计算机科学
COMPUTER SCIENCE

密集场景下基于多尺度特征聚合的人群计数方法

刘培刚, 孙洁, 杨超智, 李宗民

引用本文

刘培刚, 孙洁, 杨超智, 李宗民. 密集场景下基于多尺度特征聚合的人群计数方法[J]. 计算机科学, 2023, 50(9): 235-241.

LIU Peigang, SUN Jie, YANG Chaozhi, LI Zongmin. Crowd Counting Based on Multi-scale Feature Aggregation in Dense Scenes [J]. Computer Science, 2023, 50(9): 235-241.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于多尺度特征融合的遥感图像建筑物提取算法研究

Study on Building Extraction Algorithm of Remote Sensing Image Based on Multi-scale Feature Fusion

计算机科学, 2023, 50(9): 202-209. https://doi.org/10.11896/jsjkx.220800086

基于多编码器的多模态MRI脑肿瘤分割

Multimodal MRI Brain Tumor Segmentation Based on Multi-encoder Architecture

计算机科学, 2023, 50(6A): 220200108-6. https://doi.org/10.11896/jsjkx.220200108

基于孪生注意力网络的建设用地遥感影像变化检测

Remote Sensing Image Change Detection of Construction Land Based on Siamese AttentionNetwork

计算机科学, 2023, 50(6A): 220500040-5. https://doi.org/10.11896/jsjkx.220500040

基于多邻接图与多头注意力机制的短期交通流量预测

Short-time Traffic Flow Forecasting Based on Multi-adjacent Graph and Multi-head Attention Mechanism

计算机科学, 2023, 50(4): 40-46. https://doi.org/10.11896/jsjkx.220200079

特征增强损失与前景注意力人群计数网络

Crowd Counting Network Based on Feature Enhancement Loss and Foreground Attention

计算机科学, 2023, 50(3): 246-253. https://doi.org/10.11896/jsjkx.220100219

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220800067

密集场景下基于多尺度特征聚合的人群计数方法

刘培刚[1]　孙　洁[1]　杨超智[1]　李宗民[1,2]
1 中国石油大学(华东)计算机科学与技术学院　山东 青岛 266580
2 中国石油大学胜利学院　山东 东营 257061

摘　要　密集场景下个体尺度存在巨大差异,目标个体尺度不一导致人群计数精度不高。针对这一问题,提出了一种密集场景下基于多尺度特征聚合的人群计数方法。该方法研究不同特征层级对不同尺度个体的特征信息表示能力,通过层级连接充分获取多尺度特征;同时,提出了一个多尺度特征聚合模块,采用多列具有不同扩张率的空洞卷积,通过动态特征选择机制自动调整感受野,以有效提取不同尺度个体的特征。该方法能够在保留小尺度个体特征信息的基础上进一步扩大感受野,增强大尺度个体的检测能力,使其更好地适应人群个体的多尺度变化。在3个公共人群计数数据集上进行了实验,实验结果表明,所提模型在计数准确性上有了进一步的提高,其中在ShanghaiTech数据集Part_A上MAE为51.21,MSE为83.70。

关键词:密集场景;人群计数;空洞卷积;动态特征选择;点预测
中图法分类号　TP391.41

Crowd Counting Based on Multi-scale Feature Aggregation in Dense Scenes

LIU Peigang[1],SUN Jie[1],YANG Chaozhi[1] and LI Zongmin[1,2]
1 School of Computer Science and Technology in China University of Petroleum(East China),Qingdao,Shandong 266580,China
2 Shengli College of China University of Petroleum,Dongying,Shandong 257061,China

Abstract　Individual scales vary greatly in dense scenes,and the varying scales of target individuals lead to poor crowd counting accuracy. To address this problem,the crowd counting method based on multi-scale feature fusion in dense scenes is proposed. The method investigates the ability of different feature layers to represent feature information for individuals at different scales,with adequate access to multi-scale features through layer connections. At the same time,a multi-scale feature aggregation module is proposed,which uses multiple columns of dilated convolution with different expansion rates,and automatically adjusts the perceptual field through a dynamic feature selection mechanism to effectively extract features of individuals at different scales. The method can further expand the field of perception while preserving the information of small-scale,and improving the detection capability of large-scale individuals,making it better adapted to the multi-scale changes of the population. Experimental results on the three public population counting datasets show that the proposed model has further improved the counting accuracy,with an MAE of 51.21 and an MSE of 83.70 on the ShanghaiTech Part A dataset.

