计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190400034

基于单列多尺度卷积神经网络的人群计数

彭 贤 彭玉旭 汤 强 宋砚琪
长沙理工大学计算机与通信工程学院 长沙 410000
(827815316@qq.com)

摘 要 单张图片和监控视频中的人群计数问题在近年来受到了越来越多的关注。尺度的变化和人群遮挡等问题,导致人群计数是一项十分具有挑战性的任务,但是深度卷积神经网络被证明能有效地解决这一问题。文中提出了一种单列多尺度的卷积神经网络,该网络提供了一种数据驱动的深度学习方法,能够理解各种不同的场景,并能进行精确的计数估计。该网络模型主要由作为二维特征提取的前端与中端,和用来还原密度图的后端组成。其中,使用堆叠池代替最大池化层,在不引入额外参数的前提下增加了模型的尺度不变性。网络模型前端采用部分 VGG-16 结构;中端采用 FME(特征聚合模块),用来打破不同列之间的独立,以更好地提取多尺度特征信息;后端采用 3 列 5 层的不同扩张率的空洞卷积,在保持分辨率不变的情况下增加感受野,生成更高质量的人群密度图,并引入一种相对人数损失,以提升稀疏密度人群情况下模型的性能。该模型在两个最具挑战性的人群计数数据集上都取得了很好的效果。实验结果表明,在公开人群计数数据集 ShanghaiTech 的两个子集和 UCF_CC_50 上,该方法的平均绝对误差(MAE)和均方误差(MSE)分别是 66.2 和 103.0、8.7 和 13.4、251.0 和 329.5,性能比传统人群计数方法更好。与其他模型相比,该模型拥有更高的精度和更好的鲁棒性,对稀疏人数图像有着更好的计数效果。

关键词:卷积神经网络;人群计数;堆叠池;空洞卷积;特征聚合;相对人数损失
中图法分类号 TP391

Crowd Counting Based on Single-column Multi-scale Convolutional Neural Network

PENG Xian, PENG Yu-xu, TANG Qiang and SONG Yan-qi
School of Computer and Communication Engineering, Changsha University of Science & Technology, Changsha 410000, China

Abstract The problem of crowd counting in single images and monitoring videos has received increasing attention in recent years. Due to the scale change and crowd occlusion, crowd counting is a very challenging problem, but deep convolutional neural network has been proved to be effective in solving this problem. In this paper, a single-column multi-scale convolutional neural network is proposed, which provides a data-driven deep learning method that can understand various scenarios and perform accurate counting and estimation. The proposed network model is mainly composed of the front end and the middle end, for two-dimensional features extraction, as well as the back end, which is used to restore the density map. Stack pools are used to replace the maximum pooling layer, and scale invariance of the model is increased without introducing additional parameters. Partial vgg-16 structure is adopted at the front end of the network model, and FME (feature aggregation module) is adopted in the middle to break the independence between different columns, to better extract multi-scale feature information. At the back end, three columns and five layers of cavity convolution with different expansion rates are adopted to increase the sensing field while keeping the resolution unchanged, generating a crowd density map with higher quality. A relative population loss is introduced to improve the model performance in the case of sparse population density. This model works well on two of the most challenging crowd counting data sets. The results show that on two subsets of ShanghaiTech and UCF_CC_50, the mean absolute error (MAE) and mean square error (MSE) of the proposed method are 66.2 and 103.0, 8.7 and 13.4, 251.0 and 329.5, respectively, achieving better performance than the traditional crowd counting methods. Compared with other models, the proposed model has higher accuracy, better robustness and better counting effect for images with sparse population.

Keywords Convolutional neural networks, Crowd counting, Stacked-pooling, Dilated convolution, Feature combination, Relative head loss

到稿日期:2019-04-05 返修日期:2019-07-13 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:湖南省教育厅优秀青年项目(18B162);长沙理工大学青年教师成长计划项目(2019QJCZ014)
This work was supported by the Research Foundation of Education Bureau of Hunan Province, China(18B162) and Young Teacher Development Foundation of Changsha University of Science & Technology(2019QJCZ014).
通信作者:彭玉旭(373836911@qq.com)

