激光与光电子学进展
Laser & Optoelectronics Progress
ISSN 1006-4125,CN 31-1690/TN

《激光与光电子学进展》网络首发论文

题目：单列深度时空卷积神经网络的人群计数研究
作者：鱼春燕，徐岩，缑丽莎，南哲锋
收稿日期：2020-07-28
网络首发日期：2020-10-21
引用格式：鱼春燕，徐岩，缑丽莎，南哲锋．单列深度时空卷积神经网络的人群计数研究[J/OL]．激光与光电子学进展．
https://kns.cnki.net/kcms/detail/31.1690.TN.20201020.1311.004.html

网络首发时间：2020-10-21 11:31:55
网络首发地址：https://kns.cnki.net/kcms/detail/31.1690.TN.20201020.1311.004.html

单列深度时空卷积神经网络的人群计数研究

鱼春燕，徐岩*，缑丽莎，南哲锋

兰州交通大学电子与信息工程学院，甘肃兰州 730070

摘 要 突发性的人群聚集会给人们的人身安全带来隐患，对高风险区域有效的进行人群计数很有必要。针对多列神经网络臃肿的网络结构、产生冗余的信息及耗时长导致网络难训练的问题，提出了一种基于单列深度时空的卷积神经网络对人群进行计数，并对模型加以改进来满足视频图像计数的需要。通过在 FCN 网络结构中加入空洞卷积和跳级连接特征融合来提高网络提取特征的能力；同时在长短期记忆网络结构中加入空间变换(ST)模块来减少视频监控产生的角度畸变对计数准确性的影响，为使网络计数结果更精确将改进的 FCN 网络和关联时序的 LSTM 网络用残差连接方式进行连接，最后在 UCSD、Mall 和自建人群数据集上分别进行测试，结果表明所提算法的人群计数准确性和鲁棒性相比其他算法更优。

关键词 神经网络；人群计数；深度时空网络；空洞卷积；空间变换

中图分类号 TP391 **文献标志码** A

Research on Crowd Counting of Single Column Deep Space-Time Convolutional Neural Network

Yu Chunyan, Xu Yan*, Gou Lisha, Nan Zhefeng

School of Electronic and Information Engineering, Lanzhou Jiaotong University，

Lan zhou，Gansu 730070，China

Abstract Sudden crowd gathering will bring hidden dangers to people's personal safety. It is necessary to count people effectively in high-risk areas. Based on the bloated network structure of multi-column neural networks, the generation of redundant information, and the long time-consuming problem that makes the network difficult to train,Based on single-column depth space-time, an improved convolutional neural network is proposed to count the crowd and meet the needs of video image counting. Improve the network's ability to extract features by adding dilated convolution and hop-level connection to the FCN network structure; At the same time, a space transform (ST) module was added to the LSTM structure to reduce the influence of Angle distortion on counting accuracy caused by video surveillance. In order to make the network count results more accurate, the improved FCN network and the associated timing LSTM network are connected by residual connection.The test on the crowd counting video data set in UCSD、Mall and Self-built population data set showed that the accuracy and robustness of the proposed algorithm were better than other algorithms.

Key words neural networks;crowd counting;deep space-time network;dilated convolution; spatial transfornation

OCIS codes 100.4996；100.3008；100.2000

1 引 言

*E-mail：xuyan@mail.lzjtu.cn

现今监控设备的快速发展造成了图像和视频的大量增长，对分析视频内容有极大的需求。由于人群计数在视频监控、交通管控、应急管理等方面的潜在影响，使其在计算机视觉中得到了广泛的研究与应用。随着科技的创新和城市交通的部署，地铁的方便和快捷，已经成为了大部分人出门的首选交通工具。在出行高峰期，不可预料的人群数量会引发不可避免的人为事故，不满足智慧交通的要求同时提高了社会不安全因素，在地铁监控系统引入人群计数对安全管控很有必要，以往对人群计数的研究大多集中在基于单个图片上[1-3]，很难满足视频监控的需求。