Keywords　Intensive scenes,Crowd counting,Dilated convolution,Dynamic feature selection,Point prediction

1　引言

近年来,城市人群密集场景越来越多,恐怖事件、踩踏事件也随之增多,新冠疫情期间,人群计数技术发挥了重大作用。例如,智能监控系统在检测到某一区域的人群密度异常时便会发出安全警告,及时通知有关部门进行处理,从而提高公共安全管理的效率。然而,由于受到光线变化、人群之间相互遮挡、透视失真以及人群分布不均匀等因素影响,人群计数任务仍然面临巨大的挑战。

人群计数是计算机视觉中一项重要且具有挑战性的任务,目的是估计出给定图像或视频中的人数。现有的人群计数方法大致可以分为3类:传统计数方法、基于密度估计的方法和基于点预测的方法。早期大多数传统方法基于手工特征进行行人检测[1-3],Li等[4]通过检测图片中的行人或人的头部、肩部等局部信息进行人群计数,这类方法不适用于存在严重遮挡的密集场景。2008年,Davies等[5]发现人群总体密度和前景像素面积近似线性相关,并首次将回归方法引入人群分析,为间接计数提供了灵感。然而,前景分割本身是一项比较困难的任务,对基于回归方法的性能造成了很大影响。

近年来,采用深度神经网络进行密度估计成为了人群

到稿日期:2022-08-06　返修日期:2022-12-07
基金项目:国家重点研发计划(2019YFF0301800);国家自然科学基金(61379106);山东省自然科学基金(ZR2013FM036,ZR2015FM011)
This work was supported by the National Key R & D Program of China(2019YFF0301800),National Natural Science Foundation of China(61379106) and Shandong Provincial Natural Science Foundation(ZR2013FM036,ZR2015FM011).
通信作者:刘培刚(dongfangwy@upc.edu.cn)

计数的主流方法，研究者使用 CNN 从人群图像中提取特征，生成用于人群计数的密度图，通过对预测的密度图进行积分来获得人群计数的结果[6-13]。Zhang 等[8]提出了一个基于 CNN 的跨场景计数模型，根据目标场景特点，使用相似的训练数据来微调训练网络，以达到跨场景计数的目的。为了解决图像中个体尺度大规模变化问题，Zhang 等[14]提出了一种多列卷积神经网络 MCNN，用于捕获多尺度信息，该网络包含 3 个具有不同卷积核大小的分支，分别对应不同的感受野，能适应由于角度不同而造成的个体尺度变化。为了提升多列卷积神经网络的训练速度，Sam 等[15]提出的 Switch-CNN 网络在 MCNN 的基础上增加了一个选择分类器，能够根据输入的图像块从多个具有不同卷积核的分支中选择其最适合的分支。Sindagi 等[16]提出了一种上下文金字塔神经网络(CP-CNN)，通过结合全局和局部的上下文信息生成高质量的人群计数密度图。此外，由于注意力模型在计算机视觉任务中取得了巨大成功，一些研究[17-19]也使用注意力机制引导网络在不同的时期关注不同尺度的部分。Jiang 等[19]提出了基于注意力机制的神经网络模型 ASNet，利用密度注意网络生成不同密度区域的注意掩码，使每个密度图只关注某一密度级别的区域。以上方法可以有效避免人群遮挡带来的影响，一定程度上解决了个体尺度变化问题，但它们仍然存在一些显著的缺点：首先，基于多列的网络结构更加臃肿，导致训练困难；其次，网络结构近乎相同的不同分支会导致大量信息冗余。为了降低网络的复杂度，Li 等[20]提出将基于 CNN 的单分支网络与空洞卷积相结合，有效增加网络的感受野，但仅使用相同空洞率的空洞卷积进行堆叠会导致训练过程中的信息不连续，达不到理想的计数效果。

近期，一些研究人员开始关注个体的定位问题，Liu 等[21]提出了一种用于人群计数的点监督深度检测网络 PSDDN，该网络在训练阶段读取人头的点级标注信息，根据这些点信息首先生成一组初始化伪真值边界框，并在训练期间迭代更新，使其不断地接近真实头部边界框大小。Song 等[22]提出了一个点对点网络，用于直接接收一组带标注的头部点用于训练，并在推理过程中直接预测点。这些对个体直接进行预测的模型构造简洁，也带来了不错的性能提升。但以上工作未考虑到人群大规模尺度变化所带来的影响，网络不能同时兼顾尺度过大或过小的个体，尤其当场景包含非常密集的人群时，无法检测和定位每个人，计数结果往往不准确。

在密集场景的监控图像中，不仅在不同场景中目标的尺度差异大，而且在同一张图片中由于透视现象也会造成目标尺度不同。例如，同一张图片中大尺度个体头部所占像素数可能是小尺度个体头部的 20 倍以上，尺度的极端变化一定程度上影响了网络对人群计数的准确性。同时，基于密度估计的方法仅能大体反映出不同区域人群分布的密度情况，不能对个体进行精确定位，简单地给出一个数字远不能支持后续更高层次的人群分析任务的实际需求，不利于人群行为分析[23]、目标追踪[24]等一些下游任务的实际应用。

针对目标个体尺度不一、定位不精确的问题，本文提出了一种新的基于多尺度特征聚合的神经网络模型，在特征提取的基础网络中，研究不同的特征层级所包含的有效信息，并加以充分利用，实现精准的计数和定位效果；此外，提出了多尺度特征聚合模块，引入了一种具有共享权重参数的并行空洞卷积结构。通过学习每两个相邻尺度的相关权重来动态调整感受野，以高效利用多尺度特征表示，解决目标尺度不一的问题。在 3 个公开的人群计数数据集上进行对比实验，验证了本文方法具有更高的计数准确度。

