·图形图像处理·　　文章编号：1000-3428(2022)05-0235-07　　文献标志码：A　　中图分类号：TP391

融合注意力机制与上下文密度图的人群计数网络

吴奇元，王晓东，章联军，高海玲，赵伸豪

（宁波大学 信息科学与工程学院，浙江 宁波 315211）

摘　要：为分析商业区人群流动情况，或避免人群踩踏等公共事件的发生，通常采用人群计数方法统计监控图像中的人数信息，从而达到提前预警的效果。受目标遮挡、背景干扰、多尺度变化等因素的影响，现有的人群计数方法在统计人数信息的过程中存在误算或漏算的问题，导致准确率降低。提出一种基于注意力机制与上下文密度图融合的人群计数网络CADMFNet。以VGG16的部分卷积层作为前端网络，通过引入上采样融合模块对输入的特征图进行上下文特征融合，将不同膨胀率的膨胀卷积作为后端网络，生成高质量的中间密度图。在此基础上，采用上下文注意力模块融合不同层级的中间密度图，获得精细的人群密度图。实验结果表明，该网络在Mall数据集上的平均绝对误差和均方根误差分别为1.31和1.59，相比CSRNet、MCNN等网络，能够有效提高计数的准确度，并且具有较优的鲁棒性。

关键词：人群计数；特征融合；膨胀卷积；注意力机制；卷积神经网络

开放科学（资源服务）标志码（OSID）：

中文引用格式：吴奇元，王晓东，章联军，等. 融合注意力机制与上下文密度图的人群计数网络[J]. 计算机工程，2022，48(5)：235-241，250.

英文引用格式：WU Q Y，WANG X D，ZHANG L J，et al. Crowd counting network with attention mechanism and contextual density map[J]. Computer Engineering，2022，48(5)：235-241，250.

Crowd Counting Network with Attention Mechanism and Contextual Density Map

WU Qiyuan，WANG Xiaodong，ZHANG Lianjun，GAO Hailing，ZHAO Shenhao

（Faculty of Electrical Engineering and Computer Science，Ningbo University，Ningbo，Zhejiang 315211，China）

【Abstract】To analyze the flow of people in a business district and avoid crowd stampedes，the crowd counting method is usually used to count the number of people in a monitoring image to achieve the effect of early warning. Affected by target occlusion，background interference，multi-scale change，and other factors，existing crowd counting methods have miscalculation and omission in the process of counting the number of people，resulting in low accuracy. A crowd counting network，CADMFNet，based on attention mechanism and contextual density map fusion is proposed. Taking part of the convolution layer of VGG16 as the front-end network，the contextual features of the input feature map are fused by introducing the up-sampling fusion module，and the dilated convolution with different dilation rates are used as the back-end network to generate a high-quality intermediate density map. On this basis，the contextual attention module is used to fuse the intermediate density maps of different levels to obtain a fine population density map. The experimental results show that the average absolute error and root mean square error of the network on the Mall data set are 1.31 and 1.59，respectively. Compared with CSRNet，MCNN，and other networks，it effectively improves the counting accuracy and exhibited better robustness.

【Key words】crowd counting；feature fusion；dilated convolution；attention mechanism；Convolutional Neural Network（CNN）

DOI：10. 19678/j. issn. 1000-3428. 0063039

0　概述

人群计数的主要任务是统计指定区域内的人数，其广泛应用在公共安全管理、商业信息采集、服务资源调度等领域。目前，人群计数主流方法是通过运用计算机视觉相关知识来统计一张图片中的人

基金项目：国家自然科学基金“超高清自由视点视频感知质量模型与绘制研究”（61771269）；宁波市自然科学基金“面向自由视点视频系统的立体视频质量评价研究”（2019A610107）。

作者简介：吴奇元（1996—），男，硕士研究生，主研方向为深度学习；王晓东，教授；章联军（通信作者），实验师；高海玲、赵伸豪，硕士研究生。

收稿日期：2021-10-25　　**修回日期：**2021-12-23　　**E-mail：**zhanglianjun@nbu.edu.cn

数。与传统的人力计数方法相比，该方法在节省人力物力资源方面成效显著。人群计数作为计算机视觉领域的研究热点，仍然面临着目标遮挡、尺度变化、分布不均、背景复杂、视角透视等难题。

基于计算机视觉的人群计数方法可以分为基于检测、基于回归、基于密度图的人群计数方法。基于检测的人群计数方法通过累加检测出的单个个体，以得到图片的最终人数。该方法[1-3]适用于低密度场景，但是难以解决中高密度或者目标遮挡场景下的人群计数问题。基于回归的人群计数方法[4-6]将图像模块内的人群作为一个整体，通过特征回归算法直接得到图像模块的对应人数，然后将包含人数信息的所有图像模块结果累计加和得到图片整体人数。此类方法能够解决背景部分遮挡的问题，提高适用场景的人群密度上限，但是受视角、背景等因素的制约，在高密度场景下的准确性较低。基于密度图[7]的人群计数方法是通过生成相应的人群密度图，进而将密度图中的像素密度值累加，累加值即为最终图片人数。此类方法能够充分利用像素级别的信息，普适性较强，广泛应用于各种密度场景，是当前人群计数领域的主流方法。