1　引言

人群计数就是计算出图像或者视频帧中的人头数目，而人群密度是人群在一定时间、一定空间内的分布情况。通过准确地估计体育馆、地铁站、火车站等公共场所的人群数目，可以有效地控制和管理人流量。因此，这种统计数据在公共安全和交通管制上有着极其重要的作用。同时，通过分析大型商场的人群密度分布，可以获得顾客的购买喜好与倾向，并发掘潜在的商业价值。诸多实际需求，使得单张图片和监控视频中的人群计数问题在近年来受到了越来越多的关注。但是，在现实场景中，与其他计算机视觉问题一样，人群计数也面临着许多挑战，譬如场景遮挡、高杂波、人群分布不均匀、光照不均匀、尺度视角的变化等，这些因素使得问题的处理变得极为困难。目前还没有行之有效的人群计数算法可以应用于现实场景中，因此准确、鲁棒的人群计数算法仍是计算机视觉领域重要的研究方向之一[1]。

近年来，深度卷积神经网络在人群计数方面被广泛使用。Zhang 等[2]提出了卷积神经网络(Convolutional Neural Networks，CNN)来交替学习人群密度和人群计数。Wang 等[3]直接使用基于 CNN 的模型将图像文件映射为图像实际的人群计数值。然而，这些基于单 CNN 的算法只能提取尺度相关的特征，难以解决人群图像尺度变化的问题。本文提出了一种单列多尺度的卷积神经网络，主要贡献有：

1)本文模型采用分段式结构，前端采用类似 VGG16 的结构对图像特征进行重映射，中间采用 FME 结构提取多尺度特征，后端采用空洞卷积，在不降低分辨率的情况下增加了感受野，提升了人群密度图的质量；

2)优化了网络的最大池化层，采用堆叠池，在提高网络尺度不变性的同时，降低了大量的运算成本，且不会在模型中引入额外的参数；

3)不同于其他模型采用 L2 正则化损失，本文在采用 L2 正则化损失的基础上，引入了一种在稀疏人群密度场景下可提升计数精度的函数。

本文的模型在 ShanghaiTech 数据集与 UCF_CC_50 数据集上都有很好的效果。

2　相关工作

人群计数一直是学术界和工业界的研究热点之一，相关算法已有许多。本文将人群计数方法分为基于检测的方法、基于回归的方法和基于卷积神经网络的方法 3 种。

2.1　基于检测的方法

基于检测的方法利用给定的视觉目标检测器对人群图像中的一般人进行检测和定位[4-6]，并通过对每个被检测的人进行累加得到计数结果。然而，这些方法[7-9]需要大量的计算资源，往往受限于人为遮挡和复杂背景的限制，鲁棒性和准确性都较低。

2.2　基于回归的方法

基于回归的方法直接从图像中回归人群数量。这类方法有两个主要组成部分：低级特征提取和回归建模。低级特征包括全局特征和局部特征，提取出这些全局特征和局部特征后，利用线性回归、分段线性回归以及岭回归、高斯过程回归等不同的回归技术来学习从低级特征到行人数量的映射关系。后续工作[10-11]提出了与人群相关的更多特征，包括基于分段的特征、基于结构的特征和局部纹理特征。Lempitsky 等[12]提出了一种基于密度的算法，通过对估计的密度图进行积分得到计数。密度图提供了比标量值更多的信息。

2.3　基于卷积神经网络的方法

卷积神经网络实现了端对端训练，抛弃了前景分割、人工设计和特征提取等繁琐步骤。经过多层卷积运算之后，自动学习得到高层语义特征，然后通过反向传播修正网络参数。Zhang 等[2]提出了一种基于 CNN 的方法来计算不同场景的人群，使用相似的训练数据来微调训练网络，再测试未训练图像。此方法在大多数现有数据集上获得了良好的性能。Zhang 等[13]针对任意人群密度和任意视角的图像提出了一种多列的网络结构——多列卷积神经网络(Multi-column CNN，MCNN)。该方法构建的网络由 3 列不同的回归器组成，分别对应大、中、小 3 种不同感受野的卷积核，从而保证了对较大尺度变化的鲁棒性。Zeng 等[14]提出了一种多尺度 blob 算子，并用 blob 算子代替卷积核，在 blob 算子中采用不同感受野的卷积核来对密度图进行卷积，提取不同尺度的特征，再将这些特征融合起来。Li 等[15]基于空洞卷积提出了一种前后端网络，前端采用去除全连接层的 VGG16 网络，后端采用空洞卷积提升密度图的质量。Cao 等[16]利用类似于 Inception 架构的模块，该模块采用多个不同大小的卷积核来提取多尺度特征，最后通过反卷积来提升密度图的质量。

