计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200700107

基于跨列特征融合的人群计数方法

李佳倩　严　华
四川大学电子信息学院　成都 610065
(jiiaqian@outlook.com)

摘　要　人群计数是计算机视觉和机器学习领域中一个极具挑战性的课题。由于人群尺度变化和场景遮挡等现象会导致计数准确度不高,因此提出了一种基于跨列特征融合的人群计数方法(Cross-column Features Fusion Network,CCFNet)。该方法融合了来自多列不同接受域的特征,并且结合了拥有互质扩张率的空洞卷积,因此不仅能够增大感受野,还能保证信息的连续性,从而更好地适应人群规模的巨大变化;同时引入注意力模型引导网络聚焦于图片中的头部位置,根据注意力分数图为不同位置分配不同的权重,突出人群而弱化背景,最终得到高质量的密度图。在当前主流的人群计数数据集上的对比实验中,所提方法的平均绝对误差(Mean Absolute Error,MAE)在 ShanghaiTech 数据集的 A,B 子集上分别达到了 63.2 和 8.9,在 UCF_CC_50 数据集上达到了 222.1,在 WorldExpo'10 数据集上达到了 7.1。这表明所提方法具有更好的计数准确度,能够很好地适应不同的场景,尤其对于尺度变化较大的场景,效果优于以往的大多数算法。

关键词:人群计数;跨列特征融合;空洞卷积;注意力模型
中图法分类号　TP391

Crowd Counting Method Based on Cross-column Features Fusion

LI Jia-qian and YAN Hua
School of Electronics and Information Engineering,Sichuan University,Chengdu 610065,China

Abstract　Crowd counting is a challenging subject in computer vision and machine learning. Due to the phenomenon of crowd scale change and scene occlusion,the counting accuracy is low. A crowd counting method based on cross-column features fusion,called cross-column features fusion network(CCFNet),is proposed in this paper. CCFNet fuses features from multiple columns and different receptive fields,and combines with the dilate convolution employing coprime expansion rate. Therefore,CCFNet can not only increase the receptive field but also ensure the continuity of information,so as to adapt to the huge changes in the crowd size better. At the same time,the attention model is introduced to guide the network to focus on the head position in the images. According to the attention score graph,different weights are assigned to different positions to highlight the crowd and weaken the background. Finally,a high-quality density map is obtained. In comparative experiments on the current mainstream population counting datasets,the mean absolute error(MAE) reaches 63.2 and 8.9 on the A and B subsets of the ShanghaiTech dataset,222.1 on the UCF_CC_50 dataset,and 7.1 on the WorldExpo'10 dataset. The results show that the proposed method has better counting accuracy and can adapt to different scenes. Especially for scenes with large scale variation,its effect is better than most of the pre-vious algorithms.

Keywords　Crowd counting,Cross-column features fusion,Dilate convolution,Attention model

1　引言

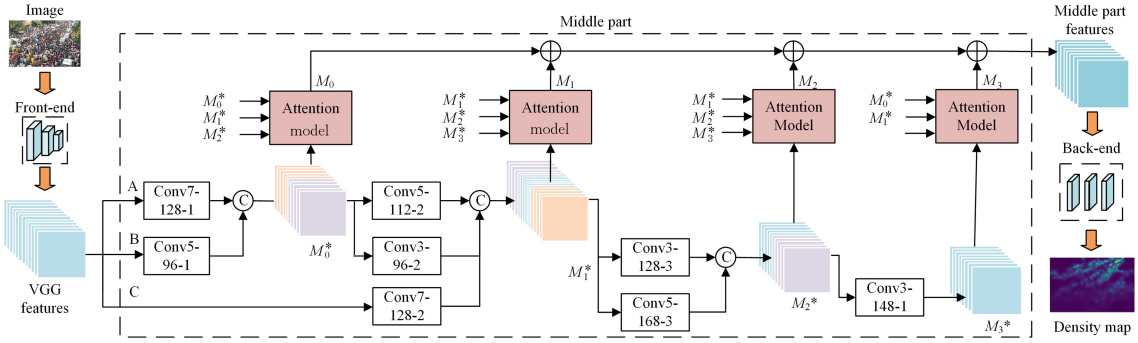
近年来,人群计数这一课题在计算机视觉领域受到了广泛关注,这是因为在公共安全、应急疏散、智慧城市规划等领域都对其有着日渐迫切的需求。目前,研究者已经对其进行了大量研究并取得了一定进展[1]。然而在现实场景中,由于存在场景遮挡、人群分布不均匀、光照不均匀、尺度视角的变化等问题,人群计数的准确度依然面临着很大的挑战。

