计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(10):1−11 [doi: 10.15888/j.cnki.csa.008126]

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

基于尺度融合的密集人群计数①

赵宏伟[1], 徐　亮[2], 王　冶[3], 安云云[4], 钱华山[5]

[1](中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
[2](北京科技大学 计算机与通信工程学院, 北京 100083)
[3](解放军 9144 部队, 青岛 266102)
[4](国网山东省电力公司 青岛市黄岛区供电公司, 青岛 266499)
[5](北京超算科技有限公司, 北京 100190)
通讯作者: 赵宏伟, E-mail: upcvagen@163.com

摘　要: 现实场景中人群尺度的巨大差异给密集人群计数算法带来了巨大的挑战, 因此提出一种基于尺度融合的密集人群计数算法. 首先对密度图构建算法进行优化, 利用多个头部检测器获取稀疏人群的部分头部尺度, 并用径向基差值进行补全, 在人群密集区域辅之以距离自适应的人群密度图生成算法, 生成更为精确的人群密度图. 其次利用移动翻转瓶颈卷积模块设计尺度融合的人群密度图回归神经网络, 并加入膨胀卷积模块进一步提升人体头部边缘特征提取能力. 最后, 通过将人群区域和非人群区域进行区分对人群密度图回归神经网络损失函数进行优化. 在实验部分, 将该算法在多个数据集上与多个同类算法进行了充分的对比实验与消融实验, 实验结果表明提出的方法能够显著提升密集人群计数算法的准确性.

关键词: 密集人群计数; 多尺度; 尺度融合; 深度学习; 密度图

引用格式: 赵宏伟,徐亮,王冶,安云云,钱华山.基于尺度融合的密集人群计数.计算机系统应用,2021,30(10):1–11. http://www.c-s-a.org.cn/1003-3254/8126.html

Crowd Counting Based on Scale Fusion

ZHAO Hong-Wei[1], XU Liang[2], WANG Ye[3], AN Yun-Yun[4], QIAN Hua-Shan[5]

[1](College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)
[2](School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China)
[3](No. 9144 Troops of PLA, Qingdao 266102, China)
[4](Qingdao Huangdao District Power Supply Company, State Grid Shandong Electric Power Company, Qingdao 266499, China)
[5](Beijing SupCompute Technology Co. Ltd., Beijing 100190, China)

Abstract: The diversity of crowd scale in reality is a great challenge to crowd counting algorithms. Therefore, a novel crowd counting algorithm based on scale fusion is proposed in this study. Firstly, the algorithm for density map generation is optimized. Multiple head detectors are used to obtain part of the head scales of the sparse crowd, and RBF interpolation is employed to complete this part of the density map. As to the dense part of crowd, the traditional distance self-adaptive algorithm is adopted to generate a more accurate density map. Secondly, the regression neural network of the density map is designed with a mobile inverted bottleneck convolution module, and a dilated convolution module is added to facilitate the extraction of head edge features. Finally, the loss function of the regression neural network is optimized by distinguishing the crowd area from the non-crowd area. In the experiment part, the algorithm is compared

① 基金项目: 国家自然科学基金 (62072469); 国家重点研发计划 (2018YFE0116700); 山东省自然科学基金 (ZR2019MF049); 中央高校基本科研业务费专项资金 (2015020031); 西海岸人工智能技术创新中心建设专项 (2019-1-5, 2019-1-6); 上海可信工业控制平台开放项目 (TICPSH202003015-ZC)
Foundation item: National Natural Science Foundation of China (62072469); National Key R & D Program of China (2018YFE0116700); Natural Science Foundation of Shandong Province (ZR2019MF049); Fundamental Research Funds for the Central Universities of China (2015020031); West Coast Artificial Intelligence Technology Innovation Center (2019-1-5, 2019-1-6); Opening Project of Shanghai Trusted Industrial Control Platform (TICPSH202003015-ZC)
收稿时间: 2021-01-11; 修改时间: 2021-02-07; 采用时间: 2021-02-23

with other similar algorithms on multiple datasets, and the results show that the proposed method can significantly improve the accuracy of crowd counting.

Key words: crowd counting; multi-scale; scale fusion; deep learning; density map

密集人群计数是计算机视觉领域的一个重要分支, 在城市规划、交通、安防等领域有着重要的研究意义. 其研究目标在于, 对于给定的密集人群图像, 能够准确地推理出其中的人员数量[1–3]. 目前主流的密集人群计数算法利用回归的思想, 利用神经网络直接根据人群图像输出人群的密度图, 然后对热度图进行积分获得图像中的人数. 在训练时人群密度图通常利用高斯核来表征图像中的人体头部, 将整个图像中所有头部高斯核进行叠加构成人群密度图[4–8].

在现实场景人群密度估计工作中, 由于不同场景下人群的密度和图像中人员的尺度差异较大, 对不同尺度的人群进行准确的密度估计在当前仍是一个极具挑战的课题. 因此, 针对性地提出了一种基于尺度融合的密集人群计数算法 (Multi-Scale Fusion Based Crowd Counting, MSFBCC), 从人群密度图生成算法、密度图回归神经网络设计和损失函数优化 3 个方面进行了改进:

(1) 人群密度图构建算法优化: 在利用高斯核表征人体头部构建人群密度图时, 由于密集人群图像中目标数量极大, 标注十分费力, 绝大多数密集人群数据集都采用以点来表征人体头部的方式[4], 因此目前人头高斯核的δ值通常由人群中头部之间的距离确定. 这种方式在人群密度较大时生成的密度图有着不错的效果, 但是在人群密度相对稀疏以及人群密度波动较大的情况下则难以对人群中的头部进行准确的表征, 其原因在于图像中人头部的大小跟头部之间的距离并不完全相关. 因此提出了一种新的人群密度图构建算法, 即在人群稀疏区域利用基于深度学习和非深度学习方法的人头检测器获取图像中部分人体头部的尺寸, 并根据边框中心坐标与人头标注点的欧氏距离与头部边框进行匹配, 利用径向基插值对此区域剩余人头的尺度进行补全, 而在人群稠密区域则采用传统的距离自适应的尺度估计方法, 从而获得所有标注人头标注点的尺度, 以此为度量创建每一个头部区域的高斯核, 生成整个图像的人群密度图.