随着图像处理器（Graphics Processing Unit，GPU）计算能力的不断提高，基于深度学习的密度图人群计数方法应运而生。该方法充分利用计算机储存和算力，极大地提高了人群计数的准确性，同时提升了人群计数的实时性，在一定程度上解决了传统人群计数方法需要人工手动选取特征的难题，从而扩大适用场景的范围。文献[8]通过3列包含不同尺寸卷积核的神经网络来提取多尺度特征，以融合提取到的多尺寸特征，从而获得最终的人群密度图。在此方法提出之后，一系列基于多列卷积神经网络的人群计数方法层出不穷。基于文献[9]在网络前端增加一个图片人群密度等级分类器，将分类成不同密度等级的图片输入到不同的列，以得到相应人群密度图。文献[10]在文献[8]方法的基础上，利用跳接操作将各列网络的特征图相融合，得到最终人群密度图。文献[11]在微调文献[8]方法的基础上，通过加入两列网络分别获取全局和局部上下文信息，从而有策略地融合多列网络特征，得到最终的人群密度图。以上所述的基于多列卷积神经网络的方法在一定程度上提高了人群计数的准确率，文献[12]提出密集场景识别网络（Congested Scene Recognition Network，CSRNet），并指出多列卷积神经网络存在一定的不足。虽然多列网络使用不同尺寸的卷积核来提取多尺寸特征，但是得到的特征图存在信息冗余，并且网络结构复杂，训练耗时久。同时包含膨胀卷积的CSRNet因使用相同的膨胀率，导致不能充分利用前端网络得到的特征图。

本文提出一种基于注意力机制与上下文密度图融合的人群计数网络CADMFNet。采用前端网络实现逐级特征融合[13]，即使用上采样融合模块（Up-sampling Fusion Module，UFM）实现高层特征和低层特征的融合，得到不同层级的多尺寸特征，将互质膨胀率[14]交替使用的膨胀卷积作为后端网络，利用上下文注意力模块（Contextual Attention Module，CAM）融合不同层级的中间人群密度图，以充分利用上下文信息，从而提升最终生成人群密度图的质量。

1 本文网络

本文提出一种基于注意力机制与上下文密度图融合的人群计数网络CADMFNet。该网络结构如图1所示。

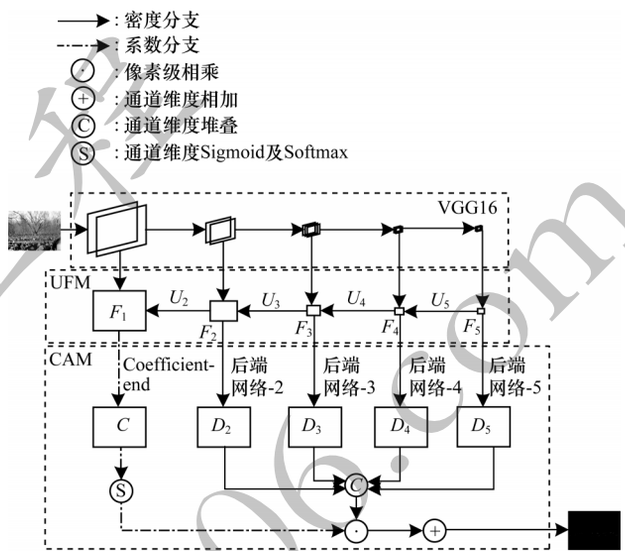


图1 本文网络结构

Fig.1 Structure of the proposed network

前端网络包含VGG16的前13层卷积层，将其按照2、2、3、3、3的层数分成5个阶段，池化层连接于下一个阶段的开头。将5个阶段的特征图依次输入到上采样融合模块（UFM）中进行上下文特征融合。其中，$U_i(i=2,3,4,5)$代表使用双线性插值的上采样操作模块，每次上采样操作后输出特征图的边长是原来输入特征图边长的2倍。$F_i(i=1,2,3,4,5)$代表特征融合后的特征图。网络在经过上采样融合模块以后，将$F_i(i=2,3,4,5)$输入到后端网络中得到中间密度图$D_i(i=2,3,4,5)$，然后将中间密度图在通道维度堆叠得到最终的中间密度图M。网络将含有最充足上下文信息的特征图F_1输入到系数网络中，以得到系数特征图C，并将其经过通道维度的Sigmoid以及Softmax操作得到最终的系数特征图X。最后采用逐个像素相乘并在通道维度相加的方法处理注意力系数X及中间密度图M，从而获得人群密度图。该网络模型除生成人群密度图的卷积层以外，其他卷积层后都跟有ReLU激活函数。