2.4　基于单列多尺度卷积神经网络的方法

单列的 CNN 只能提取尺度相关的特征，难以解决人群图像尺度变化的问题。多列网络架构在处理一幅图像中的头部尺度变化时，采用不同的列来对应不同的过滤器大小。但是，多列网络比单列网络的运算量大得多。本文设计改进了一种单列多尺度的卷积神经网络，从原始图像中学习相应尺度的密度图；采用分段式结构，前端采用类似 VGG16 的结构对图像特征进行重映射，中间采用 4 层 FME 结构(特征聚合模块)提取多尺度特征，后端采用 3 列 5 层的空洞卷积，空洞卷积的卷积核大小与扩张率都不同，其可以在不降低分辨率的情况下增加感受野，提升人群密度图的质量；使用 Stacked-pooling[17]代替最大池化层，在不引入额外参数的前提下，加强了模型的尺度不变性。通过加入相对人数损失，本文模型提升了绝对人数稀疏情况下的预测效果。

3　多尺度卷积神经网络用于人群计数

3.1　密度图的生成

基于 Zhang 等[13]的工作，本文直接在输入图像上估计人群密度图。在实际应用中，由于遮挡，几乎不可能准确地得到头部的大小，而且很难找到头部大小与密度图之间的潜在关系。但是，在拥挤的场景中，头部的大小通常与相邻两个人中心的距离有关，因此可以根据每个人与其临近的人的平均距离来自适应地确定每个人的传播参数。自适应高斯核可以不需要知道输入图像的透视图，能准确地计算出真实密度图，同时可以在高遮挡的情况下得到效果较好的人群密度图，保存更多的信息，提高模型的精度。因此，为了生成高质量的人群

密度图，采用尺度自适应高斯核是目前的最佳选择。首先，对图片中所有人的头部位置进行标注，并保存头部位置坐标；然后，将带有头部位置坐标的图片转换成人群密度图，这个人群密度图就是本文训练时需要的数据标签。如果有一个头部位置在像素点 x_i，将其表示为 $\delta(x-x_i)$，则可将有 N 个人的头部位置标记的图像表示为：

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

将式(1)与高斯核 G_δ 进行卷积，从而得到密度图 $F(x)=H(x)*G_\delta(x)$。

每个 x_i 都是 3D 场景中一个人群密度的样本，但是因为透视失真的影响，x_i 的像素跟随尺度的变化而变动。因此，为了准确地估计人群密度，需要考虑地平面与图像平面之间的透视失真。假设以头部位置代替人的位置，人群是均匀分布的，根据每人和其临近的 k 个人之间的平均距离可以得出几何失真的合理估计；而拥挤场景中几乎不可能准确地获得被遮挡的头部尺寸，只能根据与其临近的人的平均距离数据来自动确定每个人的传播参数 δ。对于给定图像中的每个头部位置，其到 k 个最近邻的人的距离为 $\{d_1^i,d_2^i,\cdots,d_k^i\}$，因此平均距离可表示为：

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

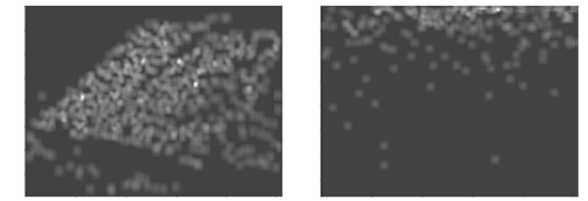
则密度 F 可表示为：

$$F(x)=\sum_{i=1}^{M}\delta(x-x_i)*G_{\sigma_i},\ \text{with}\ \sigma_i=\beta\bar{d}_i \tag{3}$$

其中，M 是图像中头部注释的总数，通过实验数据得 $\beta=0.3$ 时取得的效果最好。

图 1 为原始图像及其对应的人群密度图。通过这样的高斯卷积生成的密度图上的所有像素值之和是与原始的标注图一致的。因此，人群计数问题被转化为：先估计得到其对应的密度图，然后把获得的密度图上的每个像素的数值相加，从而得到最终的人头数。

