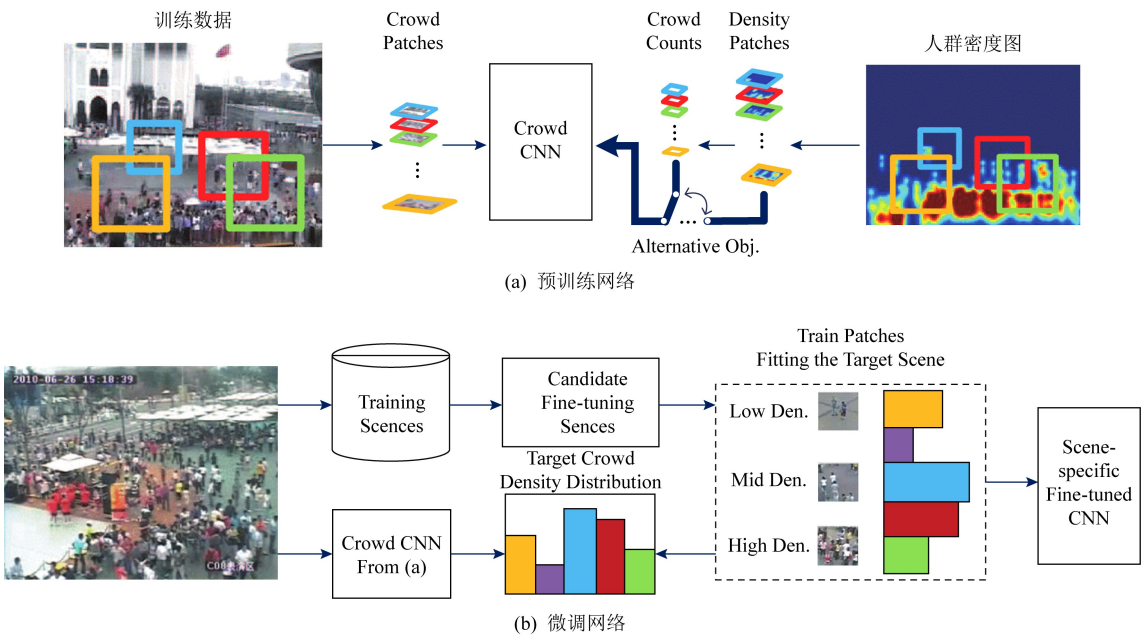


Fig. 1 The cross-scene crowd counting model proposed by Reference[24]

图 1 文献[24]提出的跨场景计数模型

其中,图 1(a)描绘了计数网络的预训练(pre-trained)过程,通过人群密度图(crowd density map)和人群计数(crowd counts)这 2 个目标任务的交替训练来优化模型.然后,算法会根据目标场景特点,选择相似场景对计数模型进行微调(fine-tuning),如图 1(b)所示,以达到跨场景计数的目的.为了提升计数准确性,作者还提出了透视图(perspective map)的概念,如图 2(a)所示,颜色越浅代表目标尺度越大.然后,通过密度图和透视图的融合,如图 2(b)所示,降低透视形变(perspective distortion)的不良影响,提升密度图质量.但是透视图较难获得,限制了该模型的推广.该工作的另一个贡献是建立了经典的人群计数数据集 WorldExpo'10,为交叉场景人群计数模型的测评提供数据.

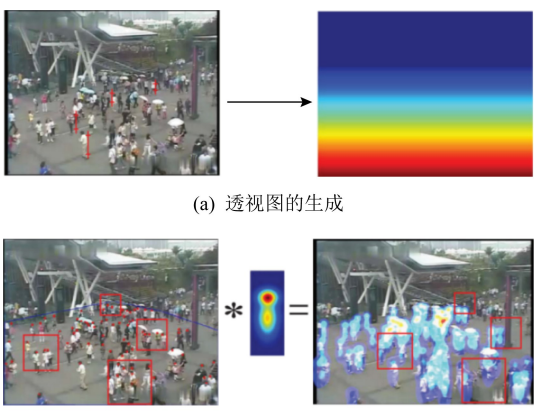


Fig. 2 Normalized crowd density map for training[24]

图 2 标准化人群密度图[24]

1.2 多分支结构计数网络

人群分布相对监控摄像头位置具有较大的不确定性,导致拍摄视角差异较大,所拍摄到的图像或视频中目标尺寸变化较大.对于人群计数任务来说,如何提高计数网络对目标尺度变化的适应性是亟待解决的问题.

为了解决多尺度问题,Boominathan 等人[34]基于 CNN 提出了一种双分支结构计数网络 CrowdNet,如图 3 所示.通过一个浅层网络(shallow network)和一个深层网络(deep network)分别提取不同尺度的特征信息进行融合来预测人群密度图.这种组合可以同时捕获高级和低级语义信息,以适应人群的非均匀缩放和视角的变化,因此有利于不同场景不同尺度的人群计数.

通过引入多路网络,使用大小不同的感受野提取不同尺度特征可以有效解决多尺度问题,由此衍生出了一系列多列卷积神经网络结构的人群计数算法.

Zhang 等人[25]受多分支深度卷积神经网络[35]的启发,提出了一种多列卷积神经网络(multi-column CNN, MCNN)用于人群计数,其结构如图 4 所示.每一分支网络采用不同大小的卷积核来提取不同尺度目标的特征信息,减少因为视角变化形成的目标大小不一导致的计数误差.MCNN 建立了图像与人群密度图之间的非线性关系,通过用全卷积层替换全连接层,使得模型可以处理任意大小的输入图片.为了进一步修正视角变化带来的影响,MCNN 在生成密度图时,没有采用固定的高斯核,而是利用自适应高斯核计算密度图,提升了密度图质量.该工作的另一贡献是收集并标注了 ShanghaiTech 人群计数数据集,该数据集由 1 198 张带标注的图像组成,包含人群分布从稀疏到密集变化的各种场景,目前该数据集已成为人群计数领域的基准数据集之一.

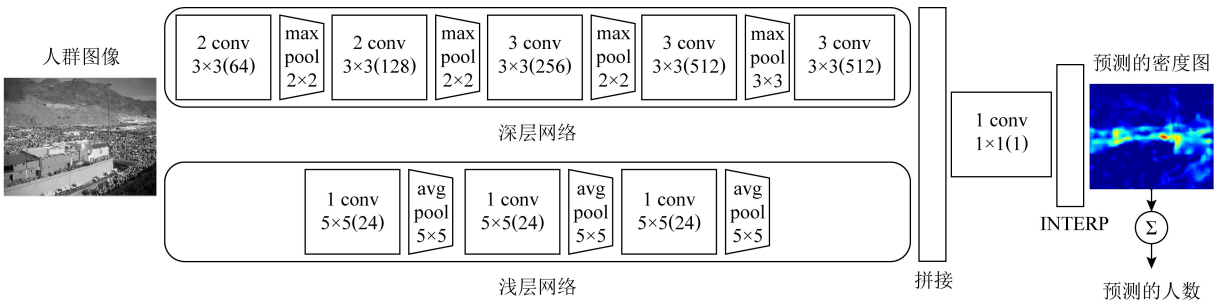


Fig. 3　The structure of two-column crowd crounting network[34]

图 3　双列人群计数网络[34]

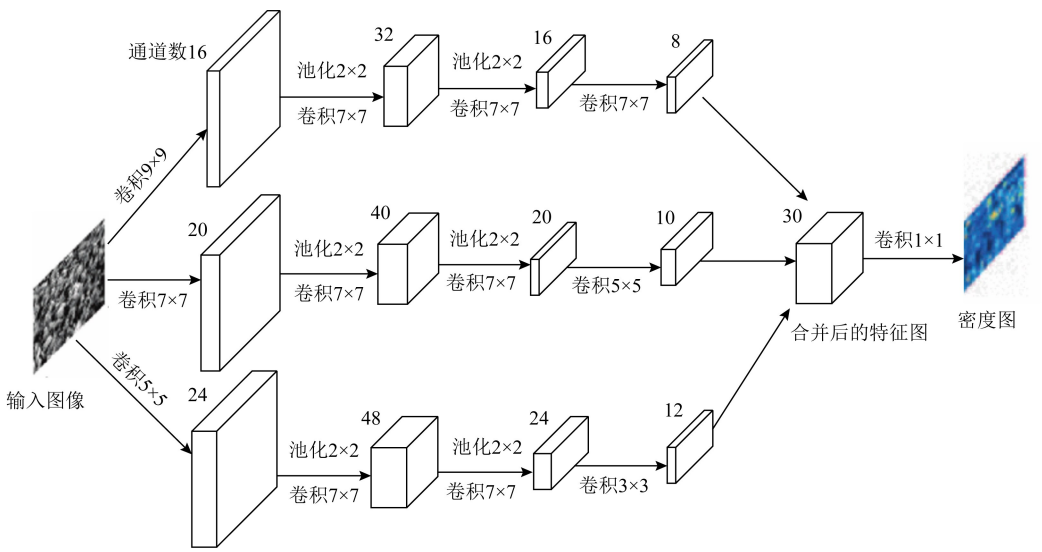


Fig. 4　The structure of the multi-column crowd counting network[25]

图 4　多列人群计数网络[25]

计数性能的好坏主要依赖于密度图的质量.为了生成更高质量的密度图,Sindagi 等人[36]提出了上下文金字塔卷积神经网络计数模型 CP-CNN,其结构如图 5 所示,通过多个 CNN 获取不同尺度的场景上下文信息,并将这些上下文信息显式地嵌入到密度图生成网络,提升密度估计的精度.CP-CNN 由 4 个部分组成,其中全局上下文估计器(global context estimator, GCE)和局部上下文估计器(local context estimator, LCE)分别提取图像的全局和局部上下文信息,即分别从全局和局部的角度预测图像的密度等级;密度估计器(density map estimator, DME)没有直接生成密度图,而是沿用了 MCNN 的多列网络结构生成高维特征图;融合卷积神经网络(fusion-CNN, F-CNN)则将前 3 个部分的输出进行

融合,生成密度图.为了弥补 DME 中丢失的细节信息,F-CNN 使用了一系列小数步长卷积层帮助重建密度图的细节.现有的 CNN 计数网络主要使用像素级欧氏距离损失函数来训练网络,这导致生成的密度图比较模糊.为此,CP-CNN 引入对抗损失(adversarial loss),利用生成对抗网络(generative adversarial net, GAN)[37]来克服欧氏距离损失函数的不足.

2017 年,Sam 等人[38]提出了一种多列选择卷积神经网络(switch convolution neural network, Switch-CNN)用于人群计数,其结构如图 6 所示.与 MCNN 不同之处在于,Switch-CNN 虽然采用多列网络结构,但是各列网络独立处理不同的区域.在送入网络之前,图像被切分成 3×3 的区域,然后对每个区域使用特定的 SWITCH 模块进行密度等级划分,并根据密度等级选择对应的分支进行计数.通过对于密度不同的人群有针对性地选用不同尺度的回归网络进行密度估计,使得最终的计数结果更为准确.Switch-CNN 也存在不容忽视的弊端,如果分支选择错误将会大大影响计数准确度.

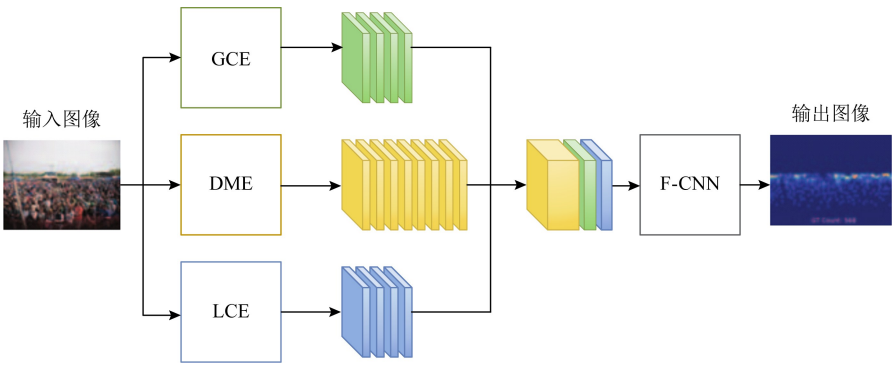


Fig. 5 Architecture of CP-CNN[36]

图 5 CP-CNN 计数模型[36]

