计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(1):241−248 [doi: 10.15888/j.cnki.csa.008892]

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

融合双注意力机制的人群计数算法①

徐晓晨, 葛　艳, 杜军威, 陈　卓

(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: 针对背景复杂、遮挡、人群分布不均等人群计数常见问题, 提出了一种结合联合损失的空间-通道双注意力机制卷积神经网络模型 (joint loss-based space-channel dual attention network, JL-SCDANet). 该网络前端进行图像粗粒度特征提取, 中间加入空间注意力机制以及通道注意力机制突出图像重点区域, 后端使用可加大感受野且不丢失图像分辨率的空洞卷积提取深层二维特征. 此外, 该模型结合联合损失函数进行训练, 以增强模型的鲁棒性. 为了验证模型的改进效果, 在 3 个公共数据集 (ShanghaiTech Part B、mall 和 UCF_CC_50) 上分别进行了对比实验, 在 ShanghaiTech Part B 数据集中平均绝对误差 (*MAE*) 和均方误差 (*MSE*) 分别达到了 8.13 和 13.13; 在 mall 数据集中 *MAE*、*MSE* 达到了 1.78 和 2.28; 在 UCF_CC_50 数据集中 *MAE*、*MSE* 分别达到了 182.12 和 210.24, 实验结果证明了该网络在提高人数统计准确率上的有效性.

关键词: 人群计数; 人群密度图; 卷积神经网络 (CNN); 注意力机制; 空洞卷积; 深度学习

引用格式: 徐晓晨,葛艳,杜军威,陈卓.融合双注意力机制的人群计数算法.计算机系统应用,2023,32(1):241–248. http://www.c-s-a.org.cn/1003-3254/8892.html

Crowd Counting Algorithm Based on Dual Attention Mechanism

XU Xiao-Chen, GE Yan, DU Jun-Wei, CHEN Zhuo

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

Abstract: Given the common problems of crowd counting with a complex background, occlusion, and uneven crowd distribution, a joint loss-based space-channel dual attention network (JL-SCDANet) is proposed. The front end of the network extracts coarse-grained features of an image, and the spatial attention mechanism and channel attention mechanism are added in the middle to highlight the key areas of the image, while the back end uses dilated convolution that can increase the receptive field without losing the image resolution to extract deep two-dimensional features. In addition, the model is trained with the joint loss function to enhance its robustness. Comparative experiments are carried out on three public data sets (i.e., ShanghaiTech Part B, mall, and UCF_CC_50) to verify the improvement effect of the model. In terms of the mean absolute error (*MAE*) and mean square error (*MSE*), the results on ShanghaiTech Part B, mall, and UCF_CC_50 reach 8.13 and 13.13, 1.78 and 2.28, and 182.12 and 210.24, respectively. The experimental results prove the effectiveness of the network in improving the accuracy of population statistics.

Key words: crowd counting; crowd density map; convolutional neural network (CNN); attention mechanism; dilated convolution; deep learning

人群密度估计在公共安全管理、公共空间设计、数据收集分析等方面的潜在影响不容忽视. 出于多种原因, 在不同的场所中都有因人群过于聚集而出现踩踏、堵塞等公共安全事故发生的机率, 因此人群密度

① 基金项目: 山东省自然科学基金 (ZR2021MF092)
收稿时间: 2022-05-16; 修改时间: 2022-06-15; 采用时间: 2022-06-29; csa 在线出版时间: 2022-11-14
CNKI 网络首发时间: 2022-11-15

估计在诸多公共安全领域中都有较高的研究价值. 通过对图像或公共视频监控信息进行人流密度估计并进行人群计数可实现对公共场所人群的有效管理, 即时输出每个场景中的人员数目, 对人口密度较高的区域进行及时警告以预判是否会导致事故发生, 从而采取适当措施降低事故发生指数. 但是在人群密度估计的实际应用中存在诸多难点, 例如物体遮挡、人员分布不规则、背景复杂以及不同的摄像视角等, 这些问题都会致使人群密度预测图与真实密度图之间相距较大, 不能准确估计人群分布以及进行人群计数, 从而使得针对某一场景的人员计数工作变得极为困难.