1.1 上采样融合模块

低层特征信息的感受野比较小，其包含小尺寸图像信息，即离摄像头位置较远的人群信息。高层特征信息的感受野较大，其包含大尺寸图像信息，即离摄像头位置较近的人群信息。上下文融合模块利

用高层语义信息和低层轮廓信息，得到较精细的人群密度图，从而获得较准确的人群计数结果。为得到包含高层语义信息和低层轮廓信息的上下文多尺寸特征，本文通过UFM得到融合后的上下文特征，并将这些不同的多尺寸上下文特征输入到后端网络，以减少特征图之间的冗余，从而更充分地利用网络提取的特征，提高计数的准确性。

在经过上采样融合模块处理之后，网络可以获得含有上下文信息的特征图$F_i(i=1,2,3,4,5)$。这些特征图的生成除了需要图1中的上采样操作以外，还需要通道堆叠以及卷积操作，其配置的详细情况如表1所示。卷积层参数表示为“（卷积核大小）-（输出通道数）-（膨胀率）”，concat为通道堆叠操作。通道堆叠和卷积操作$C_i(i=1,2,3,4,5)$作用在生成相应特征图$F_i(i=1,2,3,4,5)$之前，例如，通过C_3生成F_3主要分为2个步骤：1）在通道维度将F_4上采样后的特征图（256通道）和VGG16得到相应的特征图（256通道）进行堆叠，得到512通道的中间特征图；2）使用2个卷积操作将通道数从512先降为256，再降为128，从而得到最终的上下文特征图F_3。

表1　上采样融合模块配置信息

Table 1　Configurations information of up-sampling fusion module

卷积操作	通道堆叠	卷积参数-1	卷积参数-2
C_1	concat	3-64-1	3-64-1
C_2	concat	3-128-1	3-64-1
C_3	concat	3-256-1	3-128-1
C_4	concat	3-512-1	3-256-1
C_5	—	3-1 024-1	3-512-1

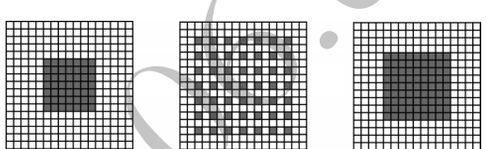
1.2　互质膨胀率的膨胀卷积

膨胀卷积能够增大感受野，且提升计数准确率。膨胀卷积通过修改卷积核结构来扩大感受野，无需进行下采样操作，不会丢失空间信息，因此其效果要优于卷积+池化+反卷积的方法。二维膨胀卷积的计算如式（1）所示：

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

其中：d为膨胀率，当$d=1$时，即为常见的普通卷积；$w(i,j)$为卷积核；$x(m+d\times i,n+d\times j)$为二维输入；$y(m,n)$为二维输出。

不同膨胀率的膨胀卷积感受野效果图如图2所示。图2中由内而外依次包含3×3卷积核1次卷积、2次卷积、3次卷积所涉及的新像素位置。从图2可以看出，膨胀率为2的膨胀卷积经过3次膨胀卷积以后，其感受野大于膨胀率采用1、2交替的膨胀卷积，更要大于膨胀率等于1的膨胀卷积。因此，膨胀卷积的膨胀率越大，经过相应膨胀卷积操作后的感受野也就越广。但是这并不意味着可以随意取膨胀卷积的膨胀率。如果膨胀率取值相同且不为1，则经过膨胀卷积之后，会出现“棋盘效应”，从图2（b）可以看出，如果连续使用膨胀率为2的膨胀卷积，就不能充分利用特征图的每个像素信息，中间会滤掉很多像素信息。此外，如果膨胀率过大，则经过膨胀卷积作用后的特征信息之间相关性较差。从图2（c）可以看出，互质的膨胀率交替使用（1、2交替）的膨胀卷积能够充分使用特征图的像素信息，以获得更准确的计数结果。因此，本文采用卷积率1、2交替的膨胀卷积作为后端网络，生成中间密度图，既可以在扩大感受野的同时，避免产生“棋盘效应”，也可以降低特征信息相关性产生的负面影响。



(a)膨胀率为1　(b)膨胀率为2　(c)膨胀率交替

图2　不同膨胀率的膨胀卷积感受野效果图

Fig.2　Effect images of dilated convolution receptive field with different dilation rates

1.3　上下文注意力模块

上下文注意力模块（CAM）主要包括系数分支以及密度图分支2个部分，如图1所示。其中，密度图分支利用不同的上下文特征信息$F_i(i=2,3,4,5)$生成相应的中间人群密度图M，系数分支将具有最全上下文信息的特征F_1作为输入，以生成系数特征图X，最终使用系数特征图给中间人群密度图打分，并将其对应的像素相加得到最终的人群密度图。

在密度图分支中，通过将不同的上下文信息输入到相应层级的多列后端网络，得到中间人群密度图。后端网络和系数网络的配置信息如表2所示。

表2　后端网络和系数网络配置信息

Table 2　Configurations information of back-end networks and coefficient-end network