2 本文方法

针对当前密集场景人群计数存在的问题，本文设计了一种多尺度特征聚合网络，该网络以个体头部的真值注释点为监督信息，在推理过程中直接预测一组个体头部的坐标点。网络模型主要由 3 部分组成：用于特征提取的主干网络、多尺度特征聚合模块、点预测模块。整体网络架构如图 1 所示。

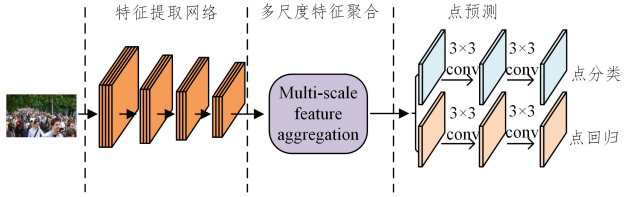


图 1 整体网络架构图

Fig. 1 Overall network architecture

2.1 特征提取网络

神经网络的不同层级特征都包含着丰富的可用信息，因此在目标检测领域 FPN[25]常被用来解决多尺度问题。虽然 FPN 通过横向连接融合了不同层级之间的特征，但也存在一些缺点：深层特征图虽然含有大尺度个体的语义信息，但丢失了其位置信息，因此深层特征不利于对大尺度个体的定位；深层特征图分辨率较小导致其并不包含小尺度个体的语义特征，因此即使采用 top-down 结构，较高分辨率的特征图中仍然无法提取到小尺度个体的语义特征，不利于小尺度个体的分类。为了充分利用不同特征层级中的有用信息，同时避免 FPN 存在的问题，本文设计了一种能够平衡大尺度个体和小尺度个体的特征提取网络。由于 VGG-16[26]具有很强的表示能力，本文采用 VGG-16_bn 中的前 13 个卷积层作为前端网络，从输入的人群图像中提取深层特征，具体分为 4 个特征级别，其下采样步长分别为{1,2,4,8}，相应的特征映射表示为$\{V_1,V_2,V_3,V_4\}$。对于最后输出的特征图 V_4，先使用 1×1 卷积减少通道数，接着采用最近邻插值进行上采样，使其大小与 V_3 一致，然后将上采样得到的结果与 V_3 进行横向连接，最后在合并后的特征图上使用 3×3 卷积和 ReLU 激活函数，以生成最终的特征图 P，如图 2 所示。

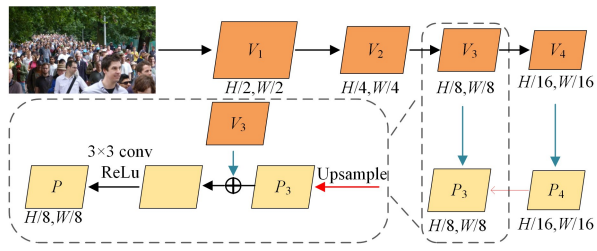


图 2 特征提取网络

Fig. 2 Feature extraction network

2.2 多尺度特征聚合

对于极度密集的场景，小尺度个体占比较大，通过加深特征层级扩大感受野，会损失小尺度个体的特征信息。为了保证小尺度个体在深层特征图中不丢失信息，需要在保证网络感受野的同时，保持特征图的大尺寸，即不使用下采样这种压缩物体信息的操作，而仅增大感受野。一种高效可行的方法是使用空洞卷积，采用多列具有不同扩张率的空洞卷积来获取多尺度特征，如图 3 所示。

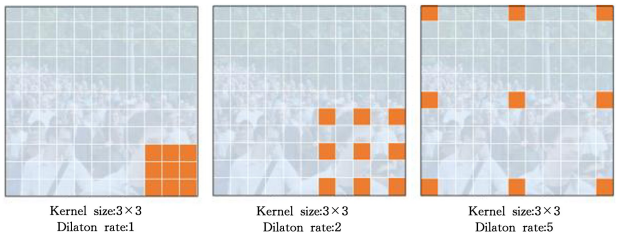


图 3 具有不同扩张率的空洞卷积

Fig. 3 Dilated convolution with different expansion rates

虽然具有不同扩张率的卷积核可以对应不同的感受野，但由于不同的图片中个体尺度差异巨大，同一张图片中不同区域的个体尺度也不尽相同，因此仅使用固定的多列具有不同扩张率的空洞卷积不能高效地处理复杂多变的尺度问题。针对这一问题，受医学图像分割中处理多尺度问题的启发[27]，本文提出的多尺度特征聚合模块包含一个动态特征选择机制，网络能够根据每两个相邻的尺度学习相关权重，自动为不同尺度的个体选择合适的感受野。