近年来, 国内外研究人员尝试使用深度学习方法[1]攻破人群计数难点. 例如: Zhang 等人[2] 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 MCNN, 该模型允许输入任意大小或分辨率的图像, 每个 CNN 列学习的特征可以适应由于透视效果或图像分辨率而导致的人或人头部大小的变化. 针对人群计数统计时存在相机透视、人群重叠、人群遮挡等众多干扰因素, 左静等人[3] 提出一种多尺度融合的深度人群计数算法. Gao 等人[4] 在传统的回归 CNN 中引入空间/通道注意模型来估计被称为“SCAR”的密度图. Xu 等人[5] 提出了一种深度信息引导的人群计数 DigCrowd 来处理拥挤的场景. 为解决公共场所中人群分布不均以及目标尺度不一而影响人数估计的问题, 袁健等人[6] 提出了基于图像视野划分的公共场所人群计数模型 IFDM. Zou 等人[7] 提出了自适应容量多尺度卷积神经网络 ACM-CNN, 可以为输入的不同部分分配不同的容量. Kong 等人[8] 通过引入弱监督人群注意网络, 提出了一种新的鲁棒人群计数方法 CWAN. 杜培德等人[9]提出了一种多尺度空间注意力特征融合网络 MAFNet来减少尺度变化和遮挡带来的影响. 杨旭等人[10] 提出两种多尺度特征融合结构: 注意力加权融合模块 AWF和自底向上融合模块 BUF. 沈宁静等人[11] 提出一种残差密集连接与注意力融合的人群计数算法. Wang 等人[12] 提出了一种混合注意网络 (HAN), 它利用渐进式嵌入尺度上下文 (PES) 信息, 使网络能够同时抑制噪声和适应头部尺度的变化. 通过并联空间注意和通道注意模块, 构建混合注意机制, 使网络更加关注头部区域, 减少背景物体的干扰. 由此可见, 不断有国内外研究者尝试在不同方面进行算法创新以获取更高的准确率, 提高人群计数精度. 虽然上述方法都在一定程度上解决了人群计数问题, 但是仍存在网络结构复杂、训练难度大等缺陷.

为了更好地在解决图像以及监控视频中所存在的遮挡、人群分布不均以及复杂环境下人群计数存在偏差等问题的同时提高训练速度、降低网络复杂度. 本文尝试利用深度学习卷积神经网络[13] 结合轻量级注意力模块进行人群计数的研究. Li 等人[14] 提出了一种称为 CSRNet 的拥挤场景识别网络来执行人群估计工作. CSRNet 人群计数网络主要由前端和后端两个部分组成, 前端利用舍弃了全连接层的深度卷积神经网络 VGG-16[15] 的前 13 层进行二维特征的提取, 在其基础上又添加了 6 个空洞卷积层[16]作为网络结构的后端回归器, 使其在提取特征时获得更大的感受野. 本文在 CSRNet 网络结构的基础上进行了部分改进, 提出了基于联合损失的空间-通道双注意力机制网络 (joint loss-based space-channel dual attention network, JL-SCDANet), 并在目前常见的公开人群计数数据集 ShanghaiTech Part B、mall 以及 UCF_CC_50 中对所改进的网络结构进行评估. 基于该 3 种数据集的平均绝对误差 *MAE* 以及均方误差 *MSE*都得到了有效的降低. 综上所述, 针对本文的人群计数网络结构所作的工作如下.

(1) 本文模型采用 3 段式结构, 前端采用 VGG-16 的前 13 层对输入图像进行二维特征提取, 中间采用空间注意力机制与通道注意力机制串联的方法对图像中人群分布不均的区域进行重点空间定位, 后端采用 6 个扩张率为 3 的空洞卷积层以扩大感受野, 并在不丢失图像分辨率的条件下提取深层二维特征. 使用该方法减小了真实人群密度图与预测人群密度图之间的差异, 提升了人群计数的准确率.

(2) 本文使用欧氏距离和改进的绝对值误差损失相结合的方式来约束预测密度图和真实密度图, 通过大量对比实验调节联合损失函数中的参数, 获得更加精确的人群计数网络训练模型.

(3) 本文所提出的人群计数网络模型在常见的 4 个数据集中都体现出了较好的性能. 在 ShanghaiTech Part B 数据集中 *MAE*、*MSE* 分别降低了 23.3% 和 17.9%, 在 mall 数据集中 *MAE*、*MSE* 分别达到了 1.78 和 2.28, 在 UCF_CC_50 数据集的 *MAE*、*MSE* 分别降低了 31.5% 和 47.1%.

1 密度图的生成

人群密度图能够反映某一图像中人群的分布情况

以及密集程度, 是模型进行人群计数的主要参照目标. 通过生成密度图的方式进行人群计数仍然是当前人群计数算法中常被使用的主流方法. 与早先人群计数方法相比, 基于密度图的人群计数方法不仅可以通过像素准确预测图像中的人员个数, 还可以提供人群空间分布情况等信息, 体现出人群分布的空间相关性. 因此, 在本文也是沿用基于密度图的方式进行人群计数工作, 旨在准确输出预测图像的人群密度图.