(2) 人群密度图回归神经网络设计: 由于目前大多数密集人群计数算法难以解决人群尺度波动大的问题, 无法结合人群多尺度特征进行准确推理, 在 EfficientNet[9] 的基础上提出了一种多尺度融合的人群密度图回归神经网络. 该网络利用 MBConv (移动反转瓶颈卷积) 模块[10] 以及 SENet[11] 压缩与激发机制构建基础特征提取模块并通过添加空洞卷积提升人群边缘特征提取的性能, 通过 BiFPN[12] 将骨架神经网络的最后 4 个模块输出的特征图进行融合, 实现了更好的密集人群尺度融合.

(3) 损失函数优化: 原始密集人群计数仅仅对图像中人群的数目进行统计, 其损失函数只涉及预测总数与真实的偏差, 没有考虑人群在图像中的分布. 更为常见的一种损失函数通过对预测密度图与真实密度图的差值对神经网络进行训练, 这样不仅可以准确地推理出图像中人群的数目, 更可以利用人群在图像中的分布进一步提升推理的准确率. 但是对于图像中的人群空白区域仍然会出现回归噪声, 即在生成密度图的无人群区域出现了非 0 数值区, 因此提出通过在损失函数增加对空白区域误差的惩罚以提升密度图回归神经网络的准确性, 即在损失函数中增加密度图人群区域像素与非人群区域像素的二元分类损失.

1 研究背景

密集人群计数对于城市公共安全和基础设施规划有着重要的意义, 但是由于图像角度和拍摄距离等原因, 人群目标的多尺度问题一直是一个极具挑战的课题. 目前主流的密集人群计数算法大多利用深度学习技术, 通过卷积神经网络将人群图像回归成人群密度图, 再通过对密度图进行积分获得图像中的人群总数.

1.1 密集人群计数

如图 1 所示, 传统的密集人群算法包括基于检测和基于回归的密集人群计数算法. 基于检测的传统密集人群计数算法分为人体整体检测和人体部件检测两种: 人体整体检测算法从人体的整体提取特征对图像中的人员进行检测并统计数目[13–16], 这种方法主要适用于人群密度较为稀疏的情况, 而密度较大的人群图

像则表现不佳, 其原因在于密集人群图像中遮挡较为严重, 难以获取完整的人体特征; 利用人体部件进行检测则更为常用[17], 如人体的头部、肩膀等, 人体头部特征更为明显而且在人群中遮挡程度相对较低, 相比对人的整体进行检测, 在准确率上有一定的提升. 而基于回归的密集人群计数方法[18–20] 特征在于: 首先提取图像中人群的各种特征, 然后利用线性回归、岭回归等方法建立起由人群特征到人群数量的映射 (人群总数或密度图).

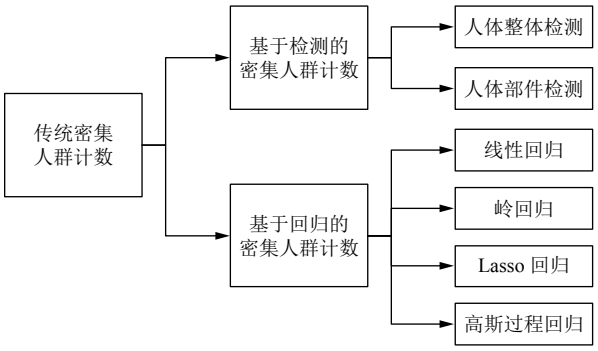


图 1 传统密集人群技术方法分类

随着深度学习技术的兴起, 卷积神经网络逐渐成为主流的密集人群计数方法. 基于深度学习的密集人群技术方法同样分为基于检测的密集人群计数和基于回归的密集人群计数. 基于检测的密集人群计数方法通常利用 YOLO[21]、SSD[22]、Faster RCNN[23] 等目标检测算法对人体的头部进行检测并统计其数量, 但同样由于人群密集区域遮挡严重并且分辨率较低, 难以达到常规目标检测的性能, 因此基于回归的密集人群计数算法更为常用一些. 基于回归的密集人群计数算法按照回归目标可以分为回归密度图和回归人群总数两种方式直接对图像中人群的总数进行回归. 由于无法获取人群在图像中的分布信息, 具有更多的不确定性, 因此 MCNN[4]、CP-CNN[8]、Switching-CNN[7]、CrowdNet[5]、CSRNet[24] 和 CTML[25] 等基于密度图回归的人群计数算法逐渐成为当前研究的主流方向[26–32].

1.2 密集人群计数中的多尺度问题

由于图像采集以及视觉透视等原因, 图像中的人员通常会有不同的尺度, 比如监控图像近大远小的透视特征会使得图像中距离摄像头不同远近的人呈现出不同的大小 (如图 2 中标注出的远近两个人头); 相同的人在不同分辨率的图像中也会有不同的尺度 (如图 3). 人群尺度的不一致性成为密集人群计数发展中的一大难题, 因此学术界对此开展了广泛的研究.