该模块完整的流程如图 4 所示。具体来说，对于特征提取网络输出的特征图 P_{in}，首先采用 3 个具有不同扩张率的空洞卷积的并行分支来分别提取不同尺度个体的特征，每个分支的卷积核大小为 3×3，扩张率分别设置为 1，2，5，得到 3 个特征图 P_1，P_2 和 P_3；然后分别连接两个相邻尺度的特征图，使其包含更多的尺度信息，考虑到相邻尺度之间的特征相关性，引入动态特征选择机制来自动为特征图选择合适的感受野。以空洞卷积的扩张率为 1 和 2 的两个分支为例，首先在通道维度上拼接 P_1 与 P_2，然后使用 3×3 卷积和 ReLU 激活函数，接着使用 1×1 卷积将通道数降低为 2，生成特征图 P_{12}。对于 P_{12}，使用 softmax 函数分别生成两个权重图 W_{x1} 和 W_{y1}，它能够反映不同尺度下空间信息的重要程度。将得到的 W_{x1} 和 W_{y1} 权重图分别与之前的特征图 P_1 和 P_2 相乘，以此实现感受野的动态选择。最后，将经过加权后得到的特征图进行融合，该过程的表达式为：

$$P'_{12}=(P_1\otimes W_{x1})\oplus(P_2\otimes W_{y1}) \tag{1}$$

其中，⊗表示元素对位相乘，⊕表示元素对位相加。该方法可以有效地获取多尺度特征，灵活选择不同尺度特征的感受野，其余分支的实现与上述过程一致。最后将相邻分支融合后的特征图进行进一步融合，为了防止小物体丢失信息，将该模块输入时的特征图 P_{in} 也一并进行融合，得到最终的输出特征图 P_{out}，该过程的表达式为：

$$P_{\text{out}}=P'_{12}\oplus P'_{23}\oplus P_{\text{in}} \tag{2}$$

2.3 点预测

针对多尺度特征聚合模块输出的特征图，本文设计了两个分支，用于最终的点预测：分类分支和回归分支。分类分支用于预测个体的类别，回归分支用于预测个体的偏移量。首先，在分类分支中，以 3×3 的卷积核进行 2 次卷积，然后使用 Softmax 归一化输出置信度分数。对于回归分支，需要预测出点坐标的偏移量，根据卷积层固有的平移不变特性，假设对于一个参考点 R_m 预测的偏移量为 $(\Delta^m_{jx},\Delta^m_{jy})$，那么该点在原图中坐标的计算式为：

$$\begin{cases}\hat{x}_j=x_m+\gamma\Delta^m_{jx}\\ \hat{y}_j=y_m+\gamma\Delta^m_{jy}\end{cases} \tag{3}$$

其中，γ 是一个标准化项，通过缩放偏移量来校正相对较小的预测。

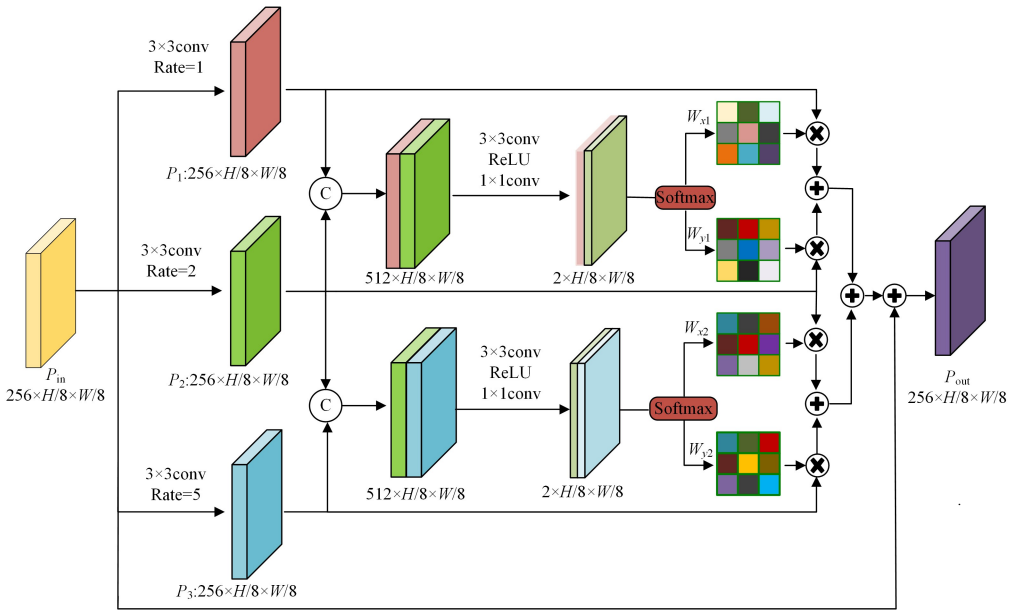


图 4 多尺度特征聚合模块

Fig. 4 Multi-scale feature aggregation module

在得到一组预测的点坐标与其对应的置信度后，参照Song等[22]的设置，本文使用一对一匹配策略将预测点与真值点进行匹配。结合预测点到真值点的距离与该预测点的置信分数来计算每一个预测点与其他真值点的距离，组成匹配代价矩阵。对于成对的代价矩阵，使用匈牙利匹配算法将预测点与真值点进行关联，将那些与真值点匹配成功的预测点视为正样本，将未匹配的点视为负样本。

