DOI:10.13718/j.cnki.xsxb.2022.08.001

基于背景辅助的高效人群计数多任务学习网络①

桑军，　刘新悦，　吴志伟，　王富森

重庆大学 大数据与软件学院，重庆 401331

摘要：在人群计数领域中，复杂背景干扰一直是一个具有挑战性的问题. 现有研究通过引入注意力机制等方式弱化背景噪声对计数的影响. 但是，随着研究的深入，人群计数网络规模不断扩大，影响了计算效率和实时应用. 为了解决复杂背景问题并提高计数效率，该文提出了一个基于背景辅助的高效人群计数多任务学习网络(BAMTLNet). 与现有网络不同，为了减少网络的参数量，只采用了 VGG-16 的前 7 层作为前端网络. 在后端网络中，为了解决复杂背景问题，我们使用了两个高度相关的人群任务：①生成估计密度图主任务，采用 3 个普通卷积层生成密度图，通过积分获得单张图片的人数. ②复杂背景分割辅助任务，采用 3 个特定的膨胀卷积层生成图片的背景分割图. 两个任务直接连接在前端网络后，没有相互交叉. 我们还设计了背景辅助多任务损失函数，通过硬参数共享的方式优化前端网络参数，向主任务传递复杂背景的高级语义信息并优化网络. 该端到端人群计数多任务学习网络仅有 10 层卷积层，参数量小，实现了网络轻量化. 在 3 个人群计数基准数据集上进行了实验，获得了令人满意的结果.

关　键　词：人群计数；背景分割；轻量化；多任务学习

中图分类号：TP391.4　　**文献标志码**：A　　**文章编号**：1000－5471(2022)08－0001－08

An Efficient Background Assistance Based on Multi-Task Learning Network for Crowd Counting

SANG Jun,　LIU Xinyue,　WU Zhiwei,　WANG Fusen

School of Big Data & Software Engineering, Chongqing University, Chongqing 401331, China

Abstract: Complex background interference is still a challenging issue in crowd counting. In the existing crowd counting methods, attention has been paid on other approaches utilized to reduce the influence of background. As the research continues, the scale of crowd counting networks is growing, which makes a negative influence on computing efficiency and real-time application. Therefore, to solve complex background problem and to improve the counting efficiency, an efficient background assistance based on multi-task learning network (BAMTLNet) has been proposed in this paper. Unlike the existing networks, the first seven layers of VGG-16 has only been used as the front-end network to reduce the number of network

① 收稿日期：2021－12－27
基金项目：面向高度透视复杂场景的人群计数研究(61971073).
作者简介：桑军，教授，博士，主要从事人工智能、计算机视觉、信息安全研究.

parameters. For the problem of complex background, two highly correlated crowd counting tasks have been utilized in the back-end network: 1) The main task of generating estimated density map, which adopts three general convolutional layers to generate a density map, and obtains the number of people in a single image by integration. 2) The auxiliary task of complex background segmentation, by which to use three specific dilated convolutional layers to generate a background segmentation map. The two tasks have directly been connected behind the front-end network with no crossing. Besides, a background-assisted multi-task loss function has been designed to optimize the front-end network parameters through hard parameters sharing, by which to transfer the high-level semantic information of complex background to the main tasks and to optimize the network. This end-to-end crowd counting multi-task learning network is able to achieve comparable performance with only ten convolutional layers and less parameters. extensive experiments have been conducted on three crowd counting benchmark datasets and obtain satisfactory results.

Key words: crowd counting; background segmentation; lightweight; multi-task learning

由于城市进程化加快，各类聚集活动猛增. 大型聚会、演唱会、体育盛会、政治活动等都伴随着人群过载、难以管控的问题. 若是调控不当，人群散乱冲撞，极易发生严重的踩踏事故. 人群聚集是一种趋势，在计算机时代，我们可以利用计算机视觉技术对人群进行分析，提前做好应急措施，避免此类事件的发生. 人群分析是计算视觉的一个高热度研究领域，包含人群计数、行人检测、行人追踪等方向. 而人群计数是人群分析中的重要课题，通过对单张图片中的行人进行计数，其模型可以应用到实时监控中，对人群管控起到很大作用.