鉴于图像中相对密集的人群通常会出现遮挡等问题, 使得对于人员的整体标记变得极为困难, 而当前常见的公共数据集所给的人群标注信息通常都以图像中人员头部中心点的位置坐标的形式给出, 这将人员的整体标记简化为点标注, 大大减少了人群计数工作的难度. 为了获得真实人群密度图, 按照文献 [2] 中所使用的密度图生成方法, 对人头中心标注点进行高斯核模糊. 假设一张人群图像中的一个人头标注坐标为$X_i(x_i,y_i)$, 则在这张图片中该人头的像素点可表示为$\delta(X-X_i)$, 其中$\delta(\cdot)$为狄柯拉函数. 若想得到带有N个人头标注点的真实密度图, 需将N个人头的像素点与高斯核$G_{\sigma_i}(X_i)$进行卷积, 具体公式为:

$$F(x)=\sum_{i=1}^{N}\delta(X-X_i)\times G_{\sigma_i}(X_i),\ \sigma_i=\beta\bar{d}_i \tag{1}$$

其中, σ_i表示高斯核大小, $\bar{d}_i$表示 K 近邻的平均距离, 假设给定某一头部位置信息, 则与其周围k个人头标记之间的距离为$\{d_1^i,d_2^i,d_3^i,\cdots,d_k^i\}$, 因此$\bar{d}_i$为:

$$\bar{d}_i=\frac{1}{k}\sum_{j=1}^{k}d_j^i \tag{2}$$

遵循文献 [2] 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.

2 JL-SCDANet 网络模型

如图 1 所示, JL-SCDANet 网络模型主要采用 3 段式结构, 前端主要使用 VGG-16 的前 13 层进行图像局部特征提取, 中间引入双注意力机制 (空间注意力机制和通道注意力机制) 进行重点空间定位并对其空间及通道相关性特征进行融合, 最后使用 VGG-16 的后端网络进行特征回归输出预测密度图并进行人数统计.

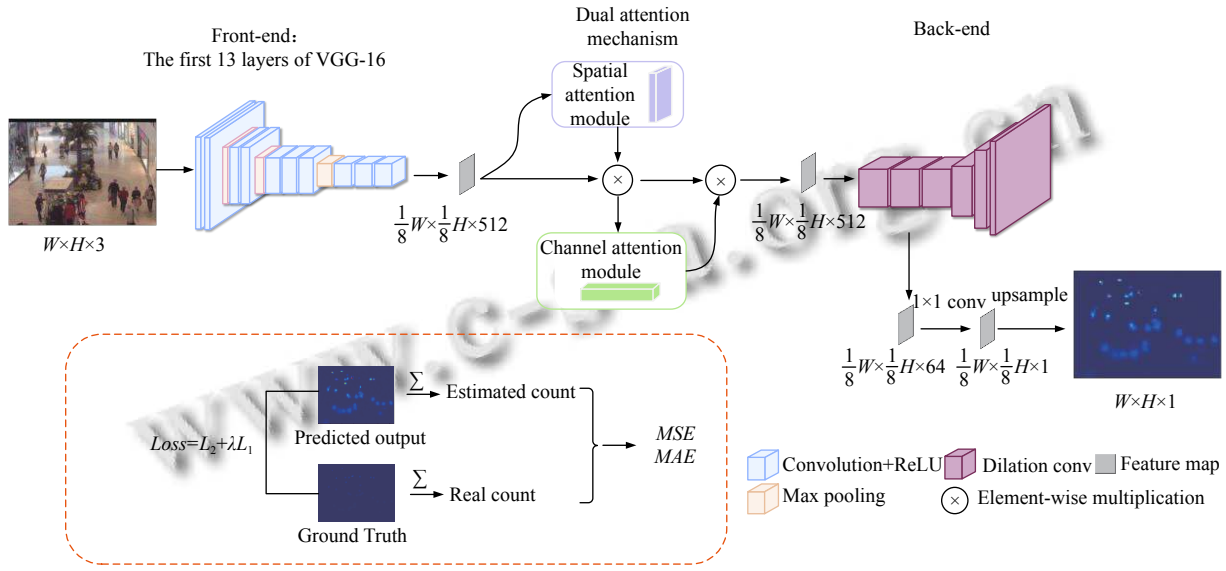


图 1 JL-SCDANet 网络结构

目前迁移学习在深度学习神经网络中被广泛使用, 迁移学习是运用已存有的知识对不同但相关领域问题进行求解的一种新的机器学习方法[17], 即利用已有的先验知识让算法学习新的知识. 鉴于强大的迁移学习能力在 VGG-16 上表现显著且其灵活性较大, 因此 JL-SCDANet 网络结构的前端采用导入了预训练模型参数的 VGG-16 的前 13 层进行图像粗粒度特征提取, 以提升训练速度. 表 1 列出了 VGG-16 的前 13 层的相关参数. 其中关于卷积层的参数表示为 conv (卷积核大小)-(通道数)-(空洞率)×(层数), 最大池化层大小全部为

2×2, 步长为 2.