2.4 损失函数

损失函数设计分为两个部分。首先，使用二分类交叉熵损失 L_{cls} 来训练网络对预测点进行分类，表达式为：

$$L_{cls}=-\frac{1}{M}\{\sum_{i=1}^{N}\log \hat{c}_{\xi(i)}+\lambda_1\sum_{i=N+1}^{M}\log(1-\hat{c}_{\xi(i)})\} \tag{4}$$

在训练初期，当预测值与真值差异过大时，L_2 损失函数对预测值的梯度非常大，导致训练不稳定，特别是当有一些离群点或者异常值时，L_2 损失就会在总损失中占据主导位置。在训练后期，如果预测值与真值的差异很小，L_1 损失将在稳定值附近波动，难以继续收敛，无法达到更高的精度。在将预测点与真值点进行匈牙利匹配时，不仅考虑了预测点与真值点之间的距离，还结合了该点的置信度得分。当某一预测点的置信度分数很高时，网络更倾向于将该预测点视为正样本，如果该预测点与真值点的距离较远，则需要预测一个较大的偏移，因为回归的目标没有明确的限制，所以可能会出现这个较大的偏移去主导整个损失函数的情况。相比 L_2 损失，*Smooth* L_1 损失对离群点、异常值更不敏感，能一定程度地防止上述情况的发生。因此在监督预测点回归时，本文选择使用更为稳健的 *Smooth* L_1 *Loss* 。*Smooth* L_1 *Loss* 的表达式为：

$$L_{loc}=\sum_{i=1}^{N}smooth\ L_1(P_i-\hat{P}_{\xi(i)}) \tag{5}$$

其中：

$$smooth\ L_1(x)=\begin{cases}0.5x^2, & \text{if } |x|<1\\ |x|-0.5, & \text{otherwise}\end{cases} \tag{6}$$

最终的损失函数 L 是上述两种损失的总和，表达式为：

$$L_{total}=L_{cls}+\lambda_2 L_{loc} \tag{7}$$

其中，λ_1 为预测为背景的加权因子，λ_2 为平衡回归损失的权重项。

3 实验结果与分析

3.1 实验设置

实验基于深度学习框架 PyTorch 实现，并在单个 NVIDIA 2080 Titan GPU 上运行。训练过程中使用在 ImageNet 上预训练好的 VGG-16_bn 作为基础网络，采取学习率更新策略，首先使用较小的学习率 0.0001，迭代 1000 轮时将学习率降低为 0.00001，使用 Adam 算法优化模型参数，批处理大小为 8，共训练 2000 个周期。在损失函数中，λ_1 设置为 0.5，λ_2 设置为 0.0002。

为了扩大训练样本以得到更加精准的网络模型，本文在数据预处理阶段使用不同的数据增强策略。首先对原始图像进行随机缩放，其缩放因子设置为[0.7, 1.3]，并保证图像的最小边不小于 128 像素。然后，使用随机水平翻转和随机裁剪，从每一张输入图像中随机裁剪出 4 个大小为 128×128 的图像块。除此之外，由于数据集中大尺度个体的样本较少，导致先前的方法对大尺度个体的检测精度不高，因此从图像底部裁剪出两个 128×128 的包含大尺度个体的固定区域，以进一步增加大尺度个体的样本数量。

3.2 评价标准

根据人群计数最常用的评价指标，本文采用平均绝对误差(MAE)和均方误差(MSE)作为评估指标来评估模型的性能。其定义如下：

$$MAE=\frac{1}{N}\sum_{j=1}^{j=N}|\hat{Q}_j-Q_j| \tag{8}$$

$$MSE=\sqrt{\frac{1}{N}\sum_{j=1}^{j=N}(\hat{Q}_j-Q_j)^2} \tag{9}$$

其中，N 是测试集中的图像数；$\hat{Q}_j$ 和 Q_j 分别表示第 j 张图像中的真实人数和估计人数。

3.3 实验对比与分析

3.3.1 在3个数据集上的效果对比

本文分别在 ShanghaiTech 数据集[14]、UCF_CC_50 数据集[28]和 UCF-QNRF 数据集[29]上进行实验验证，不同数据集的训练集与测试集样本划分详情如表 1 所列。

表 1 数据集划分详情

Table 1 Dataset segmentation details

Dataset	Total number of images	Training set	Validation set
ShanghaiTech Part_A	482	300	182
ShanghaiTech Part_B	716	400	316
UCF_CC_50	50	40	10
UCF-QNRF	1535	1201	334