当前大多数基于深度学习的人群计数方法的核心是采用标准卷积。针对图像尺度变化问题,这些方法主要是从不同分辨率的图像块中提取特征,或者采用多列架构,用不同内核大小的滤波器来提取不同尺度的特征[2-5],以此来适应尺度变化巨大的各种密集场景。然而这些方法的网络结构都相对复杂,大多都不能有效利用各分支学习的特征,并且缺乏在复杂

到稿日期:2020-07-17　返修日期:2020-10-20　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(11872069)
This work was supported by the National Natural Science Foundation of China(11872069).
通信作者:严华(yanhua@scu.edu.cn)

背景下对人群头部信息的关注，导致其不能获得令人满意的计数准确度。针对以上问题，本文提出了一种基于跨列特征融合的人群计数网络（Cross-column Features Fusion Network，CCFNet）。

（1）CCFNet 网络分为 3 部分：前端（front-end）、中段（middle part）和后端（back-end），如图 1 所示。其中，取 VGG16（Visual Geometry Group）网络的前 10 层作为前端；CCFNet 的中段采用一种跨列特征融合结构，由浅至深地将不同列的多尺度特征逐层融合；CCFNet 的后端则采用由标准卷积构成的译码器生成密度图。这样的网络结构能够增强不同列之间信息的融合，跨列学习网络特征，以提高网络对于不同尺度特征的敏感度，增强网络的鲁棒性。



注：Attention model 为注意力模块；图中卷积层以“conv（kernel size）-（filter number）-（dilation rate）”的格式进行标注；“VGG features”表示原图经过 VGG 前 10 层网络后得到的特征，该特征作为整体分别输送至 A，B，C 3 个分支；“©”表示通道拼接；蓝色方块代表经过内核大小为 3 的卷积层输出的特征；紫色方块代表经过内核大小为 5 的卷积层输出的特征；粉色方块代表经过内核大小为 7 的卷积层输出的特征；“M_i^*”表示每一层特征融合后得到的特征；“M_i”表示经过注意力模型后得到的特征

图 1 基于跨列特征融合的人群计数网络结构（电子版为彩色）

Fig. 1 Crowd counting network structure based on cross-column feature fusion

（2）CCFNet 中段网络引入具有互质扩张率的空洞卷积代替标准卷积和池化层，以保证信息的连续性，从而在尽可能不损失信息的情况下增大感受野，同时让卷积输出包含更大范围的信息。

（3）CCFNet 中段网络引入注意力机制，引导每一次特征融合后的网络聚焦于头部区域，将人群和背景进一步分离，提高了人群计数的准确度。

实验表明，CCFNet 网络模型在当前主流的人群计数数据集上都表现良好，优于现有的主要算法，能有效地提高计数准确度并增强不同场景的鲁棒性。

2 相关工作

人群计数方法分为传统方法和基于卷积神经网络的方法，下面对这两类方法作简单介绍。

2.1 传统方法

对于人群计数而言，早期的方法大多集中于基于检测的方法[6-7]，通过滑动窗口检测器检测并计数[8]或通过训练好的分类器从整个人群中提取低层特征（如 Haar 小波[9]和 HOG（Histogram of Oriented Gradient）[10]）。为了消除遮挡和复杂背景造成的影响，研究人员提出使用基于回归的方法来计数，在回归中学习从局部图像块提取的特征到计数之间的映射[11-13]。但是上述方法都忽略了空间信息，导致局部区域的结果不准确。Lempitsky 等[14]提出学习局部 patch 特征与对应对象密度图之间的线性映射，从而在学习过程中融入空间信息，通过对密度图进行积分得到最终计数结果。此后，大部分人群计数的研究都是基于该方法进行的。

2.2 基于卷积神经网络的方法

卷积神经网络在计算机视觉方面表现良好，因此目前对于人群计数这一课题的研究重心逐渐转移到深度学习方面。最初，Wang 等[15]提出了一种端到端深度 CNN（Convolutional Neural Network）回归模型，用于对极度密集的人群图像统计人数。为了解决人群尺度变化大带来的性能下降问题，研究者提出了各种多分支结构。Zhang 等[2]使用 3 列不同接受域的 CNN 来描述不同尺度的对象；Babu 等[16]从集成学习中得到启发，训练了一系列网络或回归器来处理不同的场景；Cao 等[17]采用类似 Inception[18]的模块来集成额外的分支；Shen 等[19]插入子分支以匹配跨尺度的一致性，并插入对抗性分支以减弱密度图的模糊效果。这些多分支网络虽然能够从网络的不同分支中提取多尺度特征来处理人群的尺度变化，但是在这些结构中不同分支之间的特征都缺乏信息交流，特征图之间的关联度较低，使得最终密度图的准确度不高。为了降低网络的复杂度，Li 等[20]以单列 CNN 的形式结合了空洞卷积，有效地扩展了接受域以捕获上下文信息，然而空洞卷积的特性导致了训练过程中信息不连续，使其没有达到最理想的效果；Shi 等[21]引入透视图来获取图像中人物比例变化的相关信息，进而生成高精度的密度图，但是在现实场景中透视图这一特征很难获取。此外，上述方法都是通过学习图像的密度图来进行人群计数，根据图像中头部的位置生成真实密度图，通过积分得到最终计数结果，然而头部位置的重要性并没有在其模型架构中得到明确体现。因此，在不引入其他特征的前提下，本文针对上述方法存在的问题提出了一个能够跨列融合特征的网络结构，并根据空洞卷积的特性进行了改进，同时引入了注意力机制来增加人群头部和背景之间的分离度，最终有效地提高了计数准确度。

