计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(3):210−219 [doi: 10.15888/j.cnki.csa.009777] [CSTR: 32024.14.csa.009777]

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

基于多尺度特征加权融合注意力的密集人群计数算法①

时东亮, 葛　艳, 徐慕君

(青岛科技大学 信息科学技术学院, 青岛 266061)
通信作者: 葛　艳, E-mail: geyan@qust.edu.cn

摘　要: 针对人群计数面临的人头尺寸不统一、人群密度分布不均匀、背景复杂干扰等问题, 提出一种解决多尺度变化加强关注人群区域的卷积神经网络模型 (multi-scale feature weighted fusion attention convolutional neural network, MSFANet). 该网络前端采用改进的 VGG-16 模型对输入人群图像做第 1 步的粗粒度特征提取, 中间加入多尺度特征提取模块提取图像的多尺度特征信息. 随后添加注意力模块对多尺度特征进行特征加权. 后端利用锯齿状空洞卷积模块增大感受野, 以提取图像的细节特征, 生成高质量的人群密度图. 对该模型在 3 个公开数据集上进行实验, 结果显示, 在 Shanghai Tech Part B 数据集上 *MAE* (平均绝对误差) 达到 7.8, *MSE* (均方误差) 达到 12.5. 在 Shanghai Tech Part A 数据集上 *MAE* 达到 64.9, *MSE* 达到 108.4. 在 UCF_CC_50 数据集上 *MAE* 达到 185.1, *MSE* 达到 249.8. 实验结果证实该模型有较好的准确度和鲁棒性.

关键词: 人群计数; 深度学习; 多尺度特征提取; 注意力机制; 密度图

引用格式: 时东亮,葛艳,徐慕君.基于多尺度特征加权融合注意力的密集人群计数算法.计算机系统应用,2025,34(3):210–219. http://www.c-s-a.org.cn/1003-3254/9777.html

Dense Crowd Counting Algorithm Based on Multi-scale Feature Weighted Fusion Attention

SHI Dong-Liang, GE Yan, XU Mu-Jun

(College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: In response to challenges faced in crowd counting, such as non-uniform head sizes, uneven crowd density distribution, and complex background interference, a convolutional neural network (CNN) model (multi-scale feature weighted fusion attention convolutional neural network, MSFANet) that focuses on crowd regions and addresses multi-scale changes is proposed. The front end of the network adopts an improved VGG-16 model to perform the first step of coarse-grained feature extraction on the input crowd image. A multi-scale feature extraction module is added in the middle to extract the multi-scale feature information of the image. Then, an attention module is added to weigh the multi-scale features. At the back end, a sawtooth shaped dilated convolution module is adopted to increase the receptive field, extract the detailed features of the image, and generate high-quality crowd density maps. Experiments on this model are conducted on three public datasets. The results show that on the Shanghai Tech Part B dataset, the mean absolute error (*MAE*) is reduced to 7.8, and the mean squared error (*MSE*) decreases to 12.5. On the Shanghai Tech Part A dataset, the *MAE* is reduced to 64.9, and the *MSE* decreases to 108.4. On the UCF_CC_50 dataset, the *MAE* is reduced to 185.1, and the *MSE* decreases to 249.8. These experimental results affirm that the proposed model exhibits strong accuracy and robustness.

Key words: crowd counting; deep learning; multi-scale feature extraction; attention mechanism; density map

① 收稿时间: 2024-08-06; 修改时间: 2024-08-27; 采用时间: 2024-09-10; csa 在线出版时间: 2024-12-09
CNKI 网络首发时间: 2024-12-09

近年来, 世界经济快速发展, 人口总数不断增加, 截止至 2022 年 11 月 15 日, 世界总人口达到 80 亿. 世界人口的增多使得各种人群聚集性活动的增加. 例如大型集会、示威游行、音乐现场演唱会等[1], 人群过于密集的场景下如若不加以管控限制, 将会发生严重的公共安全事故, 威胁人们的生命财产安全, 造成恶劣的影响. 所以密集人群计数显得格外重要. 人员踩踏、财产丢失等事故的发生, 大多是因为人群密度过大并且没有及时疏通超出限制的人群. 为了能够对公共场所的人群数量以及人流密度进行检测控制, 有效防止因人群密度过大导致的事故的发生. 人群计数算法的应用成为目前最为广泛使用的方法. 而利用传统监控技术获取人流量信息大多依靠人力, 例如安排安保人员进行实地巡逻或设置监控室进行实时监控等. 单纯地利用人工方法对人群进行检测和监管, 通常会有误报、漏报等情况的发生. 而利用人群计数的技术对图像进行人群计数可实现对公共场所人群更有效的管理. 切实保障人民群众的生命财产安全, 创造美好的生存环境. 在公共安全设计方面, 利用人群计数分析可以从宏观上揭示设计上存在的缺陷, 用于改进公共空间的设计, 更好地保障人民群众的出行.