ShanghaiTech 数据集由两部分组成，Part_A 包含 482 张从互联网上随机抓取的图像，Part_B 包含 716 张来自上海大都市繁忙街道的图像，这两个子集之间存在显著的人群密度差异。本文在 ShanghaiTech 数据集的两个部分分别与典型方法和最新方法进行了对比实验，结果如表 2 所列。现有的典型方法包括 Zhang 等提出的方法[8]、MCNN[14]、Switching-CNN[15]、CP-CNN[16]、CSRNet[20]、ASNet[19]，它们是基于密度估计的；现有的最新方法为 PSDDN[21] 和 P2PNet[22]，它们是基于点预测的。本文方法在该数据集 Part_A 上的 *MAE* 为 51.21，*MSE* 为 83.70。与基于密度估计的方法相比，本文方法不仅在计数准确度上有了提升，并且能够对个体进行准确定位。与基于点预测的方法 P2PNet 相比，本文在此基础上考虑了不同个体的尺度变化问题，并提出了多尺度特征聚合模块，使得计数误差进一步降低，且数据集 Part_A 中单幅图像中个体尺度变化较大，最大尺度和最小尺度个体所占像素数相差 20 倍之多。这说明本文设计的动态尺度选择机制能在保证小尺度个体特征信息的情况下有效获取大尺度个体的特征信息。对于 Part_B 中的稀疏场景，本文方法虽未取得最优精度，但也优于大多数基于密度估计的方法，通过对数据集样本及实验结果的深入分析，得出可能有以下原因：Part_B 数据集的场景范围更小，样本个体尺度变化不大，多尺度特征聚合模块在此数据集上不能更好地发挥优势。

表 2 在 ShanghaiTech 数据集上与其他方法的结果对比

Table 2 Results comparison with other methods on ShanghaiTech

Method	Part_A		Part_B	
	MAE	*MSE*	*MAE*	*MSE*
Zhang et al.[8]	181.80	277.70	32.00	49.80
MCNN[14]	110.20	173.20	26.40	41.30
Switching-CNN[15]	90.40	135.00	21.60	33.40
CP-CNN[16]	73.60	106.40	20.10	30.10
CSRNet[20]	68.20	115.00	10.60	16.00
ASNet[19]	57.78	90.13	—	—
PSDDN[21]	85.40	159.20	16.10	27.90
P2PNet[22]	52.74	85.06	**6.25**	**9.90**
Our Method	**51.21**	**83.70**	6.61	10.22

注：粗体表示最优值。

UCF_CC_50 数据集由 50 张灰色图像组成，约 63974 条注释，计数范围为 94～4543，平均每张图像有 1280 个个体。选择该数据集中 40 幅图像作为训练集，其余 10 幅图像作为测试集，按照 Idrees 等[28]的标准设置，共进行 5 倍交叉验证，取 5 次的平均值作为该数据集上的最终实验结果，在该数据集上与其他方法的对比实验结果如表 3 所列。该数据集不同图像之间的分辨率和个体数量差异巨大，并且个体尺度很小、密度极高、图片清晰度较低。从检测结果可以看出，本文方法在 UCF_CC_50 上有很大的性能提升，*MAE* 下降到170.43，*MSE* 下降到 244.73，这表明该方法充分保留了小尺度个体的特征信息，进一步验证了多尺度特征聚合模块提取小尺度个体特征的有效性。

表 3 在 UCF_CC_50 数据集上与其他方法的结果对比

Table 3 Comparison of results with other methods on the UCF_CC_50 dataset

Method	*MAE*	*MSE*
Zhang et al.[8]	467.00	498.50
MCNN[14]	377.60	509.10
Switching-CNN[15]	318.10	439.20
CP-CNN[16]	295.80	320.90
CSRNet[20]	266.10	397.50
ASNet[19]	174.80	251.60
PSDDN[21]	359.40	514.80
P2PNet[22]	172.72	256.18
Our Method	**170.43**	**244.73**

注：粗体表示最优值。

在 UCF-QNRF 数据集上与其他方法的对比结果如表 4 所列。

表 4 在 UCF-QNRf 数据集上与其他方法的结果对比

Table 4 Results comparison with other methods on UCF-QNRf dataset

Method	*MAE*	*MSE*
Zhang et al.[8]	315.00	508.00
MCNN[14]	277.00	426.00
Switching-CNN[15]	228.00	445.00
CP-CNN[16]	295.80	320.90
CSRNet[20]	122.80	207.10
ASNet[19]	91.59	159.71
P2PNet[22]	85.32	154.50
Our Method	**85.25**	**153.47**

注：粗体表示最优值。

UCF-QNRF 包含多种场景、多个视角、多种光线及密度变化的大规模已标注个体，此外还包含了建筑、植被、天空和道路等世界各地的户外真实场景，对于人群计数应用真正落地具有重要意义。本文方法在该数据集上的 *MAE* 为85.25，*MSE* 为 153.47，优于表中绝大多数的人群计数方法，具有较好的检测效果。该结果表明，采用多尺度特征聚合机制，可以有效地处理由不同相机视角和图像分辨率引起的尺度变化问题，在不同场景以及大规模变化的密集人群中具有较好的准确性与鲁棒性。

3.3.2 消融实验

本文在 ShanghaiTech 数据集的 Part_A 上进行了消融研究，以分析所提出模块的性能，表 5 列出了在该数据集上的实验结果。将本文设计的主干网络作为一个基线模型，P_3 层的预测结果 *MAE* 为 54.53，*MSE* 为 86.68。从表 5 中的第 2 行可以看出，将多尺度特征聚合模块加入到基线模型后，*MAE* 和 *MSE* 均有不同程度的下降，*MAE* 为 52.06，*MSE* 为 84.87。该结果表明，多尺度特征聚合模块在获取多尺度特征方面是有效的，进一步说明了多尺度特征聚合模块的必要性。第 3 行展示了加入 Smooth L_1 Loss 的效果，加入该损失函数后，*MSE* 有了较明显的下降，该结果说明了 Smooth L_1 Loss 的有效性，结合多尺度特征聚合模块，计数精度达到了最好的效果。

