文章编号：2095-2163(2021)05-0059-06 中图分类号：TP391.41 文献标志码：A

基于多尺度及双注意力机制的小尺寸人群计数

王良聪[1]，吴晓红[1]，陈洪刚[1]，何小海[1]，潘 建[2]，赵 威[1]
（1 四川大学 电子信息学院，成都 610065；2 中国民航局第二研究所，成都 610041）

摘 要：本文针对背景干扰、特征信息不足以及尺度剧烈变化等问题，提出了一种基于多尺度及双注意力机制（Multi-Scale and Dual Attention，MSDA）的小尺寸人群计数网络。MSDA 网络主要由空间-通道双注意力（Spatial Channel-dual Attention，SCA）模块和多尺度特征融合（Multi-scale Feature Fusion，MFF）模块构成。MFF 模块将特征送入三列拥有不同卷积核的膨胀卷积来扩大小目标的空间尺度，再通过特征级联及卷积操作进行多尺度特征融合；SCA 模块把特征送入通道注意力网络，使用空间注意力中的池化操作及逐像素相乘操作加强细节信息；最后将处理好的特征送入密度图生成模块，通过 1×1 卷积获得密度图。在 Mall 数据集和 Shanghaitech 数据集上进行了测试，取得了较好的准确率与鲁棒性。

关键词：人群计数；双注意力；特征融合；膨胀卷积

Small size crowd counting based on multi-scale and dual attention mechanism

WANG Liangcong[1], WU Xiaohong[1], CHEN Honggang[1], HE Xiaohai[1], PAN Jian[2], ZHAO Wei[1]
(1 School of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China;
2 The Second Research Institute of the Civil Aviation of China, Chengdu 610041, China)

【Abstract】Aiming at the problems of background interference, insufficient feature information, and dramatic changes in scale, This paper proposed a small-scale crowd counting network based on multi-scale and dual attention mechanism (Multi-Scale and Dual Attention, MSDA). The MSDA network was mainly composed of a Spatial Channel-dual Attention (SCA) module and a Multi-scale Feature Fusion (MFF) module. The MFF module sent the features into three columns of dilated convolutions with different convolution kernels to expand the spatial scale of small targets, and then performs multi-scale feature fusion through feature cascade and convolution operations; the SCA module input the features into the channel attention network, Then use the pooling operation in the spatial attention, and use the pixel-by-pixel multiplication operation to enhance the detailed information; Finally, send the processed features to the density map generation module, and obtain the density map through 1×1 convolution .This paper tested the proposed model on Shanghai Tech and Mall datasets, and the results show that the model achieve good accuracy and robustness.

【Key words】crowd counting; dual attention; feature fusion; dilated convolution

0 引 言

利用现代信息技术及创新成果，打造宜居、安全、便利、智能的生活环境，是社会良性发展的普遍追求。近年来，大量人口选择汇集在城市工作、安家，城市单位面积内的人口密度越来越大，因此带来了一系列的问题，这些问题是高效、有序的社会管理面临的巨大挑战。如 2020 年 1 月 7 日，伊朗高级将领苏来曼尼遭遇美方突袭不幸身亡，伊朗民众纷纷为其送葬，但送葬的过程中发生了意外，百万民众送别时发生踩踏事件，至少造成了 56 人死亡、213 人受伤。因此，提前对人群信息进行快速统计，避免严重的公共安全责任事故发生是必要的。人群密度估计，需要重点关注人群的分布信息，然而实际场景中往往面临着相似物体（如树叶，车辆）的干扰，很难从局部小区域得出判定；在此情况下，人类的做法是观察更久，同时结合其它的周围信息进行判断。受此启发，本文设计了双注意力模块来解决这样的问题。由于摄像机拍摄视角的多样性和人群位置的复杂分布，图像中的人头尺度是变化多样的，为了应对视角剧烈变换问题，设计了一个多尺度特征融合模块，来增强网络的多尺度特征提取能力，并融合多尺