3 提出方法

本文提出了一种基于跨列特征融合的网络 CCFNet（见图 1），它可以端到端地从源图像中学习并得到相应尺度的密

度图。由于 VGG[22] 具有较强的迁移学习能力和灵活的架构，因此我们选择预训练的 VGG16 的前 10 层作为 CCFNet 网络的前端；接着在中段部分部署了一个由具有互质扩张率的空洞卷积构成的跨列特征融合结构，并结合注意力机制，使其能够在尽可能大的范围内有效融合不同的尺度特征，同时强调头部位置的信息；CCFNet 网络后端部署了由标准卷积构成的译码器，可得到高质量的密度图和鲁棒性更好的网络。表 1 列出了网络的具体参数设置，卷积层的参数表示为“conv(kernel size)-(filter number)-(dilation rate)”。

表 1 CCFNet 网络结构表

Table 1 CCFNet network structure

	layer	parameters		
Front-end	1－2	conv3-64-1		
		max-pooling		
	3－4	conv3-128-1		
		max-pooling		
	5－7	conv3-256-1		
		max-pooling		
	8－10	conv3-512-1		
	layer	parameters		
Middle part	branch	A	B	C
	1	conv7-128-1	conv5-96-1	
	2	conv5-112-2	conv3-96-2	conv7-128-2
	3		conv3-128-3	conv5-168-3
	4			conv3-148-1
	layer	parameters		
Back-end	1	conv7-128-1		
	2	conv5-64-1		
	3	conv3-64-1		
	4	conv1-1-1		

3.1 跨列特征融合结构

为了处理图像中人群的尺度变化，通常需要提高网络对于尺度特征的敏感性，我们通过提取多尺度特征来解决这个问题。受文献[2]和[23]的启发，本文设计了一个跨列特征融合结构来提取和融合不同尺度的特征。

首先，CCFNet 的中段部分的起点是前端 VGG16 前 10 层网络的输出特征，我们将其作为基础来构建尺度感知特性。然后，在中段部署 3 个分支(A,B,C)通过不同内核大小的滤波器提取不同尺度的特征，由浅至深地生成 4 个包含不同语义信息的特征图。如图 1 所示，CCFNet 前端输出的特征图(图 1 中的 VGG Features)作为 A,B,C 3 个分支的输入，按照[A,B]－[A,B,C]－[B,C]－[C]的顺序分别在每一层进行列与列之间的特征融合，生成 4 个特征图(图 1 中的 M_0^*，M_1^*，M_2^* 和 M_3^*)。通过列与列之间的信息交互，可充分细化来自不同列的多尺度特征，每一个分支的滤波器内核大小如表 1 中的 Middle part 所列。对于稀疏人群，相邻人群之间的距离更大，头部尺寸通常比密集人群大。因此，我们使用带有较大滤波器的低层子网络来识别较稀疏的人群，而使用带有较小滤波器的高层子网络来识别较密集的人群。每一次特征融合的特征都来自不同尺寸的滤波器，意味着每次特征融合都是不同尺度特征之间的一次相互渗透和补充的过程，从而使得输出的密度图拥有更加丰富的信息。

与此同时，每一层经过特征融合后生成的特征图需要先经过注意力模块(注意力模块在 3.3 节进行详细介绍)，生成经过校准的 4 个特征图(图 1 中的 M_0，M_1，M_2，M_3)，再从前向后进行横向连接，即特征图上对应元素相加，就这样迭代到最后一层产生最精细的密度图。过程需要保证密度图的通道数一致，因此每次迭代前一层时需要先通过 1×1 的滤波器保证通道数一致，在本文方法中统一规定通道数为 128。这样的结构能够有效地融合语义信息丰富的深层特征和语义信息较少的浅层特征，用深层特征的语义信息充分细化浅层特征的语义信息，不同尺度特征之间的信息交流再一次被增强，使得最终的密度图能够很好地削弱尺度变化带来的影响。