随着卷积神经网络 (convolutional neural network, CNN) 的发展[2], 由于其具有强大的特征提取能力和良好的适应性得到众多研究者的青睐. 在人群计数领域发展的早期, 研究人员普遍采用基于检测的方法[3]来进行人群计数, Laradji 等人[4]和 Liu 等人[5]提出的基于检测的人群计数算法对于低密度人群有不错的计数能力但是并不能适应高密度的现代社会的人群计数. 由于基于检测的方法不能适应高度拥挤的场景, 研究者尝试使用基于回归[6]的方法来进行人群计数. 与基于检测的方法类似, 回归方法可以分为基于整体的方法和基于块的方法. 基于整体的回归方法在处理大尺度和密度变化时都会遇到困难, 而基于块的回归方法包含了图像的更多局部信息, 并且受尺度和密度变化的影响较小. 因此, 基于块的回归方法的性能通常优于基于整体的回归方法. 近年来, 随着深度学习的发展, 基于深度学习的人群计数发展迅速. Fu 等人[7]是第一个将 CNN 应用于人群计数的方法, Zhang 等人[8]开发一种能够在任意人群密度和任意视角下从单个图像中准确估计人群数量的方法. 他们提出了一个简单而有效的多列卷积神经网络 (MCNN) 体系结构, 将图像映射到其人群密度图. Babu 等人[9]提出切换卷积神经网络 (Switching-CNN), 利用图像中人群密度的变化来预测人群数量的准确性和定位, 通过对合成的密度图进行求和得到具体的预测人数. Li 等人[10]提出了一个拥挤场景识别网络 CSRNet, 以提供一种数据驱动的深度学习方法来理解高度拥挤的场景, 进行准确的计数估计. Jiang 等人[11]提出包括密度注意网络 (DANet) 和注意扩展网络 (ASNet) 的人群计数方法, 使用了一种新的自适应金字塔损失 (APLoss) 来分层计算子区域的估计损失, 从而减轻了训练偏差. 针对人群计数统计时存在相机透视、人群遮挡等众多干扰因素, 左静等人[12]以空洞卷积为基础, 构建多尺度特征提取模块, 提升人群计数的准确性. 为了解决复杂的背景干扰问题, Gao 等人[13]在传统的回归 CNN 中引入空间和通道注意模型来估计密度图. 为了解决密集人群尺度变化的问题, Song 等人[14]提出了 P2PNet 来直接预测图像中头部点坐标及其置信度. 为了解决背景干扰、训练图像标注任务繁重等问题, Savner 等人[15]提出了一种使用金字塔视觉 Transformer[16]的弱监督人群计数方法, 进行了大量的评估来验证所提方法的有效性.

目前人群计数领域仍然存在诸多挑战和问题, 例如现实场景存在背景干扰、人群分布杂乱、行人尺度变化等问题, 极大地影响了计数精度. 在图 1(a) 中, 一些背景干扰, 例如树木与密集人群特征相似, 容易对计数造成干扰. 在图 1(b) 中, 人群分布杂乱不均将对计数性能造成影响. 图 1(c) 中, 存在人群被遮挡的问题[17].

(a) 背景干扰 (b) 人群分布不均 (c) 人群遮挡

图 1 人群计数面临的一些问题

针对上述密集人群计数领域存在的问题, 提出了本文算法, 利用并列的一组具有不同空洞率的空洞卷积提取多尺度的人头目标, 解决人头尺度不统一的问题. 利用 S2-MLP v2 注意力机制解决背景干扰等问题. 后端利用多组锯齿状空洞卷积排列成的空洞卷积模块增大感受野, 进一步减小损失, 提高人群计数的精度. 在多尺度特征提取模块和 S2-MLP v2 注意力模块之间加入残差连接结构重用特征, 减少训练过程可能出现

的梯度消失、爆炸等训练问题. 本文创新点主要有以下 3 点.

(1) 提出多尺度特征提取模块对前端提取的粗粒度特征进行细化, 利用全局和局部信息, 使得上下文信息融合, 更好地提取人群信息.

(2) 加入 S2-MLP v2 注意力模块对细粒度特征进行特征加权, 突出图像人群区域, 使得模型更加关注人群区域, 提高模型的计数能力.

(3) 设计多组锯齿形状空洞卷积扩大感受野, 进一步减少特征损失, 提高计数精度, 在不损失分辨率的情况下, 生成高质量的人群密度图.

1 基于多尺度特征加权融合注意力的密集人群计数算法

本文提出一种基于多尺度特征加权融合注意力的密集人群计数算法, 该模型针对人群特征进行多尺度特征提取, 并且加入全局和局部注意力更好地融合语义和尺度不一致的特征, 更好地解决目前人群计数领域存在的人群尺度不一致的问题, 具体来说在编码阶段前端改进的 VGG-16[18]卷积神经网络对特征进行初步提取, 中间部分加入多尺度特征提取模块对不同尺度的人头特征进行提取. 利用注意力加权模块对人群区域进行特征加权, 加入残差连接部分[19], 重用特征, 解决梯度消失的问题, 解码阶段的后端采用空洞卷积[20]来获取更大的感受野以生成高质量的密度图用来计数. 网络整体模型如图 2 所示.

1.1 前端特征提取模块

在人群计数领域, 回归问题在深层网络往往会出现梯度消失的问题, 如果有新增的很多模块, 模型将很难训练, 所以目前大多数人群计数任务的前端特征提取都用改进的结构简单的 VGG-16 网络. 所以我们所提网络前端也利用改进的 VGG-16 特征提取网络, 如图 2 中模块所示, 保留原 VGG-16 模型的前 10 层卷积层, 去除不适用于人群计数的全连接层以及 *Softmax* 层, 为了防止由于过多的池化层引起的特征丢失问题, 由原来的 5 层最大池化层更改为 3 层, 可以生成更高质量密度图的同时维持平移不变性[21]. 改进的 VGG-16 特征提取网络参数如表 1 所示.