基金项目：国家自然科学基金（61891287）；四川省科技计划项目（2019YFH0034）。

作者简介：王良聪（1993-），女，硕士研究生，主要研究方向：计算机视觉、智能监控与异常行为分析；吴晓红（1970-），女，博士，副教授，主要研究方向：图像处理与模式识别；陈洪刚（1991-），男，博士，助理研究员，主要研究方向：图像处理；何小海（1964-），男，博士，教授，主要研究方向：图像处理、图像识别、图像通信；潘 建（1987-），男，学士，助理研究员，主要研究方向：智能运行控制；赵 威（1993-），男，硕士，工程师，主要研究方向：智能控制与异常行为分析。

收稿日期：2021-02-15

度信息。本文提出的基于多尺度及双注意力机制(Multi-Scale and Dual Attention,MSDA)的小尺寸人群计数网络,实现端到端的人群计数。即输入单幅图像,就可以通过对生成的密度图进行积分,得到图像中的人群数量。本文在 Shanghaitech 数据集和 Mall 数据集上进行了实验,并取得了较好的效果。本文的贡献主要有以下 3 点:

(1)受 KNN 自适应[1]标注方法的启示,根据相机的成像原理以及画面的透视畸变,提出了基于透视关系的密度图生成方法。

(2)设计了一个多尺度特征融合模块,以达到多尺度特征融合及丰富特征信息的目的。

(3)设计了空间—通道双注意力模块,来实现对无关特征的弱化,强调重要特征。

1 相关工作

早期的研究中,采用基于检测的方法[2],即使用整体或部分身体特征的检测,训练一个分类器,利用从行人中提取到的整体或局部结构来检测行人,从而进行计数。由于基于检测的方法,在背景杂乱且密度高的图像上,表现性能会大大降低,因此有人提出了基于回归的计数方法[3],该方法是学习一种从特征到人数的映射。但此方法会忽略空间信息,还会受到尺度和视角剧烈变化的影响,导致计数能力变差。

近年来,深度卷积神经网络得到了广泛应用,在人群计数方向也取得了显著的成果。例如,使用深度卷积网络直接端对端生成密度图的方法[4]。文献[1]中提出,利用 3 个具有大中小的卷积核的神经网络 MCNN,来分别提取人群中的特征,然后通过卷积层来生成密度图从而进行计数。文献[4]中,使用 3 个不同的 CNN 回归器和一个分类器来生成密度图。Wu 等[5]使用反向卷积层,自适应的分配权重给两个分支,将计数问题看做分类,从而来进行人群计数; Chen[6]等人使用像素级的注意力机制对图像进行分级,使之生成高质量的密度图。

由此可见,近年来有许多学者针对人群计数这一课题做出了努力。但是大部分的网络[7-8],虽然性能不错,但还存在一些未能很好解决的问题。如,存在特征信息提取不充足,无法从多个感受野中提取多尺度信息,也没有融合多个尺度中的特征,达到丰富细节特征的目的;并且无法排除背景中的干扰,弱化无关特征,强调重要特征,从而来提升人群计数的准确度。基于此,本文提出了一种基于多尺度及双注意力机制的小尺寸人群计数网络来解决上述问题。

2 提出方法

由于本文的任务对象主要为小尺寸密集人群,过深的网络将存在过度的冗余,并且不利于性能特征的迁移。而 VGG-16 模型深度较小,能够在保证足够源域特征的同时兼顾小尺寸目标,因此本文模型将 VGG-16 作为主干网络。将 VGG-16 与提出的多尺度特征融合模块与空间-通道双注意力模块相结合,来对图像中的小尺寸目标进行检测。

2.1 密度图生成

针对人数估计,数据集[2-3]将画面中的行人标记分别以头部某点的位置坐标(头部轮廓几何中心最佳)的形式保存,即点标注形式。采用点标注的主要原因:一是大大提高效率,不用过分地去考虑每个目标精确的尺寸问题;二是因为人体头部包含的信息较多,并且在高密度人群中,仅仅头部可见。因此使用点标注来标注头部,是人群估计中较为普遍的标注方式。

假设目标的标记坐标为 p_i, 则对图像中 n 个目标的总体标注函数为:

$$H(p) = \sum_{i=1}^{n} \delta(p - p_i) , \tag{1}$$

对于点标注,文献[2]中将每个目标的标注坐标都与二维高斯低通滤波函数 $G_\sigma(P)$ 进行卷积操作后,则将形成整体的目标密度图 $D(p)$, 即:

$$D(p) = H(p) * G_\sigma(P) = \sum_{i=1}^{n} \delta(p - p_i) * G_\sigma(P). \tag{2}$$