网络	上采样-1	卷积层参数-1	卷积层参数-2	上采样-2	卷积层参数-3	卷积层参数-4	上采样-3	卷积层参数-5
系数网络	—	3-64-1	3-32-1	—	3-16-1	—	—	3-4-1
后端网络-2	up_2	3-64-1	3-64-2	—	3-64-1	3-64-2	—	1-1-1
后端网络-3	up_2	3-128-1	3-128-2	up_2	3-128-1	3-64-2	—	1-1-1
后端网络-4	up_2	3-256-1	3-256-2	up_2	3-128-1	3-64-2	up_2	1-1-1
后端网络-5	up_4	3-512-1	3-256-2	up_2	3-128-1	3-64-2	up_2	1-1-1

在表2中，卷积层参数表示为“（卷积核大小）-（输出通道数）-（膨胀率）”，up_2表示通过上采样操作将边长变为原来2倍，up_4表示通过上采样操作将边长变为原来4倍。后端网络包含逐级上采样操作和互质膨胀率的膨胀卷积操作，能够充分利用特征图像素级别的信息，在一定程度上提高网络的准确率。网络将每列后端网络得到的中间人群密度图在通道维度进行堆叠，从而得到最终的中间人群密度图M。

系数分支通过将具有不同层级上下文信息的特征图F_1输入到系数网络，以得到中间系数特征图C，因此中间系数特征图具有全局上下文信息，进而将中间系数特征图输入到Sigmoid函数和Softmax函数中，经过相应通道间像素操作，获得最终系数特征图X。

2 实验结果与分析

2.1 实验设置

2.1.1 人群密度图的生成

人群密度图代表人在图像中出现的概率，既包含空间分布特征信息，又包含人数特征信息，通过对人群密度图中的像素密度值积分进行累加，获得对应图像或图像模块的人数。与基于回归的人群计数方法相比，基于密度图的人群计数方法优势在于能够更好地使用图像特征，进而提升计数准确度。此外，学习人头区域信息比学习人头中心点信息更容易，高斯核函数能够近似拟合不同尺寸的人头，在一定程度上模拟人群遮挡等现实场景可能出现的情况，从而解决目标遮挡等难题。因此，人群计数主流方法采用高斯核函数与冲激函数卷积生成真实人群密度图，并将其作为预处理准备工作。

假设人群标注的人头标签坐标为x_i，则整幅图像所包含的人头位置如式(2)所示：

$$H(x)=\sum_{i=1}^{N}\delta(x-x_i) \tag{2}$$

其中：N为整幅图像的总人数；x为整幅图像的像素点坐标；x_i为图像中第i个人头的坐标；$\delta(x-\mathrm{x}_i)$为冲激函数。将高斯核函数与式(2)进行卷积可得密度图函数，如式(3)所示：

$$D(x)=H(x)*G_{\sigma_i}(x) \tag{3}$$

高斯核函数$G_{\sigma_i}(x)$可表示为：

$$G_{\sigma_i}(x)=\frac{1}{2\pi\sigma_i^2}\mathrm{e}^{-\frac{(x-x_i)^2}{2\sigma_i^2}} \tag{4}$$

其中：σ_i为二维高斯分布的标准差；$G_{\sigma_i}(x)$为$D(x)$在x位置关于x_i的高斯核函数值；$D(x)$在x处的取值由所有人位置x_i的高斯核函数取值加和得到。

由高斯核函数性质可得：对应的标准差σ_i越大，第i个人在真实密度图中对应的图像区域越大，因此标准差与人头区域大小呈正相关。文献[8]提出一种适用于密集场景生成真实人群密度图的方法，使用与单个人距离最近的k个人之间的平均距离$\bar{d}_k^i$来代表σ_i，σ_i如式(5)所示：

$$\sigma_i=\beta\bar{d}_k^i \tag{5}$$

其中：$\bar{d}_k^i$为距离第i个人最近的k个人之间的平均距离。当$\beta=0.3$时[8]，利用该方法得到的人群密度图计数结果更准确。

2.1.2 损失函数

网络模型训练使用均方误差来衡量估计密度图与真实密度图之间的差异，因此，将均方误差作为损失函数来调整估计密度图的生成，如式(6)所示：

$$L_{\mathrm{Loss}}(\theta)=\frac{1}{N}\sum_{i=1}^{N}(D_p(I_i,\theta)-D_{gi})^2 \tag{6}$$

其中：θ为本文模型的网络参数；I_i为输入的第i幅图像；$D_p(I_i,\theta)$为第i幅图像得到的估计密度图；D_{gi}为第i幅图像得到的真实密度图；N为数据集包含的图像总数。