3.2 互质扩张率的空洞卷积

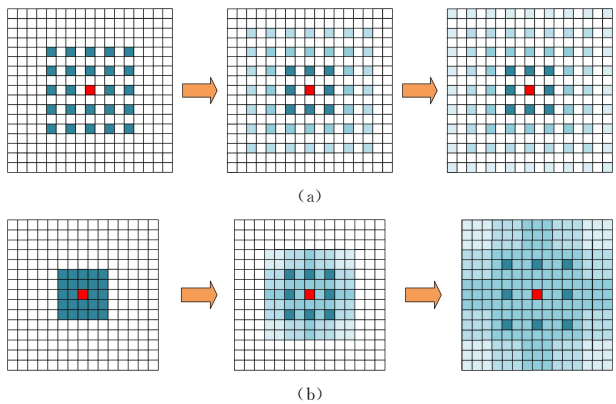
前文所述的跨列通信结构专注于对不同尺度的特征进行融合和交流，而在通常的训练过程中需要利用池化层来扩大网络的感受野。但是，图片尺寸在池化操作后会减小，难免会丢失一些信息。受文献[20]的启发，我们引入空洞卷积[24-25]来代替池化层以扩大感受野，但又不会增加参数数量和计算量，同时通过改变扩张率来保持信息的相对完整性。

一个二维空洞卷积可以被定义如下：

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

其中，r 是扩张率，如果 $r=1$ 就相当于普通卷积。当 $r>1$ 时，则表示将一个带有 $k\times k$ 滤波器的小核放大到 $k+(k-1)(r-1)$。为了保证信息的连续性，我们将多尺度跨列通信结构中的卷积层设置成表 1 所列的空洞卷积层，即每条分支的扩张率都按照层数使用一组互质数，如 A:[1,2]，B:[1,2,3]，C:[2,3,1]。这样可以实现在保持相同分辨率的情况下，充分地聚合多尺度的上下文信息。

图 2 给出了以图 1 中 B 分支为例使用不同扩张率的空洞卷积的效果对比图，图 2(a)中所有卷积层的扩张率 $r=2$，图 2(b)中 3 层卷积层的扩张率 r 分别为 1,2,3。从图 2(a)可以看出，扩张率一直为 2 的空洞卷积层叠加起来虽然感受野扩大了很多，但明显不是所有的像素都参与了计算，存在信息的缺失；从图 2(b)可以看到，采用扩张率互质的空洞卷积层叠加后不仅能扩大感受野，还能够覆盖所有像素，因此可在很大程度上保证信息的连续性。



注：像素(蓝色标记)通过 3 个内核大小分别为 5,3,3 的连续层参与中心像素(红色标记)的计算

图 2 空洞卷积扩张率的对比图(电子版为彩色)

Fig. 2 Contrast graph of expansion rate of dilate convolution

3.3 注意力模型

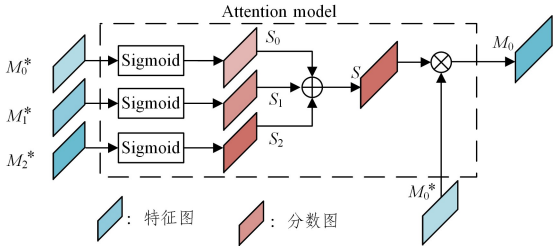
由空洞卷积构成的跨列通信结构已经能够较好地融合多尺度特征。但是我们发现，在捕捉特征的过程中，所有的像素值都拥有相对平均的权重。而在卷积网络的训练和测试过程中，计数的关键信息在于头部位置的信息，估计的密度图中头部区域会有更大的值，因此需要引导网络聚焦于头部区域，以突出头部位置的像素点。在文献[26-27]的启发下，我们利用注意力模型(attention model)来调整整个网络在密度图中对不同区域的注意程度。因此，在前述方法的基础上，我们在CCFNet中段网络的特征提取层 M_0^*，M_1^*，M_2^* 和 M_3^* 之后引入注意力模型来生成注意力分数图。注意力分数图是一个图像大小的权重图，人群区域有较高的分数值，尤其是头部区域。