1 相关工作

当前人群计数研究[1-2]大多采用卷积神经网络(CNN)生成单张图像对应的密度图，通过对密度图积分得到具体人数. 生成的密度图质量决定了计数效果的好坏. 但是在人群个体尺度变化、复杂背景干扰、人群之间严重遮挡等问题的影响下，人群计数仍然是一项具有挑战性的研究课题. 其中复杂背景干扰会使得卷积神经网络将形状类似人群头部的树叶、路灯等当作人群进行计数，造成计数误差，使模型效果变差. 为了减少复杂背景的影响，不少学者采用了注意力机制方法，将网络注意力集中到人群区域，弱化背景，提高计数效果. 文献[3]提出采用注意力机制，提高人群区域注意力，降低背景噪声影响. 但是采用注意力机制的网络大都使用了 VGG[4]或者 CSRNet[5]为基准网络，虽然计数结果不错，但网络的参数量过大，推理速度慢. 其他学者还研究了通过多任务学习的方式降低背景影响. 文献[6]将复杂背景作为网络的辅助任务，还采用了尺度变化，生成深度图等共 3 个辅助任务，但其基准网络参数量大. 文献[7]采用了人群密度分级、背景分割辅助任务增强语义信息，网络结构简单，参数量小，但计数结果一般.

由于多任务学习方法能解决复杂背景干扰问题，且选择恰当的前端网络可以使得网络参数量减少，因此，本文提出了一个基于背景辅助的高效人群计数多任务学习网络(an efficient background assistance based multi-task network for crowd counting, BAMTLNet). 与人群计数网络中大量采用 VGG[4]前 10 层作为前端网络不同，我们仅采用 VGG[4]网络的前 7 层，以减少网络的参数量. 在后端网络中，我们提出了 2 个分支，分别是生成估计密度图主任务分支和复杂背景分割辅助任务分支. 生成估计密度图的主任务分支采用 3 个卷积层生成估计密度图，用于积分得到人数. 复杂背景分割辅助任务分支采用三层空洞卷积层生成背景分割图，利用多任务学习机制，为主任务分支提供背景的语义信息，优化网络参数，以降低复杂背景对计数的影响. 此外，我们为该多任务学习网络设计了背景辅助融合多任务损失函数. 经训练，该网络能

在人群计数基准数据集上达到不错的效果，网络参数少，推理速度快.

本文主要贡献如下：

1）提出了一个基于背景辅助的高效人群计数多任务学习网络，包含生成估计密度图主任务分支和复杂背景分割辅助任务分支，减少复杂背景对计数的影响，网络参数量小.

2）针对上述网络提出了背景辅助多任务损失函数.

3）在人群计数基准数据集上与多种算法进行了对比实验[8-17]，并得到了不错的实验结果.

2　背景辅助多任务学习网络

2.1　网络结构

如图 1 所示，本文给出了网络的具体结构. 该多任务学习网络由前端网络和后端网络组成. 为了减少网络参数量，我们并没有采取常见的 VGG[4]前 10 层作为前端网络的做法. 因为后 3 层的通道数为 512，为了减少网络体积且不损失太多精度，我们采用前 7 层来提取初级特征. 为了加快训练速度，使网络快速收敛，我们在每个卷积层后加入了 BN(batch normalization)层. 在后端网络中，我们采用了多任务学习、硬参数共享的机制，分割出两个任务分支共享同一个前端网络参数. 为了获取单张图片计数，我们基于估计密度图计数方法，设计了生成估计密度图主任务分支. 该分支由带 BN 层的 Conv4 和 Conv5 卷积层组成，并通过 Conv1×1 的卷积层生成最终的估计密度图用于积分得到人数. 为了获取关于背景的高级语义信息，我们设计了复杂背景分割辅助任务分支. 该分支由 Conv6 和 Conv7 两个空洞卷积层组成，空洞率为 2，与主任务分支相似，其通过 Conv1×1 的卷积层生成最终的背景分割图. 在多任务联合学习基础下，两个任务通过共享前端网络参数，辅助任务通过反向传播过程传递背景语义信息，优化整个网络参数，降低复杂背景造成的计数影响. 与其他复杂网络相比虽没有更准确的精度，但我们的网络仅有 10 层卷积层，网络复杂度小，参数量小.