2.1.3 训练设置

由于目前公开的公共数据集中图片数量不是特别多，为防止产生过拟合现象，因此本文使用数据增强技术来扩充数据集中的图片数量。在网络训练阶段，本文对数据集图片用0.5的概率进行左右翻转，并从每张数据集图片中随机裁减出一个128×128像素的图片补丁作为网络输入；相应真实人群密度图采取相同的操作。本文使用Adam优化器，学习率被设置为固定的0.000 05，批量大小(batch size)设置为64。由于128×128像素的图片补丁可能包含的人数为0，在一定程度上可以缓解数据集图片无人负样本不足的情况，因此能够帮助模型排除复杂背景噪声的干扰。

2.2 评价指标与数据集

本文实验环境为Windows 10操作系统，配备16 GB NVIDIA Quadro RTX 5000 GPU显卡。采用CUDA 11.0版本的Pytorch深度学习框架，分别在UCF_CC_50、ShanghaiTech及Mall数据集上进行实验，初步验证了本文网络在高、中、低各类密度场景下的计数效果。

2.2.1 评价指标

本文网络模型采用平均绝对误差(MAE)和均方根误差(RMSE)作为实验结果的性能指标，反映网络计数的准确性及稳定性。

1)平均绝对误差(MAE)。计算对象为测试集图片的估计人数和标签真实人数，如式(7)所示：

$$M_{\mathrm{MAE}}=\frac{1}{N}\sum_{i=1}^{N}|\hat{C}_i-C_i| \tag{7}$$

其中：N为测试图像数；$\hat{C}_i$为第i幅图像的估计人数；C_i为第i幅图像的真实人数。

2)均方根误差(RMSE)。计算对象为测试集图片的估计人数和标签真实人数，如式(8)所示：

$$R_{\mathrm{RMSE}}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{C}_i-C_i)^2} \tag{8}$$

其中：N为测试图片数；$\hat{C}_i$为第i幅图片的估计人数；

C_i为第i幅图片的真实人数。

2.2.2　UCF_CC_50数据集

UCF_CC_50数据集[15]的图片来源于互联网，总共有50张图片，每张图片中的人数为94~4 543个人，共计63 974个人。该数据集包含的场景人群密集程度高且图片数量稀少，是目前最具挑战性的人群数据集之一。本文采用五折交叉验证[15]方法进行实验：将数据集包含的50张图片随机等分为5个图片集，依次采用一个图片集作为测试集，剩余图片集打通合并后作为训练集，最终实验结果为5次实验结果的平均值。该数据集的平均人数为1 279.5个人，属于高密度场景，因此使用MCNN方法[8]生成人群密度图。在UCF_CC_50数据集上，本文对MCNN、MSCNN[16]、ResNeXtFP[17]、PACNN[18]、CSRNet[12]等网络的评价指标进行对比，结果如表3所示。

表3　在UCF_CC_50数据集上不同网络的评价指标对比

Table 3　Evaluation indexs comparison among different networks on UCF_CC_50 dataset

网络	MAE	RMSE
MCNN网络	377.6	509.1
MSCNN网络	363.7	468.4
ResNeXtFP网络	269.6	312.3
PACNN网络	267.9	357.8
CSRNet网络	266.1	397.5
文献[19]网络	245.3	318.8
MRA-CNN网络[20]	240.8	352.6
CAT-CNN网络[21]	235.5	324.8
文献[22]网络	234.1	317.7
本文网络	238.3	363.1

从表3可以看出，在UCF_CC_50数据集上，虽然本文网络CADMFNet未取得最优的指标，但是相比CSRNet方法，其MAE下降了10.4%，RMSE下降了8.7%，说明本文网络在一定程度上可以提高人群计数的准确度。此外，由于UCF_CC_50数据集的划分方法具有一定的随机性，因此该数据集得到的计数结果虽然具备一定的参考意义，但是不具有绝对性的判断作用。

2.2.3　ShanghaiTech数据集

ShanghaiTech数据集[8]分为2个部分，共有1 198张图片，包含330 165个人。Part A中包含的图片主要来源于互联网，该部分共有482张人数范围为33~3 193个人的图片，其中，300张为训练集，182张为测试集。Part B人群密度比Part A稀疏，是上海市区繁华街道的图片，该部分包含716张人数范围为9~578个人的图片，其中，400张为训练集，316张为测试集。Part A部分平均人数为501.4个人，Part B部分平均人数为123.6个人，均属中密度场景。因此，采用文献[8]网络生成人群密度图，在ShanghaiTech数据集上不同网络的评价指标如表4所示。

表4　在ShanghaiTech数据集上不同网络的评价指标对比

Table 4　Evaluation indexs comparison among different networks on ShanghaiTech dataset

网络	ShanghaiTech Part A		ShanghaiTech Part B	
	MAE	RMSE	MAE	RMSE
MCNN网络	110.2	173.2	26.4	41.3
MSCNN网络	83.8	127.4	17.7	30.2
MRA-CNN网络	74.2	112.5	11.9	21.3
ResNeXtFP网络	69.3	104.7	14.3	21.9
CSRNet网络	68.2	115.0	10.6	16.0
文献[22]网络	66.8	100.0	11.6	18.4
PACNN网络	66.3	106.4	8.9	13.5
文献[19]网络	65.3	108.4	8.1	13.3
JCCL+Reg网络[23]	63.9	99.7	8.2	14.8
本文网络	62.2	100.6	7.9	12.1