在本文的多尺度跨列通信结构中，引入如图3所示的Attention model结构，每一次特征融合后生成的特征图 M_i^* 再经过一个Sigmoid层，利用Sigmoid激活层的门限机制，生成关于注意力分布的分数图：

$$S_i = Sigmoid(M_i^*), i \in \{0,1,2,3\} \tag{2}$$

其中，M_i^* 表示每层特征融合生成的特征图，与图1中的 M_i^* 应；S_i 表示经过Sigmoid层之后的分数图。分数图将所有像素的值映射在[0,1]区间内。为了进一步提高分数图的准确性，选取其中3层的注意力分数图汇总：

$$S = \sum_{i=0}^{2} S_i \tag{3}$$



注：本图以图1中 M_0^* 分支为例进行展示；图中虚线框中的部分与图1中的Attention model模块对应；M_0^* 表示经过跨列特征融合后的特征，与图1中的 M_0^* 对应；M_0 表示经过注意力机制的特征，与图1中的 M_0 对应

图3 Attention model结构图

Fig. 3 Attention model structure diagram

最后，用汇总后的分数图 S 再反向对之前的特征图进行校准，强调头部出现的相关区域，为这些区域分配更高的权重。M_i 表示经过注意力模型后生成的特征图，与图1中的 M_i 对应：

$$M_i = M_i^* \otimes S, i \in \{0,1,2,3\} \tag{4}$$

为了保证汇总的分数图之间有足够的关联度，对于不同的分支，选择与该分支相邻(最接近)的两张分数图以及自己对应的分数图进行汇总，具体组合如表2所列。

表2 分数图组合表

Table 2 Score graph combination

特征图(M_i^*)	分数图(S_i)
M_0^*	S_0，S_1，S_2
M_1^*	S_0，S_1，S_2
M_2^*	S_1，S_2，S_3
M_3^*	S_1，S_2，S_3

最终生成的特征图能够更突出人群部分，削弱背景对计数的影响。在后续的实验部分将通过实验证明注意力模型的有效性。

4 实验分析

4.1 训练方法

4.1.1 Ground-truth密度图生成

由于所有人群数据集都以每个人头部中心的点注释形式给出，因此需要点云到密度图的转换。基于文献[2,20]的工作，我们采用相同的策略来获得Ground-truth密度图。具体来说，对于每幅图像中的头部位置的标记，假设 x_i 表示人头中心坐标位置，可以用 $\delta(x-x_i)$ 函数表示；一张有 N 个人头标注的人群图像可以表示为：

$$H(x) = \sum_{i=1}^{N} \delta(x - x_i) \tag{5}$$

将该标注图通过一个自适应高斯内核 G_{σ_i} 转化为连续密度函数：

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) \times G_{\sigma_i} \tag{6}$$

$$\sigma_i = \beta \bar{d}_i \tag{7}$$

$$\bar{d}^i = \frac{1}{k} \sum_{k}^{j} d_j^i \tag{8}$$

其中，N 为头部标记的总数；σ_i 为高斯标准差；β 是一个定值，在本文中取0.3；d_j^i 表示图像中该人头距离其邻近人头 k 的欧氏距离和的平均，也用于表示估计的人头大小。最后对得到的密度图进行积分，得到真实的人数。

4.1.2 损失函数

本文选择欧氏距离来度量Ground-truth密度图与本文方法生成的估计密度图之间的差异，损失函数如下：

$$L(\Theta) = \frac{1}{2N} \sum_{i=1}^{N} \| F(x_i;\Theta) - F_i \|_2^2 \tag{9}$$

其中，Θ 表示参数模型，$F(x_i;\Theta)$ 表示输出模型，x_i 表示第 i 个输入图像，F_i 表示真实密度图。

4.2 评价标准

一般通过平均绝对误差(MAE)和均方误差(Mean Square Error，MSE)来评价算法的性能，其定义分别如下：

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |C_i - C_i^{GT}| \tag{10}$$

$$MSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} |C_i - C_i^{GT}|^2} \tag{11}$$

其中，N 是测试集的图片数量；C_i^{GT} 是真实密度图的人数；C_i 是估计密度图的人数，其定义如下：

$$C_i = \sum_{l=1}^{L} \sum_{w=1}^{W} z(l,w) \tag{12}$$

其中，L 和 W 表示估计密度图的长和宽，$z(l,w)$ 是估计密度图中(l,w)位置处的像素。

4.3 评估和比较

本文在3个主要数据集上展示了本文模型的性能，并与之前的方法进行了比较。图4给出了本文模型的预测效果，第一行代表需要测试的图片，第二行代表测试图片对应的Ground-truth密度图，第三行代表通过本文模型生成的密度图，每一列图片下方的文字对应该图片的实际人数和预测人数。