图 2 密集人群数据样本 (距离差异)

(a) 样本 1 (b) 样本 2

图 3 密集人群数据样本 (分辨率差异)

在人群密度图生成方面目前多采用人头标注点与二维高斯核进行卷积的方式, 使得每个人头区域的概率之和为 1, 在将所有人头区域卷积完之后进行叠加得到完整的人群密度图, 对人群密度图进行积分即可得到整体的人群数目. 因此二维高斯核的标准差σ反映着单个人头区域高斯核的分布状况, 即人头的尺度信息. 上海科技大学的张明明教授在提出 MCNN 算法时利用人头之间的距离作为判断高斯核标准差的标准[4], 并一直被后续的算法沿用, 即在确定高斯核标准差时, 利用 K 最近邻算法确定最近的几个头部标记点, 取其平均并乘以系数, 通常系数取 0.3. 但是此种方式并非完全有效, 原因在于在人群较为密集的场景中, 人群的密度和尺度是呈负相关的, 即人群中人与人紧密贴合的情况下头部标注点之间距离越远头部的尺度越大, 利用头部的尺度可以用标注点之间的距离来表征; 而在人群较为稀疏的场景中, 头部的尺度和人群密度是无关的, 因此仅依靠头部标注点之间的距离来判断头部尺度是不准确的.

针对人群的尺度多样性特征, 在密集人群计数神经网络设计上大致分为多列神经网络[33–35] 和单列神经网络[36–38] 两种. 多列神经网络以 MCNN[4] 为代表, 其

特征在于在不同神经网络分支上分别处理对应的不同尺度的特征, 再将不同分支的回归结果进行融合, 以实现密集人群尺度融合的效果. 但是多列神经网络存在参数量过大的缺陷, 相同层数的多列神经网络比单列神经网络参数多出数倍, 并且存在特征重复提取的弊端, 因此在训练成本有限的情况下目前大多利用单列神经网络的结构来进行人群密度图回归. 单列神经网络采用并行多分支模块、串行跳层连接模块等, 获取不同感受野的人群特征并将不同抽象层级的特征进行融合, 实现密集人群多尺度特征的获取. 此外, 将注意力机制[1,39] 引入密集人群计数领域, 对检测的准确率也有一定提升.

2 算法设计

为解决密集人群计数领域尺度多样性难题, 提出了基于尺度融合的密集人群计数算法, 主要包括人群密度图构建算法优化、人群密度图回归神经网络设计和损失函数优化 3 个方面.

2.1 人群密度图构建算法优化

当前人群密度图生成算法包括人头散点图标注和人头高斯核卷积两部分, 对于一张密集人群图像, 原始标注$X_i(x_i,y_i)$标注一个人头坐标, 在构建密度图时需要将人头利用狄克拉δ函数表示.

$$H(X)=\sum_{i=1}^{N}\delta(X-X_i) \tag{1}$$

然后将其与人头高斯核$G_\delta(X)$进行卷积操作获得密度图.

$$D(X)=H(X)*G_\delta(X) \tag{2}$$

但是由于视觉透视效果以及图像扭曲等原因, 人群中的人头具有不同的尺度, 因此高斯核的δ不能够用固定的值来表示. 目前在构建人头高斯核时大多利用人头之间的距离来设计尺度自适应的高斯核来对人体头部来进行模拟. 对于给定人头部标注点 x_i, 与其最近的k个头部标注点之间距离的集合为$\{d_1^i,d_2^i,d_3^i,\cdots,d_k^i\}$, 则其与周围标注点之间的平均距离可表示为:

$$\bar{d}_i=\frac{1}{m}\sum_{j=1}^{m}d_j^i \tag{3}$$

因此, 人群密度图可以更精确地表示为:

$$D(X)=H(X)*G_{\delta_i}(X),\ \delta_i=\beta\bar{d}_i \tag{4}$$

其中, β为平均距离的系数, 一般依据经验选择 β为 0.3. 但是依据经验选择的距离系数在其他数据集上不具有广泛的适用性, 并且在人群密度较低时并不平均, 最近距离法并不准确, 因此需要更为精准的头部尺度估计算法.

二维高斯核如式 (5) 所示:

$$G(x,y)=\frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{5}$$

根据正态分布$\sigma=r/3$原则, 头部高斯核像素点分布在距离头部标注点半3σ范围内的概率为 0.9974, 因此如图 4 所示, 可以将头部标注点半径3σ的区域用于表征人头区域, 人头部位高斯核标准差可由人头部位半径 r 近似表示.

$$\sigma=r/3 \tag{6}$$

图 4 人头区域半径约为3σ

为了准确地获取人头区域的半径, 结合深度学习人头检测器与非深度学习人头检测器对图像中的人头进行检测, 训练数据集包括 SCUT-HEAD 数据集[40]、HollywoodHead 数据集[41] 和 Brainwash 数据集[42], 数据集描述如表 1 所示.

表 1 人头检测器数据集

数据集		图像数	标注数
SCUT-HEAD	Part A	2000	67321
	Part B	2405	43930
HollywoodHead		224740	369846
Brainwash		91146	11917

深度学习人头检测器利用在 COCO 数据集上进行预训练的 EfficientDet 目标检测算法训练密集人群头部检测器, 将上述 3 个数据集合分别按照 5:1 的比例划分训练集和验证集, 最终训练集 266 909 张、验证集 53 382 张, 得到测试准确率为 87.28%, 图 5 为测试结果样本.

图 5 EfficientDet 人头检测效果

非深度学习方法的人头检测器利用头部的 HOG 特征和 Haar 特征进行人头检测, 具体实现为 HOG+SVM 和 Haar+AdaBoost.

在将检测框与头部标注点进行匹配时, 设计了算法 1.