经过此操作,就可将孤立的点标注扩散至贴合目标头部轮廓的置信密度分布。若假设目标头部是圆形,通过限定二维离散高斯低通滤波函数的作用区间和标准差,就可以使得单个目标在此区间内的密度积分求和为 1,从而拟合图像中的具体人数。

文献[1]中提出使用 KNN 算法自适应地估计图像中目标的尺寸,但场景的密集程度并不存在严格划分标准,难以形成一个统一、可移植的泛化方案。鉴于此,本文根据相机成像原理及图像的透视畸变问题,提出了基于透视关系的密度图生成方法。由于各成像设备的陈设一般都为水平放置,会导致在同一水平线上的人的尺度大致相同,符合远小近大的成像原理,据此关系可得出人群分布的位置与图像上的纵坐标呈正相关。

设目标头部的尺寸为 P_x, 可得出整体图像的透视关系为:

$$P_x = k * P_y + b, \tag{3}$$

其中，P_y 表示图像中的纵坐标；k 表示透射畸变因子；b 为偏移因子。k 、b 为待定系数，可根据图像中两个纵坐标位置不同的目标人头，确定整幅图像的透视关系，选择两个纵坐标不同的目标 P_{x1}、P_{x2}，可得：

$$k = \frac{P_{x1} - P_{x2}}{P_{y1} - P_{y2}}, \tag{4}$$

$$b = P_{x1} - k * P_{y1}. \tag{5}$$

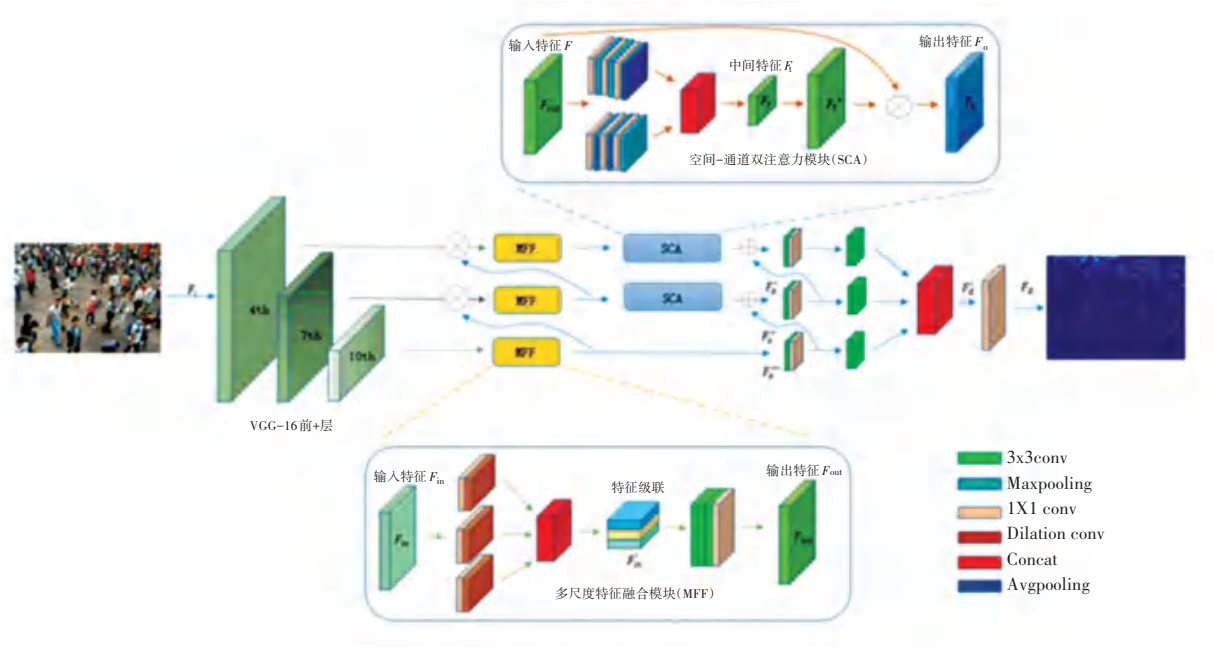

图 1　MSDA 模型

Fig. 1　Multi-scale and dual attention mechanism model

由公式(4)、式(5)可得出：

$$P_x = \frac{P_{x1} - P_{x2}}{P_{y1} - P_{y2}} * P_y + P_{x1} - \frac{P_{x1} - P_{x2}}{P_{y1} - P_{y2}} * P_{y1}. \tag{6}$$