目前的人群计数方法大致可分为基于检测方法、基于回归方法、基于密度估计方法以及基于深度学习方法等四类。基于检测的方法通过检测和跟踪方式对视频图像中的人进行计数[4-5]，检测器对单独的个人或身体部位进行逐个检测并累计叠加得到统计结果[6]，这种方法虽计数精度较高，但在密集和复杂场景下通常无法检测出实际场景中常见的小物体和被遮挡的头、身体，同时存在耗时长的缺点。基于回归的方法以人群为整体估计人群密度，能够实现大规模人群计数[7]，学习局部图像块中的特征和对应人数之间的映射[8]，回归计数的优点简化了个体检测的复杂任务。但训练需要大量标记数据使算法难度较高，不能定位人群空间位置关系。基于密度估计方法不仅解决了检测和特征回归的局限性，还能够估计图像中任何区域的目标数量，将计数问题转为估计图像密度，通过图像密度在该图像区域上的积分得到对象的数量[9]。随着卷积网络的发展深度学习在人群计数领域的应用提高了计数性能，基于深度学习方法根据网络特性的差异可分为四类：基于卷积神经网络、尺度感知模型、上下文感知模型和多任务模型。2017 年 Zhang 等[10]人提出将 FCN(Fully Convolutional Networks)网络结合 LSTM(Long Short-Term Memory networks)循环网络进行视频的车辆计数并取得有效进展，同年 Feng Xiong 等[3]从视频图像在时间上存在关联的角度，提出了一种充分利用视频序列时间信息的卷积神经网络与长短期神经网络相融合 ConvLSTM 人群密度估计模型，2018 年 Li 等[11]人将空洞卷积融入 CSRNet 模型，利用该网络能够提取高层次语义的特点准确的对密集场景的人群分布情况进行估计，虽减少模型参数但丢失很多细节信息[12]。

纵观现有人群计数的研究方法，很少有建立视频序列帧中人群数量的时间相关性模型。网络模型对视频图像中的人群特征提取较少，针对摄像机视角畸变导致无法识别物体的问题并无明显改善。对上述存在的问题，本文提出了一种基于单列深度时空的卷积神经网络人群计数模型。模型前端的密度生成模块采用单列 FCN 深度网络融入空洞卷积和跳级连接的方式，能够对细节特征进行提取，提高生成密度图的质量；模型后端的时空变换计数模块将长短期记忆网络结构和空间变换模块进行结合，改善网络对摄像机视角畸变造成物体难以识别的问题。

2 人群计数算法

2.1 基于单列深度时空计数的卷积神经网络

在监控场景下不同的人流分布、光照变化、人群遮挡以及拍摄角度畸变情况下，能够有效的计算人群数量仍然是一个极大的挑战。尽管多列网络已经取得了很大的进展，但其不同分支、相似的结构会产生大量冗余信息，使网络计算耗时长、难以训练，无法提升最终生成密度图的质量。

基于以上存在的问题，所以本文提出采用以下改进的基于单列深度时空计数的卷积神经网络，密度图生成主网络 FCN 将传统网络进行卷积化[13]，加入空洞卷积和跳跃连接提高网络提取特征的能力；时空变换部分将 LSTM 长短期记忆网络和空间变换模块引入实现图像时序关联计数，模型结构如下图 1 所示。

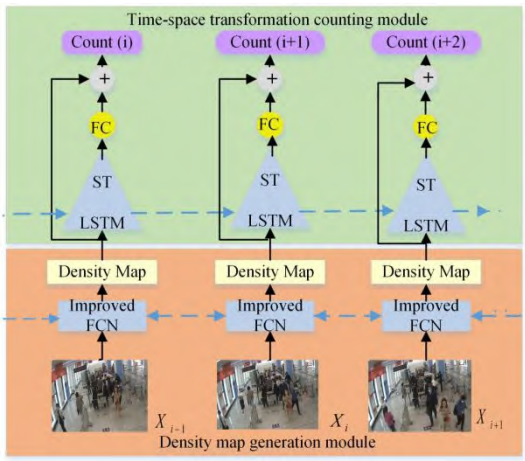

图 1 单列深时空计数的卷积神经网络

Fig.1 Convolutional neural network for single-column deep space-time counting

整个网络架构的实现主要分为两个模块，一个是密度图生成模块，这个模块主要由改进的 FCN 网络组成，另一个模块主要用 LSTM 实现的时空变换计数模块。总体网络结构通过在残差学习框架中将改进 FCN 和 LSTM 相结合来根据时序估计人群数量，改进 FCN 网络主要将像素级特征映射到人群密度中，在 LSTM 网络中插入 ST(Spatial Transformer)模块，ST 能够进行端到端的训练，将其合并到卷积神经网络并使用标准的反向传播算法进行训练，参数矩阵能确定分区的位置，以及分区的调整角度和旋转角度。

2.1.1 密度生成模块

密度生成模块的主网络采用 FCN[14] 全卷积神经卷积网络，深层网络中要实现增加感受野且降低计算量，必然要通过池化或者卷积方式进行降采样，这样虽增加了感受野但也会导致分辨率降低造成图片信息的丢失。