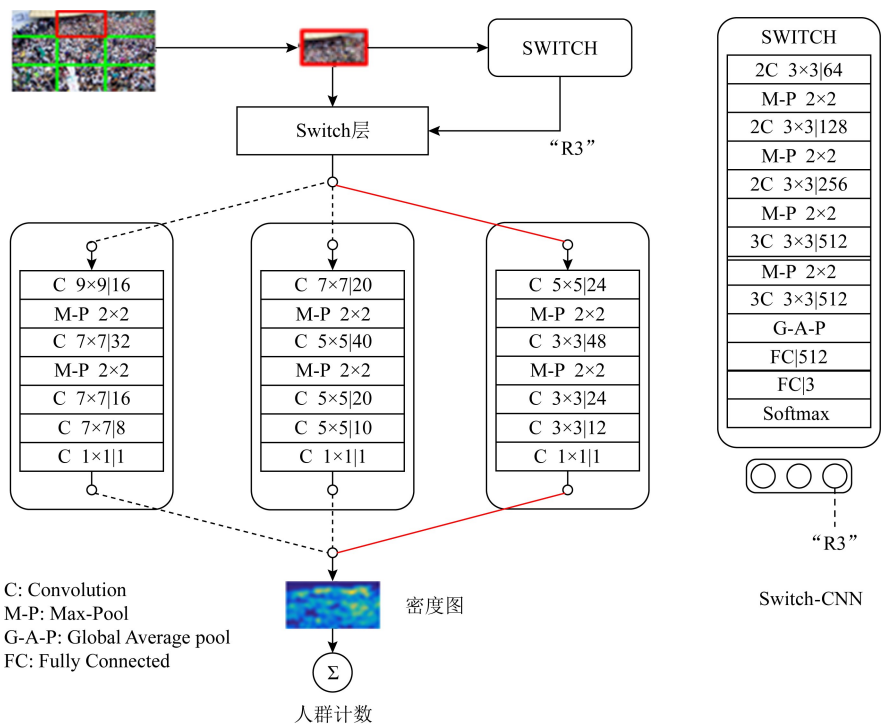


Fig. 6 Architecture of Switch-CNN[38]

图 6 Switch-CNN 计数模型[38]

Switch-CNN 根据图像块的内容选择合适的分支网络进行人群密度估计的做法,为设计多列计数网络提供了新思路.但是 Swith-CNN 将密度等级固定划分为 3 个层次,难以应对人群密度变化范围很

大的场景.为此,Sam 等人[39]对 Switch-CNN 进行改进,提出了逐步增长卷积神经网络(incrementally growing CNN, IG-CNN),其层次化训练过程如图 7 所示.从一个基础 CNN 模型(Base CNN)开始,通过不断地迭代,最后生成 1 棵 CNN 二叉树,叶子节点即为用于密度估计的回归器,其中每个回归器对应 1 种特定的密度等级.第 1 层通过聚类将训练集 D_0 划分成 D_{00} 和 D_{01} 这 2 个部分,然后 R_{00} 和 R_{01} 是由复制 R_0 而来,随后 R_{00} 和 R_{01} 分别在对应的训练集 D_{00} 和 D_{01} 上训练,其他层的构建情况相似.最终通过层次聚类,将原始训练集划分成多个子集,每个子集对应 1 个密度等级,由相应的密度估计器负责计数.测试阶段则会根据图片的密度等级选择对应的密度估计器.

在已有的人群计数模型中,通常单纯地假设场景中的人群分布是稀疏或密集的.针对稀疏场景,采用检测方法进行计数[40];而针对密集场景,则采用回归方法进行人群密度估计.这样的模型往往难以应对密度变化范围很广的人群场景的计数.为了解决这个问题,Liu 等人[41]提出了一种检测和回归相结合的人群计数模型 DecideNet,其结构如图 8 所示.该模型也是一种多列结构的计数网络,其中 RegNet 模块采用回归方法直接从图像中估计人群密度,DetNet 模块则在 Faster-RCNN 的后面添加了一个高斯卷积层(Gaussian convolution),直接将检测结果转化为人群密度图,然后 QualityNet 引入注意力模块,自动判别人群密集程度,并根据判别结果自适应地调整检测和回归这 2 种方法的权重,再根据这个权重将这 2 种密度图进行融合,以此获取

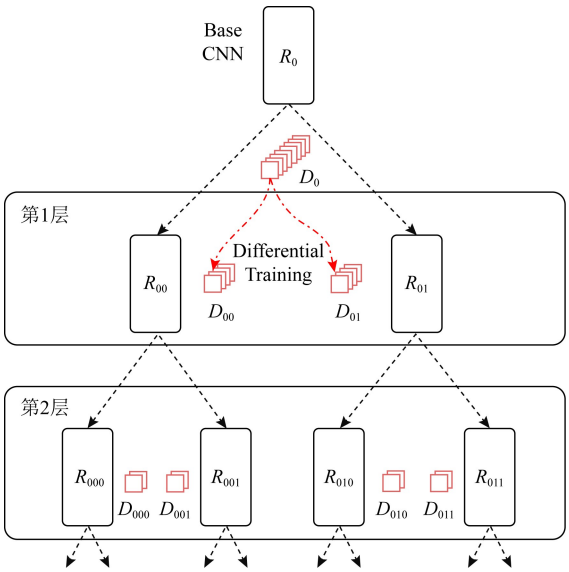


Fig. 7　Training process of IG-CNN[39]

图 7　IG-CNN 训练过程[39]

更好的最优解.但是由于 RegNet 和 DetNet 这 2 个子网络均使用了较大的感受野,模型参数过多,导致该模型的训练复杂度较高.

多列计数网络使用不同大小的卷积核提取图像的多尺度特征,其良好的效果说明多尺度表达的重要性.但是多列计数网络也引入了新的问题,首先多尺度表达的性能通常依赖于网络分支的数量,即尺度的多样性受限于分支数目,其次已有工作大多采用欧氏距离作为损失函数,假设像素之间互相独立,导致生成的密度图比较模糊.

为了解决上述问题,Cao 等人[42]提出了一种尺度聚合网络(scale aggregation network, SANet),

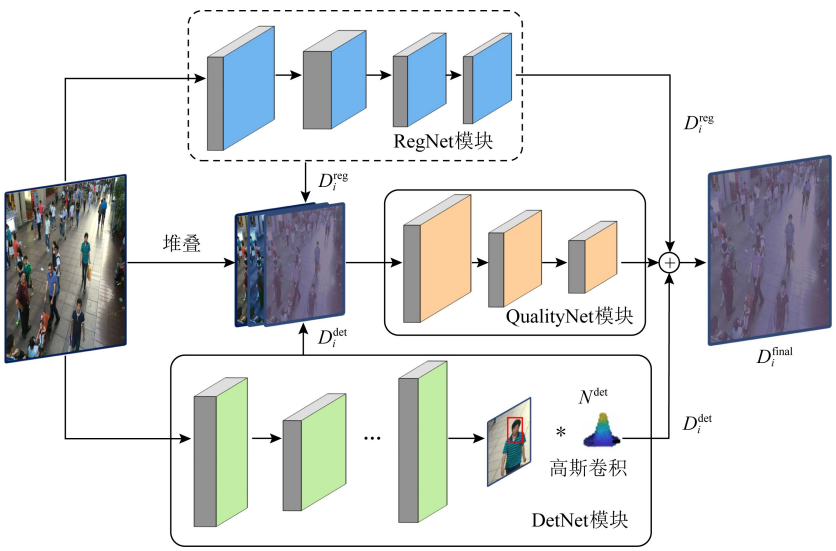


Fig. 8　Architecture of DecideNet[41]

图 8　DecideNet 网络结构[41]

其结构如图 9 所示.该模型没有采用 MCNN 的多列网络结构,而是借鉴了 Inception[43] 的架构思想,在每个卷积层同时使用不同大小的卷积核提取不同尺度的特征,最后通过反卷积生成高分辨率的密度图.整个模型由 FME(feature map encoder)和 DME(density map estimator)这 2 个部分组成,FME 聚合提取出多尺度特征,DME 融合特征生成高分辨率的密度图.度量预测的密度图与 ground-truth 的相似度时,采用 SSIM 计算局部一致性损失,然后对欧氏损失和局部一致性损失进行加权得到总损失.

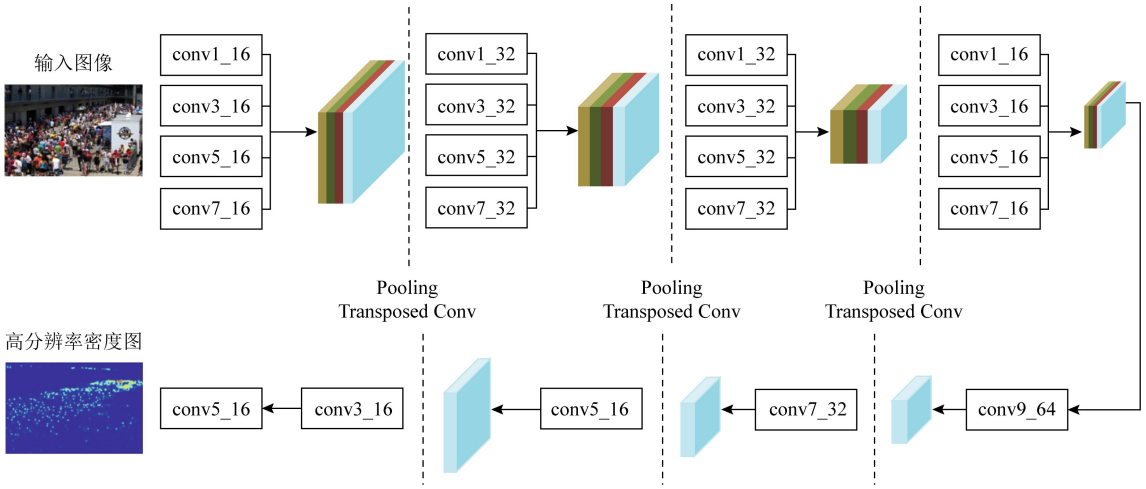


Fig. 9 Architecture of SANet[42]

图 9 尺度聚合网络结构[42]

由于“透视畸变”问题,位于不同景深的目标尺寸差异较大,对人群计数模型的建模能力提出了很高的要求.为了解决这个问题,Hossain 等人[44] 首次将注意力机制引入人群计数领域,提出了多分支的尺度感知注意力网络(scale-aware attention network, SAAN),其结构如图 10[44] 所示.该网络由 4 个模块组成,其中多尺度特征提取器(multi-scale feature extractor, MFE)负责从输入图像中提取多尺度特征图.受到 MCNN[25] 启发,MFE 被设计成包含 3 个分支的多列网络,每个分支的感受野大小不同,可以捕获不同尺度的特征;为了获得图像的全局密度信息,与 MFE 中 3 个不同尺度的分支相对应,定义了 3 个全局密度等级,然后利用全局尺度注意力(global scale attentions, GSA)模块负责提取输入图像的全局上下文信息,计算 3 个全局密度等级对应的评分,并对这 3 个分值进行归一化.GSA 只能提取图像的全局尺度信息,但在实际的人群计数图像中,不同位置往往存在密度差异,为此增加了局部尺度注意力(local scale attention, LSA)负责提取图像不同位置的细粒度局部上下文信息,并生成 3 张像素级的

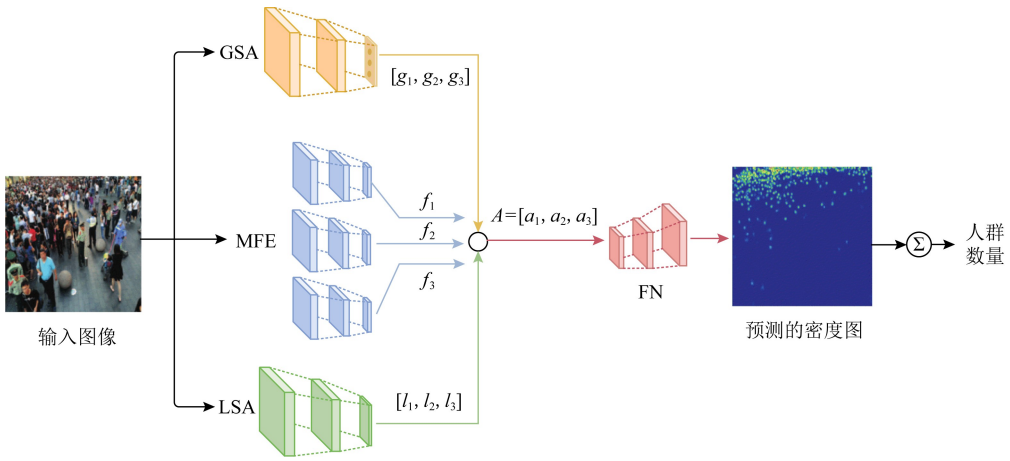


Fig. 10 Architectureof SAAN[44]

图 10 SAAN 结构[44]

注意力图,用于描述局部尺度信息;最后,根据全局和局部的尺度信息对 MFE 提取的特征图进行加权,然后将加权后的特征图输入融合网络(fusion network, FN)生成最终的密度图.

与 DecideNet[41] 相比,SAAN 通过注意力机制进行尺度选择的方式更加灵活.但是,由于 SAAN 包含 4 个子网络,且 MFE 包含多个分支,网络模型复杂、参数多、训练难度很大.

1.3　特殊结构计数网络

虽然多分支结构计数网络取得了较好的计数效果,但是多分支结构网络模型的复杂性较高,由此也带来了一些新的问题[45].首先,网络模型参数繁多、训练困难,导致计数实时性较差;其次,多分支网络的结构冗余度较高.多分支计数网络原本是想通过不同的分支采用大小不等的感受野来提取不同尺度的特征,增强特征的适用性和鲁棒性.但研究表明,不同分支学习到的特征相似度很高,并没有因为场景密集程度不同而出现明显差异.为了克服这些问题,研究人员开始尝试将一些新型 CNN 结构,例如空洞卷积网络(dilated convolutional networks)[46]、可形变卷积网络(deformable convolutional network)[47]、GAN[37] 等,引入人群计数领域,以降低计数模型复杂度,提升计数精度和人群密度图的还原度.