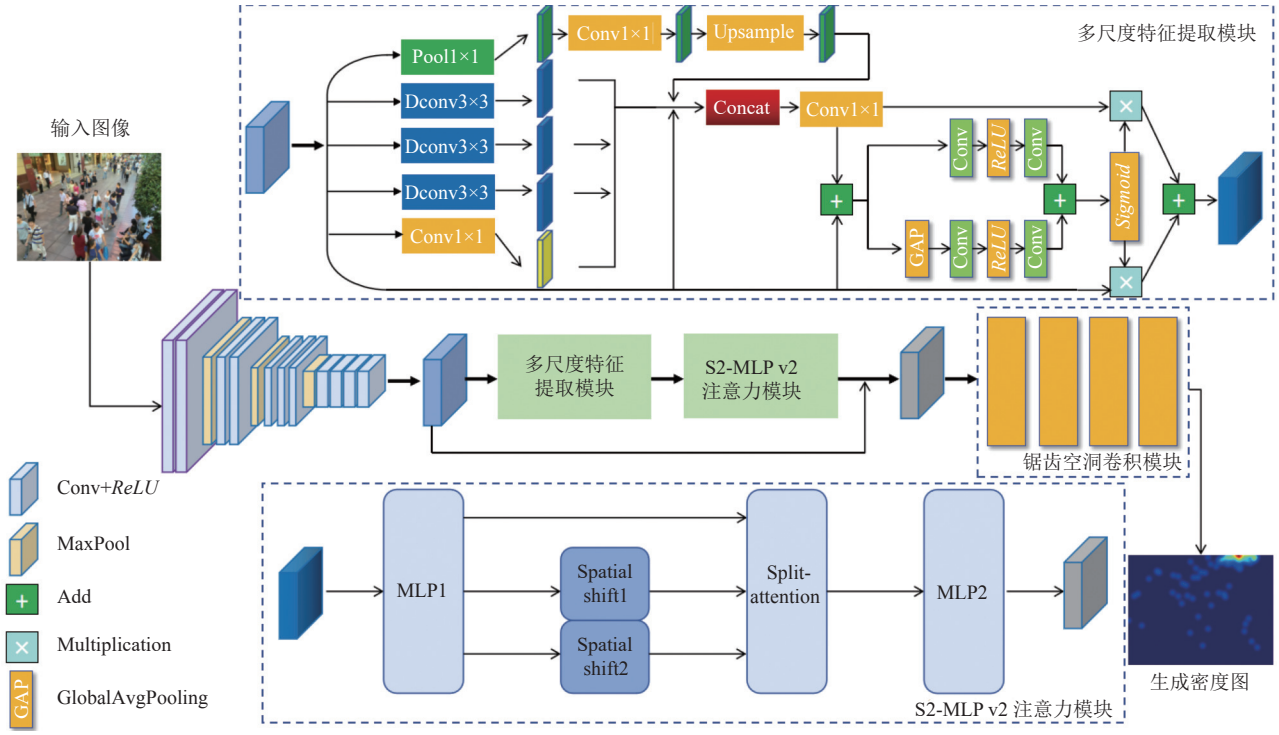

图 2 基于多尺度特征加权融合注意力的密集人群计数网络

如表 1 第 1 行所示，Conv1_1, 表示第 1 层卷积层中的第 1 个卷积核. (3_64_1) 表示卷积核尺度为 3, 卷积核个数为 64, 卷积核的步长为 1.

1.2 多尺度特征提取模块

Xu 等人[22]提出了一种深度信息引导的人群计数 DigCrowd 来应对人群密度很大的场景. 去解决人群分

布不均和目标尺寸不统一而影响人群计数准确度的问题. 由于人头尺寸在图片中存在大小, 近景的人头尺度大, 远景的人头尺度小. 为了能够充分学习到各种大小不统一的多尺度的人头信息, 需要充分利用上下文信息. 基于此, 设计了一种多尺度特征提取模块对前端网络提取到的特征进行上下文关联, 通过构建多个扩张率不同的空洞卷积捕获多尺度的人头目标特征信息. 如图 2 中模块所示. 首先, 将特征图并行送入 6 个分支进行多尺度的上下文信息提取, 其中 1×1 卷积用来保留输入特征图的空间细节信息, 如式 (1) 所示:

$$F_1 = f^{1\times1}(F) \tag{1}$$

其中, $f^{1\times1}$ 表示 1×1 卷积操作, F 表示输入特征图.

表 1 改进的 VGG-16 特征提取网络参数

序号	网络层	参数
1	Conv1_1	(3_64_1)
2	Conv1_2	(3_64_1)
3	MaxPool	—
4	Conv2_1	(3_128_1)
5	Conv2_2	(3_128_1)
6	MaxPool	—
7	Conv3_1	(3_256_1)
8	Conv3_2	(3_256_1)
9	Conv3_3	(3_256_1)
10	MaxPool	—
11	Conv4_1	(3_512_1)
12	Conv4_2	(3_512_1)
13	Conv4_3	(3_512_1)

3 个扩张率分别为 6, 12, 18 的 3×3 的空洞卷积学习不同尺度的人头目标信息, 如式 (2)–式 (4) 所示:

$$F_6 = f_D^6(F) \tag{2}$$

$$F_{12} = f_D^{12}(F) \tag{3}$$

$$F_{18} = f_D^{18}(F) \tag{4}$$

其中, f_D^6 表示空洞率为 6 的空洞卷积操作, f_D^{12} 表示空洞率为 12 的空洞卷积操作, f_D^{18} 表示空洞率为 18 的空洞卷积操作.

全局平均池化[23]分支能够保留输入图像的全局上下文信息, 为了使最后的特征图能够拼接, 使用 1 个 1×1 卷积之后对其进行双线性插值进行上采样, 如式 (5) 所示:

$$F_p = Upsample(f^{1\times1}(MaxPool(F))) \tag{5}$$

其中, $MaxPool$ 表示最大池化操作, $f^{1\times1}$ 表示 1×1 卷积操作, $Upsample$ 表示上采样操作.

通过上述操作得到 5 个分别学习了不同尺度信息的特征图, 我们对 5 个特征图进行融合, 可以得到融合多尺度的目标信息, 上下文信息的融合特征图. 该步骤是通过将特征图进行拼接实现的. 拼接之后的特征图存在通道维度过大, 计算量过大, 为了减少计算量, 设计 1×1 卷积恢复通道维度到输入时的维度, 实现从 2 560 个通道数压缩至 512 个通道数, 如式 (6) 所示:

$$F' = f^{1\times1}(Concat(F\ F_1\ F_6\ F_{12}\ F_{18}\ F_p)) \tag{6}$$

其中, $Concat$ 表示通道维度拼接.

为了更好地融合语义和尺度不一致的特征信息, 将得到的感受野较大的特征图和原始特征图进行跨尺度信息融合, 整体流程如图 3 所示, 首先将 F' 和 F 进行相加, 使其经过一组点卷积模块和一组全局平均池化模块, 分别如式 (7)、式 (8) 所示:

$$F^g = (Conv(ReLU(Conv(GAP(F+F'))))) \tag{7}$$

$$F^l = (Conv(ReLU(Conv(F+F')))) \tag{8}$$

其中, F^g 为经过全局平均池化模块的输出特征图, F^l 为经过点卷积模块的输出特征图. GAP 表示全局平均池化操作, $Conv$ 表示卷积操作, $ReLU$ 表示激活函数.