依据此方法即可对图像中的人群进行标注，从而生成密度图。如图 2 所示。其中图 2(a)为 Zhang 等[3]所提出方法的示意图，图 2 (b)为本文方法的示意图。由图中可看见，本方法自适应的匹配了人头尺寸。

(a) 固定高斯核　　(b) 本文方法

(a) Fixed Gaussian Kernel　　(b) This method

图 2　不同密度图高斯核尺寸的效果对比

Fig. 2　Comparison of the effect of Gaussian kernel size in different density maps

2.2　MSDA 模型

如图 1 所示，本文 MSDA 模型分为以下 4 个模块：VGG-16 特征提取模块，多尺度特征融合模块(MFF)，空间-通道双注意力模块(SCA)及密度图生成模块。F_i 和 F_d 为 MSDA 模型的输入与输出，以 VGG-16 中的部分卷积层及池化层作为基础结构，

在其第 4、7、10 层分别提取特征进行解码设计，而后将提取的 3 个层的特征分别送入 3 个 MFF 模块中，1×1 卷积之后将输入的特征进行转换并使得通道数统一，再对其使用膨胀卷积来扩大感受野，最后进行特征融合，并将深层融合的特征作为输入，传递给浅层，可得：

$$F_{in} = H(F_{out} \otimes F'_{in}). \tag{7}$$

式中，$H(\cdot)$ 表示卷积操作；$\otimes$ 表示逐像素相乘操作；F'_{in} 表示经过 MFF 层进行过多尺度特征融合的深层输出特征；F_{in} 是浅层特征和深层特征融合后的多尺度特征。将低层网络的融合特征 F'_{in} 与高层网络所提取的特征 F_{out} 进行逐像素相乘操作，可以使得低高层特征进行融合，得到丰富的上下文信息。

经过 SCA 模块，使用平均池化、最大池化以及卷积操作达到对无关特征的弱化及重要特征的强调，最后将特征 F_o 输入密度图生成模块。首先经过一个 3×3 卷积和一个 1×1 卷积，再与上一层的特征结合，重新送入卷积层中；然后通过 2 个 3×3 卷积，最后使用 concat 操作对特征信息进行结构化的相加

融合,加强特征之间的联系,相较于直接特征叠加,大幅减少了特征的通道数。由此可得:

$$F_d' = K(H(H(F_o''') \oplus F_o), H(H(F_o'') \oplus F'o), H(F_o''')). \quad (8)$$

其中,$H(\cdot)$ 表示卷积操作;$K(\cdot)$ 表示 concat 操作;⊕表示逐像素相加操作;F_o'、F_o''、F_o''' 分别为第 4、7、10 层经过 SCA 模块的特征。F_d' 为 3 层 concat 之后的最终特征信息层级,将 F_d' 送入 1×1 卷积层中,得到密度图 F_d。

2.2.1 多尺度特征融合模块

由于摄像机拍摄的视角和人群位置的复杂性,图像中的人头尺度是复杂多样的,因此想要更准确的进行计数,就需要进行多尺度特征提取。本模块分别在 3 个不同层中的单层特征图中提取多尺度信息,然后融合提取后的信息。如图 1 所示,在 MFF 网络中,首先使用一个 1×1 的卷积层对特征映射的通道进行压缩整合。由于低层网络的感受野较小,其语义表征能力弱,因此将整合的低层特征分别送入三个膨胀率为 1、2、3 的膨胀卷积网络中,可得:

$$F_{in}' = K(D(H(F_{in}), d = 1), D(H(F_{in}), d = 2), D(H(F_{in}), d = 3)). \quad (9)$$