2018 年,Li 等人[45] 提出了一种适用于密集人群计数的空洞卷积神经网络模型 CSRNet,其网络结构如图 11 所示.CSRNet 没有采用以往广泛使用的多分支网络结构,而是将舍弃了全连接层的 VGG-16 作为该网络的前端部分,后端则采用 6 层空洞卷积神经网络,构成一个单通道计数网络,大幅削减了网络参数量,降低了训练难度.同时,借助空洞卷积可以在保持分辨率的同时扩大感受野的优势,保留了更多的图像细节信息,使得生成的人群分布密度图质量更高.CSRNet 后端有 A,B,C,D 这 4 组不同的配置,其中 B 组方案在 ShanghaiTech Part A 数据集上的表现最优.

CSRNet 的成功为密集人群计数提供了新的思路,随后许多学者开始效仿采用空洞卷积进行人群计数研究[48].

多分支计数网络的不同分支之间缺少相互协作,每个分支只是试图通过最小化欧氏损失优化自己的估计.由于每个分支只在特定尺度上表现良好,导致平均各分支结果后生成的密度图较模糊,同时由于在网络中使用池化层,大大降低了密度图的分辨率,使得最终的计数结果产生误差.此外,存在跨尺度统计不一致问题,一个图像分割成多份分别输入网络得到的总人数和将输入整张图像计算得出的人数存在差异.

CSRNet配置			
A	*B*	*C*	*D*
输入(分辨率不固定的彩色图像)			
前端 (fine-tuned from VGG-16)			
conv3-64-1 conv3-64-1			
最大池化			
conv3-128-1 conv3-128-1			
最大池化			
conv3-256-1 conv3-256-1 conv3-256-1			
最大池化			
conv3-512-1 conv3-512-1 conv3-512-1			
后端(4种不同的配置)			
conv3-512-1 conv3-512-1 conv3-512-1 conv3-256-1 conv3-128-1 conv3-64-1	conv3-512-2 conv3-512-2 conv3-512-2 conv3-256-2 conv3-128-2 conv3-64-2	conv3-512-2 conv3-512-2 conv3-512-2 conv3-256-4 conv3-128-4 conv3-64-4	conv3-512-4 conv3-512-4 conv3-512-4 conv3-256-4 conv3-128-4 conv3-64-4
conv1-1-1			

Fig. 11　Configuration of CSRNet[45]

图 11　CSRNet 配置[45]

为解决这些问题,受 GAN 在图像翻译方面[49] 成功应用的启发,文献[50]提出了一种基于 GAN 的跨尺度人群计数网络(adversarial cross-scale consistency pursuit network, ACSCP),其结构如图 12[50] 所示.对抗损失的引入使得生成的密度图更加尖锐,U-Net 结构[51] 的生成器保证了密度图的高分辨率,同时跨尺度一致性正则化约束了图像间的跨尺度误差.因此,该模型最终能生成质量好、分辨率高的人群分布密度图,从而获得更高的人群计数精度.

利用 GAN 来提高人群计数精度的方法,开启了一种新的思路.在 SFCN[52] 计数网络中,使用了改进的 Cycle GAN[53] 产生数据集风格相似的图片,并贡献了 GCC 数据集.DACC[54] 中也使用 Cycle GAN 进行风格迁移.

基于深度神经网络的人群计数解决方案虽然取得了显著成果,但在高度拥挤嘈杂场景中,计数效果仍然会受到背景噪音、遮挡和不一致的人群分布的严重影响.为了解决这个问题,Liu 等人[55] 提出了一种融合了注意力机制的可形变卷积网络

ADCrowdNet 用于人群计数.如图 13[55]所示,该网络模型主要由2个部分串联而成,其中注意力图生成器(attention map generator, AMG)用于检测人群候选区域,并估计这些区域的拥挤程度,为后续人群密度图的生成提供精细化的先验知识.通过注意力机制,可以过滤掉复杂背景等无关信息,使得后续工作只关注人群区域,降低各种噪声的干扰.密度图估计器(density map estimator, DME)是一个多尺度可形变卷积网络,用于生成高质量的密度图.由于注入了注意力,可形变卷积添加了方向参数,卷积核在注意力指导下在特征图上延伸,可以对不同形状的人群分布进行建模,很好地适应了真实场景中摄像机视角失真和人群分布多样性导致的畸变,保证了拥挤场景中人群密度图的准确性.

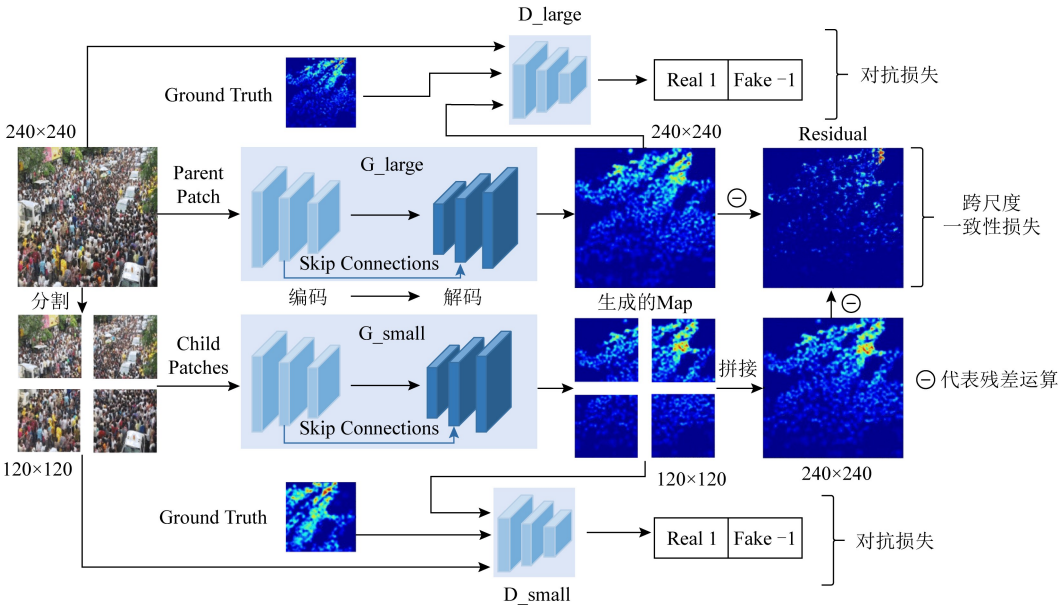


Fig. 12 Architecture of ACSCP[50]

图 12 ACSCP 网络结构[50]

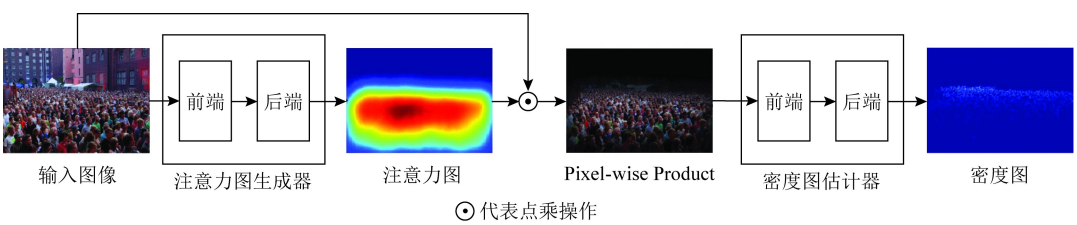


Fig. 13 Architecture of ADCrowdNet[55]

图 13 ADCrowdNet 架构[55]

注意力图生成器 AMG 的网络结构如图 14 所示,采用了 VGG-16 网络前 10 个卷积层作为前端(front end),用来提取图像的底层特征,后端(back end)架构类似 Inception 结构[43],采用多个空洞率不同的空洞卷积层[56]扩大感受野,应对不同尺度的人群分布.后端输出 2 通道的特征图 $\boldsymbol{F}_c$ 和 $\boldsymbol{F}_b$,分别代表前景(人群)和背景.然后,通过对特征图取全局平均池化 GAP 获得相应的权重 $\boldsymbol{W}_c$ 和 $\boldsymbol{W}_b$,再对其结果用 softmax 进行分类获取概率 $\boldsymbol{P}_c$ 和 $\boldsymbol{P}_b$.最后,对特征图和概率进行点乘获得注意力图.

密度图估计器 DME 的网络结构如图 15 所示,前端依然使用 VGG-16,后端架构依然类似 inception 结构,但是采用了更适合拥挤嘈杂场景的多尺度可形变卷积,以适应人群分布的几何形变.

同年,DADNet[57]也同样使用可形变卷积进行人群计数,取得了较好的计数效果.

背景噪声会对人群计数算法的性能带来重大影响.为了减少背景噪声干扰,许多学者进行了尝试,例如 ADCrowdNet 通过注意力机制,过滤掉背景,让模型只关注人群区域.此外,也有学者试图将图像分割技术 MASK R-CNN[58]应用于人群计数领域,以去除背景噪声.

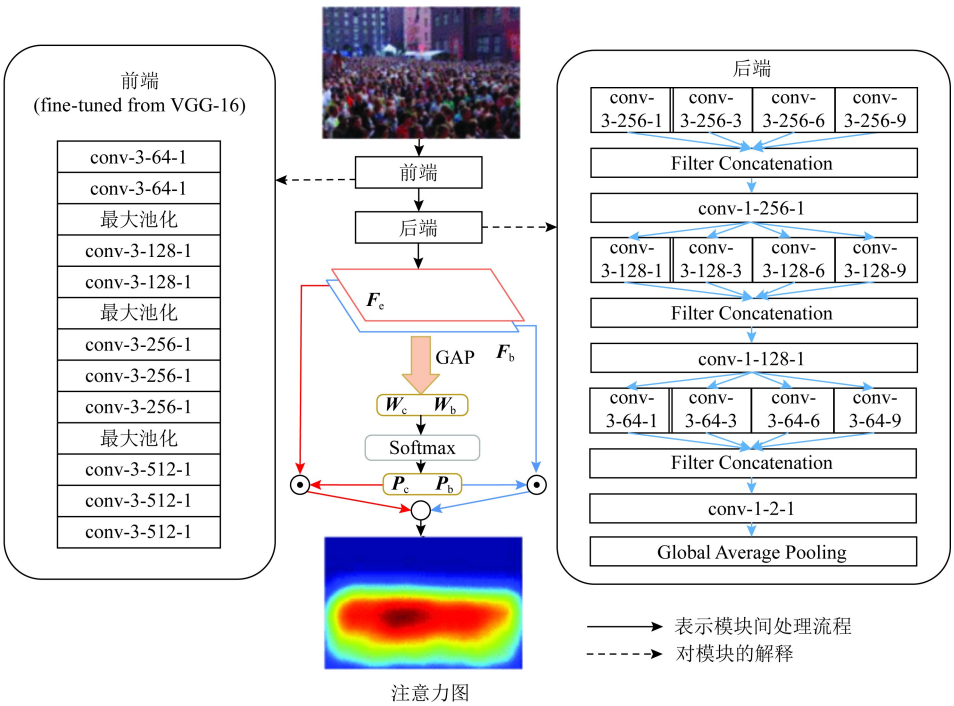


Fig. 14　Architecture of attention map generator

图 14　注意力图生成器

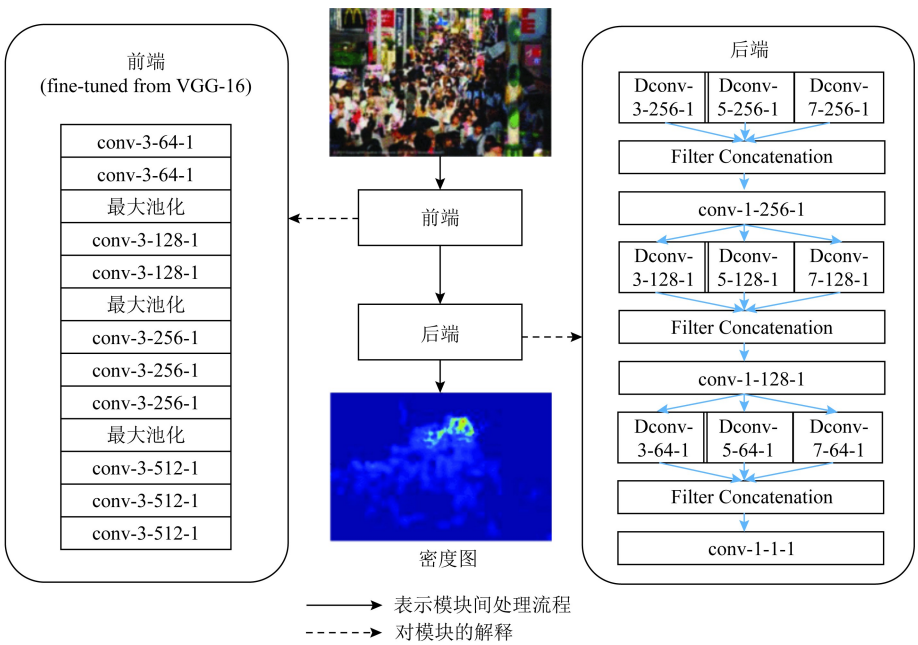


Fig. 15　Architecture of density map estimator

图 15　密度图估计器

实现背景和人群分割的难点在于如何制作用于分割的 ground truth.为此,研究者们进行了各种尝试,SFANet[59]采用了将原本的坐标点 ground truth 进行固定高斯核大小的高斯模糊,再选取一定阈值对其进行 0 和 1 的二值化,由此形成分割 ground truth;MAN[60]采用了固定高斯核对原本坐标点 ground truth 进行处理,并将非 0 值全置为 1,形成分割 ground truth;W-Net[61]则采用 SANet[42]中的归一化高斯核方法对坐标点图进行高斯模糊,再设置一定的阈值进行二分类;SGANet[62]采用每个人头使用 25×25 的方格表示,以此制作 ground truth.