(a)原始图像



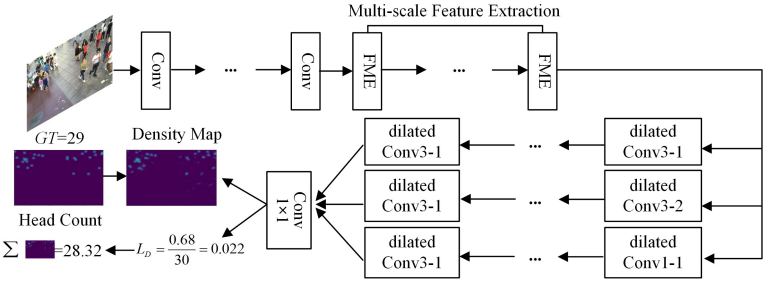
(b)人群密度图

图 1 人群图像及其密度图

Fig. 1 Crowd images and ground truth density maps

3.2 模型架构

人群图像一般由多个不同大小的经过透视失真的人物像素组成。单列的网络架构很难通过相同大小的内核组合来提取多尺度的组合特征，而多列的网络架构往往又比单列架构的运算量大很多。基于此，本文设计了一个多尺度卷积神经网络来从原始图像中学习尺度相关的密度图。通过在单列的网络架构中使用多尺度的 FME 模块来提取各种尺度的特征信息，再将其聚合在下一个阶段进行统一计算[18]。此模块分支分别提取大、中、小 3 种不同尺度的特征信息以及原始信息，再通过拼接操作将 4 个通道的特征信息融在一起。这样的架构既能提取多尺度信息，又能极大地降低计算量。本文网络架构如图 2 所示，分为前端、中端、后端 3 个部分，分别用于特征映射、多尺度特征提取、密度图回归。



注：FME 为特征聚合模块；Dilated Conv 为空洞卷积，其中 3-2 代表卷积核大小为 3 * 3，扩张率为 2；Density Map 为密度图；GT 为真值；L_D 为相对人数损失

图 2 用于人群计数的单列多尺度卷积神经网络

Fig. 2 Single-column multi-scale convolutional neural network for crowd counting

VGG 网络一文中提到网络的深度是算法优良性能的关键部分[19]。2 个 3 * 3 的堆叠卷积层的有限感受野是 5 * 5，3 个 3 * 3 的堆叠卷积层的感受野是 7 * 7，因此可以通过小尺寸卷积层的堆叠替代大尺寸卷积层，并且感受野大小不变。本文在不影响感受野的情况下，采用 3 * 3 的卷积核堆叠，以加深网络深度，增加模型的非线性性。多个 3 * 3 的卷积层拥有的参数比一个大尺寸的 filte 更少，假设卷积层的输入和输出的特征图大小同为 C，那么 3 个 3 * 3 的卷积层参数的个数为 $3*(3*3*C*C)=27C^2$，而一个 7 * 7 的卷积层参数为 $49C^2$，因此可以把 3 个 3 * 3 卷积核看成一个 7 * 7 的卷积核的分解(中间层有非线性的分解，并且起到隐式正则化的作用)。这种替换可以使得模型的参数更少，并且网络单层神经元个数足够多，生成的特征抽象程度足够高，提高了网络的精确度与运行速度，同时拥有更少的网络参数。网络的前 4 个卷积层是 3 * 3 的卷积核，通过其对图像特征进行重映射。为了保证模型的非线性特性，每个卷积层之后都加上一个非线性激活函数层，这里使用的是修正线性单元(Rectified Linear Units，ReLU)，ReLU 可以加速网络的收敛[20]。

FME 模块是一个类似 Inception 模型的结构，用来提取不同尺度规模的特征。Stacked-pooling 是一种堆叠池，一种

特殊的池化层，对特征进行压缩并提取主要特征。多层空洞卷积可以在不影响分辨率的情况下增加感受野，通过空洞卷积可以得到质量更高的密度图，并在每个卷积层后都添加 ReLU 作为激活函数。由于密度图的值始终为正，需要在最后一层卷积层后添加 ReLU 激活函数以加强密度图的回归。表 1 列出了详细的参数设置，其中所有卷积层都通过填充保持原来的大小。卷积层的参数表示为“conv(kernel size)-(filter number)-(dilation rate)”，堆叠池的参数表示为内核集。

表 1　单列多尺度 CNN

Table 1　Single-column multi-scale CNN

SCM-CNN		
Input		
Conv3-64-1		
Conv3-64-1		
Stacked-pooling(2,2,3)		
Conv3-128-1		
Conv3-128-1		
FME-128-1		
FME-128-1		
Stacked-pooling(2,2,3)		
FME-256-1		
FME-256-1		
Conv3-256-1	Conv3-256-2	Conv1-256-1
Conv3-256-1	Conv3-256-2	Conv1-256-1
Conv3-128-1	Conv3-128-2	Conv1-128-1
Conv3-64-1	Conv3-64-2	Conv1-64-1
Conv3-32-1	Conv3-32-2	Conv1-32-1
Conv1-1-1		

3.3　特征聚合模块

由于存在透视失真，图像通常包含大量不同大小的头部，因此具有相同大小感受野的滤波器不太可能捕捉到不同尺度下人群密度的特征。为了更精确地估算不同图像的人群密度，需要利用不同尺度进行整合计算。特征聚合模块(Feature Map Encoder，FME)通过拼接操作来打破不同列之间的独立，如图 3 所示。