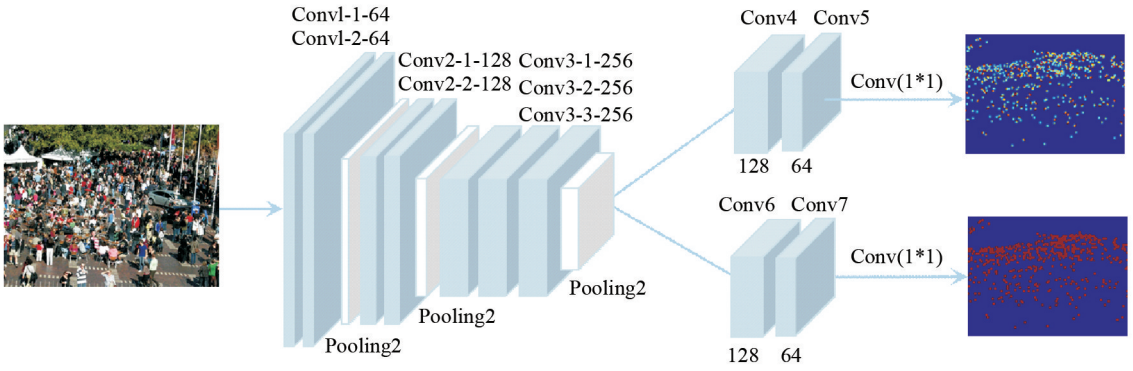

图 1　基于背景辅助的高效人群计数多任务学习网络(BAMTLNet)详细结构

2.2　损失函数

为了更好地训练上述多任务学习网络，我们设计了背景辅助多任务损失函数. 针对生成估计密度图的主任务，我们采用欧几里得损失计算真实密度图标签与估计密度图之间的差值. 针对复杂背景分割辅助任务，我们同样采用欧几里得损失计算真实背景分割图标签与估计分割图之间的差值，两个损失如下：

$$L^D=\frac{1}{N}\sum_{i=1}^{N}F(D^{gt}(X_i)-D^{est}(X_i;\Theta)) \tag{1}$$

$$L^S=\frac{1}{N}\sum_{i=1}^{N}F(S^{gt}(X_i)-S^{est}(X_i;\Theta)) \tag{2}$$

其中：N 表示在一个训练批次中输入图片的数目，$F(g)$ 是欧几里得距离，D^{gt} 表示真实密度图，D^{est} 表示

网络生成的估计密度图，S^{gt} 表示真实背景分割图标签，S^{est} 表示估计背景分割图，X_i 为输入图片，Θ 为网络参数.

对于该网络，我们采用了多任务联合学习的方式训练，最终的损失函数为：

$$L = L^D + L^S \tag{3}$$

3 实验细节

3.1 标签生成

在网络训练之前，我们需要制作真实密度图与背景分割图标签. 我们采用了人群计数中常见的真实密度图标签生成方法[8]，利用几何自适应高斯核与人群头部标注点图生成真实密度图：

$$D(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_{\sigma_i}(x),\ \sigma_i = \beta \bar{d}^i \tag{4}$$

其中：x_i 是每个行人头部位置点标注，$\delta(x - x_i)$ 表示一个行人，σ_i 表示第 i 个高斯核的标准差，$G_{\sigma_i}(x)$ 是高斯核函数，$\bar{d}^i$ 表示该行人头部与其 k 个邻居的平均距离，根据 MCNN[8] 模型所采用的设定，β 为定值取 0.3，k 取值 3.

在真实密度图的基础上，我们对图中的非零像素值取 1，否则取值 0，可以得到背景分割图标签：

$$S_i(p) = \begin{cases} 0 & \text{if } D_i(p) = 0 \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

其中 p 为真实密度图 $D_i(p)$ 中的第 i 个像素.

3.2 数据集

我们在实验中使用了 3 个人群计数基准数据集，分别为 ShanghaiTech[8]，UCF_CC_50[9] 和 UCF_QNRF[17].

ShanghaiTech 数据集[8]分为 Part A 和 Part B 两部分. 每部分都有训练集和验证集. 每张图片有对应的点标签，包含 1198 张图片，330 165 个注释头部. 它是目前使用最为广泛的人群计数基准数据集.

UCF_CC_50 数据集[9]包含 50 张图像，人群数量变化大，十分具有挑战性. 由于其数据集太小，我们在训练时采用了五折交叉验证的方法.

UCF_QNRF 数据集[17]是一个从网络收集的大规模数据集，包含透视场景和复杂背景的 1 535 张高分辨率图像，其人数从 50 到 12 000 不等.