从表4可以看出，在ShanghaiTech数据集上，相比CSRNet网络的结果，本文网络在ShanghaiTech Part A的MAE下降了8.8%，RMSE下降了12.5%；在ShanghaiTech Part B上的MAE下降了25.5%，RMSE下降了24.4%。

2.2.4　Mall数据集

Mall数据集[5]来源于安装在购物中心的监控摄像头拍摄得到的图片，总共包含2 000张图片，共有62 325个人。在这2 000张图片中，前800张划分到训练集，后1 200张划分到测试集。该数据集包含人数为13~53个人的图片，平均人数为31.2个人，属于低密度场景。此外，该数据集包含遮挡、光照、镜子成像等干扰。本文根据透视图生成真实密度图，透视图包含人群远近信息权重，主要采用式(9)的方法来生成二维高斯分布的标准差($\sigma_{i,j}$)，要求满足$7\leqslant\sigma_{i,j}\leqslant25$且向上取整：

$$\sigma_{i,j}=\frac{25}{\sqrt{x_{i,j}}},x_{i,j}>0 \tag{9}$$

其中：$x_{i,j}$为透视图中第i行、j列的像素值；$\sigma_{i,j}$为图片中第i行、j列的人头对应的二维高斯标准差。在Mall数据集上，本文对文献[5]、IFDM[24]、CNN-Boosting[25]、DRSAN[26]、E3D[27]、DecideNet[28]等网络的评价指标进行对比，结果如表5所示。

表5　在Mall数据集上不同网络的评价指标对比

Table 5　Evaluation indexs comparison among different networks on Mall dataset

网络	MAE	RMSE
文献[5]网络	3.15	15.70
IFDM网络	2.45	3.20
CNN-Boosting网络	2.01	—
DRSAN网络	1.72	2.10
E3D网络	1.64	2.13
DecideNet网络	1.52	1.90
LA-Batch网络[29]	1.34	1.60
SAAN网络[30]	1.28	1.68
本文网络	1.31	1.59

从表5可以看出，在Mall数据集上本文网络的评价指标基本取得了最优的结果。测试集平均每张图片对应的人数绝对误差为1.31，对应的人数均方根误差为1.59。

由以上结果分析可知，本文网络适用于高密度场景、中密度场景和低密度场景的人群计数。

2.3 实验结果

在不同数据集上本文网络的可视化结果如图3所示。图3中第1列为每个数据集中的原图，第2列为真实人群密度图，包含真实人数，第3列为本文网络的估计人群密度图，包含估计人数。在UCF_CC_50数据集上本文网络的真实人数为1 037个，估计人数为1 372.6个。在ShanghaiTech Part A数据集上本文网络的真实人数为502个，估计人数为432.6个。在ShanghaiTech Part B数据集上本文网络的真实人数为181个，估计人数为166.2个。在Mall数据集上本文网络的真实人数为33个，估计人数为32.5个。

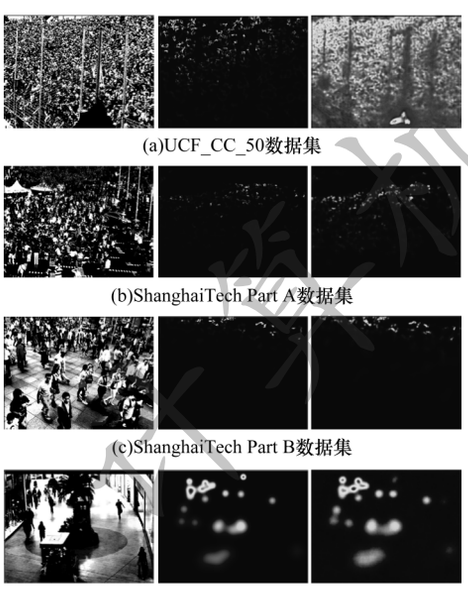


图3 本文网络的可视化结果

Fig.3 Visualization results of the proposed network

不同列密度图的可视化结果如图4所示，网络输入为ShanghaiTech Part A数据集中的图片，经过CADMFNet后，不同列对应的密度图输出为密度图_i(i=2, 3, 4, 5)。其中，密度图_i对应后端网络Back-end-i输出的密度图与相应系数特征图相乘，得到的特征图作为网络的输出。密度图_2对应小尺寸的特征信息，重点关注离摄像头较远处的信息；密度图_3对应尺寸稍大的特征信息，重点关注较小尺寸的信息；密度图_4对应比较全局的信息，重点关注较大尺寸的信息；密度图_5对应全局信息，重点关注大尺寸信息。将上述密度图叠加后得到最终密度图。本文网络估计的密度图人数为218.2个，而真实密度图人数为223个。因此。本文网络利用上采样融合模块(UFM)和上下文注意力模块(CAM)一定程度上可以解决多列网络中不同列提取的特征图之间冗余信息的问题，从而提高网络的准确率。