总之,如何降低背景噪声干扰仍然是人群计数

领域未来需要重点关注的问题.除了以上结合分割算法的人群计数算法以外,CFF[63]将分割任务、分类任务、计数任务结合,为我们提供了多任务结合的思路.

由分析可知,随着研究的深入,计数模型的结构在不断发生变化.为了解决多尺度问题,计数网络从最初简单的单分支结构演变为复杂的多分支结构,使得计数准确性得到了提升.但是多分支结构会带来了网络参数量大、计算复杂度高等问题,导致计数模型的效率低下.为了克服这些问题,研究人员在设计时又试图重新回归简单的单分支网络结构,通过引入各种新型 CNN 技术来降低模型复杂度,同时提升计数精度.因此,减少分支数量,让计数模型既简单又有效,将是未来模型网络结构的设计方向.

此外,从分析中可知,注意力机制、空洞卷积、对抗生成网络、可形变卷积等 CNN 技术可以解决计数领域存在的多尺度、复杂背景干扰等问题,帮助提升密度图质量.因此,未来在设计网络时,可以考虑结合这些技术提升计数精度.

2 人群计数损失函数

损失函数的作用是评价模型的预测值与真实值 ground-truth 的一致程度,是模型训练中不可缺少的一部分.损失函数值越小,说明预测值越接近真实值,则模型的计数性能越好.在人群计数任务中,通过定义损失函数,可以将人群密度图的映射关系学习转化为一个最优化问题.常用的人群计数损失函数包括欧氏损失、结构相似性损失等.神经网络训练的目的就是是找到使损失函数值最小的网络参数值.

2.1 欧氏距离损失

早期绝大多数基于密度图进行人群计数的工作,例如跨场景计数模型[24]、MCNN[25]、CrowdNet[34]、Switch-CNN[38]、CSRNet[45]等方法,均采用像素级的欧氏距离作为模型损失函数,度量估计密度图与真实密度图之间的差距:

$$L(\theta)=\frac{1}{N}\sum_{i=1}^{N}\|F(X_i;\theta)-F_i\|^2, \tag{1}$$

其中,$F(X_i;\theta)$是参数为 θ 的映射函数,它将输入图像 X_i 映射到预测密度图,F_i 是真实密度图,N 为训练样本个数.

由于欧氏距离损失简单、训练速度快,且计数效果较好,早期得到了较为广泛的应用.但是欧氏距离损失的鲁棒性较差,很容易因为个别像素点的极端情况而影响整体的计数效果.此外,欧氏距离损失是取所有像素点的平均,并不关注图片的结构化信息.对于同一张图片,容易出现人群密集区域预测值偏小,而人群稀疏区域预测值偏大的问题,但是最终的平均结果却没有体现这些问题,从而导致生成的密度图模糊、细节不清晰.

2.2 结构相似性损失

由于欧氏距离损失不足以表达人的视觉系统对图片的直观感受,导致生成的密度图质量不高.为了克服欧氏距离损失的不足,SANet[42]提出了以结构相似性指标(structual similarity index)[31]为基础的结构相似性损失来度量密度图的质量.结构相似性指标是由 Wang 等人[31]提出的一种图像质量评价标准,记为 $SSIM$.不同于基于像素的误差评价标准,$SSIM$ 从图像的亮度、对比度和结构这 3 个方面度量图像相似性,并通过均值、方差、协方差 3 个局部统计量计算 2 张图像之间的相似度.$SSIM$ 的取值范围在$-1\sim1$之间,$SSIM$ 值越大,说明相似度越高.结构相似性指标 $SSIM$ 的计算方法为

$$SSIM=\frac{(2\mu_e\mu_g+c_1)(2\sigma_{eg}+c_2)}{(\mu_e^2+\mu_g^2+c_1)(\sigma_e^2+\sigma_g^2+c_2)}, \tag{2}$$

其中,g,e 分别代表真实密度图和生成密度图,μ_e 和 σ_e^2 表示生成密度图的均值和方差,μ_g 和 σ_g^2 表示真实密度图的均值和方差,σ_{eg}表示真实密度图和生成密度图之间的协方差.为了避免分母为 0 出现异常,平滑系数 c_1, c_2 设为极小的常数值.结构相似性损失 L_{SSIM}的计算方法为

$$L_{SSIM}=1-\frac{1}{N}\sum_{x\in X}SSIM(x), \tag{3}$$

其中,N 代表密度图的像素点数量,X 是生成密度图与真实密度图相同像素点位置对应的图像块集合.

实验表明,结构相似性损失确实可以提高生成密度图质量,相比于关注像素间差异的欧氏距离损失,结构相似性损失能够更好地关注图像间对应局部块的差异,从而更好地生成密度图.在后续的研究中,计数模型 SFCN[52]也采用了类似的做法.

为了进一步提高计数精度,许多学者对结构相似性损失进行改进.DSSINet[64]将空洞卷积融入结构相似性度量中,构建了一个空洞卷积网络 DMS-SSIM 用于计算结构相似性损失 L_{SSIM}.通过扩大 $SSIM$ 指标的感受野,每个像素点可以融合多尺度信息,使得在不同尺度下,可以输出局部区域的高质量密度图.

2.3　生成对抗损失

基于密度图的人群计数方法通常以单张静态的人群图像作为输入,然后输出1张与输入图像对应的人群密度图,这一目标本质上可视作一个图像转换问题(image-to-image translation). GAN[37]为解决图像转换问题提供了一个可行的思路,即可以通过生成网络和判别网络的不断博弈,进而使生成网络学习人群密度分布,生成密度图的质量逐渐趋好;判别网络也通过不断训练,提高本身的判别能力.损失函数作为生成对抗网络的关键,对于生成对抗网络训练、求解最优值的过程尤为重要.在人群计数领域,可以使用对抗损失函数,通过对抗的方式对生成图片进行矫正,由此避免出现密度图模糊问题.

CP-CNN[36]网络在欧氏距离损失的基础上,增加了生成对抗损失,提高了预测密度图的质量,其损失函数为

$$L_T=L_E+\lambda_a L_A, \tag{4}$$

$$L_E=\frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\|\varphi(X^{w,h})-(Y^{w,h})\|_2, \tag{5}$$

$$L_A=-\log(\varphi_D(\varphi(X))), \tag{6}$$

其中,L_T 是总损耗,L_E 是生成密度图与对应的真实密度图之间的像素级欧氏损失,λ_a 是权重因子,L_A 是对抗性损失,X 是尺寸为 $W\times H$ 的输入图像,Y 是 ground truth 密度图,φ 是由 DME 和 F-CNN 组成的网络,φ_D 是用于计算对抗损失的鉴别子网络.

在之后的人群计数算法研究中,对抗损失屡见不鲜.ACSCP[50]网络采用 U-Net 作为密度图生成器,并使用了对抗损失,可定义为

$$L_A(G,D)=E_{x,y\sim Pdata(x,y)}[\log D(x,y)]+E_{x\sim Pdata(x)}[\log(1-D(x,G(x)))], \tag{7}$$

其中,x 表示训练块,y 表示相应的 ground truth. G 是生成网络,D 是判别网络,G 试图最小化这个目标函数,而 D 试图将其最大化,通过判别网络与生成网络的一种联合训练得到最终的模型.RPNet[65]采用了一种对抗结构来提取拥挤区域的结构特征.

对抗损失对于密度图质量的提升有着显著作用,但对抗损失也有着难以训练的缺点.除这3种损失外,人群计数任务使用的损失函数还有很多,例如人群统计损失,但是每个损失函数各有优缺点,因此实际应用中,常常会联合多种损失,共同构建一个综合性的损失函数.

对于人群计数任务来说,密度图质量的优劣将直接影响计数性能.现有的损失函数虽然可以生成密度图,但是仍有许多亟待改进的地方.未来如何定义新的损失函数,以生成高质量的密度图也是该领域的一个研究重点.

3　ground-truth 密度图生成方法

为了训练计数网络,需要对人群图片中的目标进行标注.常见的做法是为图片中的每个人头标注中心坐标,然后再利用高斯核将坐标图转化为 ground-truth 人群密度图.ground-truth 密度图质量的高低,直接影响网络的训练结果.优质的 ground-truth 能使网络更好地学习到人群图片特征,计数网络的鲁棒性和适应性也会更好.近年来对 ground-truth 生成方法的研究从未停止过,ground-truth 密度图生成的关键在于如何选择高斯核,设置不同的高斯核对网络性能的影响很大,常用的3种高斯核设置方法为:

1) 几何自适应法

由于存在透视效应,在人群图片中远近景目标的尺寸差异较大,不同位置人头对应着不同大小的像素区域.因此要想生成更精确的人群密度图,就需要考虑透视畸变的影响,大人头应采用大尺寸高斯核,小人头则正好相反.MCNN[25]认为在拥挤的场景中,头部大小通常与相邻 k 个人中心点的距离有关.因此根据每个人与其 k 个邻居的平均距离来自适应地确定每个人的头部尺寸,也就是高斯卷积核的方差,然后将所有人头卷积后的结果进行累加,生成人群密度图.这种方法虽然考虑了多尺度差异,但是对于近处目标来说,人头间距远大于人头的实际尺寸,导致高斯核尺寸过大,近处人群的密度图会因为高斯核函数的值过小而消失.如图16[25]所示,密度图中只能看到远处有人群,而近处的人群极不明显.

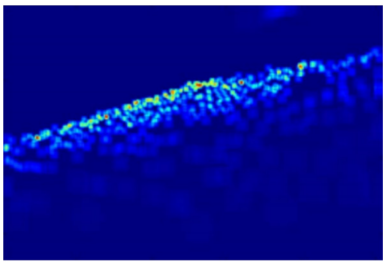
2) 固定高斯核法

该方法忽略了人头尺寸差异,以及自身与邻居的相似性,无论图片中哪个位置的人头均采用方差大小固定的高斯核对每个人头进行高斯模糊,采用固定高斯核的算法有 CP-CNN[36],其生成的 ground-truth 密度图如图17[36]所示.固定高斯核法解决了几何自适应法中的近处人头消失的问题,但是由于高斯核大小固定,对于远处人头来说,高斯核尺寸可能过大,使得远处人头出现重叠,降低了密度图质量.

3) 内容感知标注法

为解决方法1)2)存在的问题,Oghaz 等人[66]提出了一种通过内容感知标注技术生成密度图的方法.首先,用暴力最近邻(brute-force nearest neighbor)

(a) 原图

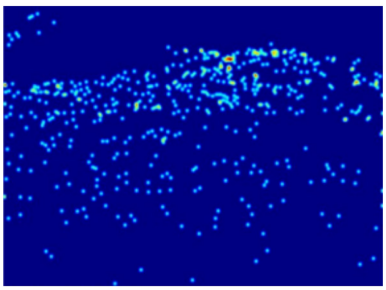


(b) ground-truth密度图

Fig. 16 Geometric adaptive method[25]

图 16 几何自适应法[25]

(a) 原图



(b) ground-truth密度图

Fig. 17 Fixed Gaussian kernel method[36]

图 17 固定高斯核法[36]

算法定位最近的头部,再用无监督分割算法 Chan-Vese 分割出头部区域,然后依据邻居头部的大小计算高斯核尺寸,其生成的密度图如图 18[66] 所示.该方法也是根据邻居情况灵活确定高斯核大小,但是与几何自适应法相比,它采用 brute-force 最近邻算法替代 k-d 树空间划分法(k-d tree space partitioning approach)来寻找最近邻,这样能确保寻找结果与实际相符.

(a) 来自ShanghaiTech数据集的样本

(b) 对应的内容感知密度图

Fig. 18 Content-aware annotation method[66]

图 18 内容感知标注法[66]

总之,高质量密度图是人群计数算法成功的基础和关键,因此 ground-truth 的生成方法将是人群计数领域未来的一个研究重点.

4 评价指标

为了对不同模型的准确率以及鲁棒性进行测评,需要有合适的评价指标.在人群计数领域,常用的评价指标有均方误差(mean squared error, *MSE*)、平均绝对误差(mean absolute error, *MAE*)和均方根误差(root mean squared error, *RMSE*),具体定义为

$$MSE=\frac{1}{N}\sum_{i=1}^{N}(R_i-R_i^{\mathrm{GT}})^2, \tag{8}$$

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|R_i-R_i^{\mathrm{GT}}|, \tag{9}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(R_i-R_i^{\mathrm{GT}})^2}, \tag{10}$$

其中,N 表示样本数量,R_i 表示预测划分比例,R_i^{GT} 表示真实划分比例.