表 5 加入不同模块的效果

Table 5 Results of adding different modules

Baseline	Multi-scale feature	Smooth L_1 Loss	*MAE*	*MSE*
√			54.53	86.68
√	√		52.06	84.87
√		√	53.76	85.29
√	√	√	**51.21**	**83.70**

注：√ 表示加入该模块，粗体表示最优值。

此项实验研究在多尺度特征聚合模块设置不同扩张率的空洞卷积对计数精度的影响，表 6 列出了在 ShanghaiTech Part_A 上的实验结果。将不同分支中空洞卷积的扩张率设置为(1,2,3)时，*MAE* 和 *MSE* 有较小幅度的下降，设置为(1,2,5)时达到了最好的效果，*MAE* 为 52.06，*MSE* 为 84.87。此外，将扩张率设置为(2,3,5)时计数误差没有降低，这可能是因为较大的扩张率在获取大尺度个体特征的同时一定程度地丢失了小尺度个体的信息，不利于小尺度个体的检测。

表 6 使用不同扩张率的空洞卷积预测的结果

Table 6 Predicted results using dilated convolution with different expansion rates

different expansion rates	*MAE*	*MSE*
Baseline	54.53	86.68
Baseline+(1,2,3)	53.83	85.32
Baseline+(1,2,5)	**52.06**	**84.87**
Baseline+(2,3,5)	54.68	86.40

注：粗体表示最优值，(m,n,s) 表示增加不同扩张率的空洞卷积。

3.3.3 可视化效果及分析

图 5 给出了本文方法的预测结果与基于点预测的方法(P2PNet[22])以及真实标注点的比较结果。这 4 幅图像分别来自上述 4 个不同的数据集，真值以及预测值分别在图像的左下角显示。可以观察到，所有预测的图片都可以对个体

进行定位。由 ShanghaiTech Part_A 和 UCF-QNRF 数据集上的检测结果可以看出，与 P2PNet 相比，本文方法不仅在计数准确度上优于 P2PNet，对于图像底部较大尺度的个体，P2PNet 并未检测出，而本文方法均可成功检测到，这进一步证明了本文提出的多尺度特征聚合模块在捕获大尺度个体上具有的优异性能。由 UCF_CC_50 上的检测结果可以看出，本文方法在高度密集的场景下也能较好地发挥作用，充分提取极小尺度个体的特征信息，未造成信息丢失，且能够较为精准地检测来自不同场景与不同密度分布的图像，这些优异的结果证明了本文方法的有效性和鲁棒性。此外，选取 ShanghaiTech Part_A 中图片的检测结果与 P2PNet 模型进行对比，如图 6 所示，真值以及预测值分别在图像的右下角显示。由图 6 中红色框的个体可以看出，在距离摄像机较近的区域，其个体尺度较大，P2PNet 未能检测到这部分个体，而本文方法则可以成功检测到。其次，与真值注释相比，P2PNet 不仅漏检了较大尺度的个体，对其他小尺度个体也存在误检的情况，导致预测点与真值点人数相差较大，而本文方法不仅能够检测到大尺度个体，对小尺度个体的检测也具有更高的准确度。

图 5 在不同数据集上与其他方法的可视化效果对比

Fig. 5 Visualization results comparison with other methods on different datasets

图 6 与其他方法的可视化效果对比(电子版为彩图)

Fig. 6 Visualization results comparison with other methods

结束语 针对个体尺度不一、定位不准确这一问题，本文提出了一种新的基于多尺度的端到端的人群计数模型。为了有效利用不同特征层级间的多尺度特征，增加特征层级间的融合策略，本文进一步提出多尺度特征聚合模块，使模型对尺度相差很大的个体也能同时兼顾，在 3 个公开的人群计数数据集上进行了实验，证明了本文方法有效且优于其他人群计数相关方法。目前小尺度个体的定位准确度仍有待提高，在未来的工作中，将进一步研究对于不同尺度的个体，包含其语义信息和位置信息最为丰富的特征层级，训练网络自适应地选取最适合的特征层级进行预测，进一步提高计数和定位精度。

参 考 文 献

[1] DALAL N, TRIGGS B. Histograms of Oriented Gradients for Human Detection [C]// IEEE Computer Society Conference on Computer Vision & Pattern Recognition. IEEE, 2005.

[2] ENZWEILER M, GAVRILA D M. Monocular Pedestrian Detection: Survey and Experiments[J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2009, 31: 2179-2195.

[3] LEE M H, CHUNG K H, CHOI G K, et al. Measurement of Sr-90 in Aqueous Samples Using Liquid Scintillation Counting with Full Spectrum DPM Method [J]. Applied Radiation and Isotopes, 2002, 57(2): 257-263.