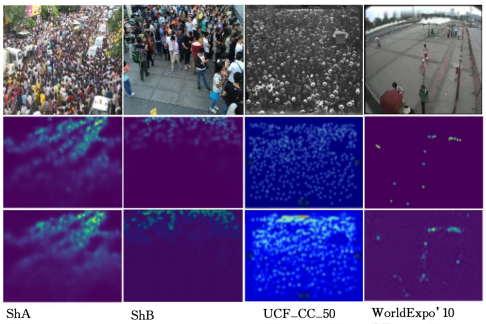


图 4 3 个主要数据集上的测试结果

Fig. 4 Test results on three main datasets

由于本文的跨列通信结构能够很好地提取并融合多尺度特征,得到了对人群特点更全面的描述,因此提高了人群计数的精度。而其结合注意力模型能够使网络更关注人群的头部信息,削弱背景的影响,进一步提升了网络的性能。从图 4 可知,本文模型生成的密度图与 Ground-truth 密度图非常接近,在人群极度密集的情况下预测结果与实际结果的差距也保持在 30 人以内,充分证明了本文方法的有效性。

4.3.1 ShanghaiTech 数据集

ShanghaiTech 人群统计数据集[2]包含 1 198 张带注释的图片,共计 330 165 人。该数据集由两部分组成,A 部分包含从互联网上随机下载的高度拥挤场景的 482 张图片,B 部分包含在上海街道拍摄的相对稀疏的人群场景的 716 张图片。与其他数据集相比,ShanghaiTech 数据集的大部分图像分辨率较低,因此我们保持其分辨率进行训练和测试。将本文方法与近期提出的 6 种方法进行了比较,结果如表 3 所列。与其他方法相比,本文方法在 A 部分获得了最佳的 *MAE* 和 *MSE*(最高的精度),在 B 部分也获得了相对不错的结果,表明在人口密度分布不均匀、人群中人头的尺度高度多样化的情况下,本文方法也可以相对准确地定位每个人的位置。

表 3 ShanghaiTech 数据集上的评估结果

Table 3 Evaluation results on ShanghaiTech dataset

NET	ShanghaiTechA		ShanghaiTechB	
	MAE	*MSE*	*MAE*	*MSE*
MCNN[2]	110.2	173.2	26.4	41.3
MSCNN[3]	83.8	127.4	17.7	30.2
MRA-CNN[27]	74.2	112.5	11.9	21.3
CSRNet[20]	68.2	115.0	10.6	16.0
PACNN[21]	66.3	106.4	**8.9**	**13.5**
CAT-CNN[28]	66.7	101.7	11.2	20.0
Ours	**63.2**	**97.3**	**8.9**	14.3

4.3.2 UCF_CC_50 数据集

UCF_CC_50 数据集[29]包括 50 幅角度和分辨率都不同的图像。每张图片上注释的人数从 94 到 4 543 不等,平均人数为 1 280,有限的图像数量使其成为颇具挑战的数据集。我们按照文献[29]中的标准设置进行 5 倍交叉验证,将所有图像以 10 张为一组分成 5 组。然后选择 4 组进行训练,剩下的 1 组进行测试。最后,取 5 组结果的平均值作为最终结果。*MAE* 和 *MSE* 的结果如表 4 所列。

表 4 UCF_CC_50 数据集上的评估结果

Table 4 Evaluation results on UCF_CC_50 dataset

NET	*MAE*	*MSE*
MCNN[2]	377.6	509.1
MSCNN[3]	363.7	468.4
MRA-CNN[27]	240.8	352.6
CSRNet[20]	266.1	397.5
PACNN[21]	267.9	357.8
CAT-CNN[28]	235.5	324.8
Ours	**222.1**	**293.6**

可以看出,本文方法在这个尺度变化极大的数据集上,相比其他方法达到了最好的效果,能够很好地适应不同规模的人群场景。

4.3.3 WorldExpo'10 数据集

WorldExpo'10 数据集[30]有 3 980 个注释帧,其来自由 108 个不同的监控摄像机捕获的 1 132 个视频序列。按照文献[30]中的标准协议,我们从 103 个场景中选取 3 380 个注释帧作为训练集,从剩下的 5 个场景中选取 600 帧作为测试集。该数据集的人群相对稀疏,在测试时,我们只测量给定感兴趣区域(RoI)下的人群数量。表 5 列出了每个场景的 *MAE* 及其平均值。本文方法在 3 个场景中均获得了最低的 *MAE*,整体也获得了最低的平均 *MAE*。实验结果表明,本文方法在处理稀疏和稠密人群时都具有很好的通用性和有效性。

表 5 WorldExpo'10 数据集上的评估结果

Table 5 Evaluation results on WorldExpo'10 dataset

Net	S1	S2	S3	S4	S5	Ave
MCNN[2]	3.4	20.6	12.9	13.0	3.7	11.6
CP-CNN[15]	2.9	14.7	10.5	10.4	5.8	8.8
CSRNet[20]	2.9	11.5	8.6	16.6	3.4	8.6
PACNN[21]	2.3	12.5	9.1	**11.2**	3.8	7.8
MRA-CNN[27]	2.4	11.4	9.3	10.5	3.7	7.5
CAT-CNN[28]	2.2	**9.8**	10.2	**11.2**	2.5	7.2
Ours	**1.3**	10.3	**8.0**	13.5	**2.4**	**7.1**