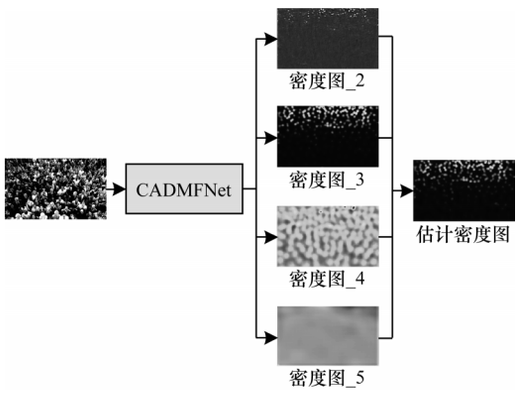


图4 不同列密度图的可视化结果

Fig.4 Visualization results of different column density maps

CADMFNet网络取得较优的准确性与鲁棒性的同时，也存在一定的不足，比如参数量过大、运行速度较慢。CSRNet方法的参数量为62 MB，CADMFNet方法的参数量达到了145 MB。因此，CADMFNet网络适用于对准确率要求比较高但对实时性相对较低的场景，如学校、商场和景区的人流统计等。

2.4 消融实验

本文在ShanghaiTech Part A数据集上进行消融实验，验证互质膨胀率交替使用(Alternating Dilation Rate，ADR)、上采样融合模块(UFM)、注意力模块(Attention Module，AM)和后端网络中逐级上采样(Step by step Up-sampling，SU)的有效性。其中，AM和SU包含上下文注意力模块(CAM)，验证其有效性相当于验证了上下文注意力模块的有效性。注意力模块主要表现为图1中Coefficient-end所在的那一列。Baseline为CADMFNet去掉ADR、UFM、AM、SU之后的网络(dilation rate=2)。在ShanghaiTech Part A数据集上的消融实验结果如表6所示。

表6 在ShanghaiTech Part A数据集上的消融实验结果

Table 6 Ablation experimental results on ShanghaiTech Part A dataset

网络	MAE	RMSE
Baseline网络	69.7	116.0
Baseline + ADR网络	68.4	103.2
Baseline + ADR + AM网络	68.0	109.0
Baseline + ADR + AM + UFM网络	64.5	111.3
CADMFNet网络	62.2	100.6

从表6可以看出，本文提出的ADR、UFM、AM和SU使得人群计数的MAE下降，具有提升计数准确性的效果。在Baseline基础上增加相关模块后，虽然人群计数的RMSE产生一定的波动，但是其相差不大且都比Baseline的RMSE小。因此，上述模块能够有效增强人群计数的稳定性。CADMFNet同时使用以上4个模块，能够提高网络的准确性和稳定性，使整体效果达到最优。相比Baseline网络，CADMFNet的MAE降低了10.8%，RMSE降低了13.3%。

3　结束语

本文提出一种基于注意力机制与上下文密度图融合的人群计数网络CADMFNet。采用上采样融合模块、上下文注意力模块来减少多列网络中不同列之间的冗余信息，通过不同膨胀率的膨胀卷积生成高质量的中间密度图，以提高人群计数的准确度。在此基础上，将上述互补密度图相加得到最终的人群密度图。实验结果表明，相比CSRNet网络，本文网络能够提高人群计数的准确性与稳定性，其在Mall数据集上的平均绝对误差和均方根误差分别为1.31和1.59。后续将从同比例减少网络通道数角度优化人群计数方法，以提高网络的运行效率。

参考文献

[1] ZOU L H，LIU Y C. A new algorithm of counting human based on segmentation of human faces in color image[C]// Proceedings of International Conference on Computational Intelligence and Security. Washington D. C.，USA：IEEE Press，2009：505-509.

[2] HOU Y L，PANG G K H. People counting and human detection in a challenging situation[J]. IEEE Transactions on Systems，Man，and Cybernetics-Part A：Systems and Humans，2011，41(1)：24-33.

[3] IDREES H，SOOMRO K，SHAH M. Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2015，37(10)：1986-1998.

[4] RYAN D，DENMAN S，FOOKES C，et al. Crowd counting using multiple local features[C]//Proceedings of Digital Image Computing：Techniques and Applications. Washington D. C.，USA：IEEE Press，2009：81-88.

[5] CHEN K，LOY C C，GONG S G，et al. Feature mining for localised crowd counting[EB/OL]. [2021-09-22]. http://www. eecs. qmul. ac. uk/~sgg/papers/ChenEtAl_BMVC2012. pdf.

[6] CHAN A B，VASCONCELOS N. Bayesian poisson regression for crowd counting[C]//Proceedings of the 12th International Conference on Computer Vision. Washington D. C.，USA：IEEE Press，2009：545-551.