MSE 和 *RMSE* 可以反映模型的鲁棒性,而 *MAE* 可以反映模型的准确性.通过对各个人群计数模型的评价指标 *MSE*,*MAE*,*RMSE* 的比较,可以评定各个计数模型的性能.

由于上述评价指标存在一定的局限性,很多研究人员进行了不同的改进,以适应不同的评价需求.例如,原始的 *MSE*,*MAE*,*RMSE* 只能度量全局鲁棒性和准确性,无法评价局部区域的计数性能,因此 Tian 等人[67]将 *MAE* 和 *RMSE* 扩展成块平均绝对误差(patch mean absolute error, *PMAE*),和块均方误差(patch mean squared error, *PMSE*),用于评价局部区域的计数效果.此外,对于基于密度图的人群计数算法来说,密度图质量高低对算法性能优

劣具有决定性作用,因此也可以采用已有的图像质量评价指标来衡量计数模型的性能.

5　人群计数数据集

随着人群计数算法研究的不断推进,该领域数据集的丰富性和针对性在逐步提高,图片数量以及质量也在进一步提升.表 1 按照时间顺序列举了具有代表性的人群计数数据集,不仅包括早期创建的经典人群计数数据集,也包括近年来新出现的数据集.这些数据集在拍摄视角、场景类型、平均分辨率、图像数量、每张图像所标注的目标数量等方面各有不同,总体呈现多样化特点.分 2 个部分对数据集进行简要介绍.

Table 1　Crowd Counting Datasets

表 1　人群计数数据集

数据集	年份	特点	平均分辨率	图像数量	标注总数	最少人数	最多人数	平均人数
UCSD[68]	2008	单一场景	158×238	2 000	49 885	11	46	25
Mall[69]	2012	单一场景	480×640	2 000	62 325	13	53	31
UCF_CC_50[21]	2013	多场景、密集	2 101×2 888	50	63 974	94	4 543	1 279
WorldExpo'10[24,70]	2015	多场景	576×720	3 980	199 923	1	253	50
ShanghaiTech PartA[25]	2016	多场景	589×868	482	241 677	33	3 139	501
ShanghaiTech PartB[25]	2016	多场景	768×1 024	716	88 488	9	578	123
CityUHK-X[71]	2017	多场景	384×512	3191	106 783			33
UCF-QNRF[72]	2018	多场景	2 013×2 902	1 535	1 251 642	49	12 865	815
SmartCity[73]	2018	多场景	1080×1920	50	369	1	14	7
Fudan-ShanghaiTech[74]	2019	视频人群计数	1 080×1 920	15 000	39 4081			27
Beijing-BRT[75]	2019	单一场景	640×360	1 280	16 795			13
DroneCrowd[76]	2019	多场景	1 080×1 920	33 600	4 864 280			145
DLR-ACD[77]	2019	多场景航拍图像	3 619×5 226	33	226 291	285	24 368	6 857
NWPU-Crowd[78]	2020	多场景	2 191×3 209	5 109	2 133 375	0	20 033	418
JHU-CROWD++[79]	2020	多场景	1 430×910	4 372	1 515 005	0	25 791	346
DISCO[80]	2020	视听,极端条件	1 080×1 920	1 935	170 270	1	709	88

5.1　经典人群计数数据集

本节主要介绍早期的经典人群计数数据集,包括 WorldExpo'10[24,70],ShanghaiTech[25],UCSD[68],Mall[69],UCF_CC_50[21],它们经常被看作是验证计数算法有效性的基准数据集,在近几年的人群计数算法研究中应用最为广泛[81].其中,UCSD,Mall,WorldExpo'10,ShanghaiTech PartB 主要针对人群稀疏场景,UCF_CC_50 和 ShanghaiTech PartA 则主要针对人群密集场景;在数据量方面,WorldExpo'10,UCSD,Mall 的数据量较大;UCSD,Mall,WorldExpo'10,ShanghaiTech PartB 数据集图片的分辨率是固定的,其他 2 个数据集中的图像分辨率是随机变化的.

数据集 UCSD 和 Mall 中的图像均来自相同的视频序列,在图像之间不存在视角变化.而其他经典数据集的图像样本来自不同的视频序列,视角和人群尺度的变化较大.表 2～7 通过度量准确性的 *MAE* 和度量鲁棒性的 *MSE* 这 2 个评价指标,比较了不同计数算法在各种经典人群计数数据集上的表现,分析了算法表现优劣的原因.所有实验数据均来自算法相关的参考文献.

UCSD 数据集[68]是最早创建的人群计数数据集之一.包含 2 000 帧从人行道视频监控中采集的图像,每帧的分辨率为 238×158.每隔 5 帧人工标注 1 次,其余帧中的行人位置则使用线性插值方式创建,最终标注了 49885 个行人目标.该数据集的人群密度相对较低,平均 1 帧约 15 人,由于数据是从一个位置采集的,场景和透视角度单一.

表 2 列出了不同计数网络在 UCSD 数据集上的实验结果,由表可知,随着时间推移,算法性能在不断提升.评价指标 *MAE* 和 *MSE* 排名前 3 的算法分别是 E3D[82],PACNN[83],PaDNet[67].其中,PaDNet

提出了针对不同密度人群进行计数的泛密度计数方法;E3D 中最主要的创新是结合了 3D 卷积核来编码局部时空特征,该网络主要针对视频中的人群计数,但在图像数据集上依然取得了良好的性能;PACNN 将透视信息集成到密度回归中,以方便融合目标比例变化相关的特征.其次,考虑了局部注意力的网络 ADCrowdNet 以及考虑尺度多样性的计数网络 MCNN,SANet,ACSCP 等性能表现均较好.由此分析可知,对于较为稀疏的人群场景,场景的尺度多样性是最应该考虑的要素,而且将局部信息作为额外的辅助信息,将有助于提升计数性能.

Table 2 Comparison of Crowd Counting Networks on UCSD

表 2 不同计数网络在 UCSD 数据集上的性能对比

年份	期刊/会议	方法	*MAE*	*MSE*
2015	CVPR	文献[24]	1.6	3.31
2016	ECCV	Hydra-CNN[84]	1.65	
	ECCV	CNN-Boosting[85]	1.1	
	CVPR	MCNN[25]	1.07	1.35
2017	ICCV	ConvLSTM-nt[86]	1.73	3.52
	CVPR	Switch-CNN[38]	1.62	2.1
	ICCV	ConvLSTM[86]	1.3	1.79
	ICCV	Bidirectional ConvLSTM[86]	1.13	1.43
2018	CVPR	CSRNet[45]	1.16	1.47
	CVPR	ACSCP[50]	1.04	1.35
	ECCV	SANet[42]	1.02	1.29
	TIP	BSAD[87]	1	1.4
2019	WACV	SPN[88]	1.03	1.32
	ICCV	SPANet+SANet[89]	1	1.28
	CVPR	ADCrowdNet(DME)[55]	0.98	1.25
	BMVC	E3D[82]	0.93	1.17
	CVPR	PACNN[83]	0.89	1.18
	TIP	PaDNet[67]	0.85	1.06

Mall 数据集[69]是由安装在购物中心的监控摄像头采集而来,共包含 2 000 帧分辨率为 320×240 的图像样本,标注了行人目标 6 000 个,前 800 帧用于训练,剩余 1 200 帧用于测试.该数据集场景复杂,人群密度以及光照条件差异较大,而且图像存在严重的透视畸变,目标的表观特征和尺度差异较大.与 UCSD 数据集相比,Mall 数据集的人群密度相对较高,然而这 2 个数据集由于都在固定地点拍摄,所以均存在场景单一的问题,无法反应室内场景的实际状况.此外,该数据集还存在由场景对象,例如摊位、植物等,引起的严重遮挡,这一特性增加了人群计数的难度.

表 3 列出不同计数网络在 Mall 数据集上的运行结果.其中按照 *MAE* 和 *MSE* 排名,表现最好的算法包括 DecideNet[41], DRSAN[90], E3D[82], SAAN[44].其中,SAAN 网络利用了多尺度注意力机制;E3D 考虑了局部时空特征;DecideNet 中有检测分支,更加关注局部信息;DRSAN 主要是通过区域精细化过程自适应地解决了可学习的空间变换模块中的 2 个问题,来更好地适应摄像机的不同视角变化,这种方法很好地考虑到了图片中不同人群的尺度特征.

Table 3 Comparison of Crowd Counting Networks on Mall

表 3 不同计数网络在 Mall 数据集上的性能对比

年份	期刊/会议	方法	*MAE*	*MSE*
2012	BMVC	文献[69]	3.15	15.7
2016	ECCV	CNN-Boosting[85]	2.01	
2017	ICCV	ConvLSTM-nt[86]	2.53	11.2
	ICCV	ConvLSTM[86]	2.24	8.5
	ICCV	Bidirectional ConvLSTM[86]	2.1	7.6
2018	CVPR	DecideNet[41]	1.52	1.9
	IJCAI	DRSAN[90]	1.72	2.1
2019	BMVC	E3D[82]	1.64	2.13
	WACV	SAAN[44]	1.28	1.68

相较于其他数据集,Mall 与 UCSD 这 2 个数据集的人群密度均较小.由这 2 个数据集中各模型的实验结果可得,对于较为稀疏的场景,我们应该更关注人群局部特征和多尺度特征,而空洞卷积在稀疏场景的效果并没有特别突出.

MCNN 网络在提出多阵列网络结构的同时,还创建了人群计数数据集 ShanghaiTech.该数据集包含 1 198 张图片,分为 part *A* 和 part *B* 这 2 个部分,共标注了 330 165 个头部位置.人群分布较为密集的 Part *A* 包含 300 张训练图片,182 张测试图片,图像分辨率是变化的;人群分布较为稀疏的 Part *B* 包含 400 张训练图片,316 张测试图片,图像分辨率固定不变.总体上看,在 ShanghaiTech 数据集上进行精确计数是具有挑战性的,因为该数据集无论是场景类型,还是透视角度和人群密度都变化多样.

表 4 和表 5 为各计数网络在 ShanghaiTech Part *A* 和 Part B 上的运行结果.在 Part *A* 上,性能表现较好的网络包括 SPANet+SANet,S-DCNet,PGCNet,ADSCNet.其中,SPANet 将空间上下文融入人群计数,并与考虑尺度特征的 SANet 相结合,

Table 4　Comparison of Crowd Counting Networks on ShanghaiTech Part *A*

表 4　不同计数网络在 ShanghaiTech Part *A* 数据集上的性能对比

年份	期刊/会议	方法	*MAE*	*MSE*
2016	CVPR	MCNN[25]	110.2	173.2
2017	AVSS	CMTL[91]	101.3	152.4
	CVPR	Switch-CNN[38]	90.4	135
	ICIP	MSCNN[92]	83.8	127.4
	ICCV	CP-CNN[38]	73.6	106.4
2018	AAAI	TDF-CNN[93]	97.5	145.1
	WACV	SaCNN[73]	86.8	139.2
	CVPR	ACSCP[50]	75.7	102.7
	CVPR	D-ConvNet-v1[94]	73.5	112.3
	CVPR	IG-CNN[39]	72.5	118.2
	CVPR	L2R[95](Multi-task, Query-by-example)	72	106.6
	IJCAI	DRSAN[90]	69.3	96.4
	ECCV	ic-CNN(one stage)[96]	69.8	117.3
	ECCV	ic-CNN(two stages)[96]	68.5	116.2
	CVPR	CSRNet[45]	68.2	115
	ECCV	SANet[42]	67	104.5
2019	CVPRW	GSP(one stage, efficient)[97]	70.7	103.6
	AAAI	GWTA-CCNN[98]	154.7	229.4
	ICASSP	ASD[99]	65.6	98
	ICCV	CFF[63]	65.2	109.4
	CVPR	SFCN[52]	64.8	107.5
	ICCV	SPN+L2SM[100]	64.2	98.4
	CVPR	TEDnet[101]	64.2	109.1
	CVPR	ADCrowdNet (AMG-bAttn-DME)[55]	63.2	98.9
	CVPR	PACNN[83]	66.3	106.4
	CVPR	PACNN+CSRNet	62.4	102
	CVPR	CAN[48]	62.3	100
	TIP	HA-CCN[102]	62.9	94.9
	ICCV	BL[103]	62.8	101.8
	WACV	SPN[88]	61.7	99.5
	ICCV	DSSINet[64]	60.63	96.04
	ICCV	MBTTBF-SCFB[104]	60.2	94.1
	ICCV	RANet[105]	59.4	102
	ICCV	SPANet +SANet[42]	59.4	92.5
	TIP	PaDNet[67]	59.2	98.1
	ICCV	S-DCNet[106]	58.3	95
	ICCV	PGCNet[107]	57	86
2020	AAAI	DUBNet[108]	64.6	106.8
	CVPR	ADSCNet[109]	55.4	97.7