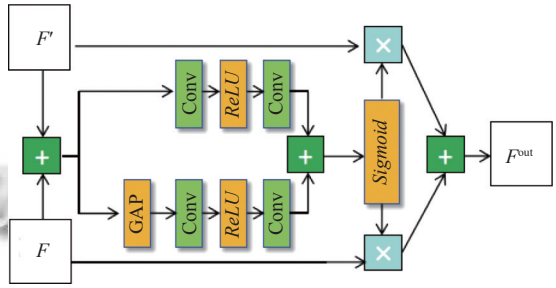

图 3 多尺度特征上下文信息融合

之后 F^g 和 F^l 相加, 经过 $Sigmoid$ 激活函数后得到计算权重, 使其与输入特征进行注意力操作得到最终的输出特征图, 如式 (9) 所示:

$$F^{\text{out}} = Sigmoid(F^g + F^l)\times F + (1 - Sigmoid(F^g + F^l))\times F' \tag{9}$$

其中, $Sigmoid$ 表示激活函数, F^{out} 表示最终输出的特征图, 该特征图是多尺度特征模块所提取的特征再次经过全局平均池化和局部卷积加权并且激活后的特征图.

1.3 S2-MLP v2 注意力模块

杜培德等人[24]提出了一种多尺度空间注意力机制

特征融合网络 MAFNet 来减少尺度变化带来的影响, 同时利用注意力机制去除背景干扰带来的影响. 在人类处理图像视觉数据时, 通常会关注我们感兴趣的部位, 自动忽略掉我们不感兴趣的部位, 这就是人类视觉的注意力机制. 注意力机制可以突出学习到我们感兴趣的特征, 利用这一点, 本文引入 S2-MLP v2[25]注意力模块, 如图 2 中模块所示. 注意力机制可分为通道注意力机制、空间注意力机制、混合域注意力机制, 其中代表有 SENet[26]、CA[27]、CBAM[28]等. S2-MLP v2 注意力是一种区别于 CNN 和 Transformer 的新架构, 它无需卷积与自注意力. S2-MLP v2 的注意力机制仅依赖于在空域或者特征通道上重复实施的多层感知机来实现, 仅依赖于基础矩阵乘操作、数据排布变换 (比如 reshape、transposition) 以及非线性层. 具体实现是将输入 $C\times H\times W$ 的特征图 $F\in R^{C\times H\times W}$ 经过 1 个全连接的MLP, 对特定位置信息进行交流, 第 1 个MLP的输出的通道维度是输入的 3 倍, 公式如式 (10) 所示:

$$F_{m1}=MLP_1(F)\in R^{3C\times H\times W} \tag{10}$$

其中, MLP_1 是第 1 个全连接的MLP, $F_{m1}\in R^{3C\times H\times W}$ 是经过第 1 个MLP后的特征图.

拓展后的特征图 F_{m1} 沿着通道维度平均分割成 3 部分, 分别用来做接下来的空间信息 Spatial-shift 操作的 3 组输入, 3 组分割分别如式 (11)–式 (13) 所示:

$$F_1=F_{m1}[:,:,1:C] \tag{11}$$

$$F_2=F_{m1}[:,:,C+1:2C] \tag{12}$$

$$F_3=F_{m1}[:,:,2C+1:3C] \tag{13}$$

Spatial-shift 操作是一个固定的空间位置移动的操作, 不需要额外的参数, 如图 4 所示, 在通道维度分成 4 组, 相当于在 4 个方向上的空间移动操作.

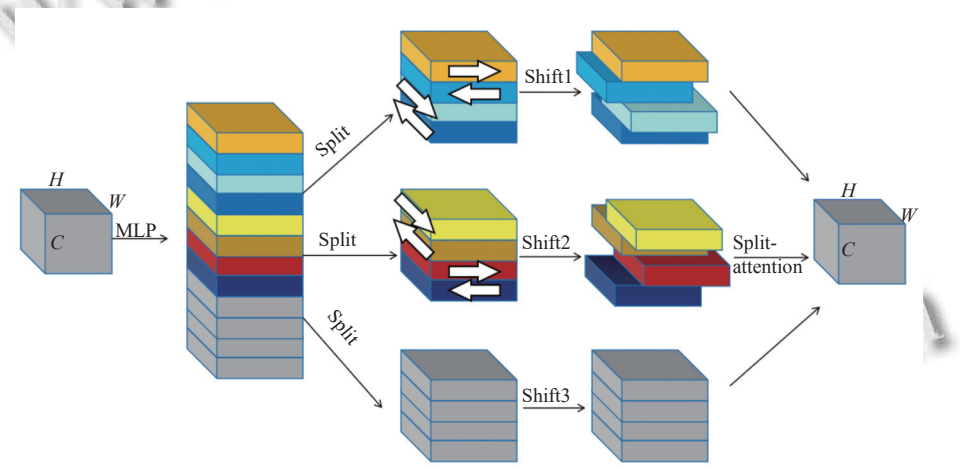

图 4 S2-MLP v2 注意力模块 Spatial-shift 操作图

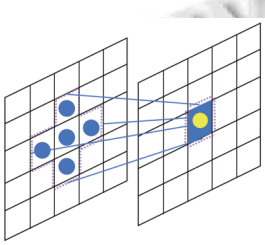

图 5 S2-MLP v2 注意力模块感受野图

空间位移会增大感受野, 通过向不同的方向进行空间移动, 我们得知在形如位置 (x,y) 处会增加来自 4 个方向 $(x-1,y)$, $(x,y-1)$, $(x+1,y)$, $(x,y+1)$ 的特征图空间信息, 再加上第 3 组不变的原始特征图, 最终的感受野是菱形的. 如图 5 所示, 实现了不同尺度和特征通道维度的信息通信. 空间位移操作后将调整了的特征图送入Split-attention模块中, Split-attention模块来捕获输入特征图在通道维度上的关键信息, 对来自不同操作的多组特征图进行增强, 将这些信息与原始特征图进行加权组合, 提升模型的表达能力.