卷积神经网络运算通常是通过混合跨通道和空间信息来提取特征, 当图像背景较为复杂时无法提取重要特征信息, 这很容易使网络在训练的过程中忽略人群主要信息而降低预测人群密度图的准确率. 为了解决这类问题, 本文在网络结构中加入了注意力机制, 使用注意力机制告诉网络应该把注意力集中在哪里, 最终使网络有效关注重要特征而抑制非重要特征[18], 同时使用空间注意力机制与通道注意力机制两个轻量级注意力模块串联的形式融合图像空间特征信息与通道特征信息, 在空间和通道两个维度集中重要特征并有效传递特征信息进入后端网络结构, 下面分别介绍每个注意力机制模块以及所使用的联合损失函数.

表 1 Front-end 网络参数表

层号	网络
Cov1–2	Conv3-64-1×2
Cov3	Max pooling
Cov4–5	Conv3-128-1×2
Cov6	Max pooling
Cov7–9	Conv3-256-1×3
Cov10	Max pooling
Cov11–13	Conv3-512-1×3

2.1 空间注意力机制

为了解决复杂背景下对于人头的特征提取, 需要重点强调图像中的人群特征. 由此, 本文在 VGG-16 前 13 层之后引入空间注意力机制, 有效获取复杂环境中的人头关键信息, 增强图像空间信息权重, 舍弃类似环境等与人群计数工作内容不相关的信息. 空间注意力机制如图 2 所示.

经过 VGG-16 前 13 层进行人群图像粗粒度特征提取后得到$C\times W\times H$的中间特征图F, 将F作为空间注意力模块的输入特征图. 首先沿通道轴对F分别进行全局最大池化 (global max pooling) 以及全局平均池化 (global average pooling) 生成两个二维特征图$F^s_{\max}\in\mathbb{R}^{1\times W\times H}$和$F^s_{\text{avg}}\in\mathbb{R}^{1\times W\times H}$, 其中分别表示输入特征图的最大池化特征与平均池化特征; 然后将所生成的特征图进行通道拼接, 此时通道数变为原来的 2 倍; 接着将拼接后的特征图放入卷积核为7×7的卷积层中进行降维, 将通道数还原为单通道, 用以消除通道域特征信息对空间信息的影响, 并通过激活函数 Sigmoid 将空间特征信息进行归一化处理, 生成空间注意力特征图M_s; 最后将空间特征信息M_s与输入特征信息F对应位置逐元素相乘得到最终的空间特征融合信息F_s. 空间注意力模块的具体计算公式如式 (3) 所示:

$$\begin{aligned}M_S(F)&=\sigma(f^{7\times7}([AvgPool(F);MaxPool(F)]))\\&=\sigma(f^{7\times7}(F^s_{\max};F^s_{\text{avg}}))\end{aligned}\tag{3}$$

空间特征融合信息的计算公式如式 (4) 所示:

$$F_s=F\otimes M_s\tag{4}$$

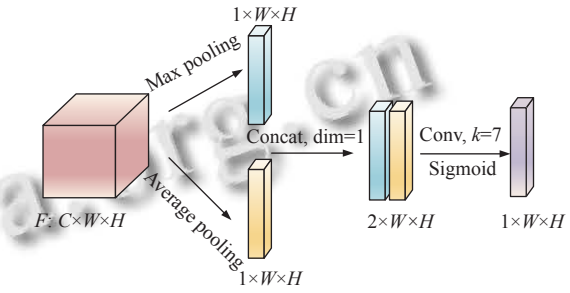


图 2 空间注意力模块

2.2 通道注意力机制

卷积神经网络默认每个通道是同等重要的, 而在实际情况中, 不同通道的重要性是有所不同的, 有的通道对最终的分类结果影响较大[19]. 鉴于其在处理多通道图像信息时所采取的平等对待策略会大大降低网络模型处理低频或高频信息的灵活度, 在跨特征通道上的学习能力较差, 无法有效获取深层次网络的准确特征, 因此使用卷积神经网络进行特征提取时通常需要使用通道注意力机制对图像通道特征附加相关权重, 以增强主要通道的特征信息. 通道注意力机制如图 3 所示.