算法 1. 头部检测框与头部标注点匹配算法

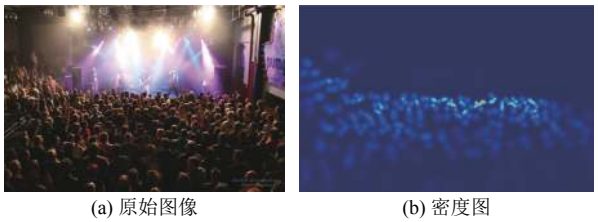
输入: 人群头部标注点与头部检测框
输出: 与头部标注点相匹配的检测框列表

```
begin
  for i=1 to n do
    计算检测框的中心坐标
end
将所有头部标注点标记为未匹配状态
for i=1 to n do
begin
  选择任意头部标注点为距离检测框 box_i 中心最近的头部标注点 head_closest, 两者之间的欧式距离定义为 dist_min
  for j=1 to n do
  begin
    if 头部标注点 headj 为未比对状态 then
      计算检测框 box_i 中心点与头部标注点 head_j 的欧氏距离 dist(i, j)
      if dist(i, j) < 最小欧氏距离 dist_min then
      最近的头部标注点 head_closest 更新为 head_j
      将最小欧氏距离 dist_min 更新为 dist(i, j)
  将检测框 box_i 与头部标注点 head_closest 进行匹配
  将头部标注点 head_closest 标注为已比对状态
  end
end
```

经由算法 1 可以将深度学习方法与非深度学习方法检测到的头部标注框与头部标注点分别进行匹配, 然后利用匹配结果获得检测区域头部的尺度, 计算公式如式 (7) 所示, 其中r_i为第i个人头标注点的尺度, $height_i$和$width_i$分别为第i个匹配检测框的高度和宽度.

$$r_i = (height_i + width_i)/2 \tag{7}$$

由此可以获得图像中部分头部的尺度, 根据近大远小的透视特点, 利用径向基插值 (RBF 插值) 方法对未匹配头部的尺度进行补全. 实验过程中发现无论深度学习方式还是非深度学习方式, 在人群密度较大、遮挡严重且头部较小的区域无法检测出有效的检测框, 因此根据头部标注点的分布, 在无检测框却分布均匀的头部密集区域采用根据头部之间的距离估算人头尺度的距离自适应算法, 如式 (2)–式 (4) 所示, 将两种方式获得的密度图进行拼接, 最终获得人群密度图如图 6 所示.



(a) 原始图像 (b) 密度图

图 6 原始图像与生成密度图

2.2 MSFBCC 神经网络设计

EfficientNet[9] 和 EfficientDet[12] 分别是图像分类和目标检测领域里程碑式的网络结构, 受其启发, 在人群密度图回归网络设计上以 EfficientNet-B7 为骨架特征提取网络并进行优化以适应多尺度人群特征图融合的需求. 到目前为止, MSFBCC 是首个将 EfficientNet 应用在密集人群技术领域的算法.

在网络结构的设计上, 以在 MobileNetV2[10] 中提出的 MBConv 模块为网络主体, 引入 SENet 的压缩-激发机制对网络进行优化. MBConv 模块结构如图 7 所示, 其主要构成为深度可分离卷积和 SENet, 输入特征图首先利用 1×1 卷积进行升维, 再经过深度可分离卷积与 SENet 进行特征提取, 最后经由 1×1 卷积降维之后与原始输入特征图进行相加.

如图 8 所示, MSFBCC 网络主要由 Stem 层、7 个特征提取区块和特征融合网络组成, 其中, 7 个特征提取区块分别由多个 MBConv 模块组成, MBConv 的个数分别为{5, 8, 8, 11, 11, 13, 5}, 每个 MBConv 模块对应输出特征图维度设置与 EfficientNet-B7 一致, 7 个特征提取区块输出特征图的维度分别对应为{32, 48, 80, 160, 244, 384, 640}.

密集人群中的人体头部为多小目标且更需要关注其边缘特征, 鉴于膨胀卷积在密集人群头部特征提取中发挥的显著成效[24], 在 7 个特征提取区块的前 4 个加入膨胀卷积模块, 膨胀率均为 2. 对于 MSFBCC 的后 4 个模块生成的特征图, 利用 BiFPN 进行不同尺度等级特征图的融合, 在融合时需要利用上采样进行尺度对齐, 融合方式如图 8 所示, 在将特征图融合之后通过反卷积层对特征图进行放大得到输出密度图.

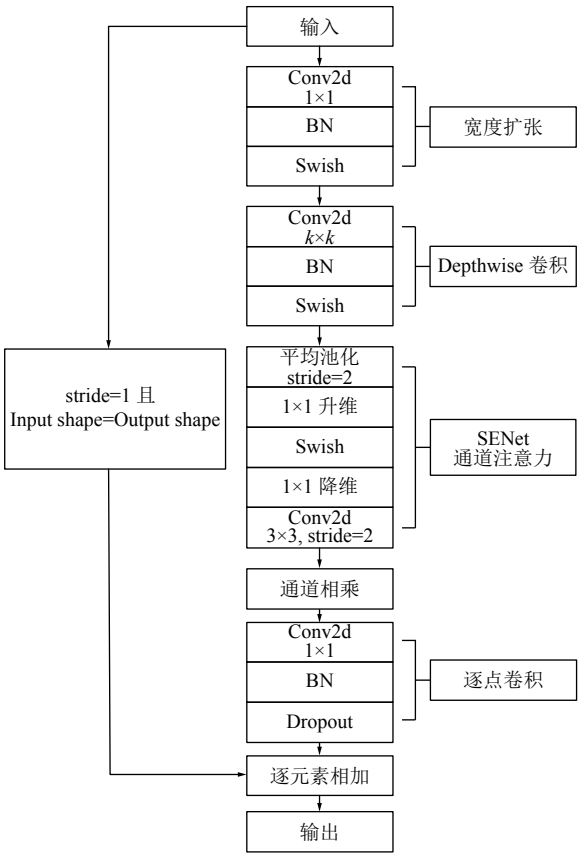


图 7 MBConv 模块

在训练时将 EfficientDet 在 COCO 数据集上训练的模型作为训练模型的部分初始权重, 用于初始化除特征提取网络前 4 层空洞卷积模块之外的所有可训练参数, 特征提取网络中的 4 个空洞卷积模块则采用随机初始化.