1.4 锯齿空洞卷积模块

为了获取高质量的密度图, 通过利用多组空洞卷积作为生成密度图前的卷积, 以此来扩大感受野并在不丢失分辨率的情况下提取到深层特征. 空洞卷积的公式如式 (14) 所示:

$$y(m,n)=\sum_{i=1}^{M}\sum_{j=1}^{N}x(m+r\times i,n+r\times j)w(i,j) \tag{14}$$

其中, $y(m,n)$是$x(m,n)$和滤波器$w(i,j)$输入空洞卷积操作后的输出, 其长度和宽度分别为 m 和 n, 参数 r 是空洞率 (dilation rate).

空洞卷积是一种使用特殊核的卷积, 使用的是可以自定义空洞率的稀疏核, 这种稀疏核可以实现网络参数量不增加的情况下扩大感受野, 同样增加多层普通卷积层也可以增大感受野, 但是会增大计算量, 增加模型的参数量. 图 6 是 3×3 卷积在不同空洞率下的感受野大小. 图 7 是 3×3 卷积在空洞率为 2 情况下的感受野大小.

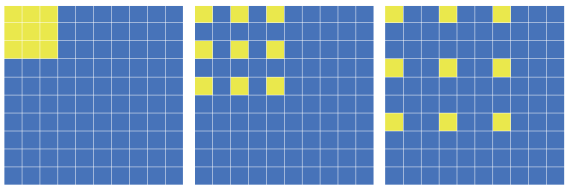

图 6 空洞卷积感受野大小图

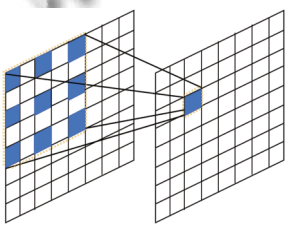

图 7 3×3 空洞卷积在空洞率为 2 时的感受野图

使用空洞率相同的多个空洞卷积会产生局部信息丢失的问题, 这是由于空洞卷积某一层的卷积结果来自上一层的独立集合, 没有相互依赖关系, 并且由于空洞卷积的稀疏采样使得远距离卷积得到的信息之间的相关性大大减少.

通常可以设置规则使得多个空洞卷积同时使用时避免出现上述问题, 基于文献[29]提到的方法, 可以设置连续空洞卷积的空洞率分别为 1、2、3, 这称之为锯齿状空洞卷积. 这样设计使得上层卷积可以从更广泛的像素范围内提取信息, 它们位于与原始配置相同的区域. 这个过程在所有的卷积层中都重复进行, 从而使上层的感受野保持不变.

基于以上分析, 本文设计 6 个空洞卷积, 空洞率采用锯齿形状排布, 分别为 1、2、3、2、2、2, 卷积核个数分别为 512, 512, 512, 256, 128, 64,使之逐渐恢复到输入时的通道数, 最后通过 1 个 1×1 卷积来使得通道数降为 1, 用来生成高质量的密度图.

1.5 密度图的生成

人群计数领域一般采用将图片生成密度图来进行训练和计算损失, 最终的预测也是将预测图片通过训练好的网络生成对应的人群密度图进行积分来得到最终的人数结果. 密度图可以衡量图片中人群分布情况和人群的密集程度. 我们对密度图的生成采用文献[8]提出的方法, 一张图片在像素点$X_i(x,y)$ 位置处有一个人的头部, 我们可以将其用函数$\delta(x-X_i)$来表示, 则一幅含有 N 个人头部位置标记的图像可表示为式 (15):

$$H(x)=\sum_{i=1}^{N}\delta(x-X_i) \tag{15}$$

为了将其转化为连续的密度函数形式, 便于人数计数进行积分这里利用一个高斯核函数$G_\sigma(x)$与该函数进行卷积操作, 从而生成一个连续的密度图函数如式 (16) 所示:

$$F(x)=H(x)*G_\sigma(x) \tag{16}$$

其中，*表示卷积操作.

然而, 我们所拍摄的照片由于焦距不同、失真等问题导致人头部位会失真, 或者人头部大小不一等问题. 根据文献[10]提出的解决方法, 可以在密集人群中通过计算每个人头部的位置与其相邻人头部的平均距离自适应地确定每个人的参数σ, 具体的高斯核参数σ计算过程如下所示, 对于一副确定的人群图片, 我们得到每个人头位置的坐标x_i, 定义与其相近的头部的距离为$\{d_{i1},d_{i2},d_{i3},d_{i4},\cdots,d_{ik}\}$, 则$k$个相邻头部到头部$x_i$的平均距离可以定义为式 (17) 所示:

$$d_i=\frac{1}{k}\times\sum_{j=1}^{k}d_{ij} \tag{17}$$

所以, 高斯核函数的参数σ可以由平均距离d_i来确定, 所以最终的密度图函数F定义为式 (18):

$$F(x)=\sum_{i=1}^{N}\delta(x-x_i)G_{\sigma_i}(x),\ \text{with}\ \sigma_i=\beta d_i \tag{18}$$

其中, N代表图像中人头的总个数, x_i代表图像中一个人头部的坐标位置, 根据文献[10]中的相关配置, 当 $\beta=0.3$ 且 $k=3$ 时密度图较为准确, 针对不同的数据集, 可通过改变高斯核平均头部大小生成所有人头标注.