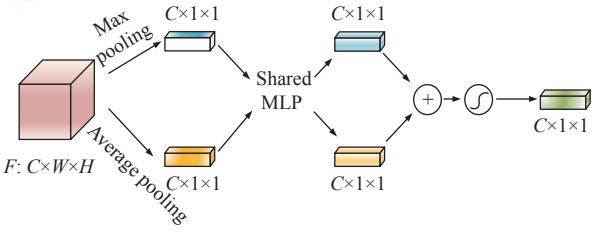


图 3 通道注意力模块

通道注意力机制同样使用全局平均池化以及全局最大池化将尺寸为$C\times W\times H$输入特征图F进行空间信息聚合, 分别生成两个不同的空间上下文特征信息图$F^c_{\text{avg}}\in\mathbb{R}^{C\times1\times1}$和$F^c_{\max}\in\mathbb{R}^{C\times1\times1}$; 接着将它们分别转发至一个两层的共享神经网络 MLP 中以生成所需的通道注意力特征图M_c, 其中第 1 层的神经元个数为C/r

(r为缩减率), 第 2 层的神经元个数为C, 激活函数均使用 ReLU 函数; 最后将得到的两种通道特征信息图进行逐元素加和操作, 并使用 Sigmoid 函数进行激活以融合所输出的特征向量. 通道注意力模块的计算公式如式 (5) 所示:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F^c_{\text{avg}})) + W_1(W_0(F^c_{\max}))) \end{aligned} \tag{5}$$

其中, σ表示 Sigmoid 激活函数, $W_0 \in \mathbb{R}^{C/r \times C}$和$W_1 \in \mathbb{R}^{C \times C/r}$分别表示共享神经网络 MLP 各层的权重. 通道特征融合信息F_c的计算过程如式 (6) 所示:

$$F_c = M_c \otimes F \tag{6}$$

2.3 联合损失函数

目前在常用的人群计数网络中通常只使用欧氏距离作为损失函数进行网络的训练, 而本文则使用欧式距离损失函数联合改进的绝对值误差损失函数对预测密度图与真实密度图之间的差异进行约束. 欧式距离损失函数如式 (7) 所示:

$$L_2 = \frac{1}{2N}\sum_{i=1}^{N}\left\|Z(X_i;\theta) - Z_i^{GT}\right\|_2^2 \tag{7}$$

其中, N为一个训练批的图像个数; $Z(X_i;\theta)$为网络参数为θ时对输入图像X_i的预测密度图; Z_i^{GT}为输入图像X_i真实密度图.

欧式距离损失函数利用预测密度图与真实密度图作差后平方的方式计算两者之间的误差, 其值越小, 预测密度图越接近真实密度图. 但是, 倘若单独使用L_2, 当预测值与真实值相差较大时, 由于使用了先作差后平方, 这就会放大误差, 即当出现一个离群点时, 会使误差变大, 训练模型将会赋予其更高的权重, 这会降低模型的整体性能. 当所给数据集图像中人群分布变化较大时将非常不利于模型训练出较好的性能. 因此, 本文在使用欧式距离的基础上联合使用改进的绝对值误差损失函数以最小化离群点对于整个模型的影响, 降低L_2对离群点的敏感度, 这有利于当图像中出现人群分布不均等相关问题时提高检测准确率, 增强模型对于离群点的鲁棒性. 使用L_1可以降低图像中人群分布不均匀时的数据对整个模型预测结果的影响, 使用L_2可以使模型在训练的过程中更快的收敛. 改进的绝对值误差损失函数如式 (8) 所示:

$$L_1 = \frac{1}{2N}\sum_{i=1}^{N}\left|Z(X_i;\theta) - Z_i^{GT}\right| \tag{8}$$

其中, N为一个训练批的图像个数; $Z(X_i;\theta)$为网络参数为θ时对输入图像X_i的预测密度图; Z_i^{GT}为输入图像X_i真实密度图.

最终所使用的联合损失函数为:

$$Loss = L_2 + \lambda L_1 \tag{9}$$

其中, λ为超参数, 用于权衡L_1与L_2之间的权重, 具体数值通过对比实验得出.

3 实验与分析

3.1 环境与参数设置

本实验在训练的过程中主要的操作系统环境为 Ubuntu 18.04.5, GPU 为 NVIDIA GeForce GTX 1080 Ti, 所使用的实验框架为 PyTorch 1.10+Python 3.6+cuda 11.4+anaconda 3. 由于 Adam 算法计算效率高且对于内存的需求较低, 因此在模型的优化上主要使用 Adam 优化器, 同时设置初始学习率为 0.000 01, 且动量设为 0.9. 本文所得到的实验结果均在以上环境及参数中获得.