其中,$H(\cdot)$ 表示卷积操作;$K(\cdot)$ 表示 concat 操作;$D(\cdot)$ 表示膨胀卷积操作及其中的 d 为膨胀率。F_{in} 经过三列膨胀卷积操作,使用 concat 操作以及特征级联进行多尺度的特征融合,再经过 3 个 3×3 卷积扩大感受空间,从更广的视野非线性判断各位置的特征取舍;再经过一个 1×1 卷积层将特征进行转换并使得通道数统一,得到 F_{in}';MFF 模块以此来扩大低层特征中的感受野,将语义表征能力增强。

2.2.2 空间-通道双注意力模块(SCA)

一般的注意力模块只能将原始图片中的空间信息变换到另一个空间中,并保留关键信息或解决信息超载问题,而无法在空间和通道上关注特征和加强联系。鉴于此,本文设计了一个空间-通道双注意力模块,使用通道注意力网络,学习各通道的依赖程度,并根据依赖程度对不同的特征图进行调整,再结合使用空间注意力。此举不仅弥补了通道注意力的某些不足之处,还可以强调重要特征信息并忽略了无关特征信息。本模块构成如图 1 中 SCA 模块所示。首先将输入的特征 F_f, 分别送进 2 个不同的通道,然后进行 1×1 的卷积操作来整合特征,再分别在 2 个通道中使用最大池化层和平均池化层。可得:

$$F_f = K(M((A(H(F_i),2)^2),2), A((M(H(F_i),2)^2),2)). \quad (10)$$

其中,$H(\cdot)$ 表示卷积操作;$K(\cdot)$ 表示 concat 操作;$M(\cdot)$ 表示最大池化操作;$A(\cdot)$ 表示平均池化操作;公式(10)中的 2 表示 $pool = 2$。使用最大池化层 $M(\cdot)$ 可以收集目标中更细节的线索,而平均池化层 $A(\cdot)$ 可以将特征进行压缩,此时就实现了在通道上关注人群特征。将经过处理的特征快速进行不同于上一次的平均池化以及最大池化,加上空间注意力。

最后将特征 F_f 进行上采样,将其与原始特征 F_i 进行逐像素相乘操作,得到输出特征 F_o。则有:

$$F_o = \mathrm{Upsample}(F_f, 6) \otimes F_i. \quad (11)$$

其中,$Upsample(\cdot)$ 表示上采样操作,⊗表示逐像素相乘操作。

3 实验与分析

3.1 模型训练

在训练阶段,人群密度研究工作中一般都将欧几里得损失当做训练损失,损失函数定义如下:

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \| gt_i - gt(X_i;\theta) \|^2. \quad (12)$$

其中,$g\,t_i$ 表示输入的第 i 张地面真实密度图;$gt(X_i;\theta)$ 表示预测估计的密度图;X_i 表示输入的第 i 张图像;θ 是计数网络中可学习的参数。

因 Adam[8] 具备计算效率高、内存要求低等优点,本文将其作为优化器;设置初始学习率(Learning rate)为0.000 01;同时为了使梯度下降方向更稳定、准确、防止震荡,令每次训练输入所选取的样本数(batch size)等于 4,并随机打乱每次样本的输入顺序。

3.2 评价指标

本文采用的评价指标:平均绝对误差(Mean Absolute Error, MAE)、均方误差(Mean Squared Error,MSE)。其定义如下:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |gt_i - et_i|, \quad (13)$$

$$MSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (gt_i - et_i)^2}. \quad (14)$$

其中,n 代表测试集的样本总数;gt_i 表示第 i 张测试图的实际人数值;而 et_i 表示对第 i 张测试图的估计值。MAE 主要是考量真实值与估计值之间的误差平均,反映的是估计的准确性,而 MSE 作为方差指标,反映的是算法鲁棒性。由于这 2 项指标为误