本文对原 FCN 网络通过增加空洞卷积[15]解决这一问题，提高人群计数精度。空洞卷积是由文献[16]提出的，通过在非零滤波器抽头之间插入孔来滤除上采样，最后将特征响应双线性插值回原始图像大小，在不增加参数个数的情况下有效地扩大了感受野。在几个空洞卷积层输出之后，我们将其与第二个最大池化层进行结合，减少特征损失并获取更多特征；最后连接两个反卷积进行上采样，将第一个反卷积的上采样得到的 1/8 分辨率结果热图与模型正向卷积操作得到的 1/8 分辨率的特征图进行融合操作(Fuse Operation)，减少上采样过程信息的丢失获取更多的特征。网络模型的正向卷积前两层卷积池化操作获得特征较低级不利于对人群密度特征的抽象，因此直接从 1/8 图像分辨率通过上采样得到输入图像原分辨率大小。网络最后的一个 1×1 卷积核作为回归器，将特征映射到人群密度中，实现人群密度图的生成，网络结构如下图 2 所示。

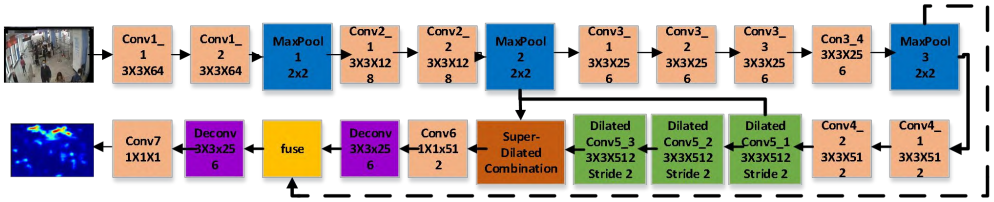

图 2 改进 FCN 网络

Fig.2 Improved FCN network

2.1.2 时空变换计数模块

本文受文献[10]对视频监控车辆计数的启发，对原网络进行改进最后实现对视频人群的计数如下图 3 所示。摄像机的视角会造成成像中人尺度和旋转角度的巨大变化，影响最后人群计数的准确度。本文为解决这个问题在最后的 LSTM 模块中插入一个空间变换网络。在第 i 次迭代中，首先用空间变换器网络确定要细化的图像区域。

$$a_i, g_i = \mathrm{LSTM}(a_{i-1}, g_{i-1}), \tag{1}$$

其中 a_i 是当前迭代的内存单元，g_i 为隐状态，LSTM 获取密度图的过去信息以及变化信息。空间变换网络 ST 由 Jaderberg[17]等人提出，该网络能够在人群密度图中动态地定位一个注意区域，对其进行尺度变换和旋转，变换矩阵（2）式所示:

$$\boldsymbol{T}_i = \begin{bmatrix} \theta_{11}^i & \theta_{12}^i & \theta_{13}^i \\ \theta_{21}^i & \theta_{22}^i & \theta_{23}^i \end{bmatrix}, \tag{2}$$

本文利用隐状态 g_i 计算具有全连接层的变换矩阵 $\boldsymbol{T}_i$ 的参数，其中 θ^i 为变化参数，然后根据变换矩阵 $\boldsymbol{T}_i$ 从完整密度图 M_{i-1} 中提取区域密度图 d_i ，表达式如下：

$$d_i = ST(M_{i-1}, \boldsymbol{T}_i), \tag{3}$$

式中 ST 代表空间变换器，区域密度图 d_i 会通过过双线性插值调整到给定的大小 $w*h$。本文将经过 ST 网络变换的区域密度图 d_i 与采样相同大小的真值密度图进行对比，计算区域损失函数 L_{ST} ，通过这个损失函数来优化 ST 网络的参数，网络结构如下图 3 所示。