3.2 评价指标

实验中所采用的评价指标与目前常用的人群计数网络的评价指标相同, 为均方误差 (mean square error, *MSE*) 和平均绝对误差 (mean absolute error, *MAE*), 公式如下:

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|C_i - C_i^{GT}\right|^2} \tag{10}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|C_i - C_i^{GT}\right| \tag{11}$$

其中, N为一个训练批次的图像数, C_i表示模型的预测人数, C_i^{GT}为输入图像的真实人数.

3.3 实验结果对比及分析

目前, 有不少研究机构公开了用于人数统计研究的数据集, 这些数据集中不仅包含了人群图像及其标注文件, 还存在基于视频监控下的人群视频数据及其标注文件. 本文主要选择 3 个常见的公共数据集进行模型的训练以及测试工作. 数据集的详细信息如表 2 所示.

3.3.1 ShanghaiTech Part B 数据集

ShanghaiTech 数据集共包含 1 198 张带有标注文件的图像, 其中总人数为 330 165 人. 该数据集由 Part A

与 Part B 两部分组成, 本论文主要使用 Part B 数据集训练和测试网络性能. Part B 一共含有 716 张来源于上海街头的人群图像, 其中包含训练图像 400 张, 测试图像 316 张. 该部分图像人群分布较为稀疏, 平均每幅图像的人数为 123 人. 为了验证本文改进算法在 ShanghaiTech Part B 数据集中所表现的性能, 将其与其他 4 个目前的主流人群计数算法进行对比, 结果如表 3 所示.

表 2 标准数据集参数

数据集名称	数据集总量 (张)	分辨率 (像素)	平均每幅人数
ShanghaiTech Part B[2]	716	768×1024	123
Mall[20]	2 000	640×480	31
UCF_CC_50[21]	50	非统一	1 279

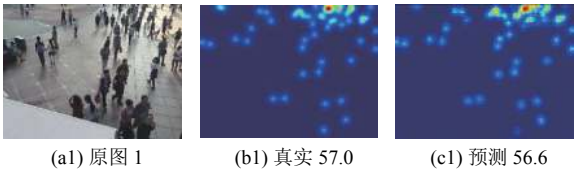
表 3 ShanghaiTech Part B 数据集上的对比结果

算法	*MAE*	*MSE*
MCNN (2016)[2]	26.4	41.3
CSRNet (2018)[12]	10.6	16.0
SCAR (2019)[4]	9.5	15.2
MSFNet (2020)[3]	9.6	14.3
RDCAF (2022)[20]	8.51	14.2
JL-SCDANet	**8.13**	**13.13**

在该数据集中随机选择两张图片的测试结果如图 4 所示, 从图中可以看出, 图 4(a1) 中存在人群分布不均等问题, 图 4(a2) 存在部分遮挡以及远处目标较小等问题. 图 4(a1) 的真实人数为 57 人, 通过 JL-SCDANet 网络模型预测的人数为 56.6 人, 仅存在 0.7% 的差距; 图 4(a2) 的真实人数为 138 人, 通过 JL-SCDANet 网络模型预测的人数为 125 人, 同时仅存在 9% 的差距. 可见, 本模型可有效解决图像部分遮挡与人员分布不均这两个问题.

3.3.2 Mall 数据集

该数据集主要取自于国外某商场的监控视频, 通过将监控视频按帧截图抽取其中的 2 000 帧视频图像用于人群计数工作的研究. 该 2 000 帧视频图像场景较为固定, 人群分布相对稀疏, 平均每幅图像的人数约为 31 人, 主要的标注点为人头中心. 该数据集中存在较多的物体遮挡与环境干扰, 如图 5(a) 所示, 因此在该数据集上的人群计数工作较为困难. 为了验证本文模型的抗遮挡和抗干扰能力, 本文使用了该数据集的前 800 帧图像截图用于模型训练, 后 1 200 帧图像截图用于模型测试, 并与其他 4 个主流模型进行了对比实验, 对比结果如表 4 所示.



(a1) 原图 1 (b1) 真实 57.0 (c1) 预测 56.6

(a2) 原图 2 (b2) 真实 138.0 (c2) 预测 125.0

图 4 ShanghaiTech Part B 数据集测试结果

(a) 原图 (b) 真实 46.0 (c) 预测 44.2

图 5 Mall 数据集测试结果

表 4 Mall 数据集上的对比结果

算法	*MAE*	*MSE*
DigCrowd (2019)[5]	3.21	16.4
ACM-CNN (2019)[7]	2.3	3.1
CWAN (2020)[8]	2.06	2.90
IFDM (2021)[6]	2.45	3.2
RDCAF (2022)[20]	1.79	2.32
JL-SCDANet	**1.78**	**2.28**

在该数据集中随机挑选一张图片的测试结果如图 5 所示, 从图 5(a) 中看出, 原图存在背景较为复杂, 人员相对稀疏等问题. 原图的真实人数为 46 人, 通过 JL-SCDANet 网络模型预测出的人数为 44.2 人, 仅存在 3% 的差距, 准确率较高. 且用 JL-SCDANet 网络模型输出的密度图也较为清晰, 由此可见, 本模型在环境较为复杂的图片中仍能保持较好的性能.