差度量，因此算法的 *MAE*、*MSE* 值越小越好。

3.3　数据集

3.3.1　Shanghaitech 数据集

该数据集分为两部分：Shanghaitech Part_A（简称 SHA）和 Shanghaitech Part_B（简称 SHB），并分别将图像中人体头部的中心区域某点的位置坐标保存在文件中。SHA 源于互联网照片，由训练集中的 300 张图像和测试集中的 182 张图像构成，存在少量的灰度样本，共计241 677个标注点，照片质量不一，且绝大多数图像都拥有高密度人群；该数据集样本风格各异，且存在很多相似目标的干扰，某些极其拥挤的场景甚至人眼也难以准确计数，适合检验算法对远景高密度人群的估计能力。SHB 拍摄于上海繁华的街道，由 400 张训练图像和 316 张测试图像构成，场景不固定，共计88 488个标注点；该数据集更贴近具体街道应用场景，训练集、测试集中存在很多相似场景，画面中人员特征信息较多，能够检验算法在城市监控场景中的表现能力。

3.3.2　Mall 数据集

该数据集获取于国外某购物中心，由拍摄视频中抽取的 2 000 帧图片构成，且场景固定，并尽量以头部中心点的坐标作为标注，共计 62 325 个标注点。数据集的难点是远处目标模糊以及植物的干扰，其代表的是稀疏、固定场景。参照文献[2]，选择前 800 帧图像训练，其余 1 200 帧图像进行测试。

3.4　实验结果

3.4.1　在 Mall 数据集的实验结果比较

在 Mall 数据集上，将本文的网络与其它网络进行了比较，比较结果见表 1。结果表明，提出方法的性能有所改进；与 2020 年提出的 CWAN[9] 网络相比，MAE 和 MSE 分别提高了 0.56 和 0.87。在 Mall 数据集上的人群密度效果如图 3 所示。

表 1　Mall 数据集的实验结果对比

Tab. 1　Comparison of experimental results in the mall dataset

方法	MAE	MSE
Boosting CNN[10]	2.01	N/A
IFDM[11]	2.45	3.2
CWAN[9]	2.06	2.90
ours	1.5	2.13

3.4.2　在 Shanghaitech 数据集上的实验结果比较

在 ShanghaiTech partA & partB 两个数据集上，将本文网络和其它网络进行了比较，结果见表 2。结果表明，本文提出的方法性能有显著改进。在 SHA 中，本文方法与网络 CSRNet[12] 相比，MAE/MSE 分别提高了 4.5/12.7，与 2020 年所提网络 HANG[13] 相比也提高了 1.6/4.1；在 SHB 部分，MAE/MSE 比 CSRNet 提高了 2.18/2.71，与 HANG 相比提高了 1.58/4.31。其效果如图 3 所示。

图 3　数据集中估计的密度图（左为 SHA，中为 SHB，右为 Mall）

Fig. 3　The estimated density map in the dataset（SHA on the left，SHB in the middle，Mall on the right）

表 2　Shanghaitech 数据集实验结果对比

Tab. 2　Comparison of experimental results in the Shanghaitech dataset

方法	SHA		SHB	
	MAE	MSE	MAE	MSE
Switch-CNN[4]	90.4	135.0	21.6	33.4
CSRNet[12]	68.2	115.0	10.6	16.0
SANet[7]	67.0	104.5	8.4	13.6
LMCNN[6]	69.3	106.4	11.1	14.4
HAGN[13]	65.3	106.4	10.0	17.6
ASD[5]	65.6	98.0	8.5	13.7
MCNN[1]	110.2	173.2	26.4	41.3
ours	62.1	98.19	8.3	12.49

3.5　网络结构分析

为了验证本文所提出的 MFF 模块及 SCA 模块的有效性，在 ShanghaiTech 数据集上进行了验证，验证结果如表 3 所示。SHA 部分的 MAE/MSE 提高了 7.24/10.25；SHB 部分的 MAE/MSE 也同样提高了 0.55/1.74，MFF 模块中使用 1×1 卷积核进行特征信息整合，可在不影响感受野的情况下增强决策函数的非线性，并且结合膨胀卷积之后可在不损失图像分辨率与尺寸情况下有效扩大感受野，减小参数量。MFF 模块增加了多尺度特征融合，提取不同人头的细节信息并且使得高低层特征融合，联合上下文信息。

表 3　网络结构分析结果

Tab. 3　Network structure analysis results

	SHA		SHB	
Module	MAE	MSE	MAE	MSE
基本框架	75.34	111.98	9.33	14.75
+MFF	68.1	101.73	8.78	13.02
+MFF+SCA	62.1	98.19	8.3	12.49