Table 5　Comparison of Crowd Counting Networks on ShanghaiTech Part *B*

表 5　不同计数网络在 ShanghaiTech Part *B* 数据集上的性能对比

年份	期刊/会议	方法	*MAE*	*MSE*
2016	CVPR	MCNN[25]	26.4	41.3
2017	ICIP	MSCNN[92]	17.7	30.2
	AVSS	CMTL[91]	20	31.1
	CVPR	Switch-CNN[38]	21.6	33.4
	ICCV	CP-CNN[38]	20.1	30.1
2018	TIP	BSAD[87]	20.2	35.6
	WACV	SaCNN[73]	16.2	25.8
	CVPR	ACSCP[50]	17.2	27.4
	CVPR	CSRNet[45]	10.6	16
	CVPR	IG-CNN[39]	13.6	21.1
	CVPR	D-ConvNet-v1[94]	18.7	26
	CVPR	DecideNet[41]	21.53	31.98
	CVPR	DecideNet+R3	20.75	29.42
	CVPR	L2R[95](Multi-task, Query-by-example)	14.4	23.8
	CVPR	L2R(Multi-task, Keyword)	13.7	21.4
	IJCAI	DRSAN[90]	11.1	18.2
	AAAI	TDF-CNN[93]	20.7	32.8
	ECCV	ic-CNN[96](one stage)	10.4	16.7
	ECCV	ic-CNN(two stages)	10.7	16
	ECCV	SANet[42]	8.4	13.6
2019	CVPRW	GSP[97](one stage, efficient)	9.1	15.9
	WACV	SPN[88]	9.4	14.4
	ICCV	PGCNet[107]	8.8	13.7
	ICASSP	ASD[99]	8.5	13.7
	CVPR	TEDnet[101]	8.2	12.8
	TIP	HA-CCN[102]	8.1	13.4
	TIP	PaDNet[67]	8.1	12.2
	ICCV	CAN[48]	7.9	12.9
	CVPR	RANet[105]	7.8	12.2
	CVPR	ADCrowdNet (AMG-attn-DME)[55]	7.7	12.9
	AAAI	DUBNet[108]	7.7	12.5
	CVPR	ADCrowdNet (AMG-DME)[55]	7.6	13.9
	CVPR	SFCN[52]	7.6	13
	CVPR	PACNN[83]	8.9	13.5
	CVPR	PACNN+CSRNet	7.6	11.8
	ICCV	BL[103]	7.7	12.7
	ICCV	CFF[63]	7.2	12.2
	ICCV	SPN+L2SM	7.2	11.1
	ICCV	DSSINet[64]	6.85	10.34
	ICCV	S-DCNet[106]	6.7	10.7
	ICCV	SPANet[89]+SANet[42]	6.5	9.9
2020	CVPR	ADSCNet[109]	6.4	11.3

得到的模型拥有很好的鲁棒性;S-DCNet 是一种空间分而治之的网络,通过获取局部特征来实现图片整体的计数;PGCNet 克服了由于透视效应而产生的场景尺度变化,获得了较好的计数性能;ADSCNet 提出了一种具有自我校正监督的自适应空洞网络计数算法,对空洞卷积进行改进,使其可以根据图片场景及尺度变换而自适应地选择不同的空洞卷积. Part *B* 部分去除了 PGCNet 网络,增加了 DSSINet 网络的比较.该网络引入了基于空洞卷积的结构化损失,能更好地体现图片中的局部损失.

由数据对比可知,稀疏场景的人群计数效果明显优于密集场景的人群计数效果.因此,在未来的研究中,密集场景人群计数将依然是该领域的研究重点.

UCF_CC_50 数据集是第 1 个真正意义上具有挑战性的大规模人群计数数据集.包含了 50 张不同分辨率的图片,内容涵盖了音乐会、抗议活动、体育场和马拉松比赛等不同场景.整个数据集中共标注了 63 075 个头部位置,其中每张图片包含的人数从 94 到 4 543 不等,密度等级变化极大.

表 6 是不同计数网络在 UCF_CC_50 数据集上的运行结果.在性能指标 *MAE* 和 *MSE* 上排名前 4 的方法包括 PaDNet,SPN+L2SM,ASD,CAN,其中 PaDNet 表现最好,其采用的融合图像不同密度的泛密度方法恰好适用于 UCF_CC_50 这种人群密度变化范围较广的数据集;SPN 提出了一个比例金字塔网络(SPN),该网络采用共享的单个深列结构,并通过尺度金字塔模块提取高层的多种尺度信息,其与 L2SM 结合,更加关注于人群多尺度信息;ASD 是一个场景自适应框架,能够更好地对可变人群场景进行计数;CAN 采用了空间金字塔池化结构处理人群多尺度特征,在此数据集上获得了较好的鲁棒性.

由表 6 和分析可得,空洞卷积和多尺度网络在此数据集上的表现效果更好.相比 UCSD,Mall,ShanghaiTech,UCF_CC_50 这 4 个数据集的效果,Switch-CNN 网络的性能提升明显,而 UCF_CC_50 数据集的场景更为复杂,由此可得,Switch 结构增加了模型的鲁棒性,多阵列模型的效果明显好于单列计数网络模型.

早期的人群计数方法主要关注单一场景的计数问题,导致模型跨场景计数性能较差,为此 Zhang 等人构建了采集于上海世界博览会的人群计数数据集 WorldExpro'10.该数据集由 108 个监控探头采集的 1 132 个视频序列组成,通过从不同位置的摄像头采集数据,确保了场景类型的多样性.其中,3 980 帧图像进行了人工标注,每帧的分辨率为 576×720,

Table 6 Comparison of Crowd Counting Networks on UCF_CC_50

表 6 不同计数网络在 UCF_CC_50 数据集上的性能对比

年份	期刊/会议	方法	*MAE*	*MSE*
2013	CVPR	文献[21]	468	590.3
2015	CVPR	文献[24]	467	498.5
2016	ACM MM	CrowdNet[34]	452.5	
	CVPR	MCNN[25]	377.6	509.1
	ECCV	CNN-Boosting[85]	364.4	
	ECCV	Hydra-CNN[84]	333.73	425.26
2017	ICIP	MSCNN[92]	363.7	468.4
	AVSS	CMTL[91]	322.8	397.9
	CVPR	Switch-CNN[38]	318.1	439.2
	ICCV	CP-CNN[36]	298.8	320.9
	ICCV	ConvLSTM-nt[86]	284.5	297.1
2018	TIP	BSAD[87]	409.5	563.7
	AAAI	TDF-CNN[93]	354.7	491.4
	WACV	SaCNN[73]	314.9	424.8
	CVPR	IG-CNN[39]	291.4	349.4
	CVPR	ACSCP[50]	291	404.6
	CVPR	L2R(Multi-task, Query-by-example)[95]	291.5	397.6
	CVPR	L2R(Multi-task, Keyword)	279.6	388.9
	CVPR	D-ConvNet-v1[94]	288.4	404.7
	CVPR	CSRNet[45]	266.1	397.5
	ECCV	ic-CNN[96](two stages)	260.9	365.5
	ECCV	SANet[42]	258.4	334.9
	IJCAI	DRSAN[90]	219.2	250.2
2019	AAAI	GWTA-CCNN[98]	433.7	583.3
	WACV	SPN[88]	259.2	335.9
	CVPR	ADCrowdNet[55](DME)	257.1	363.5
	TIP	HA-CCN[102]	256.2	348.4
	CVPR	TEDnet[101]	249.4	354.5
	CVPR	PACNN[83]	267.9	357.8
	CVPR	PACNN+CSRNet	241.7	320.7
	ICCV	RANet[105]	239.8	319.4
	ICCV	MBTTBF-SCFB[104]	233.1	300.9
	ICCV	BL[103]	229.3	308.2
	ICCV	DSSINet[64]	216.9	302.4
	CVPR	SFCN[52]	214.2	318.2
	CVPR	CAN[48]	212.2	243.7
	ICCV	S-DCNet[106]	204.2	301.3
	ICASSP	ASD[99]	196.2	270.9
	ICCV	SPN+L2SM	188.4	315.3
	TIP	PaDNet[67]	185.8	278.3

总共标注了 199 923 个目标位置.该数据集被划分为 2 个部分,来自 103 个场景的 1 127 个视频序列作为训练集,其余 5 个场景的数据作为测试集.每个测试场景由 120 个标记帧组成,观众数量从 1～220 不等.虽然尝试捕捉不同密度级别的场景,但在测试集中,多样性仅限于 5 个场景,人群数量最大被限制在 220 个.因此,该数据集不足以评估为极端密集场景设计的人群计数算法.

表 7 列出了不同计数网络在 WorldExpo'10 数据集上的 *MAE* 值.其中,采用融入空洞率的结构性损失的网络 DSSINet 的平均性能最好;融合了图像上下文信息的 CP-CNN 和 CAN 网络对于多角度、多尺度场景的效果较好;在 S_2,S_3,S_5 场景中,空洞卷积的表现都是最好;此外,包含空洞卷积和可形变卷积的 ADCrowdNet 在 S_4 场景下得到了很好的计数效果;加入透视引导卷积(PGC)的网络 PGCNet 在场景 S_3 上获得很好的效果,可见尺度信息对于场景 S_3 的重要性.由分析可知,在人群相对稀疏的场景下,空洞卷积可以在不同场景下取得很好的效果,结构性损失在多个场景的计数中都表现良好.

Table 7　Comparison of Crowd Counting Networks on WorldExpo'10

表 7　不同计数网络在 WorldExpo'10 数据集上的性能对比

年份	期刊/会议	方法	场景 *MAE* 值					5 组场景 *MAE* 的平均值
			S_1	S_2	S_3	S_4	S_5	
2015	CVPR	Zhang 2015[24]	9.8	14.1	14.3	22.2	3.7	12.8
2016	CVPR	MCNN[25]	3.4	20.6	12.9	13.0	8.1	11.6
2017	ICIP	MSCNN[92]	7.8	15.4	14.9	11.8	5.8	11.1
	ICCV	ConvLSTM-nt[86]	8.6	16.9	14.6	15.4	4.0	11.9
	CVPR	Switch-CNN[38]	4.4	15.7	10.0	11.0	5.9	9.4
	ICCV	CP-CNN[36]	2.9	14.7	10.5	10.4	5.8	8.9
2018	AAAI	TDF-CNN[93]	2.7	23.4	10.7	17.6	3.3	11.5
	CVPR	IG-CNN[39]	2.6	16.1	10.15	20.2	7.6	11.3
	ECCV	ic-CNN[96]	17.0	12.3	9.2	8.1	4.7	10.3
	CVPR	DecideNet[41]	2.0	13.14	8.9	17.4	4.75	9.2
	CVPR	CSRNet[45]	2.9	11.5	8.6	16.6	3.4	8.6
	WACV	SaCNN[73]	2.6	13.5	10.6	12.5	3.3	8.5
	ECCV	SANet[42]	2.6	13.2	9.0	13.3	3.0	8.2
	CVPR	ACSCP[50]	2.8	14.05	9.6	8.1	2.9	7.5
2019	ICCV	PGCNet[107]	2.5	12.7	8.4	13.7	3.2	8.1
	CVPR	TEDnet[101]	2.3	10.1	11.3	13.8	2.6	8.0
	CVPR	PACNN[83]	2.3	12.5	9.1	11.2	3.8	7.8
	CVPR	ADCrowdNet(AMG-bAttn-DME)[55]	1.7	14.4	11.5	7.9	3.0	7.7
	CVPR	ADCrowdNet(AMG-attn-DME)[55]	1.6	13.2	8.7	10.6	2.6	7.3
	CVPR	CAN[48]	2.9	12.0	10.0	7.9	4.3	7.4
	CVPR	CAN(ECAN)[48]	2.4	9.4	8.8	11.2	4.0	7.2
	ICCV	DSSINet[64]	1.57	9.51	9.46	10.35	2.49	6.7

5.2　其他人群计数数据集

本节主要介绍近几年新出现的人群计数数据集,包括 DISCO[80],NWPU-Crowd[78],UCF-QNRF[72],JHU-CROWD++[79]等.这些数据集的出现在一定程度上缓解了经典数据集存在的场景单一、图像质量不高、数据规模过小等问题.

CityUHK-X[71]是由香港城市大学 VISAL 实验室创建的人群计数数据集,包含来自 55 个场景的 3 191 张图片,其中训练集由来自 43 个场景的 2 503 张图片构成,共标注了 78 592 个实例;测试集则由来自 12 个场景的 688 张图片构成,共标注了 28 191 个实例.该数据集的特色在于将拍摄角度和高度作为场景上下文辅助信息,然后卷积核权重随之自适应变化,以提升计数准确性.

UCF-QNRF[72]数据集具有场景丰富,视角、密度以及光照条件均变化多样的特点,是一个非常

具有挑战性的人群计数数据集.它共包含 1 535 张密集人群场景图片的数据集,其中训练集 1 201 张图像,测试集 334 张图像,共有 1 251 642 个目标被标注,由于标注数量众多,该数据集适合采用深度卷积神经网络进行训练.此外,该数据集图片的分辨率很高,因此在训练过程中可能出现内存不足.

SmartCity 数据集[73]主要用于验证计数模型在人群稀疏场景中的有效性.现有的人群计数数据集主要采集自人群密集场景,基于密集场景数据集训练出来的网络难以保证对稀疏场景的泛化性.为此,腾讯优图从 10 种不同城市场景中,采集了 50 张图片.这些图像包括室内和室外 2 种场景,均采用了很高的视角拍摄,图像中行人稀少,平均数量只有 7.4 个.