3.3.3 UCF_CC_50 数据集

相较于以上两个数据集, UCF_CC_50 数据集具有极大的挑战性, 该数据集主要来源于多种场景, 像是音乐会、群众抗议、马拉松等. 与此同时, 数据集中的每张图片分辨率不一, 人群密度与拍摄视角也各不相同, 图像之间差异较大. 整体来看人群密度较高, 平均每幅图中标注了数千人. 但由于该数据集中所包含的图片较少, 仅包含 50 张图片, 因此需要对其进行数据扩充.

目前利用该数据集进行人群密度研究中, 为了充分利用该数据集, 通常使用五折交叉验证的方法进行

模型的训练, 而本文利用此数据集进行训练前先对数据集进行数据扩充, 以保证数据集的充分利用. 数据扩充的主要方式为旋转与裁剪, 首先将数据集中的每张图片按照逆时针分别旋转 90°和 180°; 然后对原图以及旋转之后的图片以(x,y)为左上点进行定点裁剪, 裁剪所使用的宽与高均为原图的 1/2. 倘若(x,y)分别选择 (45, 45)、(85, 85)、(125, 125) 进行裁剪, 则得到的图片如图 6(a), 图 6(b), 图 6(c) 所示, 以同样的方式对旋转之后的图片裁剪过程不再赘述.

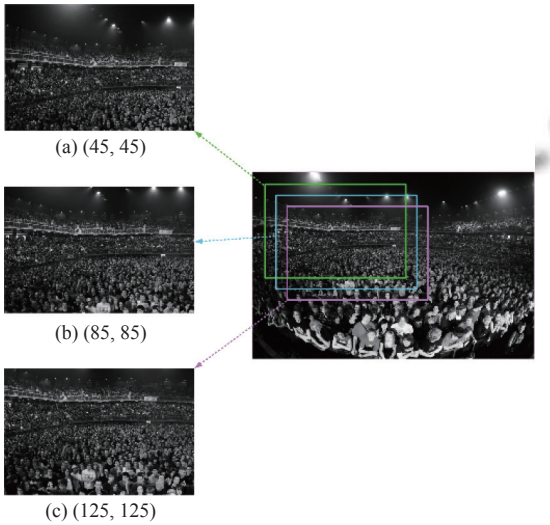


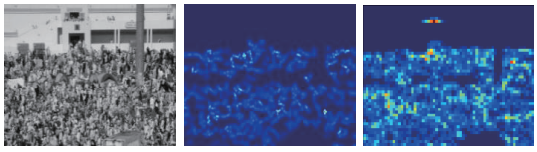
图 6 UCF_CC_50 单张图片裁剪示例图

经过对 UCF_CC_50 的数据集扩充过程后, 最后我们选择 540 幅图片进行训练, 360 幅图片进行测试, 并与常见的其他 4 种人群计数算法进行对比实验, 对比实验结果如表 5 所示.

该数据集中随机选择一张图片的测试结果如图 7 所示, 从原图中看出, 该图像中人员分布相对集中且密度较大, 存在小部分遮挡, 使用 JL-SCDANet 网络模型相较于所对比的其他模型而言, 准确率有所提高. 该图片的真实人数为 465.9 人, 预测人数为 352.5 人. 由此可见, 本模型在密集人群图像中也可表现出良好的性能.

表 5 UCF_CC_50 数据集上的对比结果

算法	*MAE*	*MSE*
MCNN (2016)[2]	377.6	509.1
CSRNet (2018)[14]	266.1	397.5
MAFNet (2021)[9]	196.7	293.3
ResNet50+BUF (2022)[10]	242.6	359.5
HANet (2022)[21]	195.2	268.6
JL-SCDANet	**182.12**	**210.24**



(a) 原图 (b) 真实 465.9 (c) 预测 352.5

图 7 UCF_CC_50 单张图片测试效果图

3.4 消融实验

在第 2.3 节中提出了本文所使用的联合损失函数, 其中超参数λ用来权衡L_1与L_2之间的权重, 为了验证不同超参数对实验效果的影响, 本文在 ShanghaiTech Part B 数据集上针对不同的λ做消融实验, 实验结果如表 6 所示.