4.4 消融实验

本文在 ShanghaiTechA 数据集和 UCF_CC_50 数据集上进行了消融实验,以证实拥有空洞卷积的跨列通信结构和引入注意力模型的优势。

本文在 ShanghaiTechA 数据集上对比了纯粹的多列网络、单列结构结合空洞卷积的网络、多列结构结合空洞卷积的网络以及本文的跨列通信结构结合空洞卷积这 4 种方法的效果。从表 6 可以看出,在该数据集上将多列结构和空洞卷积结合之后,其性能有所提高,当加入跨列通信结构后,不同列之间的信息交流增加,该方法的性能得到了进一步的提高,计数效果显著提升。

表 6 ShanghaiTechA 数据集的消融实验

Table 6 Ablation experiments on ShanghaiTechA dataset

Model	*MAE*	*MSE*
MCNN[3]	110.2	173.2
CSRNet[20]	68.2	115.0
Ours-multi-columns	66.3	110.8
Ours	**63.2**	**97.3**

另外,为了证明注意力机制在本文方法中的有效性,我们

在 UCF_CC_50 数据集上进行了验证。在其他条件都保持不变的情况下，分别验证了不引入注意力模型以及分别叠加了 1 层、2 层、3 层、4 层注意力模型的结果，如表 7 所列，表中 Ours-attention 后的数字表示叠加注意力模型的层数，0 代表不引入注意力模型。从表 7 可以看出，通过使用单一注意力模型，本文网络的性能有一定的提高，其 *MAE* 和 *MSE* 比不使用注意力模型时分别降低了 18.8 和 28.2（Ours-atention1）。但是，单一的注意力模型不能利用更全面的信息，并且反向校准的力度较小。多层注意力模型叠加可以融合更丰富的信息，通过扩大像素对应的分数值，进一步强调头部的位置。然而一些得分相对较高的背景元素的分数也被同步扩大，反向校准时可能会降低人群与背景的分离程度，因此需要选择最佳叠加层数。实验结果表明，叠加 3 层注意力模型能够达到最好的效果，Ours-atention3 的 *MAE* 和 *MSE* 比不使用注意力模型时分别降低了 37.5 和 41.1。叠加 3 层注意力模型时，每张特征图选择要叠加的分数图的不同组合如表 8 所列。实验证明，该特征图选择与其相邻的两层特征图的分数图与自己的分数图叠加能够产生最好的效果。

表 7 UCF_CC_50 数据集的消融实验

Table 7 Ablation experiments on UCF_CC_50 dataset

Model	*MAE*	*MSE*
Ours-atention0	259.6	334.7
Ours-atention1	240.8	306.5
Ours-atention2	236.8	315.4
Ours-atention3	**222.1**	**293.6**
Ours-atention4	230.9	298.9

表 8 分数图组合实验

Table 8 Score figure combination experiment

序号		不同组合的结果			
1	特征图(M_i^*)	M_0^*	M_1^*	M_2^*	M_3^*
	分数图(S_i)	S_0,S_1,S_2	S_0,S_1,S_2	S_0,S_1,S_2	S_1,S_2,S_3
	MAE	231.4			
	MSE	301.1			
2	特征图(M_i^*)	M_0^*	M_1^*	M_2^*	M_3^*
	分数图(S_i)	S_0,S_1,S_2	S_1,S_2,S_3	S_1,S_2,S_3	S_1,S_2,S_3
	MAE	228.5			
	MSE	305.9			
3	特征图(M_i^*)	M_0^*	M_1^*	M_2^*	M_3^*
	分数图(S_i)	S_0,S_1,S_2	S_0,S_1,S_2	S_1,S_2,S_3	S_1,S_2,S_3
	MAE	**222.1**			
	MSE	**293.6**			
4	特征图(M_i^*)	M_0^*	M_1^*	M_2^*	M_3^*
	分数图(S_i)	S_0,S_1,S_3	S_0,S_1,S_3	S_0,S_2,S_3	S_0,S_2,S_3
	MAE	232.4			
	MSE	309.1			
5	特征图(M_i^*)	M_0^*	M_1^*	M_2^*	M_3^*
	分数图(S_i)	S_0,S_0,S_0	S_1,S_1,S_1	S_2,S_2,S_2	S_3,S_3,S_3
	MAE	230.0			
	MSE	299.3			