1.6 损失函数

人群计数领域大多采用欧氏距离作为损失函数来约束网络的训练学习过程, 本文网络也采用欧氏距离作为损失函数, 用来衡量模型预测值与真实值之间的

差异, 衡量预测密度图与真实密度图之间的差异. 如果预测密度图与真实密度图非常接近, 则损失值较小, 如果差异较大, 则损失值较大, 其数学含义为计算预测值与真实值之间的直线距离, 以此来量化模型预测错误的程度, 从而对其进行约束, 使模型更好的学习, 欧氏距离损失函数的定义如式 (19) 所示:

$$L=\frac{1}{2N}\sum_{i=1}^{N}\left\|Z(X_i;\theta)-Z_i^{\mathrm{GT}}\right\|_2^2 \tag{19}$$

其中, N为 1 个训练批次的图片数量, $Z(X_i,\theta)$是网络参数当前为θ时对输入的图片X_i做出的预测密度图, Z_i^{GT}是对应的输入图片X_i对应的标签文件, 即为真实密度图.

2 实验数据

2.1 核函数参数的选择

由于不同数据集人群分布差距较大, 对应的标签文件生成时参数的选择也有所不同, 根据文献[8]的设计, 对于不同数据集生成密度图文件采用的高斯核α的选择如表 2 所示.

表 2 不同数据集生成密度图文件采用的高斯核 α

数据集	高斯核α
Shanghai Tech Part A[8]	自适应高斯核
Shanghai Tech Part B[8]	固定高斯核α=15
UCF_CC_50[30]	自适应高斯核

2.2 评价标准

人群计数领域常用的评价标准是平均绝对误差 (mean absolute error, *MAE*) 和均方误差 (mean squared error, *MSE*), 本文采用 *MAE* 和 *MSE* 作为衡量算法性能的标准. 计算公式如式 (20)、式 (21) 所示:

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|z_i-\hat{z}_i| \tag{20}$$

$$MSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(z_i-\hat{z}_i)} \tag{21}$$

其中, N为测试图片的数量, z_i代表测试图片对应的真是密度图的值, $\hat{z}_i$代表测试图片经过网络预测的密度图的值.

3 结果与分析

3.1 模型训练平台

本文所有实验训练过程统一环境为 Ubuntu 20.04, GPU 为 NVIDIA GeForce RTX 3090, 实验框架为 PyTorch 1.12.0+Python 3.7+Cuda 11.3+anaconda 3. 网络训练过程由于数据集存在尺寸不统一的问题, 故采用批量为 1 对其进行训练, 使用 SGD (stochastic gradient descent) 优化器优化训练过程, 初始学习率 (learning rate) 设置为 1E–7, 动量 (momentum) 设置为 0.95, 训练轮次设置为 400 轮. 前端特征提取模块使用在 Image Net 数据集上预训练的 VGG-16 的参数, 其余网络层均采用均值为 0, 标准差为 0.01 的高斯分布随机初始化.

3.2 实验结果

Shanghai Tech[8]数据集是人群计数领域最常用的数据集, 总共包含 1 198 张人群分布各不相同的密集人群图片, 总计 330 165 人. 根据人群分布的密度化差异被分为 Part A 和 Part B 两部分, 其中 Part B 是从上海街头进行人工拍摄所获取的 716 张较为稀疏的人群图片, 场景相对较简单且密集度较低, 其中包含人数从 9–578 不等, 分为训练集和测试集, 分别有 400 张图片和 316 张图片. 表 3 展示了本文算法与近些年来人群计数领域优秀算法的对比结果. 结果显示本文算法在 Shanghai Tech Part B 数据集上达到了最好的 *MAE* (7.8) 和 *MSE* (12.5), 结果表明本文所提出的算法可以在人群分布相对稀疏的人群图片中有很好的计数能力.

表 3 在 Shanghai Tech Part B 数据集上不同网络模型对比

网络	*MAE*	*MSE*
MCNN[8]	26.4	41.3
Switching-CNN[9]	21.6	33.4
CP-CNN[31]	20.1	30.1
CSRNet[10]	10.6	16.0
CFFNet[32]	9.6	17.7
MSFNet[12]	9.6	14.3
RDCAF[33]	8.5	14.2
SANet[34]	8.4	13.6
本文	**7.8**	**12.5**

图 8 展示了本文算法在 Shanghai Tech Part B 数据集上部分可视化预测结果, 其中 Est 表示预测人数, Gt 表示真实人数. 可以看出本文所提算法在人群分布不均匀的情况下仍能保持较高的计数准确度. 在人头尺度不统一的情况下得到了较好的计数结果, 说明本文所提算法具有良好的多尺度特征的提取能力.

Part A 部分是从互联网上爬取的 482 张较为密集的人群图片组成的, 由于密集度大幅度变化, 难度较高, 是人群计数算法的挑战之一. 其中包含人数从 33–3 139 不等, 分为训练集和测试集两部分, 分别有 300 张图片

和 182 张图片. 表 4 展示了本文所提出的算法在 Shanghai Tech Part A 数据集上达到了最好的 *MAE* (64.9) 和较好的 *MSE* (108.4), 表明本文所提出的算法可以在人群分布较为密集的人群图片中实现很好的计数能力.

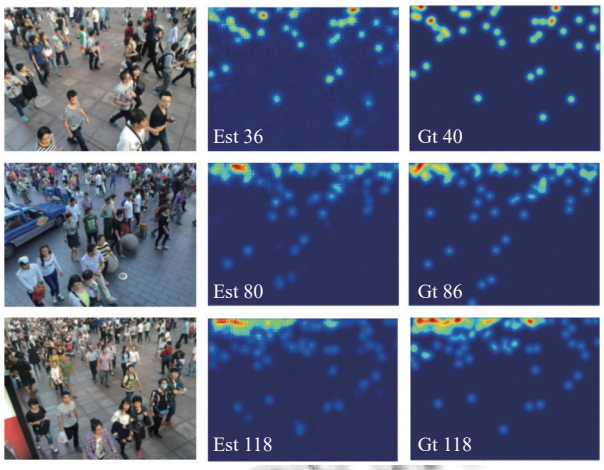

图 8 Shanghai Tech Part B 数据集部分可视化结果

表 4 在 Shanghai Tech Part A 数据集上不同网络模型对比

模型	*MAE*	*MSE*
文献[35]	181.8	277.7
MCNN[8]	110.2	173.2
Switching-CNN[9]	90.4	135.0
CP-CNN[31]	73.6	**106.4**
ADCrowdNet[36]	70.9	115.2
CSRNet[10]	68.2	115.0
本文	**64.9**	108.4

图 9 是本文所提算法在 Shanghai Tech Part A 数据集上部分图片的可视化结果. 结果表明本文所提算法能够应对多角度密集人群照片, 并且能取得较好的计数精度. 在面对人群遮挡、背景复杂等问题时, 也表现出较好的计数准确度. 本文所涉及的注意力机制模块发挥了作用, 提高了模型对人群部位的关注程度.