Fudan-ShanghaiTech 数据集[74]为进行基于视频的人群计数算法的研究提供了数据.已有的数据集主要面向基于图像的人群计数,为了更好地推动基于视频的人群计数算法的研究,研究人员从 13 个不同场景中捕获了 100 个视频,这些视频包含 150 000 帧图片,共标注了 394 081 个实体.其中训练集包含 60 个视频,共 9 000 帧图像;测试集包含剩余的 40 个视频,共 6 000 帧图像.

Beijing-BRT[75]是一个智能交通领域的人群计数数据集,包含 1 280 张从北京快速公交(bus rapid transit, BRT)采集的图片,其中 720 张用于训练,560 张用于测试.每张图片像素大小为 640×360,共标注了 16 795 个行人目标.该数据集与实际情况比较相符,涵盖了各种光照条件,而且时间跨度比较大,从白天到夜晚均有图像数据,因此基于该数据集训练出来的计数模型泛化能力较强.

DroneCrowd[76]数据集是由天津大学机器学习和数据挖掘实验室的 AISKYEYE 团队通过无人机拍摄创建,由 288 段视频剪辑和 10 209 张静态图像构成.数据集图像涵盖不同的地理位置、标注目标类型以及密集程度,变化范围广泛,很具有代表性.不仅可以用于视频或图像的目标检测和跟踪任务的研究,也可以用于人群计数任务的研究.

DLR-ACD[77]是一个包括 33 张航拍图像的人群计数数据集,数据集图片来自不同的城市场景,包括运动会、露天集会、庆典等存在大量人员聚集的场合,采用安装在直升机上的摄像头直接拍摄,所得到图片的空间分辨率在 4.5cm/pixel～15cm/pixel 之间变化.对图片中的每个人进行了手工标注,共标注了 226 291 个实例.

NWPU-Crowd[78]是目前人群计数领域最大的数据集,拥有 5 109 张图片和 2 133 238 个标注实体,而且单张图片的标注实体数量变化范围非常大,对计数任务来说虽然挑战极大,但也有助于提升训练模型的泛化性;该数据集的图片分辨率较高,有利于计数准确性的提升.此外,部分图片的目标标注数量为 0,这些负样本的加入有助于提升训练模型的鲁棒性.该数据集还提供了一个平台,供研究人员进行计数模型的性能比较.

JHU-CROWD++[79]也是一个非常具有挑战性的大规模人群计数数据集,包含 4 372 张图像,共计 151 万个标注,所有图像采集于各种不同的场景和环境条件,甚至包括一些基于恶劣天气变化和光照变化的图像,覆盖面很广.此外,该数据集与 NWPU-Crowd 类似,引入负样本,增强训练模型的鲁棒性,同时对人头采用了多种标注方式,包括点、近似边界框、模糊级别等,为不同计数算法的训练提供支撑条件.

DISCO[80]是一个极具特色的大规模人群计数基准数据集,包含 1 935 张图片和 170 270 个带标注的实体,每张图片对应一段时长为 1 s 的音频剪辑.最终通过声音和图像的共同作用,实现视听人群计数.

5.3 讨 论

随着人群计数领域受关注程度的提高和研究的深入,人群计数数据集也逐渐增多,主要呈现 5 个特点:

1) 在场景方面,由早期的单一化向多样化演变,部分数据集甚至包含极端条件下的场景图像,由此训练出来的模型跨场景迁移能力更强.

2) 在图像分辨率方面,早期场景图像分辨率较低,图像质量较差,人群特征不明显,不利于模型训练.随着视频设备发展,图像分辨率不断增强,计数的准确率不断攀升.

3) 在视角和尺度方面,变化范围更广,更贴近现实情况,有助于提升计数模型的泛化性和实用性.

4) 数据规模不断增强,更加适合采用深度学习方法进行训练.此外,数据规模的增强降低了模型的过拟合风险.

5) 样本类型更加丰富.早期人群计数数据集中每张图片均有人,标注数量至少为 1,无人负样本的加入可以帮助模型过滤噪声,提升鲁棒性.

此外,分析实验数据可知,采用了注意力机制、空洞卷积以及额外辅助信息的网络往往性能较好.主要是由于注意力机制可以帮助计数网络专注于有效信息,排除噪声干扰;空洞卷积可以在不增加模型

参数和计算量的前提下,扩大感受野,捕获多尺度信息,保留图像更多细节;而额外的辅助信息,例如视角,可以辅助处理多尺度问题.

目前,虽然已经构建了各种人群计数数据集,为验证计数算法的有效性提供了数据支撑,但是在场景多样性、标注准确性以及视图多样性等方面依然无法满足实验需求,这些也将是今后构建数据集时,需要重点考虑的问题.对于某些场景来说,采集图像非常困难且无法实现准确标注,此时可以考虑通过人工合成的方法生成图片,例如 GCC[52] 通过生成对抗网络人工合成了大量图片,为构建数据集提供了新思路.

6　总结与展望

近年来人群计数算法研究,尤其是基于深度学习的人群计数算法研究已经取得了明显进展,但是要在智能视频监控系统中真正应用并普及仍然面临许多挑战[110],例如相互遮挡、透视扭曲、照明变化以及天气变化等因素,都会影响计数的准确性.今后可以针对这些问题,从 3 个方面开展工作:

1) 遮挡条件下的人群计数.随着人群密度增大,人与人之间会产生遮挡,下一步可以研究在遮挡条件下如何进行人群计数同时获取人群分布等细节信息.

2) 特殊天气条件下的人群计数.现实中天气变化多样,不仅有风和日丽,亦有风雨交加.特殊天气下的数据采集和标注较困难[111],研究相对较少.下一步可以重点关注特殊天气条件下的人群计数问题,同时构建相应的数据集.

3) 昏暗光照条件下的人群计数.在光照不足的环境中,摄像头拍摄的图片往往较模糊,人头无法清晰辨认,下一步可以研究昏暗光照条件下人群计数问题的处理方法.

本文针对近年来人群计数领域的相关论文进行调研,在简单回顾传统人群计数算法之后,对基于深度学习的人群计数方法进行了系统性的总结和介绍,并给出了这个方向未来的研究趋势,希望可以给相关研究人员提供一些参考.

作者贡献声明:余鹰负责综述选题确定、文章主体撰写和修订等工作,并指导和督促完成相关文献资料的收集整理以及论文初稿的写作;朱慧琳和钱进参与文献资料的分析、整理和论文初稿的写作;潘诚参与了文献资料的收集以及部分图表数据的绘制;苗夺谦负责提出论文修改意见,指导论文写作.

参　考　文　献

[1] Zhou Bolei, Tang Xiaoou, Wang Xiaogang. Learning collective crowd behaviors with dynamic pedestrian-agents [J]. International Journal of Computer Vision, 2015, 111: 50-68

[2] Saxena S, Brémond F, Thonnat M, et al. Crowd behavior recognition for video surveillance [C] //Proc of the Int Conf on Advanced Concepts for Intelligent Vision Systems. Berlin: Springer, 2008: 970-981

[3] Ko T. A survey on behavior analysis in video surveillance for homeland security applications [C] //Proc of the Applied Imagery Pattern Recognition Workshop. Piscataway, NJ: IEEE, 2008: DOI:10.1109/AIPR.2008.4906450

[4] Huang Lida, Chen Tao, Wang Yan, et al. Congestion detection of pedestrians using the velocity entropy: A case study of Love Parade 2010 disaster [J]. Physica A: Statistical Mechanics and its Applications, 2015, 440: 200-209

[5] Li Weixin, Mahadevan V, Vasconcelos N. Anomaly detection and localization in crowded scenes [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36(1): 18-32

[6] Chaker R, Aghbari Z A, Junejo I N. Social network model for crowd anomaly detection and localization [J]. Pattern Recognition, 2017, 61: 266-281

[7] Benabbas Y, Ihaddadene N, Djeraba C. Motion pattern extraction and event detection for automatic visual surveillance [J/OL]. EURASIP Journal on Image and Video Processing. 2011 [2020-12-23], https://doi.org/10.1155/2011/163682

[8] Abdelghany A, Abdelghany K, Mahmassanic H, et al. Modeling framework for optimal evacuation of large-scale crowded pedestrian facilities [J]. European Journal of Operational Research, 2014, 237(3): 1105-1118

[9] Almeida J E, Rowwetti R J F, Coelho A L. Crowd simulation modeling applied to emergency and evacuation simulations using multi-agent systems [EB/OL]. (2013-03-15) [2020-12-23]. https://arxiv.org/abs/1303.4692

[10] Shao Jing, Kang Kai, Loy C C, et al. Deeply learned attributes for crowded scene understanding [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 4657-4666

[11] Zhou Bolei, Wang Xiaogang, Tang Xiaoou. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2012: 2871-2878

[12] Marsden M, Mcguinness K, Little S, et al. People, penguins and petri dishes: Adapting object counting models to new visual domains and object types without forgetting [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8070-8079

[13] Guerrero-Gómez-Olmedo R, Torre-Jiménez B, López-Sastre R, et al. Extremely overlapping vehicle counting [G] //LNCS 9117: Proc of the Iberian Conf on Pattern Recognition and Image Analysis. Berlin: Springer, 2015: 423-431

[14] Gao Can, Zhou Jie, Miao Duoqian, et al. Three-way decision with co-training for partially labeled data [J]. Information Sciences, 2021, 544: 500-518

[15] Gavrila D M, Philomin V. Real-time object detection for "smart" vehicles [C] //Proc of the 7th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 1999: 87-93

[16] Topkaya I S, Erdogan H, Porikli F, Counting people by clustering person detector outputs [C] //Proc of the 11th IEEE Int Conf on Advanced Video and Signal Based Surveillance. Piscataway, NJ: IEEE, 2014:313-318

[17] Li Min, Zhang Zhaoxiang, Huang Kaiqi, et al. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection [C] // Proc of the 19th IEEE Int Conf on Pattern Recognition. Piscataway, NJ: IEEE, 2008: DOI: 10.1109/ICPR.2008.4761705

[18] Leibe B, Seemann E, Schiele B. Pedestrian detection in crowded scenes [C] //Proc of the IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2005: 878-885

[19] Enzweiler M, Gavrila D M. Monocular pedestrian detection: Survey and experiments [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(12): 2179-2195

[20] Chan A B, Vasconcelos N. Bayesian poisson regression for crowd counting [C] //Proc of the 12th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2009: 545-551

[21] Idrees H, Saleemi I, Seibert C, et al. Multi-source multi-scale counting in extremely dense crowd images [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2013: 2547-2554

[22] Chen Ke, Gong Shaogang, Tao Xiang, et al. Cumulative attribute space for age and crowd density estimation [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2013: 2467-2474

[23] Lempitsky V, Zisserman A. Learning to count objects in images [C] //Proc of the 23rd Annual Conf on Neural Information Processing Systems. New York: ACM, 2010: 1324-1332

[24] Zhang Cong, Li Hongsheng, Wang Xiaogang, et al. Cross-scene crowd counting via deep convolutional neural networks [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 833-841

[25] Zhang Yingying, Zhou Desen, Chen Siqin, et al. Single-image crowd counting via multi-column convolutional neural network [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 589-597

[26] Ge Weina, Collins R T. Marked point processes for crowd counting [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2009: 2913-2920

[27] Rodriguez M, Laptev I, Sivic J, et al. Density-aware person detection and tracking in crowds [C] //Proc of the Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2011: 2423-2430

[28] Chen Sheng, Fern A, Todorovic S. Person count localization in videos from noisy foreground and detections [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1364-1372

[29] Zhang Hao, Wu Jianxin. A survey on unsupervised image retrieval using deep features [J]. Journal of Computer Research and Development, 2018, 55(9): 1829-1842 (in Chinese)
(张皓, 吴建鑫. 基于深度特征的无监督图像检索研究综述 [J]. 计算机研究与发展, 2018, 55(9): 1829-1842)

[30] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11):2278-2324

[31] Wang Zhou, Bovik A C, Sheikh H R, et al. Image quality assessment: From error visibility to structural similarity [J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612

[32] Wang Chuan, Zhang Hua, Yang Liang, et al. Deep people counting in extremely dense crowds [C] //Proc of the 23rd ACM Int Conf on Multimedia. New York: ACM, 2015: 1299-1302

[33] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C] // Proc of the 25th Int Conf on Neural Information Processing systems. Cambridge, MA: MIT Press, 2012: 1097-1105

[34] Boominathan L, Kruthiventi S S S, Babu R V, et al. CrowdNet: A deep convolutional network for dense crowd counting [C] //Proc of the 24th ACM Int Conf on Multimedia. New York: ACM, 2016: 640-644

[35] Ciregan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2012: 3642-3649

[36] Sindagi V A, Patel V M. Generating high-quality crowd density maps using contextual pyramid CNNs [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 1861-1870

[37] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C] //Proc ofthe 27th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 2672-2680

[38] Sam D B, Surya S, Babu R V, et al. Switching convolutional neural network for crowd counting [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition Honolulu. Piscataway, NJ: IEEE, 2017: 4031-4039

[39] Sam D B, Sajjan N N, Babu R V, et al. Divide and grow: Capturing huge diversity in crowd images with incrementally growing CNN [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 3618-3626

[40] 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

[41] Liu Jiang, Gao Chenqiang, Meng Deyu, et al. DecideNet: Counting varying density crowds through attention guided detection and density estimation [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 5197-5206