3.3 训练细节

本次训练采用 Tesla K80 显卡. 针对不同的数据集我们设置了不同的高斯核标准差，对于比较稀疏的 ShanghaiTech PartB，我们设置了 σ 为 15.0，对其他较为密集的数据集我们设置了 σ 为 4.0. 为了提高网络的泛化性能，我们采用了数据增强的方法. 我们首先对图片进行随机裁剪，之后进行左右翻转、灰度化以及增强图像对比度. 实验中，我们设置了 4 000 个 epoch，batch size 的大小为 64，学习率为 5×10^{-6}，采用的优化器为 Adam.

3.4 评价指标

在所有实验中，我们采用了常见的平均绝对误差(MAE) 和均方根误差($RMSE$) 对计数精度进行评价. 两个评价指标定义如下：

$$MAE = \frac{1}{N} \sum_{i=1}^{N} | c_i - \bar{c}_i | \tag{6}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} | c_i - \bar{c}_i |^2} \tag{7}$$

其中：N 表示测试图像的数量，c_i 表示第 i 张图像的真实人数，$\bar{c}_i$ 表示第 i 张估计密度图的人数.

4　结果分析

4.1　消融实验

本文提出了基于背景辅助的高效人群计数多任务学习网络(BAMTLNet). 为了减少网络参数量，我们使用了 VGG 的前 7 层作为前端网络提取初级特征. 为了降低复杂背景对计数的影响，后端使用多任务网络硬参数共享机制，即生成估计密度图的主任务分支与复杂背景分割辅助任务分支共享前端网络参数. 为了更好地训练网络，我们提出了背景辅助多任务损失函数. 为了证明复杂背景分割辅助任务分支的作用，我们针对网络分支做了消融实验. 我们在 3 个人群计数基准数据集上与其他研究进行了对比实验. 实验证明，我们的网络在精度与网络轻量化上都达到了不错的效果.

为检验 BAMTLNet 中，背景分割辅助任务能降低复杂背景对计数精度的影响，我们对复杂背景分割辅助任务分支进行了消融. 如表 1 所示，我们在 ShanghaiTech Part A 上分别做了包含辅助任务分支与不包含辅助任务分支的实验. 可以看出，加上辅助任务后，计数精度提高了 1.3，表明加入背景分割辅助任务后，网络学习到了背景的语义信息，提高了计数性能.

表 1　BAMTLNet 在 ShanghaiTech Part A 上的消融实验结果

方　法	*MAE*	*RMSE*
VGG-7＋主任务	68.4	106.0
VGG-7＋主任务＋辅助任务	67.1	108.2

4.2　对比实验

为检验 BAMTLNet 在不同数据集上的性能，我们在 ShanghaiTech Part A，ShanghaiTech Part B，UCF_CC_50 和 UCF_QNRF 上进行了实验，并与其他人群计数网络进行了对比.

表 2 展示了 BAMTLNet 在 ShanghaiTech 数据集上的性能. 在对比结果表格中，我们的网络在ShanghaiTech 数据集上取得了不错的结果. Part A 的结果是 67.1，虽然没有比上 ADCrowdNet[3] 的结果，但是我们的网络参数量更小. 在 Part B 上，我们得到了更好的 *MAE* 值，为 7.8. 相比于 ADCrowdNet[3] 提高了 4.9%.

表 2　ShanghaiTech 数据集上的对比实验结果

方法	Part A		Part B	
	MAE	*RMSE*	*MAE*	*RMSE*
MCNN[8]	110.2	152.4	26.4	41.3
CMTL[10]	101.3	152.4	20.0	31.1
CP-CNN[11]	73.6	106.4	20.1	30.1
Switching-CNN[12]	90.4	135.0	21.6	33.4
DecideNet[13]	—	—	20.75	29.42
BSAD[14]	—	—	20.2	35.6
ACSCP[15]	75.7	102.7	17.2	27.4
PCC Net[7]	73.5	124.0	11.0	19.0
CSRNet[5]	68.2	115.0	10.6	16.0
ADCrowdNet[3]	63.2	98.9	8.2	15.7
BAMTLNet	67.1	108.2	7.8	13.0

表 3 展示了 UCF_CC_50 数据集上的对比实验. 其中，我们的实验结果 *MAE* 为 241.7，比 PCC Net[7] 低 1.7. 可能是由于该数据集图片数量太少，做五折交叉验证时划分 5 个图片集的划分方式导致我们的实验结果略低.