UCF_CC_50[30]是人群计数领域首个密集人群图像数据集. 数据集总共含有 50 张不同分辨率的灰度图像, 密度极大, 场景复杂, 对网络计数能力要求较高, 是一个非常具有挑战性的数据集. 图片中人数最少的有 94 人, 人数最多的有 4 543 人, 人数密集程度波动特别大, 对于网络的泛化能力要求很高, 50 张图片的人群标注平均为 1 280 个, 总的标注数为 63 974 个. 根据文献[30]的检测标准, 通常采用五折交叉验证来对该数据集上的模型进行验证性能. 表 5 展示了本文所提出的算法在 UCF_CC_50 数据集上达到了最好的 *MAE* (185.1) 和较好的 *MSE* (249.8). 结果表明本文所提出的算法在人群分布非常密集的人群图片中实现很好的计数能力, 本文所提网络对于超密集人群有不错的泛化能力.

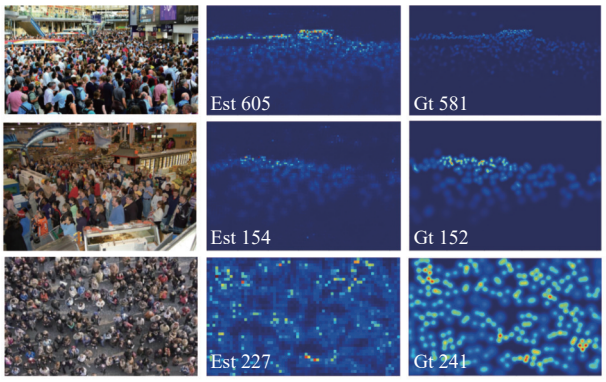

图 9 Shanghai Tech Part A 数据集部分可视化结果

表 5 在 UCF_CC_50 数据集上不同网络模型对比

模型	*MAE*	*MSE*
MCNN[8]	377.6	509.1
Switching-CNN[9]	318.1	439.2
ACSCP[37]	291.0	404.6
CSRNet[10]	266.1	397.5
SANet[34]	258.4	334.9
TEDNet[38]	249.4	354.5
MSEN[39]	226.7	310.6
MFFBSNet[40]	196.2	274.6
本文	**185.1**	**249.8**

图 10 是本文所提算法在 UCF_CC_50 数据集上的部分可视化预测结果. 结果显示, 本文所提算法在超密集人群图片中可以得到理想的计数结果. 在人头尺度变化极大的图片中可以得到较高的计数精度, 面对人群遮挡、视角畸变、复杂前景背景干扰仍能保持良好的计数能力.

3.3 消融实验

为验证本文所使用模块的有效性, 本文在 Shanghai Tech PartB 数据集上进行消融实验, 实验结果如表 6 所示. 其中, VGG 表示只有前端改进的 VGG-16 特征提取模块, Dconv 表示后端空洞卷积模块, MSF 指本文提出的 MSFANet 多尺度特征提取模块, S2 指本文提出的 S2-MLP v2 注意力机制模块. VGG＋Dconv 表示使用前端改进的 VGG-16 特征提取模块和后端空洞卷积模块; VGG＋Dconv＋MSF 表示使用了除 S2-MLP v2 注意力机制模块外的所有模块. 由实验结果可知, 同时使用多尺度特征提取模块、S2-MLP v2 注意力机制模块、空洞卷积模块时, 实验效果最佳, 验证了所提模块的有效性, 大幅度提高了网络的学习能力以及泛化能力.

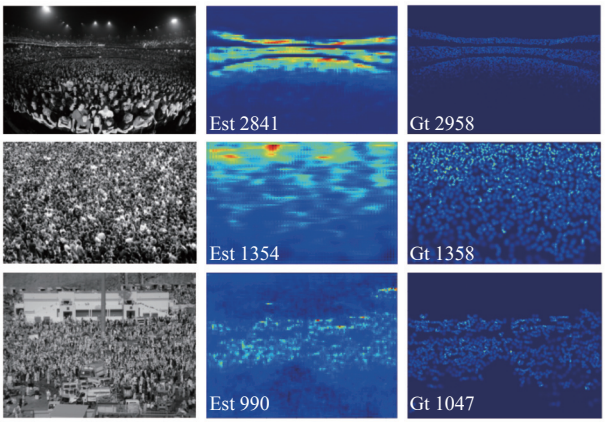

图 10 UCF_CC_50 数据集部分可视化结果

表 6 在 Shanghai Tech PartB 数据集上消融实验结果

模型	*MAE*	*MSE*
VGG	12.2	19.0
VGG＋Dconv	11.1	17.0
VGG＋Dconv＋MSF	9.2	15.7
本文 (VGG＋Dconv＋MSF+S2)	**7.8**	**12.5**

4 结论

本文提出了一种多尺度特征加权融合注意力的密集人群计数算法. 该算法从 VGG-16 改进提取初步特征, 将初步特征进行多尺度特征的进一步提取, 利用 S2-MLP v2 注意力机制进行特征加权, 抑制背景干扰, 使用锯齿状空洞卷积加大网络的感受野并保持高分辨率, 最终输出高分辨率的密度图, 实现更高精度的计数, 降低计数误差. 最后的实验结果表明了本文所提算法在面对复杂背景干扰、人头尺度不统一、人群分布不均匀的情况下表现出很好的计数能力, 并且有很好的泛化能力, 能够在多种人群分布极不相同的数据集中表现出较好的计数精度. 同时本文所提算法具有多尺度特征提取的能力, 能够使模型更加关注人群区域, 降低计数误差, 提高计数能力, 在一定程度上解决了人群计数领域存在的一些问题.