[7] LEMPITSKY V，ZISSERMAN A. Learning to count objects in images[C]//Proceedings of the 23rd International Conference on Neural Information Processing Systems. Cambridge，USA：MIT Press，2010：1324-1332.

[8] ZHANG Y Y，ZHOU D S，CHEN S Q，et al. Single-image crowd counting via multi-column convolutional neural network[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2016：589-597.

[9] SAM D B，SURYA S，BABU R V. Switching convolutional neural network for crowd counting[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2017：4031-4039.

[10] 陆金刚，张莉. 基于多尺度多列卷积神经网络的密集人群计数模型[J]. 计算机应用，2019，39(12)：3445-3449.
LU J G，ZHANG L. Crowd counting model based on multi-scale multi-column convolutional neural network[J]. Journal of Computer Applications，2019，39(12)：3445-3449.(in Chinese)

[11] SINDAGI V A，PATEL V M. Generating high-quality crowd density maps using contextual pyramid CNNs[C]// Proceedings of IEEE International Conference on Computer Vision. Washington D. C.，USA：IEEE Press，2017：1879-1888.

[12] LI Y H，ZHANG X F，CHEN D M. CSRNet：dilated convolutional neural networks for understanding the highly congested scenes[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2018：1091-1100.

[13] SONG Q Y，WANG C G，WANG Y B，et al. To choose or to fuse？ scale selection for crowd counting[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence.[S. l.]：AAAI Press，2021：2576-2583.

[14] 李佳倩，严华. 基于跨列特征融合的人群计数方法[J]. 计算机科学，2021，48(6)：118-124.
LI J Q，YAN H. Crowd counting method based on cross-column features fusion[J]. Computer Science，2021，48(6)：118-124.(in Chinese)

[15] IDREES H，SALEEMI I，SEIBERT C，et al. Multi-source multi-scale counting in extremely dense crowd images[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2013：2547-2554.

[16] ZENG L K，XU X M，CAI B L，et al. Multi-scale convolutional neural networks for crowd counting[C]// Proceedings of IEEE International Conference on Image Processing. Washington D. C.，USA：IEEE Press，2017：465-469.

[17] KALYANI G，JANAKIRAMAIAH B，PRASAD L V N，et al. Efficient crowd counting model using feature pyramid network and ResNeXt[J]. Soft Computing，2021，25(15)：10497-10507.

[18] SHI M J，YANG Z H，XU C，et al. Revisiting perspective information for efficient crowd counting[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2019：7271-7280.

[19] ZHUGE J C，DING N N，XING S J，et al. An improved deep multiscale crowd counting network with perspective awareness[J]. Optoelectronics Letters，2021，17(6)：367-372.

[20] ZHANG Y M，ZHOU C L，CHANG F L，et al. Multi-resolution attention convolutional neural network for crowd counting[J]. Neurocomputing，2019，329：144-152.

[21] CHEN J W，SU W，WANG Z F. Crowd counting with crowd attention convolutional neural network[J]. Neurocomputing，2020，382：210-220.

[22] 翟强，王陆洋，殷保群，等. 基于尺度自适应卷积神经网络的人群计数算法[J]. 计算机工程，2020，46(2)：250-254，261.
ZHAI Q，WANG L Y，YIN B Q，et al. Crowd counting algorithm based on scale adaptive convolutional neural network[J]. Computer Engineering，2020，46(2)：250-254，261.(in Chinese)

（下转第250页）

（上接第241页）

[23] JIANG M Y, LIN J Z, WANG Z J. A smartly simple way for joint crowd counting and localization[J]. Neurocomputing, 2021, 459: 35-43.

[24] 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[J]. 计算机应用研究，2021，38(4)：1256-1260，1280.
YUAN J, WANG S S, LUO Y W. Public place crowd counting model based on image field division [J]. Application Research of Computers, 2021, 38(4): 1256-1260, 1280. (in Chinese)

[25] WALACH E, WOLF L. Learning to count with CNN boosting[C]// Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 660-676.

[26] LIU L B, WANG H J, LI G B, et al. Crowd counting using deep recurrent spatial-aware network[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. New York, USA: ACM Press, 2018: 849-855.

[27] ZOU Z K, SHAO H L, QU X Y, et al. Enhanced 3D convolutional networks for crowd counting [EB/OL]. [2021-09-22]. https://arxiv. org/abs/1908. 04121.

[28] LIU J, GAO C Q, MENG D Y, et al. DecideNet: counting varying density crowds through attention guided detection and density estimation [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 5197-5206.

[29] ZHOU J T, ZHANG L, DU J W, et al. Locality-aware crowd counting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 65: 99-103.

[30] HOSSAIN M, HOSSEINZADEH M, CHANDA O, et al. Crowd counting using scale-aware attention networks[C]// Proceedings of IEEE Winter Conference on Applications of Computer Vision. Washington D. C., USA: IEEE Press, 2019: 1280-1288.

编辑 薛晋栋