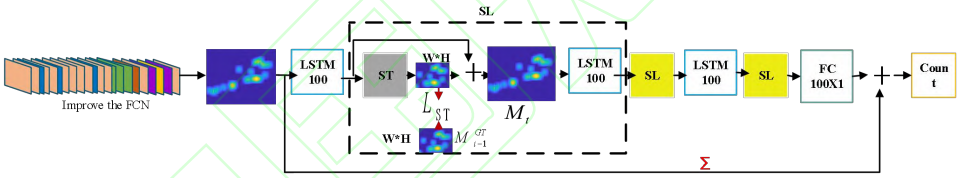

图 3 时空变换计数网络

Fig.3 Time-space transformation counting network

2.2 网络模型训练

2.2.1 更新函数

LSTM 模型在任意时刻 t 的更新方程如下：

$$\begin{aligned} \mathrm{i}_\mathrm{t} &= \sigma_i(W_{xi}x_t + W_{hi}h_{t-1} + w_{ci} * c_{t-1} + b_i) \\ f_\mathrm{t} &= \sigma_f(W_{xf}x_t + W_{hf}h_{t-1} + w_{cf} * c_{t-1} + b_f) \\ o_\mathrm{t} &= \sigma_o(W_{xo}x_t + W_{ho}h_{t-1} + w_{co} * c_t + b_o) \\ c_\mathrm{t} &= f_t * c_{t-1} + i_t * \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ h_t &= \sigma_t * \tanh(c_t) \end{aligned}, \tag{4}$$

改进网络 FCN 最后输出的密度映射变为一维向量 x_t，h_{t-1} 代表 t-1 时刻隐藏层的状态值，c_{t-1} 代表 t-1 时刻记忆单元状态值。LSTM 隐藏层的激活函数用 sigmoid 函数[18]，记忆单元状态值和隐藏层状态值的更新激活函数用双曲正切函数。门控参数 i_t、f_t 和 o_t 通过门控

网络的权重参数W_{xi}、W_{hi}、W_{xf}、W_{hf}、W_{xo}、W_{ho}，以及偏置参数b_i、b_f、b_o并与输入的x_t和前一时刻隐藏层状态值h_{t-1}结合求得。

2.2.2 密度图

人群计数获得视频图像的人群密度图，积分值表示图像的人数。若x_i表示图像x的第i个人头标记点的位置，函数$\delta(x-x_i)$表示它的密度图，图像x若包含N个人头标记点，则其密度图如下式：

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

将高斯核滤波器G_δ与上式进行卷积，得到密度图$M(x)=H(x)*G_\delta(x)$。根据透视原理，三维的人群图片场景在用二维彩色图像反映的过程中，标记点的每个像素点代表人头所占的面积会发生变化，上式的密度图生成方式会导致结果不够准确。为了处理透视失真带来的视角扭曲，本文采用自适应高斯滤波器与（5）式卷积，得到密度图公式如下：

$$M(x)=\sum_{i=1}^{N}\delta(x-x_i)*G_{\delta_i(x)},\delta_i(x)=\varphi\,\bar{d}_i\,, \tag{6}$$

其中，$\bar{d}_i$表示标记点x_i与其邻近的k_0个人头间的平均距离，通过大量的实验得知当φ=0.3时得到的密度图质量最好。

2.2.3 损失函数

本文采用欧式距离对估计人群密度与地面真值的误差进行测量。密度图估计的损失函数由如下公式定义：

$$L_{\text{Density}}=\frac{1}{2N}\sum_{i=1}^{N}\sum_{p=1}^{p}\left\|\hat{M}_i(p;\theta_{\text{FCN}})-M_i(p)\right\|_2^2\,, \tag{7}$$

$$L_{\text{ST}}=\frac{1}{2N}\sum_{i=1}^{N}\left\|\hat{M}_i^{\text{ST}}(w*h)-M_i(w*h)\right\|_2^2\,, \tag{8}$$

式中N为训练单批样本的数量(Batch Size)，$\hat{M}_i(p;\theta_{\text{FCN}})$为密度估计网络FCN输出的第$i$帧中人群密度图的像素$p$对应的人群密度估计值，$M_i(p)$为第$i$帧的像素$p$对应的人群真值密度图，$\theta_{\text{FCN}}$为FCN网络中需学习的参数，$\hat{M}_i^{\text{ST}}(w*h)$为第$i$帧经过ST空间变换网络之后生成的$w*h$大小的密度图，$M_i(w*h)$为其对应$w*h$大小的真值密度图。针对人群计数任务包括基本计数和残差计数：（1）基本计数对整个图像的密度图进行整合；（2）残差计数通过LSTM学习获得，本文将两者相加得到最后估计的人群数量：

$$\hat{C}_i=\sum_{p=1}^{p}\hat{M}_i(p)+G(\hat{M}_i;\gamma,\eta)\,, \tag{9}$$

式中为$\hat{M}_i$第i帧中改进FCN网络生成的人群密度图，$G(\hat{M}_i;\gamma,\eta)$为估计的残差数量，γ是LSTM中需学习的参数，η是全连接层需学习的参数。人群计数估计损失函数如下：

$$L_{\text{count}}=\frac{1}{2N}\sum_{i=1}^{N}(\hat{C}_i-C_i)^2\,, \tag{10}$$

式中C_i为第i帧中乘客计数真实值，$\hat{C}_i$为本文网络得出第i帧估算得出的人群计数值。综上几个函数表达式，网络总损失函数定义为：

$$L_{\text{Global}} = \lambda(L_{\text{Density}} + \beta L_{\text{ST}}) + L_{\text{count}}, \quad (11)$$