表3 UCF_CC_50数据集上的对比实验结果

方法	UCF_CC_50	
	MAE	*RMSE*
MCNN[8]	377.6	509.1
CP-CNN[11]	295.8	320.9
Switching-CNN[12]	318.1	439.2
ACSCP[15]	291.0	404.6
DDCN[16]	286.2	479.6
PCC Net[7]	240.0	315.5
BAMTLNet	241.7	323.8

表4展示了UCF_QNRF数据集上的对比实验结果. 我们的网络在该数据集上表现不错. 其中*MAE*为101.5，*RMSE*为167.8. 在数据集足够的情况下，我们的网络也可以在密集数据集上达到很好的性能.

表4 UCF_QNRF数据集上的对比实验结果

方法	UCF_QNRF	
	MAE	*RMSE*
MCNN[8]	277	426
Switching-CNN[12]	228	445
CMTL[10]	252	514
CL[17]	132	191
BAMTLNet	101.5	167.8

为了证明BAMTLNet网络参数量小，我们对网络参数量进行了计算比较，见表5. 可以看出BAMTLNet网络参数量为2.47×10^6，低于3×10^6但未低至1×10^6以下. 我们认为在一定程度上，参数量的增加可以带来一定的精度提升. 而参数量太小，网络学习到的特征越少，精度也会下降. 我们在精度与轻量化中取舍，保证一定的精度又适当地降低网络的参数量，使网络达到不错的性能. 一方面，模型参数量一定程度上反映了模型的计算量和所需时间. 另一方面，在具体比较计算时间时，因实验环境要求非常严格，需要保证完全相同的环境参数. 因此，本文未对比各模型在不同数据集上的计算时间. 然而，在网络模型参数量方面，我们的模型达到了2.47 M的参数轻量化，能在网络训练阶段减少训练时间. 同时，在人群计数精度方面，我们在ShanghaiTech Part A上获得了较为令人满意的结果.

表5 网络参数量对比

网络	MCNN[8]	Switching-CNN[12]	ACSCP[15]	CSRNet[5]	BAMTLNet
参数量/$\times10^6$	0.13	15.1	5.1	16.2	2.47

4.3 可视化结果

图2展示了BAMTLNet的可视化结果. 第一列表示原图像；第二列表示真实标签密度图；第三列表示网络所生成的估计密度图. 我们在3个数据集上对真实密度图与估计密度图做出了评估，并将其可视化. 可以看出，BAMTLNet生成的估计密度图与真实密度图标签相差不大，但在个别密集的区域会有些许的偏差. 因为我们的网络更关注图片的背景而非极度密集场景. 总体来说，BAMTLNet能完整生成图片的密度图并估计人数，性能表现优秀.