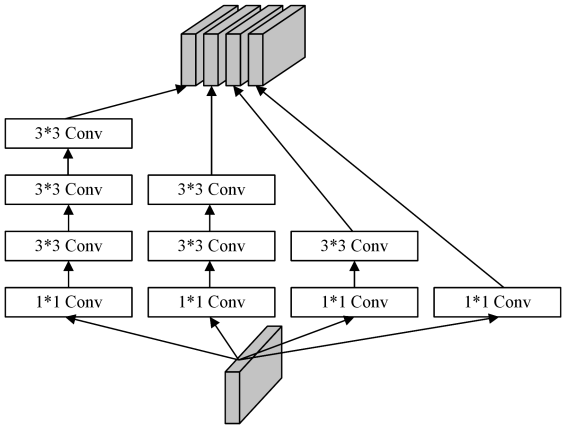


图 3　特征聚合模块

Fig. 3　Feature map encoder

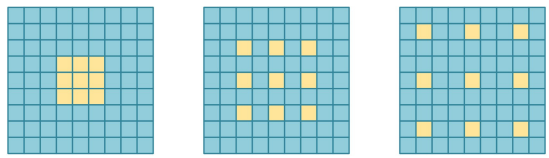
FME 模块可以扩展到任意数量的分支，本文采用 4 个分支，但只采用 1 * 1 与 3 * 3 的卷积核。此模块的第一个分支只采用 1 * 1 的卷积核，是为了保留上一层的特征尺度来覆盖小目标；其余 3 个分支都采用了 3 * 3 卷积核的堆叠来模仿大卷核的感受野(3 * 3，5 * 5，7 * 7)。但是，它们前面增加了一层 1 * 1 的卷积来对特征进行降维，使得新特征图的特征表达更佳。FME 模块的 4 个分支对应 4 种不同的感受野，可以学习 4 种不同尺度的特征信息，并且采用 3 * 3 卷积核的堆叠实现大感受野，大幅度地减少了参数量。同时，为了简略起见，每一个分支的通道数都设置为相等的，并在每个卷积核后增加了一个 ReLU。直观地来说，FME 模块就是一个不同大小感受野的集成，此模块可以捕捉到密集人群中人群的多尺度外观，有利于人群计数。

3.4　空洞卷积

本文模型中的一个关键组件就是空洞卷积层。2-D 空洞卷积的定义如下：

$$y(m,n)=\sum_{i=1}^{M}\sum_{j=1}^{N}x(m+r\times i,n+r\times j)w(i,j) \tag{4}$$

其中，$y(m,n)$是 $x(m,n)$作为输入时，加上一个过滤器 $w(i,j)$，分别以 m 和 n 为长和宽进行空洞卷积时的输出；参数 r 为扩张率，如果 $r=1$，则空洞的卷积就是普通的卷积。空洞卷积是池化层的一个很好的替代，在分割任务中使准确率有了显著的提高[21-23]。虽然池化层(如最大池化层)被广泛用于保持不变性和控制过拟合，但它们极大地降低了空间分辨率，这意味丢失了着特征图的空间信息。反卷层[24-25]可以减少信息的丢失，但是额外的复杂性和执行延迟可能并不适用于所有的情况。空洞卷积是更好的选择，它使用稀疏内核(如图 4 所示)来交替池化和卷积层。在保持特征图分辨率方面，与卷积＋池化＋反卷积的方案相比，空洞卷积具有明显的优势；且实验证实，当扩张率为 2 时其取得的效果最好。



(a)扩张率为 1　(b)扩张率为 2　(c)扩张率为 3

图 4　内核大小为 3 * 3 时不同扩张率的空洞卷积

Fig. 4　Dilated convolution with different expansion rate when kernel size is 3 * 3

3.5　堆叠池

在深度 CNN 中，通常使用 $k=2$ 等小池内核，因为较大的池内核可能会过度丢弃原始特征图信息。然而，不同大小步长的池内核能够为 CNN 提供不同范围的尺度不变性，在人群计数中的体现就是不同尺度的图像区域通常具有较高的视觉相似性。因此，本文采用了一组具有不同大小步长的内核，用以提高网络模型的尺度不变性。

堆叠池其实就是池化层的堆叠，除一个池内核外，它的池操作是在向下采样的特征映射上计算的，其中中间的特征映射连续计算为：

$$\downarrow_{s_i'}Y_i'=Y_{i-1}'*\rho_{k_i}^{(s_i')} \tag{5}$$

具体地，$Y_0'=X$ 是输入特征映射。核大小 k_i'对应于 k_i 的某个变换。步长 $s_{i=1}'=s$，$s_{i>1}'=1$。根据式(6)，可以得到堆叠池的输出连接中间特征映射。

$$\downarrow_s Y_{\text{stacked}}=\frac{1}{n}\sum_{i=1}^{n}\downarrow_{s_i'}Y_i' \tag{6}$$