2.3 损失函数设计

密集人群计数神经网络最后的输出为密度图, 将输入图像转化为密度图实际是一种回归问题, 因此传统的密集人群计数损失函数为 *MSE* 损失函数, 定义如式 (8), 其中, N 为一个批次中样本图像的数目, M 为密度图像素总数, $D(X_i;\theta)$为预测密度图, D_i为真实密度图, y_j和$\hat{y}_j$分别为密度图中像素点真实值和预测值.

$$L_R(\theta)=\frac{1}{N}\sum_{i=1}^{N}\|D(X_i;\theta)-D_i\|_2^2=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_j-\hat{y}_j\right)^2 \quad (8)$$

在复现其他同类算法工作时发现在不存在人的空白区域会出现非 0 数值区, 即在回归中出现了噪声. 为进一步提升回归的准确性, 需要对人群区域和非人群区域进行区分, 对非人群区域的噪声进行抑制, 采用了如式 (9) 所示的二值交叉熵辅助损失函数, 其中c_j和$\hat{c}_j$分别为图像中像素点j真实值和预测值的标签. 因此, 将密度图回归损失函数与像素分类损失函数相结合, 最终经过优化的密集人群计数损失函数如式 (10) 所示.

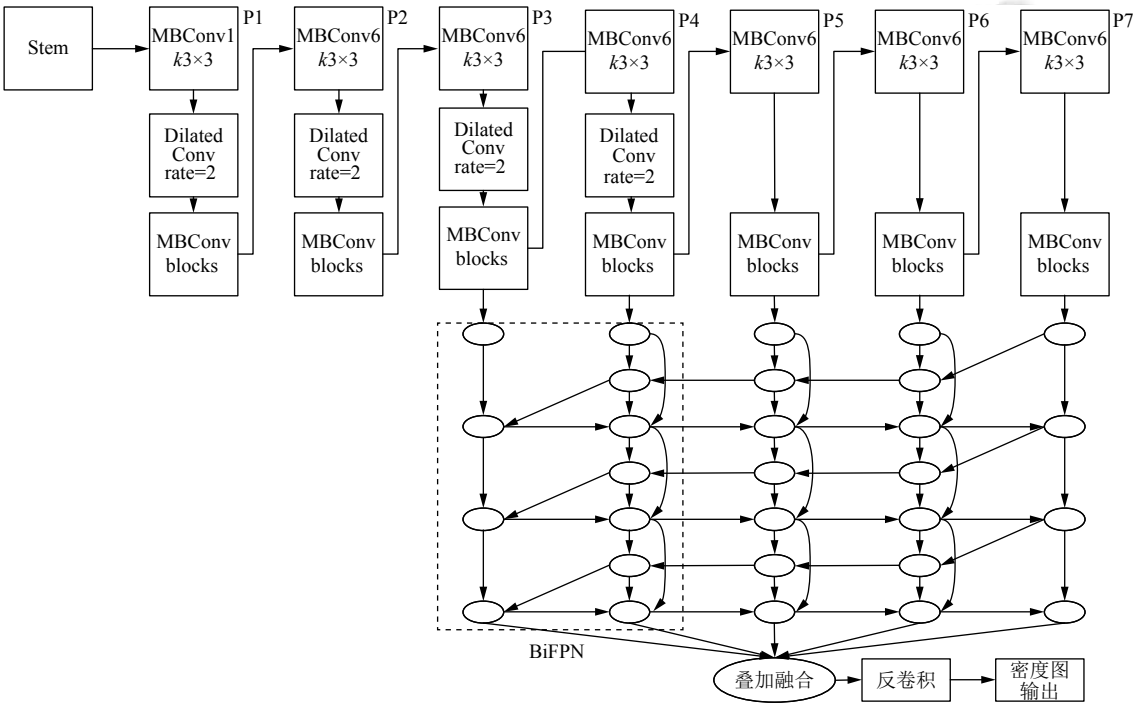


图 8 MSFBCC 网络结构图

$$L_R(\theta) = -\sum_{j=1}^{M}\left[c_j \cdot \ln\left(\hat{c}_j\right) + \left(1-c_j\right)\cdot \ln\left(1-\hat{c}_j\right)\right] \quad (9)$$

$$L(\theta) = L_R(\theta) + L_C(\theta)$$
$$= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[\left(y_j-\hat{y}_j\right)^2 - c_j \cdot \ln\left(\hat{c}_j\right) - \left(1-c_j\right)\cdot \ln\left(1-\hat{c}_j\right)\right] \quad (10)$$

3 实验分析

为了验证基于尺度融合的密集人群计数算法的有效性和泛化性, 在多个数据集上进行了充分的对比实验, 实验结果表明所提出的方法能够超越绝大多数密集人群计数算法.

3.1 环境配置

所进行实验均在 Ubuntu 20.04 系统下进行, 显卡驱动版本为 455.23.05, CUDA 版本为 11.1, 采用 Python 3.6.5, PyTorch 1.6. 硬件环境配置 CPU 为 Intel® Core™ i7-10700K CPU @ 3.80 GHz × 16, 内存 32 GB, GPU 为 Nvidia GeForce RTX 3090, 显存为 24 GB.

3.2 数据集

实验部分采用的数据集包括 Mall 数据集[20]、UCSD 数据集[43]、ShanghaiTech 数据集[4]、UCF_CC_50 数据集[44] 和 UCF-QNRF 数据集[45], 数据集详细描述和数据集样本分别如表 2 和图 9 所示.