表 6 ShanghaiTech Part B 数据集上的消融实验

方法	λ	*MAE*	*MSE*
CSRNet+L_2	—	10.6	16.0
JL-SCDANet+联合损失	0.0	8.47	14.65
JL-SCDANet+联合损失	**0.2**	**8.13**	**13.13**
JL-SCDANet+联合损失	0.3	8.88	13.70

通过对 ShanghaiTech Part B 数据集在 CSRNet 的测试结果以及在不同超参数λ下 JL-SCDANet+联合损失的测试结果进行对比可看出, 当λ=0.2 时, *MAE* 相较于 CSRNet 下降了 23.3%, *MSE* 下降了 17.9%, 达到最低. 因此, 在使用 ShanghaiTech Part B 数据集进行模型测试时, 将联合函数的超参数设置为 0.2 可获得最佳测试效果. 同时也看出, 使用联合损失函数进行模型的训练, 有利于降低人群计数的误差, 提高计数的准确率.

经过研究不同数据集以及不同超参数测试结果后发现, 在不同数据集上需设置不同的λ值, 表 7 列出了针对不同的数据集所采用的λ大小.

表 7 各数据集上所使用的λ值

数据集	λ
ShanghaiTech Part B	0.2
UCF_CC_50	0.2
Mall	0.0

4 结束语

针对目前人群计数工作中常见的图像背景复杂、遮挡以及人群分布不均等问题, 本文提出一种基于联合损失的空间-通道双注意力机制网络 JL-SCDANet. 该网络在前端与后端网络的基础上串连加入了空间注意力机制和通道注意力机制, 通过双注意力机制根据

图像不同区域的重要程度让网络学习人群图像的关键信息, 有效解决了图像中遮挡、环境复杂等常见难题. 同时, 该网络配合联合损失函数进行模型的训练, 并在 3 个数据集上调节不同的超参数得到最低误差. 实验结果表明, 本方法与所对比的其他人群计数方法相比, 所预测的密度图与真实密度图相差较小, 准确率有所提高, 以此可证明本模型的泛化能力与鲁棒性. 但是本模型还存在明显的不足, 即目前针对视频数据只能先截图后测试, 下一步的工作是研究如何对视频数据中的人群数目进行统计, 以实现动态数据人群计数目标.

参考文献

1 蒋妮, 周海洋, 余飞鸿. 基于计算机视觉的目标计数方法综述. 激光与光电子学进展, 2021, 58(14): 43–59.

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

3 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): 307–315.

4 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]

5 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]

6 袁健, 王姗姗, 罗英伟. 基于图像视野划分的公共场所人群计数模型. 计算机应用研究, 2021, 38(4): 1256–1260, 1280. [doi: 10.19734/j.issn.1001-3695.2020.02.0076]

7 Zou ZK, Cheng Y, Qu XY, *et al*. Attend to count: Crowd counting with adaptive capacity multi-scale CNNs. Neurocomputing, 2019, 367: 75–83. [doi: 10.1016/j.neucom.2019.08.009]

8 Kong XY, Zhao MM, Zhou H, *et al*. Weakly supervised crowd-wise attention for robust crowd counting. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020. 2722–2726.

9 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]

10 杨旭, 黄进, 秦泽宇, 等. 基于多尺度特征融合的人群计数算法. 计算机系统应用, 2022, 31(1): 226–235. [doi: 10.15888/j.cnki.csa.008250]

11 沈宁静, 袁健. 基于残差密集连接与注意力融合的人群计数算法. 电子科技, 2022, 35(6): 6–12.

12 Wang FS, Sang J, Wu ZY, *et al*. Hybrid attention network based on progressive embedding scale-context for crowd counting. Information Sciences, 2022, 591: 306–318. [doi: 10.1016/j.ins.2022.01.046]

13 LeCun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278–2324. [doi: 10.1109/5.726791]

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

15 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

16 Zhang Q, Cui ZP, Niu XG, *et al*. Image segmentation with pyramid dilated convolution based on ResNet and U-Net. Proceedings of the 24th International Conference on Neural Information Processing. Guangzhou: Springer, 2017. 364–372.

17 庄福振, 罗平, 何清, 等. 迁移学习研究进展. 软件学报, 2015, 26(1): 26–39. [doi: 10.13328/j.cnki.jos.004631]

18 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 3–19.

19 Wu HP, Zou ZX, Gui J, *et al*. Multi-grained attention networks for single image super-resolution. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(2): 512–522. [doi: 10.1109/TCSVT.2020.2988895]

20 Chen K, Loy CC, Gong SG, *et al*. Feature mining for localised crowd counting. Proceedings of the British Machine Vision Conference. Surrey: BMVA Press, 2012. 1–11.

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

(校对责编: 孙君艳)