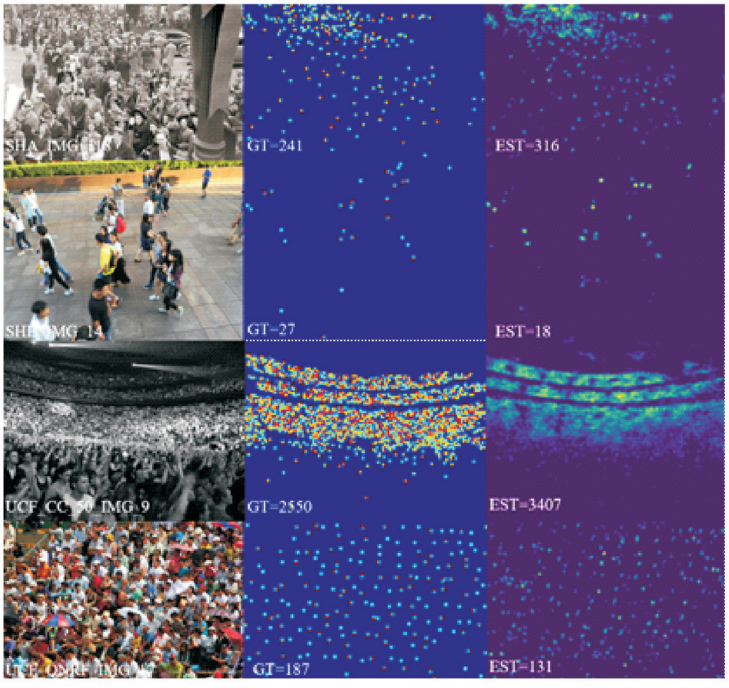

图 2　BAMTLNet 的可视化结果

5　结论

本文提出了基于背景辅助的高效人群计数多任务学习网络(BAMTLNet). 为了减少网络参数量，我们使用了 VGG 的前 7 层作为前端网络提取初级特征. 为了降低复杂背景对计数的影响，后端使用多任务网络硬参数共享机制，采用生成估计密度图的主任务分支与复杂背景分割辅助任务分支共享前端网络参数. 为了更好训练网络，我们提出了背景辅助多任务损失函数. 为了证明复杂背景分割辅助任务分支的作用，我们针对网络分支做了消融实验. 我们在 3 个人群计数基准数据集上与其他研究进行了对比实验. 实验证明，我们的网络在精度与网络轻量化上都达到了不错的效果.

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，2018，107：3-16.

[2] GAO G S，GAO J Y，LIU Q J，et al. CNN-Based Density Estimation and Crowd Counting：A Survey [EB/OL]. (2003-08-06)[2021-12-01]. https：//arxiv. org/abs/2003. 12783

[3] LIU N，LONG Y C，ZOU C Q，et al. ADCrowdNet：an Attention-Injective Deformable Convolutional Network for Crowd Understanding [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York：IEEE Computer Society Press，2019：3220-3229.

[4] SIMONYAN K，ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition [EB/OL]. (2014-05-06)[2021-12-05]. https：//arxiv. org/abs/1409. 1556.

[5] LI Y H，ZHANG X F，CHEN D M. CSRNet：Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes [J]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York：IEEE Computer Society Press，2018：1091-1100.

[6] ZHAO M M，ZHANG J，ZHANG C Y，et al. Leveraging Heterogeneous Auxiliary Tasks to Assist Crowd Counting [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York：IEEE Computer

Society Press, 2019: 12728-12737.

[7] GAO J Y, WANG Q, LI X L. PCC Net: Perspective Crowd Counting via Spatial Convolutional Network [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(10): 3486-3498.

[8] ZHANG Y Y, ZHOU D S, CHEN S Q, et al. Single-Image Crowd Counting via Multi-Column Convolutional Neural Network [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Computer Society Press, 2016: 589-597.

[9] IDREES H, SALEEMI I, SEIBERT C, et al. Multi-Source Multi-Scale Counting in Extremely Dense Crowd Images [C]// 2013 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Computer Society Press, 2013: 2547-2554.

[10] SINDAGI V A, PATEL V M. CNN-Based Cascaded Multi-Task Learning of High-Level Prior and Density Estimation for Crowd Counting [C]// 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). New York: IEEE Computer Society Press, 2017.

[11] SINDAGI V A, PATEL V M. Generating High-Quality Crowd Density Maps Using Contextual Pyramid CNNS [C]// 2017 IEEE International Conference on Computer Vision. New York: IEEE Computer Society Press, 2017: 1879-1888.

[12] SAM D B, SURYA S, BABU R V. Switching Convolutional Neural Network for Crowd Counting [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Computer Society Press, 2017: 4031-4039.

[13] LIU J, GAO C Q, MENG D Y, et al. DecideNet: Counting Varying Density Crowds through Attention Guided Detection and Density Estimation [J]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Computer Society Press, 2018: 5197-5206.

[14] HUANG S Y, LI X, ZHANG Z F, et al. Body Structure Aware Deep Crowd Counting [J]. IEEE Transactions on Image Processing, 2018, 27(3): 1049-1059.

[15] SHEN Z, XU Y, NI B B, et al. Crowd Counting via Adversarial Cross-Scale Consistency Pursuit [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Computer Society Press, 2018: 5245-5254.

[16] WANG L Y, YIN B Q, TANG X, et al. Removing Background Interference for Crowd Counting via De-Background Detail Convolutional Network [J]. Neurocomputing, 2019, 332: 360-371.

[17] IDREES H, TAYYAB M, ATHREY K, et al. Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds [EB/OL]. (2018-08-16)[2021-12-05]. http://www.cs.ucf.edu/~haroon/datafiles/Idrees_Counting_ECCV_2018.pdf.

责任编辑 张枸