表 2 密集人群计数数据集

数据集		图片数	人头标注数	平均人头数	最小人头数	最大人头数	分辨率
Mall		2000	62 325	31	13	53	320×640
UCSD		2000	49 885	25	11	46	158×238
UCF_CC_50		20	63 974	1279	94	4543	不统一
ShanghaiTech	Part A	482	241 677	501	33	3139	不统一
	Part B	716	88 488	124	9	716	1024×768
UCF-QNRF		1535	1251 642	815	94	12 865	不统一

(a) Mall 数据集 (b) UCF_CC_50 数据集 (c) ShanghaiTech 数据集 Part A (d) ShanghaiTech 数据集 Part B (e) UCSD 数据集 (f) UCF-QNRF 数据集

图 9 数据集样本

3.3 评估指标

密集人群计数算法的评估指标包括 *MAE* (绝对平均误差) 和 *MSE* (均方误差). *MAE* 和 *MSE* 的计算如下所示, 其中N为图像总数, $Count_i^{pre}$为第i张图像预测的人群人数, 与之相对的, $Count_i^{gt}$为第i张图像真实的人群人数.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|Count_i^{pre} - Count_i^{gt}\right| \quad (11)$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|Count_i^{pre} - Count_i^{gt}\right|^2} \quad (12)$$

3.4 对比实验

为验证基于尺度融合的密集人群计数算法的准确性, 在 5 个数据集上与多个同类人群计数算法进行了对比实验, 实验结果如表 3–表 6 所示. 从实验结果可以看出, 相对于绝大多数同类算法, 基于尺度融合的密集人群计数算法具有更好的准确性和稳定性, 并在多个数据集上取得了最高的准确率.

表 3 在 ShanghaiTech Part A & B 和 UCF_CC_50 数据集上的算法效果对比

算法	ShanghaiTech Part A		ShanghaiTech Part B		UCF_CC_50	
	MAE	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*
MCNN[4]	110.2	173.2	26.4	41.3	377.6	509.1
Switching CNN[7]	90.4	135.0	21.6	33.4	318.1	439.2
CP-CNN[8]	73.6	106.14	20.1	30.1	295.8	320.9
LBP+RR[20]	303.2	371.0	59.1	81.7	—	—
CSRNet[24]	—	—	—	—	266.1	397.5
CMTL[25]	101.3	152.4	20.0	31.1	322.8	397.9
MSCNN[35]	83.8	127.4	17.7	30.2	363.7	468.4
SAAN[39]	—	—	—	—	271.6	391.0
Cross-scene[46]	181.8	277.7	32.0	49.8	467.0	498.5
ic-CNN[47]	—	—	—	—	260.9	365.5
PSDDN[48]	85.4	159.2	16.1	27.9	359.4	514.8
MSFBCC	**70.3**	**101.5**	**13.5**	**20.4**	**251.1**	**308.4**

表 4 在 Mall 数据集上的算法效果对比

算法	*MAE*	*MSE*
MCNN[4]	2.24	8.5
Faster RCNN[23]	5.91	6.60
GP[43]	3.72	20.1
CA-RR[49]	3.43	17.7
Count-Forest[50]	2.50	10.0
DRSAN[51]	1.72	**2.1**
AT-CFCN[52]	2.28	2.90
ACSPNet[53]	1.76	2.24
MSFBCC	**1.57**	2.42

表 5 在 UCSD 数据集上的算法效果对比

算法	*MAE*	*MSE*
MCNN[4]	1.07	1.35
Switching CNN[7]	1.62	2.10
CSRNet[24]	1.16	1.47
GP[27]	2.24	7.97
Cross-scene[46]	1.60	3.31
CA-RR[49]	2.07	6.89
ACSPNet[53]	1.02	1.28
SANet[54]	1.02	1.29
PaDNet[55]	0.85	**1.06**
SPN[56]	1.03	1.32
MSFBCC	**0.81**	1.21

表 6 在 UCF-QNRF 数据集上的算法效果对比

算法	*MAE*	*MSE*
Switching CNN[7]	224.0	445.0
CMTL[25]	252.0	514.0
SAA-Net[57]	97.5	167.8
SFANet[58]	100.8	174.5
DUBNet[59]	116.0	178.0
DSNet[60]	91.4	160.4
TEDnet[61]	113.0	188.0
PCC Net[30]	132.0	191.0
LSC-CNN[31]	120.5	218.2
S-DCNet[32]	104.4	176.1
MSFBCC	**90.2**	**159.6**

4 消融实验

为探究基于尺度融合的密集人群计数算法各个部分的有效性, 在 ShanghaiTech 数据集 B 部分上分别针对人群密度图生成算法、神经网络设计和损失函数优化设计了消融实验, 实验相关算法包括 MCNN、MSCNN、Switching CNN 和 CMTL. 实验结果表明, 所提出的人群密度图生成算法、在特征提取网络中加入膨胀卷积的策略和改进的损失函数均能够显著提升密集人群计数算法的准确率.

4.1 人群密度图生成算法

分别利用固定尺度和人群之间的距离来预估人群中头部的尺度 (如图 10(a) 和图 10(b)), 与图 10(c) 相比, 可以看出改进的人群密度图能够更好地估计人群中头部的尺度.

(a) 固定值表征头部尺度

(b) 利用人群之间的距离估计头部尺度

(c) 基于尺度融合的密集人数计数估计头部尺度

图 10 不同方式估计头部尺度

分别将所提出的改进人群密度图生成算法与原始的距离自适应人群密度图生成算法得到的密度图作为训练标签, 对上述 4 种算法以及 MSFBCC 进行训练, 以验证改进的人群密度图生成算法的有效性. 表 7 为利用提出的人群密度图生成算法对与原始距离自适应的密度图生成算法的对比实验, 从表中可以看出, 与原始的距离自适应人群密度图生成算法相比, 所提出的改进人群密度图生成算法能够普遍提升人群密度图回归网络的准确率, 其中对于 Switching CNN 和 MSCNN 的提升最为显著, *MAE* 和 *MSE* 分别降低了 5.9 和 7.2.