由表 3 数据可以看出，添加了 SCA 模型之后，

在 SHA 和 SHB 两个数据集上都有明显的提高。如 SHA 上的 MAE/MSE 分别提高了 6.0/3.54；在 SHB 上也提高了 0.48/0.53，SCA 模块中对输入的特征分别施加 3 层的平均池化、最大池化后进行特征叠加，为之后的操作提供更多的选择及降低参数量；并且 SCA 模块使用取舍权重因子与原输入特征相乘得到输出特征，以实现对输入特征的弱化；因此，SCA 模块可以有效地弱化无关特征，强调目标信息。

4 结束语

本文提出的基于多尺度和双注意力机制的人群计数网络模型，使用膨胀卷积、最大池化和平均池化设计了 MFF 模块和 SCA 模块。MFF 模块扩大了小目标的尺度空间，使浅层特征和深层特征进行了多尺度特征融合，改善了小目标中的特征信息缺乏问题和尺度剧烈变化问题。SCA 模块使用池化层将注意力放在小目标上，可排除无关干扰信息，提取有效特征、减小参数量。在数据集上的测试结果表明，此方法比现有许多方法都有效，可应用于景区游客统计，反映游客的实时分布，安排相应的旅游服务；也可对拥挤的现象进行预警，实时的检测人群密度，以发现异常聚集或逃离事件的发生，从而及时协调医疗、警员力量。

参考文献

[1] ZHANG Yingying, ZHOU Desen, CHEN Siqin, et al. Single-Image Crowd Counting via Multi-Column Convolutional Neural Network [C]//IEEE Conference on Computer Vision and Pattern Recognition. 2016: 589-97.

[2] CHEN Ke, LOY Chen Change, GONG Shaogang, et al. Feature Mining for Localised Crowd Counting [C]//British Machine Vision Conference. 2012: 3.

[3] CHAN A B, LIANG Z S, VASCONCELOS N. Privacy preserving crowd monitoring: Counting people without people models or tracking [C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. 2008: 1-7.

[4] Sam Deepak Babu, Surya Shiv, Babu R. Venkatesh. Switching Convolutional Neural Network for Crowd Counting [J]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4031-4039.

[5] WU Xingjiao, ZHENG Yingbin, YE Hao, et al. Adaptive Scenario Discovery for Crowd Counting [C]//IEEE International Conference on Acoustics, Speech and Signal Processing. 2019: 2382-2386.

[6] 陈美云，王必胜，曹国，等. 基于像素级注意力机制的人群计数方法 [J]. 计算机应用，2020，40(1)：56-61.

[7] CAO Xinkun, WANG Zhipeng, ZHAO Yanyun, et al. Scale aggregation network for accurate and efficient crowd counting [C]//Proceedings of the European Conference on Computer Vision. 2018: 734-750.

[8] DAI K J, R-FCN Y L. Object detection via region-based fully convolutional networks. arxiv preprint [M]. arXiv preprint. 2016.

[9] KONG Xiyu, ZHAO Muming, ZHOU Hao, et al. Weakly Supervised Crowd-Wise Attention For Robust Crowd Counting [C]//IEEE International Conference on Acoustics, Speech and Signal Processing. 2020: 2722-2726.

[10] Walach Elad, Wolf Lior. Learning to Count with CNN Boosting [C]//European Conference on Computer Vision. 2016: 660-76.

[11] 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.

[12] LI Yuhong, ZHANG Xiaofan, CHEN Deming. CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes [C]//computer vision and pattern recognition. 2018: 1091-1100.

[13] DUAN Zuodong, XIE Yujun, DENG Jiahao. HAGN: Hierarchical Attention Guided Network for Crowd Counting [J]. IEEE Access, 2020, 8(3): 6376-6385.

(上接第 58 页)

[9] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition [C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 6450-6459.

[10] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.

[11] WORRALL D E, GARBIN S J, TURMUKHAMBETOV D, et al. Harmonic networks: Deep translation and rotation equivariance [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5028-5037.

[12] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2. 2015: 2017-2025.

[13] ROCCO I, ARANDJELOVIC R, SIVIC J. Convolutional neural network architecture for geometric matching[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 6148-6157.