[4] MIN L, ZHANG Z, HUANG K, et al. Estimating the Number of People in Crowded Scenes by MID Based Foreground Segmentation and Head-shoulder Detection [C]// The 19th International Conference on Pattern Recognition. IEEE, 2009.

[5] DAVIES A C, JIA H Y, VELASTIN S A. Crowd Monitoring Using Image Processing[J]. Electronics & Communication Engineering Journal, 1995, 7(1): 37-47.

[6] MIN F, PEI X, LI X, et al. Fast Crowd Density Estimation with Convolutional Neural Networks[J]. Engineering Applications of Artificial Intelligence, 2015, 43(aug.): 81-88.

[7] WANG C, HUA Z, LIANG Y, et al. Deep People Counting in Extremely Dense Crowds [C]// The 23rd ACM International Conference. ACM, 2015.

[8] ZHANG C, LI H, WANG X, et al. Cross-scene Crowd Counting Via Deep Convolutional Neural Networks [C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2015: 833-841.

[9] ARTETA C, LEMPITSKY V, NOBLE J A, et al. Interactive Object Counting [C]// European Conference on Computer Vision. Cham: Springer, 2014.

[10] PENG X, PENG Y X, TANG Q, et al. Crowd Counting Based on Single-column Multi-scale Convolutional Neural Network[J]. Computer Science, 2020, 47(4): 150-156.

[11] PHAM V Q, KOZAKAYA T, YAMAGUCHI O, et al. COUNT

Forest：Co-voting Uncertain Number of Targets Using Random Forest for Crowd Density Estimation [C]// 2015 IEEE International Conference on ComputerVision(ICCV). IEEE，2015.

[12] WALACH E，WOLF L. Learning to Count with CNN Boosting [C]// European Conference on Computer Vision. Cham：Springer，2016.

[13] LI J Q，YAN H. Crowd Counting Method Based on Cross-column Features Fusion[J]. Computer Science，2021，48(6)：118-124.

[14] ZHANG Y，ZHOU D，CHEN S，et al. Single-image Crowd Counting Via Multi-column Convolutional Neural Network [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE，2016.

[15] SAM D B，SURYA S，BABU R V. Switching Convolutional Neural Network for Crowd Counting [C]// Computer Vision & Pattern Recognition. IEEE，2017：5744-5752.

[16] SINDAGI V A，PATEL V M. Generating High-quality Crowd Density Maps Using Contextual Pyramid CNNs [C]// 2017 IEEE International Conference on Computer Vision (ICCV). IEEE，2017.

[17] HOSSAIN M，HOSSEINZADEH M，CHANDA O，et al. Crowd Counting Using Scale-aware Attention Networks [C]// 2019 IEEE Winter Conferenceon Applications of Computer Vision (WACV). IEEE，2019.

[18] ZHANG A，SHEN J，XIAO Z，et al. Relational Attention Network for Crowd Counting [C]// 2019 IEEE/CVF International Conference on Computer Vision(ICCV). IEEE，2020.

[19] JIANG X，ZHANG L，XU M，et al. Attention Scaling for Crowd Counting [C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). IEEE，2020.

[20] LI Y，ZHANG X，CHEN D. CSRNet：Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE，2018.

[21] LIU Y，SHI M，ZHAO Q，et al. Point in，Box out：Beyond Counting Persons in Crowds [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE，2019.

[22] SONG Q，WANG C，JIANG Z，et al. Rethinking Counting and Localization in Crowds：A Purely Point-based Framework[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE，2021：3345-3354，

[23] JING S，CHEN C L，WANG X. Scene-independent Group Profiling in Crowd [C]// Computer Vision & Pattern Recognition. IEEE，2014.

[24] ZHU F，WANG X G. Crowd Tracking by Group Structure Evolution[J]. IEEE Trans on Circuits and Systems for Video Technology，2016，28(3)：772-786.

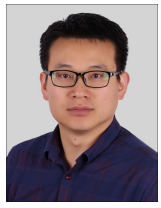
[25] LIN T Y，DOLLAR P，GIRSHICK R，et al. Feature Pyramid Networks for Object Detection [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE Computer Society，2017：2117-2125.

[26] SIMONYAN K，ZISSERMAN A. Very Deep Convolutional Networks for Large-scale Image Recognition [J/OL]. Computer Science，2014. https：//doi. org/10. 48550/arXiv. 1409. 1556.

[27] WU H，WANG W，ZHONG J，et al. SCS-Net：A Scale and Context Sensitive Network for Retinal Vessel Segmentation [J]. Medical Image Analysis，2021，70(10)：102025.

[28] IDREES H，SALEEMI I，SHAH M. Multi-source Multi-scale Counting in Dense Crowd Images [C]// Computer Vision and Pattern Recognition. IEEE，2013：2547-2554.

[29] DEB D，VENTURA J. An Aggregated Multicolumn Dilated Convolution Network for Perspective-free Counting[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops(CVPRW). IEEE，2013：308-317.



LIU Peigang，born in 1979，Ph.D，postgraduate supervisor，is a member of China Computer Federation. His main research interests include graphical image processing and data science and applications.

（责任编辑：喻藜）