参考文献

1 余鹰, 朱慧琳, 钱进, 等. 基于深度学习的人群计数研究综述. 计算机研究与发展, 2021, 58(12): 2724–2747. [doi: 10.7544/issn1000-1239.2021.20200699]

2 Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 1980, 36(4): 193–202. [doi: 10.1007/BF00344251]

3 Topkaya IS, Erdogan H, Porikli F. Counting people by clustering person detector outputs. Proceedings of the 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Seoul: IEEE, 2014. 313–318.

4 Laradji IH, Rostamzadeh N, Pinheiro PO, *et al*. Where are the blobs: Counting by localization with point supervision. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 560–576.

5 Liu YT, Shi MJ, Zhao QJ, *et al*. Point in, box out: Beyond counting persons in crowds. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 6462–6471.

6 Ryan D, Denman S, Sridharan S, *et al*. An evaluation of crowd counting methods, features and regression models. Computer Vision and Image Understanding, 2015, 130: 1–17. [doi: 10.1016/j.cviu.2014.07.008]

7 Fu M, Xu P, Li XD, *et al*. Fast crowd density estimation with convolutional neural networks. Engineering Applications of Artificial Intelligence, 2015, 43: 81–88. [doi: 10.1016/j.engappai.2015.04.006]

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

9 Babu Sam D, Surya S, Venkatesh Babu R. Switching convolutional neural network for crowd counting. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4031–4039.

10 Li YH, Zhang XF, Chen DM. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1091–1100.

11 Jiang XH, Zhang L, Xu ML, *et al*. Attention scaling for crowd counting. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 4705–4714.

12 左静, 巴玉林. 基于多尺度融合的深度人群计数算法. 激光与光电子学进展, 2020, 57(24): 241502.

13 Gao JY, Wang Q, Yuan Y. SCAR: Spatial-/channel-wise attention regression networks for crowd counting. Neurocomputing, 2019, 363: 1–8. [doi: 10.1016/j.neucom.2019.08.018]

14 Song QY, Wang CA, Jiang ZK, *et al*. Rethinking counting and localization in crowds: A purely point-based framework. Proceedings of the 2021 IEEE/CVF International Conference

on Computer Vision. Montreal: IEEE, 2021. 3345–3354.
15 Savner SS, Kanhangad V. CrowdFormer: Weakly-supervised crowd counting with improved generalizability. Journal of Visual Communication and Image Representation, 2023, 94: 103853. [doi: 10.1016/j.jvcir.2023.103853]
16 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 6000–6010.
17 余鹰, 潘诚, 朱慧琳, 等. 融合通道与空间注意力的编解码人群计数算法. 计算机科学与探索, 2022, 16(11): 2547–2556. [doi: 10.3778/j.issn.1673-9418.2104122]
18 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556v6, 2015.
19 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.
20 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122v3, 2016.
21 Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe: ACM, 2012. 1097–1105.
22 Xu ML, Ge ZY, Jiang XH, *et al*. Depth information guided crowd counting for complex crowd scenes. Pattern Recognition Letters, 2019, 125: 563–569. [doi: 10.1016/j.patrec.2019.02.026]
23 Lin M, Chen Q, Yan SC. Network in network. arXiv:1312.4400v3, 2014.
24 杜培德, 严华. 基于多尺度空间注意力特征融合的人群计数网络. 计算机应用, 2021, 41(2): 537–543. [doi: 10.11772/j.issn.1001-9081.2020060793]
25 Yu T, Li X, Cai YF, *et al*. S^2-MLP: Spatial-shift MLP architecture for vision. Proceedings of the 2021 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2021. 3615–3624.
26 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.
27 Hou QB, Zhou DQ, Feng JS. Coordinate attention for efficient mobile network design. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 13708–13717.
28 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 3–19.
29 Wang PQ, Chen PF, Yuan Y, *et al*. Understanding convolution for semantic segmentation. Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe: IEEE, 2018. 1451–1460.
30 Idrees H, Saleemi I, Seibert C, *et al*. Multi-source multi-scale counting in extremely dense crowd images. Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 2547–2554.
31 Sindagi VA, Patel VM. Generating high-quality crowd density maps using contextual pyramid CNNs. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 1879–1888.
32 邹敏, 黄虹, 杜浸, 等. 基于特征融合编解码的人群计数和密度估计. 计算机工程与设计, 2023, 44(7): 2110–2117. [doi: 10.16208/j.issn1000-7024.2023.07.025]
33 Chen K, Loy CC, Gong SG, *et al*. Feature mining for localised crowd counting. Proceedings of the 2012 British Machine Vision Conference. Surrey, 2012. 1–11.
34 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 (ECCV). Munich: Springer, 2018. 757–773.
35 Zhang C, Li H, Wang X, *et al*. Cross-scene crowd counting via deep convolutional neural networks. Proceedings of the 2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2015. 833–841.
36 Liu N, Long YC, Zou CQ, *et al*. ADCrowdnet: An attention-injective deformable convolutional network for crowd understanding. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3220–3229.
37 Shen Z, Xu Y, Ni BB, *et al*. Crowd counting via adversarial cross-scale consistency pursuit. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5245–5254.
38 Jiang XL, Xiao ZH, Zhang BC, *et al*. Crowd counting and density estimation by trellis encoder-decoder networks. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 6126–6135.
39 徐涛, 段仪浓, 杜佳浩, 等. 基于多尺度增强网络的人群计数方法. 电子与信息学报, 2021, 43(6): 1764–1771. [doi: 10.11999/JEIT200331]
40 赵佳彬, 徐慧英, 朱蓉, 等. 基于多尺度特征融合与背景抑制的 MFFBSNet 网络人群计数算法. 计算机工程与科学, 1–13. http://kns.cnki.net/kcms/detail/43.1258.TP.20240627.1334.002.html. (2024-06-29)[2024-08-27].

(校对责编: 王欣欣)