表 7 改进的密度图生成算法效果对比

算法	原始算法		改进算法	
	MAE	*MSE*	*MAE*	*MSE*
MCNN[4]	26.4	41.3	24.3	36.2
Switching CNN[7]	21.6	33.4	15.7	30.9
CMTL[25]	20.0	31.1	18.2	15.3
MSCNN[35]	17.7	30.2	16.3	23.0
MSFBCC	17.3	22.1	13.5	20.4

4.2 神经网络设计

为验证 MSFBCC 神经网络的有效性, 将基础特征网络与利用膨胀卷积模块改进的特征提取网络进行对比实验. 表 8 为利用 EfficientNet 与利用膨胀卷积模块改进的 EfficientNet 作为特征提取网络的实验效果, 表 3 与表 8 的实验结果表明, 将 EfficientNet 应用在密集人群计数领域具有较高的准确性, 利用膨胀卷积模块对特征提取网络进行改进则会进一步提高计数的准确率.

表 8 特征提取网络效果对比

算法	*MAE*	*MSE*
EfficientNet	18.3	31.5
EfficientNet+膨胀卷积模块	13.5	20.4

4.3 损失函数优化

为了验证优化损失函数的有效性, 将改进的密度图回归损失函数在消融实验中的 5 个密集人群计数算法中进行应用, 表 9 为利用原始损失函数和优化过的损失函数的准确率对比, 从表中可以看出, 所提出的改进的损失函数能够一定程度上提升密集人群计数算法的准确率.

表 9 改进的损失函数应用在不同算法上的效果对比

算法	原始算法		改进后的算法	
	MAE	*MSE*	*MAE*	*MSE*
MCNN[4]	26.4	41.3	20.3	33.9
Switching CNN[7]	21.6	33.4	17.6	30.1
CMTL[25]	20.0	31.1	17.5	31.6
MSCNN[35]	17.7	30.2	15.2	29.3
MSFBCC	15.5	26.3	13.5	20.4

5 结论与展望

针对传统密集人群计数算法难以适应人群尺度多样性的问题, 提出了一种基于尺度融合的密集人群计数算法 MSDBCC, 分别从人群密度图生成算法、密度图回归神经网络设计和损失函数优化 3 个方面展开了研究. 充分的对比实验和消融实验表明, MSFBCC 能够准确地推理出密集人群图像中的人员数量并超越同类的大多数算法, 并且所提出的 3 个策略对计数算法的准确率有明显的提高作用. 但是 MSFBCC 特征提取网络参数量较大, 训练比较费时并且实际部署需要较多计算资源, 因此未来的研究应当聚焦于缩小网络规模, 提高网络的推理效率.

参考文献

1 张友梅. 基于注意力卷积神经网络的人群计数算法研究 [博士学位论文]. 济南: 山东大学, 2019.

2 周成博, 陶青川. 基于景区场景下的人群计数. 现代计算机, 2016, (5): 52–57. [doi: 10.3969/j.issn.1007-1423.2016.05.012]

3 陈训敏, 叶书函, 詹瑞. 基于多任务学习及由粗到精的卷积神经网络人群计数模型. 计算机科学, 2020, 47(11A): 183–187, 208.

4 Zhang YY, Zhou DS, Chen SQ, *et al*. Single-image crowd counting via multi-column convolutional neural network. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 589–597.

5 Boominathan L, Kruthiventi SSS, Babu RV. CrowdNet: A deep convolutional network for dense crowd counting. Proceedings of the 24th ACM International Conference on Multimedia. New York: ACM, 2016. 640–644.

6 Liu J, Gao CQ, Meng DY, *et al*. DecideNet: Counting varying density crowds through attention guided detection and density estimation. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5197–5206.

7 Sam DB, Surya S, Babu RV. Switching convolutional neural network for crowd counting. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4031–4039.

8 Sindagi VA, Patel VM. Generating high-quality crowd density maps using contextual pyramid CNNs. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 1861–1870.

9 Tan MX, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. arXiv: 905.11946, 2020.

10 Sandler M, Howard A, Zhu ML, *et al*. MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4510–4520.

11 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of 2018 IEEE/CVF Conference on Computer

Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.
12 Tan MX, Pang RM, Le QV. EfficientDet: Scalable and efficient object detection. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10781–10790.
13 Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 886–893.
14 Leibe B, Seemann E, Schiele B. Pedestrian detection in crowded scenes. Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 878–885.
15 Enzweiler M, Gavrila DM. Monocular pedestrian detection: Survey and experiments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(12): 2179–2195. [doi: 10.1109/TPAMI.2008.260]
16 Tuzel O, Porikli F, Meer P. Pedestrian detection via classification on riemannian manifolds. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(10): 1713–1727. [doi: 10.1109/TPAMI.2008.75]
17 Wu B, Nevatia R. Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. International Journal of Computer Vision, 2007, 75(2): 247–266. [doi: 10.1007/s11263-006-0027-7]
18 Chan AB, Vasconcelos N. Bayesian Poisson regression for crowd counting. Proceedings of 2009 IEEE 12th International Conference on Computer Vision. Kyoto: IEEE, 2009. 545–551.
19 Ryan D, Denman S, Fookes C, *et al*. Crowd counting using multiple local features. Proceedings of 2009 Digital Image Computing: Techniques and Applications. Melbourne: IEEE, 2009. 81–88.
20 Chen K, Loy CC, Gong SG, *et al*. Feature mining for localised crowd counting. Proceedings of the British Machine Vision Conference. Leeds: BMVA Press, 2012. 1–11.
21 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788.
22 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. Proceedings of the 14th European Conference on Computer Vision. Cham: Springer, 2016. 21–37.
23 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]
24 Li YH, Zhang XF, Chen DM. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1091–1100.
25 Sindagi VA, Patel VM. CNN-Based cascaded multi-task learning of high-level prior and density estimation for crowd counting. Proceedings of 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance. Lecce: IEEE, 2017. 1–6.
26 时增林, 叶阳东, 吴云鹏, 等. 基于序的空间金字塔池化网络的人群计数方法. 自动化学报, 2016, 42(6): 866–874.
27 覃勋辉, 王修飞, 周曦, 等. 多种人群密度场景下的人群计数. 中国图象图形学报, 2013, 18(4): 392–398. [doi: 10.11834/jig.20130405]
28 付倩慧, 李庆奎, 傅景楠, 等. 基于空间维度循环感知网络的密集人群计数模型. 计算机应用, 2021, 41(2): 544–549.
29 吴晓燕. 基于回归模型的全卷积网络人群计数算法. 计算机工程与设计, 2020, 41(10): 2867–2871.
30 Gao JY, Wang Q, Li XL. PCC Net: Perspective crowd counting via spatial convolutional network. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(10): 3486–3498. [doi: 10.1109/TCSVT.2019.2919139]
31 Sam DB, Peri SV, Sundararaman MN, *et al*. Locate, size and count: Accurately resolving people in dense crowds via detection. arXiv: 1906.07538, 2020.
32 Xiong HP, Lu H, Liu CX, *et al*. From open set to closed set: Counting objects by spatial divide-and-conquer. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 8362–8371.
33 严芳芳, 吴秦. 多通道融合分组卷积神经网络的人群计数算法. 小型微型计算机系统, 2020, 41(10): 2200–2205. [doi: 10.3969/j.issn.1000-1220.2020.10.029]
34 唐斯琪, 陶蔚, 张梁梁, 等. 一种多列特征图融合的深度人群计数算法. 郑州大学学报 (理学版), 2018, 50(2): 69–74.
35 Zeng LK, Xu XM, Cai BL, *et al*. Multi-scale convolutional neural networks for crowd counting. Proceedings of 2017 IEEE International Conference on Image Processing. Beijing: IEEE, 2017. 465–469.
36 鱼春燕, 徐岩, 缑丽莎, 等. 基于单列深度时空卷积神经网络的人群计数. 激光与光电子学进展. 2021, 58(08): 143–151.