图 5 展示了堆叠池的一个示例，其中内核集 $K=\{2,2,3\}$，步长 $S=\{2,1,1\}$。实验显示，这种配置在本文模型中起到了很好的效果。

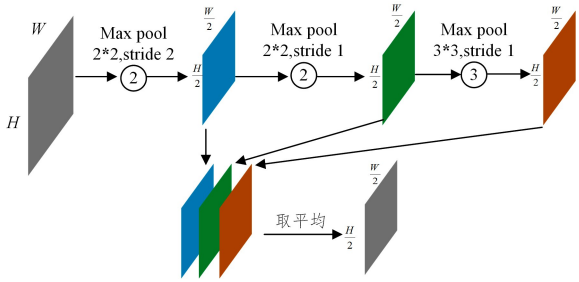


图 5 内核为{2,2,3}的堆叠池

Fig. 5 Stacked-pooling with {2,2,3} kernel

3.6 损失函数

本文模型训练首先采用欧几里得损失来测量估计的密度图与真值之间的差值。

$$L(\Theta)=\frac{1}{2N}\sum_{i=1}^{N}\| F(X_i;\Theta)-F_i \|_2^2 \tag{7}$$

其中，Θ 表示参数模型，$F(X_i;\Theta)$ 代表输出模型，X_i 和 F_i 分别为第 i 个输入图像和真值密度图。

但是，大多数代表性的方法在行人稀少的人群场景中表现不佳。这个问题不能通过式(7)这个损失函数来解决。因为相对于密集人群，稀疏人群的绝对行人数量通常不是很大[26]。为了解决这个问题，本文引入一个相对人数损失：

$$L_D(\Theta)=\frac{1}{N}\left\| \frac{F_D(X_i;\Theta)-D_i}{D_i+1} \right\|_2^2 \tag{8}$$

其中，$F_D(X_i;\Theta)$ 代表预测的人数，D_i 表示图像中实际的人数，分母为 D_i+1 是为了防止分母为零。此损失函数将集中学习预测误差较大的样本，当绝对人数非常稀疏时在网络中使用相对人数损失，使结果得到了显著的提升。

相对人数损失，是真值与预测结果的直接差值与真值做比。在模型初步收敛后，加入相对人数损失再进行训练。在人数非常稀疏时，相对人数损失的值会比较高，可以加强模型对稀疏人数场景的拟合效果。在常规或高密度人数场景时，相对人数损失的值会特别小，基本不会影响模型的性能。同时，相对人数损失在训练时的权重占比也很小。综合这两点，相对人数损失对模型具备优化效果，却不影响其他场景下的性能。

训练过程中首先采用欧氏损失将模型训练到收敛，再加入相对人数损失进行再一步的训练。其训练权重为 $L=\alpha_1 * L(\Theta)+\alpha_2 * L_D(\Theta)$。

4 实验

本节在 ShanghaiTech 数据集上对单列多尺度卷积神经网络进行评估。所有的卷积神经网络都是基于 Tensorflow 进行训练的。实验结果表明，该方法在稀疏密度条件下计数的精度和鲁棒性均较好。

4.1 评价标准

采用平均绝对误差(MAE)和平均均方根误差(MSE)来评价算法的性能。MAE 和 MSE 的定义如下：

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i| \tag{9}$$

$$MSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} \tag{10}$$

其中，N 是测试集的图像数目；y_i 和 $\hat{y}_i$ 分别表示对于第 i 张测试图像，标注图的人头数和预测得到的人头数。

MAE 用于衡量人群计数算法的准确率，MSE 用来评价算法的鲁棒性。

4.2 ShanghaiTech 数据集

ShanghaiTech 数据集是一个大规模人群计数数据集，其中包含 1198 张带标注的图片，共 330165 人。数据集由两部分组成：A 部分是从互联网上抓取的 482 张图片，B 部分是通过摄像头从街道上抓取的 716 张图片。它们都被分为包含 300 幅图像的训练集和一个包含其余图像的测试集。

4.2.1 模型训练

为了保证有足够数量的数据进行模型训练，本文从每个图像中裁剪 9 块并翻转它们来执行数据增强。将 9 个裁剪点固定为顶部、中部和底部，并结合左侧、中部和右侧，每块都是原始大小的 90%。网络整体设置类似于 VGG，但是中端和后端用 FME 和空洞卷积代替，设置如表 1 所列。