结束语 本文提出了一种基于跨列特征融合的人群计数方法，从增强跨列通信和引入注意力模型两个方面来处理人群尺度的巨大变化。该方法将跨列通信结构与注意力分数图相结合，增强了网络对于尺度变化的敏感性，突出了人群在图片中的位置，同时对引入的空洞卷积进行了改进，保证了训练过程中信息的连续性。与现有的方法相比，本文方法不需要对图片本身做任何处理，也不用引入透视图等其他特征，极大地提高了适用性。通过实验证明，CCFNet 在现有的主流数据集上都显示了其优越的性能，尤其对于尺度变化较大的图片有较高的计数准确率和鲁棒性。本文方法在设计过程中采用欧氏距离作为损失函数，最终生成的密度图仅仅是对实际概率密度图的一个粗略逼近，准确度很难保证。由于贝叶斯损失函数能够根据数据集提供的点进行标注期望值意义上的回归估计，在后续的研究中可以将其作为损失函数来提高密度图的准确度。

参 考 文 献

[1] SINDAGI V A, PATEL V M. A survey of recent advances in cnn-based single image crowd counting and density estimation [J]. Pattern Recognition Letters, 2018, 107: 3-16.

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

[3] ZENG L, XU X, CAI B, et al. Multi-scale convolutional neural networks for crowd counting[C]// 2017 IEEE International Conference on Image Processing(ICIP). IEEE, 2017: 465-469.

[4] JIANG X, XIAO Z, ZHANG B, et al. Crowd counting and density estimation by trellis encoder-decoder networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6133-6142.

[5] SINDAGI V A, PATEL V M. Generating high-quality crowd density maps using contextual pyramid cnns[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 1861-1870.

[6] WU B, NEVATIA R. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors[C]// Tenth IEEE International Conference on Computer Vision (ICCV'05). IEEE, 2005, 1: 90-97.

[7] WANG M, WANG X. Automatic adaptation of a generic pedestrian detector to a specific traffic scene[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2011: 3401-3408.

[8] DOLLAR P, WOJEK C, SCHIELE B, et al. Pedestrian detection: An evaluation of the state of the art[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 34(4): 743-761.

[9] VIOLA P, JONES M J. Robust real-time face detection[J]. International Journal of Computer Vision, 2004, 57(2): 137-154.

[10] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005: 886-893.

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

[12] CHEN K, LOY C C, GONG S, et al. Feature mining for localised

crowd counting[C]// British Machine Vision Conference. 2012:3.
[13] RYAN D, DENMAN S, FOOKES C, et al. Crowd counting using multiple local features[C]// 2009 Digital Image Computing: Techniques and Applications. IEEE, 2009:81-88.
[14] LEMPITSKY V, ZISSERMAN A. Learning to count objects in images[C]// Advances in Neural Information Processing Systems. 2010:1324-1332.
[15] WANG C, ZHANG H, YANG L, et al. Deep people counting in extremely dense crowds[C]// Proceedings of the 23rd ACM International Conference on Multimedia. 2015:1299-1302.
[16] BABU S D, SAJJAN N N, VENKATESH B R, et al. Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:3618-3626.
[17] CAO X, WANG Z, ZHAO Y, et al. Scale aggregation network for accurate and efficient crowd counting[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 734-750.
[18] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:1-9.
[19] SHEN Z, XU Y, NI B, et al. Crowd counting via adversarial cross-scale consistency pursuit[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5245-5254.
[20] LI Y, ZHANG X, CHEN D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:1091-1100.
[21] SHI M, YANG Z, XU C, et al. Revisiting perspective information for efficient crowd counting[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 7279-7288.
[22] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409. 1556, 2014.
[23] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.
[24] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[J]. arXiv: 1511. 07122, 2015.
[25] WANG P, CHEN P, YUAN Y, et al. Understanding convolution for semantic segmentation[C]// 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018: 1451-1460.
[26] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:7132-7141.
[27] ZHANG Y, ZHOU C, CHANG F, et al. Multi-resolution attention convolutional neural network for crowdcounting[J]. Neurocomputing, 2019, 329:144-152.
[28] CHEN J, SU W, WANG Z. Crowd counting with crowd attention convolutional neuralnetwork[J]. Neurocomputing, 2020, 382:210-220.
[29] IDREES H, SALEEMI I, SEIBERT C, et al. Multi-source multi-scale counting in extremely dense crowd images[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013:2547-2554.
[30] ZHANG C, LI H, WANG X, et al. Cross-scene crowd counting via deep convolutional neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:833-841.

LI Jia-qian, born in 1996, postgraduate. Her main research interests include computer vision and deep learning.

YAN Hua, born in 1971, Ph. D, professor. His main research interests include Intelligent information system and so on.