37 马皓, 殷保群, 彭思凡. 基于特征金字塔网络的人群计数算法. 计算机工程, 2019, 45(7): 203–207.

38 Sindagi VA, Patel VM. Multi-level bottom-top and top-bottom feature fusion for crowd counting. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 1002–1012.

39 Hossain M, Hosseinzadeh M, Chanda O, *et al*. Crowd counting using scale-aware attention networks. Proceedings of 2019 IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2019. 1280–1288.

40 Peng DZ, Sun ZK, Chen ZR, *et al*. Detecting heads using feature refine net and cascaded multi-scale architecture. Proceedings of 2018 24th International Conference on Pattern Recognition. Beijing: IEEE, 2018. 2528–2533.

41 Vu TH, Osokin A, Laptev I. Context-aware CNNs for person head detection. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 2893–2901.

42 Stewart R, Andriluka M, Ng AY. End-to-end people detection in crowded scenes. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2325–2333.

43 Chan AB, Liang ZSJ, Vasconcelos N. Privacy preserving crowd monitoring: Counting people without people models or tracking. Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage: IEEE, 2008. 1–7.

44 Idrees H, Saleemi I, Seibert C, *et al*. Multi-source multi-scale counting in extremely dense crowd images. Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 2547–2554.

45 Idrees H, Tayyab M, Athrey K, *et al*. Composition loss for counting, density map estimation and localization in dense crowds. Proceedings of the 15th European Conference on Computer Vision. Cham: Springer, 2018. 532–546.

46 Zhang C, Li HS, Wang XG, *et al*. Cross-scene crowd counting via deep convolutional neural networks. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 833–841.

47 Ranjan V, Le H, Hoai M. Iterative crowd counting. Proceedings of the 15th European Conference on Computer Vision. Cham: Springer, 2018. 270–285.

48 Liu YT, Shi MJ, Zhao QJ, *et al*. Point in, box out: Beyond counting persons in crowds. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 6469–6478.

49 Chen K, Gong SG, Xiang T, *et al*. Cumulative attribute space for age and crowd density estimation. Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 2467–2474.

50 Pham VQ, Kozakaya T, Yamaguchi O, *et al*. COUNT forest: Co-voting uncertain number of targets using random forest for crowd density estimation. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 3253–3261.

51 Liu LB, Wang HJ, Li GB, *et al*. Crowd counting using deep recurrent spatial-aware network. arXiv: 1807.00601, 2018.

52 Zhao MM, Zhang J, Zhang CY, *et al*. Leveraging heterogeneous auxiliary tasks to assist crowd counting. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 12736–12745.

53 Ma JJ, Dai YP, Tan YP. Atrous convolutions spatial pyramid network for crowd counting and density estimation. Neurocomputing, 2019, 350: 91–101. [doi: 10.1016/j.neucom.2019.03.065]

54 Cao XK, Wang ZP, Zhao YY, *et al*. Scale aggregation network for accurate and efficient crowd counting. Proceedings of the 15th European Conference on Computer Vision. Cham: Springer, 2018. 734–750.

55 Tian YK, Lei YM, Zhang JP, *et al*. PadNet: Pan-density crowd counting. IEEE Transactions on Image Processing, 2019, 29: 2714–2727.

56 Chen XY, Bin YR, Sang N, *et al*. Scale pyramid network for crowd counting. Proceedings of 2019 IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2019. 1941–1950.

57 Varior RR, Shuai B, Tighe J, *et al*. Scale-aware attention network for crowd counting. arXiv: 1901.06026, 2019.

58 Zhu L, Zhao ZJ, Lu C, *et al*. Dual path multi-scale fusion networks with attention for crowd counting. arXiv: 1902.01115, 2019.

59 Oh M, Olsen PA, Ramamurthy KN. Crowd counting with decomposed uncertainty. Proceedings of the AAAI Conference on Artificial Intelligence, 34(07), 11799–11806.

60 Dai F, Liu H, Ma YK, *et al*. Dense scale network for crowd counting. arXiv: 1906.09707, 2019.

61 Jiang XL, Xiao ZH, Zhang BC, *et al*. Crowd counting and density estimation by trellis encoder-decoder networks. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 6133–6142.