[42] Cao Xinkun, Wang Zhipeng, Zhao Yanyun, et al. Scale aggregation network for accurate and efficient crowd counting [G] //LNCS 11209: Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 734-750

[43] Szegedy C, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1-9

[44] Hossain M, Hosseinzadeh M, Chanda O, et al. Crowd counting using scale-aware attention networks [C] //Proc of the IEEE Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2019: 1280-1288

[45] Li Yuhong, Zhang Xiaofan, Chen Deming. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 1091-1100

[46] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions [EB/OL]. (2016-04-30) [2020-12-23]. https://arxiv.org/abs/1511.07122

[47] Dai Jifeng, Qi Haozhi, Xiong Yuwen, et al. Deformable convolutional networks [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 764-773

[48] Liu Weizhe, Salzmann M, Fua P. Context-aware crowd counting [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 5099-5108

[49] Isola P, Zhu Junyan, Zhou Tinghui, et al. Image-to-image translation with conditional adversarial networks [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 1125-1134

[50] Shen Zan, Xu Yi, Ni Bingbing, et al. Crowd counting via adversarial cross-scale consistency pursuit [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 5245-5254

[51] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation [C] //Proc of Int Conf on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015: 234-241

[52] Wang Qi, Gao Junyu, Lin Wei, et al. Learning from synthetic data for crowd counting in the wild [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 8198-8207

[53] Zhu Junyan, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 2242-2251

[54] Gao Junyu, Han Tao, Wang Qi, et al. Domain-adaptive crowd counting via inter-domain features segregation and gaussian-prior reconstruction [EB/OL]. (2019-12-13) [2020-12-23]. https://arxiv.org/abs/1912.03677

[55] Liu Ning, Long Yongchao, Zou Changqing, et al. ADCrowdNet: Anattention-injective deformable convolutional network for crowd understanding [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 3225-3234

[56] Wei Yunchao, Xiao Huaxin, Shi Honghui, et al. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7268-7277

[57] Guo Dan, Li Kun, Zha Zhengjun, et al. DADNet: Dilated-Attention-Deformable ConvNet for crowd counting [C] //Proc of the 27th ACM Int Conf on Multimedia. New York: ACM, 2019: 1823-1832

[58] He Kaiming, Gkioxari G, Dollár P, et al. Mask R-CNN [C] //Proc of IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 2980-2988

[59] Zhu Liang, Zhao Zhijian, Lu Chao, et al. Dual path multi-scale fusion networks with attention for crowd counting [EB/OL]. (2019-02-04) [2020-12-23]. https://arxiv.org/abs/1911.07990

[60] Jiang Shenqin, Lu Xiaobo, Lei Yinjie, et al. Mask-aware networks for crowd counting [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(9): 3119-3129

[61] Valloli V K, Mehta K. W-Net: Reinforced U-Net for density map estimation [EB/OL]. (2019-05-29) [2020-12-23]. https://arxiv.org/abs/1903.11249

[62] Wang Qian, Breckon T P. Crowd counting via segmentation guided attention network and curriculum loss [EB/OL]. (2020-08-03) [2020-12-23]. https://arxiv.org/abs/1911.07990

[63] Shi Zenglin, Mettes P, Snoek C. Counting with focus for free [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 4199-4208

[64] Liu Lingbo, Qiu Zhilin, Li Guanbin, et al. Crowdcounting with deep structured scale integration network [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 1774-1783

[65] Yang Yifan, Li Guorong, Wu Zhe, et al. Reverse perspective network for perspective-aware object counting [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 4374-4383

[66] Oghaz M M, Khadka A R, Argyriou V, et al. Content-awaredensity map for crowd counting and density estimation [EB/OL]. (2019-06-17) [2020-12-23]. https://arxiv.org/abs/1906.07258

[67] Tian Yukun, Lei Yiming, Zhang Junping, et al. PaDNet: Pan-density crowd counting [J]. IEEE Transactions on Image Processing, 2020, 29: DOI:10.1109/TIP.2019.2952083

[68] Chan A B, Liang Z S J, Vasconcelos N. Privacy preserving crowd monitoring: Counting people without people models or tracking [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2008: DOI:10.1109/CVPR.2008.4587569

[69] Chen Ke, Loy C C, Gong Shaogang, et al. Feature mining for localised crowd counting [C] //Proc of the British Machine Vision Conf. Guildford, UK: BMVA, 2012: 21.1-21.11

[70] Zhang Cong, Kang Kai, Li Hongsheng, et al. Data-driven crowd understanding: A baseline for a large-scale crowd dataset [J]. IEEE Transactions on Multimedia, 2016, 18 (6): 1048-1061

[71] Kang Di, Dhar D, Chan A B. Incorporating side information by adaptive convolution [J]. International Journal of Computer Vision, 2020, 128: 2897-2918

[72] Sagar A. Bayesian multi scale neural network for crowd counting [EB/OL]. (2020-08-12) [2020-12-23]. https://arxiv.org/abs/2007.14245

[73] Zhang Lu, Shi Miaojing, Chen Qiaobo. Crowd counting via scale-adaptive convolutional neural network [C] //Proc of the IEEE Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2018: 1113-1121

[74] Fang Yanyan, Zhan Biyun, Cai Wandi, et al. Locality-constrained spatial transformer network for video crowd counting [C] //Proc of the IEEE Int Conf on Multimedia and Expo. Piscataway, NJ: IEEE, 2019: 814-819

[75] Ding Xinghao, Lin Zhirui, He Fujin, et al. A deeply-recursive convolutional network for crowd counting [C] // Proc of the IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 1942-1946

[76] Zhu Pengfei, Wen Longyin, Du Dawei, et al. Vision Meets Drones: Past, Present and Future [EB/OL]. (2020-07-16) [2020-12-23]. https://arxiv.org/abs/ 2001.06303

[77] Bahmanyar R, Vig E, Reinartz P. MRCNet: Crowd counting and density map estimation in aerial and ground imagery [EB/OL]. (2019-09-27) [2020-12-23]. https://arxiv.org/abs/1909.12743

[78] Wang Qi, Gao Junyu, Lin Wei, et al. NWPU-Crowd: A large-scale benchmark for crowd counting [J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020, https//doi.org/10.1109/TPAMI.2020.3013269

[79] Sindagi V, Yasarla R, Patel V M M. JHU-CROWD++: Large-scale crowd counting dataset and a benchmark method [J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020 [2020-12-23]. https//doi. org/10. 1109/TPAMI.2020.3035969

[80] Hu Di, Mou Lichao, Wang Qingzhong, et al. Ambient sound helps: Audiovisual crowd counting in extreme conditions [EB/OL]. (2020-05-16) [2020-12-23]. https://arxiv.org/abs/2005.07097

[81] 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, 2018, 107: 3-16

[82] Zou Zhikang, Shao Huiliang, Qu Xiaoye, et al. Enhanced 3D convolutional networks for crowd counting [C] //Proc of the 30th British Machine Vision Conf. Guildford, UK: BMVA, 2019: 250.1-250.13

[83] Shi Miaojing, Yang Zhaohui, Xu Chao, et al. Revisiting perspective information for efficient crowd counting [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 7279-7288

[84] Oñoro-Rubio D, López-Sastre R J. Towards perspective-free object counting with deep learning [G] //LNCS 9911: Proc of the European Conf on Computer Vision. Berlin: Springer, 2016: 615-629

[85] Walach E, Wolf L. Learning to count with CNN boosting [G] //LNCS 9906: Proc of the European Conf on Computer Vision. Berlin: Springer, 2016: 660-676

[86] Xiong Feng, Shi Xingjian, Yeung D Y. Spatiotemporal modeling for crowd counting in videos [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 5151-5159

[87] Huang Siyu, Li Xi, Zhang Zhongfei, et al. Body structure aware deep crowd counting [J]. IEEE Transactions on Image Processing, 2017, 27(3): 1049-1059

[88] Chen Xinya, Bin Yanrui, Sang Nong, et al. Scale pyramid network for crowd counting [C] //Proc of the IEEE Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2019: 1941-1950

[89] Cheng Zhiqi, Li Junxiu, Dai Qi, et al. Learning spatial awareness to improve crowd counting [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 6152-6161

[90] Liu Lingbo, Wang Hongjun, Li Guanbin, et al. Crowd counting using deep recurrent spatial-aware network [C] // Proc of the 27th Int Joint Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 849-855

[91] Sindagi V A, Patel V M. CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting [C] //Proc of the 14th IEEE Int Conf on Advanced Video and Signal Based Surveillance. Piscataway, NJ: IEEE, 2017: 1-6

[92] Zeng Lingke, Xu Xiangmin, Cai Bolun, et al. Multi-scale convolutional neural networks for crowd counting [C] //Proc of the IEEE Int Conf on Image Processing. Piscataway, NJ: IEEE, 2017: 465-469

[93] Sam D B, Babu R V. Top-down feedback for crowd counting convolutional neural network [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018: 7323-7330

[94] Zhang Le, Shi Zenglin, Cheng Ming-Ming, et al. Nonlinear regression via deep negative correlation learning [J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019 [2020-12-23]. https://doi. org/10. 1109/TPAMI.2019.2943860

[95] Liu Xialei, van de Weijer J, Bagdanov A D. Leveraging unlabeled data for crowd counting by learning to rank [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7661-7669

[96] Ranjan V, Le H, Hoai M. Iterative crowd counting [G] // LNCS 11211: Proc of the European Conf on Computer Vision. Berlin: Springer, 2018: 278-293

[97] Aich S, Stavness I. Global sum pooling: A generalization trick for object counting with small datasets of large images [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2019: 73-82

[98] Sam D B, Sajjan N N, Maurya H, et al. Almost unsupervised learning for dense crowd counting [C] //Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2019: 8868-8875

[99] Wu Xingjiao, Zheng Yingbin, Ye Hao, et al. Adaptive scenario discovery for crowd counting [C] //Proc of the IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2019: 2382-2386

[100] Xu Chenfeng, Qiu Kai, Fu Jianlong, et al. Learn to scale: Generating multipolar normalized density maps for crowd counting [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 8381-8389

[101] Jiang Xiaolong, Xiao Zehao, Zhang Baochang, et al. Crowd counting and density estimation by trellis encoder-decoder networks [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 6126-6135

[102] Sindagi V A, Patel V M. HA-CCN: Hierarchical attention-based crowd counting network [J]. IEEE Transactions on Image Processing. 2020, 29: 323-335

[103] Ma Zhiheng, Wei Xing, Hong Xiaopeng, et al. Bayesian loss for crowd count estimation with point supervision [C] // Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 6141-6150

[104] Sindagi V, Patel V. Multi-level bottom-top and top-bottom feature fusion for crowd counting [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 1002-1012

[105] Zhang Anran, Shen Jiayi, Xiao Zehao, et al. Relational attention network for crowd counting [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 6788-6797

[106] Xiong Haipeng, Lu Hao, Liu Chengxin, et al. From open set to closed set: Counting objects by spatial divide-and-conquer [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 8361-8370

[107] Yan Zhaoyi, Yuan Yuchen, Zuo Wangmeng, et al. Perspective-guided convolution networks for crowd counting [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 952-961

[108] Oh M, Olsen P, Ramamurthy K N. Crowd counting with decomposed uncertainty [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2020: 11799-11806

[109] Bai Shuai, He Zhiqun, Qiao Yu, et al. Adaptive dilated network with-self-correction supervision for counting [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 4593-4602

[110] Gao Guangshuai, Gao Junyu, Liu Qingjie, et al. CNN-based density estimation and crowd counting: A survey [EB/OL]. (2020-03-28) [2020-12-23]. https://arxiv.org/abs/2003.12783

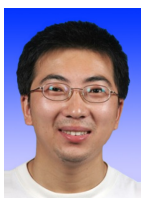
[111] Xu Jing, Liu Peng, Liu Jiafeng, et al. A detection algorithm for rain-affected moving objects [J]. Journal of Computer Research and Development, 2009, 46(11): 1885-1892 (in Chinese)
(徐晶, 刘鹏, 刘家锋, 等. 一种受雨滴影响的运动目标检测方法[J]. 计算机研究与发展, 2009, 46(11): 1885-1892)

Yu Ying, born in 1979. PhD, associate professor, master supervisor. Member of CCF. Her main research interests include machine learning, computer vision, and granular computing.

余　鹰, 1979 年生. 博士, 副教授, 硕士生导师. CCF 会员. 主要研究方向为机器学习、计算机视觉、粒计算.

Zhu Huilin, born in 1996. Master. Her main research interests include machine learning and computer vision.

朱慧琳, 1996 年生. 硕士. 主要研究方向为机器学习和计算机视觉.



Qian Jin, born in 1975. PhD, professor, master supervisor. Member of CCF. His main research interests include granular computing, big data mining, and machine learning.

钱　进, 1975 年生. 博士, 教授, 硕士生导师. CCF 会员. 主要研究方向为粒计算、大数据挖掘和机器学习.

Pan Cheng, born in 1995. Master candidate. His main research interests include machine learning and computer vision.

潘　诚, 1995 年生. 硕士研究生. 主要研究方向为机器学习和计算机视觉.

Miao Duoqian, born in 1964. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include machine learning, granular computing, and pattern recognition.

苗夺谦, 1964 年生. 博士, 教授, 博士生导师. CCF 高级会员. 主要研究方向为机器学习、粒计算、模式识别等.