在模型优化中，本文使用了 RMSprop 优化和指数衰减法，初始学习率设为 1×10^{-4}，衰减系数为 0.9，衰减速度为 20。相对人数损失参数设置为 $L=1.0*L(\Theta)+0.1*L_D(\Theta)$。

4.2.2 实验结果

将本文方法与现有的 5 种在 ShanghaiTech 数据集上测试过的方法进行对比。Zhang 等[2]设计了一种卷积神经网络，从原始图像中回归密度图，并进行人群计数。Zhang 等[13]提出了一种 MCNN，采用多列不同尺度的卷积网络从原始图像中回归人群密度图并计算人群计数值。Zeng 等[14]提出了一种采用多尺度 blob 的卷积网络(Multi-Scale CNN，MSCNN)。Li 等[15]提出了一种前后端网络，前端采用 VGG 结构，后端采用空洞卷积以提升密度图质量。Cao 等[16]利用类似于 Inception 架构的模块，该模块采用多个不同大小的卷积核提取多尺度特征，最后通过反卷积提升密度图质量。

表 2 的结果表明，本文模型在 PartA 与 PartB 两个部分的测试效果都很好，在 ShanghaiTech 数据集上实现了最先进的性能；在人群密集与人群稀疏场景的精准性和鲁棒性都表现良好。

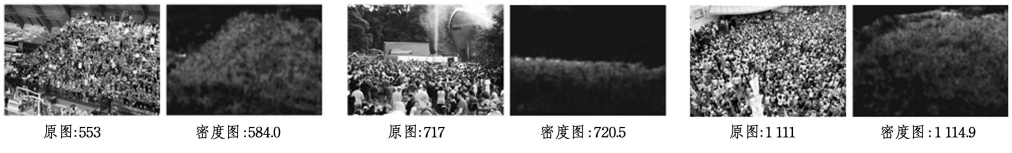
表 2 各方法在 ShanghaiTech 数据集上的性能

Table 2 Performance of each method in ShanghaiTech dataset

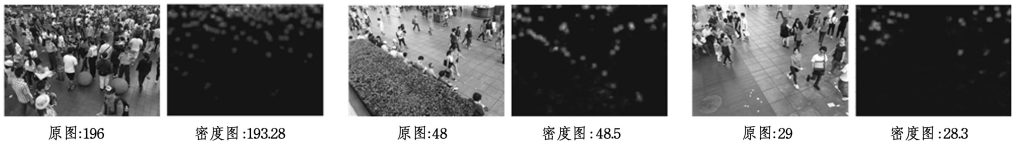
Method	Part_A		Part_B	
	MAE	*MSE*	*MAE*	*MSE*
Zhang et al.[2]	181.8	277.7	32.0	49.8
MCNN[13]	110.2	173.2	26.4	41.3
MSCNN[14]	83.8	127.4	17.7	30.2
CSRNet[15]	68.2	115.0	10.6	16.0
SANet[16]	67.0	104.5	**8.4**	13.6
SCM-CNN(ours)	**66.2**	**103.0**	8.7	**13.4**

图 6 给出了本文模型在 PartA 与 PartB 两个部分上的实

际预测效果。真值和估计的人数值分为位于原始图像和相应的密度图下。可以看出，本文模型在稀疏人数的场景下也有很好的表现，在只有二十几人场景下的误差也不足 1 人；而其余模型在人群稀疏图像上的相对误差就比较大。



(a)PartA



(b)PartB

图 6　在 ShanghaiTech 数据集上的测试结果

Fig. 6　Test results on the ShanghaiTech dataset

4.3　UCF_CC_50 数据集

UCF_CC_50 数据集[26]一共包括 50 张不同的灰度图片。这一数据集最大的特点是人群数变化很大，从 94 人到 4 543 人不等，平均每张图上有 1 280 人。

4.3.1　模型训练

在训练前，将 UCF_CC_50 数据集进行五重交叉验证，每个部分中包含 10 张图，每张图裁剪为 9 块，同时对其相应的人群密度图也进行同样的操作。这种数据增广操作可以极大地提升训练集的规模。训练期间，模型初始化与 ShanghaiTech 数据集上的实验几乎一样，除了学习速率固定为 1×10^{-6} 以保证模型收敛，相对人数损失参数设置为 $L=0.9*L(\Theta)+0.1*L_D(\Theta)$。

4.3.2　实验结果

将本文方法与现有的 7 种在 UCF_CC_50 数据集上测试过的方法进行对比，其具体结果如表 3 所列。Lempitsky 等[12]与 Rodriguez 等[27]都是通过手工提取特征从输入图片中还原密度图；其余 5 种都是基于 CNN 的方法，利用单列/多列网络对 UCF_CC_50 数据集进行测试。由表 3 可以得出，本文方法在 UCF_CC_50 数据集上仍然有着很好的性能，与其他方法相比，本文方法的精准性和鲁棒性都更高。

表 3　各方法在 UCF_CC_50 数据集上的性能

Table 3　Performance of each method on UCF_CC_50 dataset

Method	MAE	MSE
Rodriguez et al.[27]	665.7	697.8
Lempitsky et al.[12]	493.4	487.1
Zhang et al.[2]	467.0	541.6
MCNN[13]	377.6	509.1
MSCNN[14]	363.7	468.4
CSRNet[15]	266.1	397.5
SANet[16]	258.4	334.9
SCM-CNN(ours)	**251.0**	**329.5**

结束语　本文提出了一种基于多尺度全卷积网络的方法来估计静止图像的人群密度，通过将估计的人群密度图进行简单求和，得到最终对人群数量的估计值。与一般基于卷积神经网络计算不同场景下人群数量的方法相比，本文方法不需要在训练场景和测试场景上使用透视图，有效解决了实际应用中透视图难以获得的问题，极大地提高了适用性。实验结果表明，使用单列多尺度全卷积网络的组合可以有效地解决对不同人群规模以及高密度人群计数困难的问题，同时加入新的损失函数及池化层，对绝对人数较少的图像也有很好的鲁棒性。本模型在人数密集时也有高准确率和鲁棒性。

本文方法是在 ShanghaiTech 数据集与 UCF_CC_50 数据集上进行训练的，数据集图片都较少，未来可以考虑引入一个辅佐训练集来提升训练集的大小，以进一步提高人群计数的准确性。

参 考 文 献

[1] QU J, SHI Z L, YE Y D. Unbalanced crowd density estimation based on convolutional features[J]. Computer Science, 2018, 45(8): 236-241.

[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] 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. ACM, 2015: 1299-1302.

[4] LIN S F, CHEN J Y, CHAO H X. Estimation of number of people in crowded scenes using perspective transformation[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2001, 31(6): 645-654.

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

[6] WANG M, WANG X. Automatic adaptation of a generic pedestrian detector to a specific traffic scene[C]// CVPR 2011. IEEE, 2011: 3401-3408

[7] GE W, COLLINS R T. Marked point processes for crowd counting[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 2913-2920.

[8] 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.

[9] LIN Z, DAVIS L S. Shape-based human detection and segmentation via hierarchical part-template matching[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(4): 604-618.

[10] CHAN A B, VASCONCELOS N. Bayesian poisson regression for crowd counting[C]// 2009 IEEE 12th International Conference on Computer Vision. IEEE, 2009: 545-551.

[11] CHEN K, LOY C C, GONG S, et al. Feature mining for localised crowd counting[C]// BMVC. 2012: 3.

[12] LEMPITSKY V, ZISSERMAN A. Learning to count objects in images[C]// Advances in Neural Information Processing Systems. 2010: 1324-1332.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] HUANG S, LI X, CHENG Z Q, et al. Stacked pooling: Improving crowd counting by boosting scale invariance[J]. arXiv: 1808.07456, 2018

[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] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409.1556, 2014.

[20] NAIR V, HINTON G E. Rectified linear units improve restricted boltzmann machines[C]// Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010: 807-814.

[21] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[J]. arXiv: 1511.07122, 2015.

[22] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.

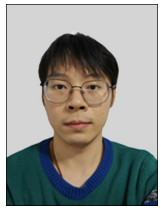
[23] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv: 1706.05587, 2017.

[24] ZEILER M D, KRISHNAN D, TAYLOR G W, et al. Deconvolutional networks[C]// 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010: 2528-2535.

[25] NOH H, HONG S, HAN B. Learning deconvolution network for semantic segmentation[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1520-1528.

[26] ZHANG L, SHI M, CHEN Q. Crowd counting via scale-adaptive convolutional neural network[C]// 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018: 1113-1121.

[27] RODRIGUEZ M, LAPTEV I, SIVIC J, et al. Density-aware person detection and tracking in crowds[C]// 2011 International Conference on Computer Vision. IEEE, 2011: 2423-2430.



PENG Xian, born in 1994, master. His main research area is deep learning.

PENG Yu-xu, born in 1977, Ph.D, associate professor, CCF member, mainly focuses on signal and information processing.