式中，λ 为人群密度估计损失函数的权重，β 是局部密度图的损失权重，调整 λ 、β 来提升模型整体性能，损失函数用批处理 *Adam* 优化算法和反向传播算法进行优化。

3 实验设计及结果分析

3.1 实验环境及参数设置

本文算法实验在深度学习工作站上开展训练和测试，并测试了训练好的网络模型架构对自己数据集的检测计数情况，实验中的两个任务是联合训练的，这样可以减少使用不必要的参数，同时模型能够得到更好的训练结果。操作系统为 Windows，CPU 为锐龙 3700x，GPU 用 rtx2060，实验框架为 CUDA10+anaconda3+python3.6+pytorch，训练网络模型过程中，批大小(Batch Size)设置为 30，LSTM 时间步长设为 10，网络最初的学习率(Learning Rate)设为 10^{-4}，迭代次数设为10^5，公式（11）中总损失函数的 λ 参数设为 0.1，β 设为 0.001。

3.2 数据集训练

3.2.1 评价指标

为了评估模型在人群密度估计和计数准确性上的性能指标，平均绝对误差 (MAE)和平均平方误差 (MSE)评估模型，其定义如下：

$$f_{\text{MAE}} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x}_i|, \quad f_{\text{MSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}, \quad (12)$$

式中 N 是所有的测试视频序列的帧总数，x_i 和 $\hat{x}_i$ 分别表示第 i 图片的实际人数和模型估计出的预测人数。

3.2.2 UCSD 数据集

UCSD 数据集来自校园内监控摄像机拍摄的 2000 帧图像，238x158 像素的帧分辨率，10fps 的帧速率，每帧有 11 到 46 不等的人数。本文与文献[19]相同的实验设置，选用 601 到 1400 做训练帧，数据集中剩余的 1200 帧做测试帧。下表显示了不同方法在 UCSD 数据集上的准确性，图 4 是每帧图片预测值和真实值之间的分布，从表 1 和图 4 中可以看出，在 UCSD 数据集上本文方法也优于多列网络计数的性能。

表 1 UCSD 数据集比较

Table 1 Comparison of UCSD data sets

Method	f_{MAE}	f_{MSE}
ConvLSTM[3]	1.30	1.79
Bidirectional ConvLSTM[3]	1.13	1.43
Gaussian Process Regression[19]	2.24	7.97
Ridge Regression[20]	2.25	7.82
FCN-rLSTM[10]	1.54	3.02
Methods in literature[21]	2.07	6.86
Paper network	1.05	1.59

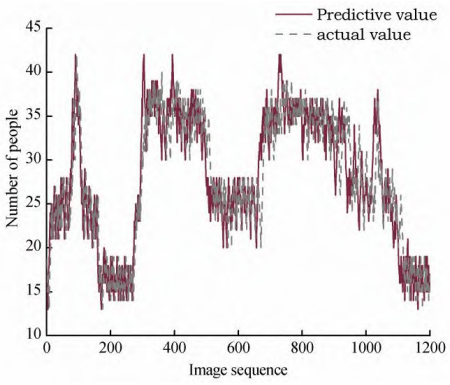

图 4 UCSD 数据集本文方法计数结果

Fig. 4 UCSD data sets counting results of this paper

从表 1 和图 4 中能够看出，（1）在 UCSD 数据集上所提方法对人群计数表现良好，究其原因可能是数据集中人群数量较少且遮挡少，证明改进方法对计数准确性提高有效；（2）本文方法相比之前最优方法在平均绝对误差和均方误差上分别降低 7.08%和 11.2%。

3. 2. 3 Mall 数据集实验与分析

Mall 数据集的特点是人群密度变化大、人群活动方式多包括目标的静止和运动，透视畸变和遮挡较严重。它是由监控摄像机安装在购物中心得到的数据集[20]，目标总人数在 6000 左右，640x480 像素，实验将前 1-800 帧作为训练，其余 1200 帧作测试集。在此数据集中，本文所提方法实验以及测试的结果如图 5 所示，与不同方法性能指标对比结果如表 2 所示。

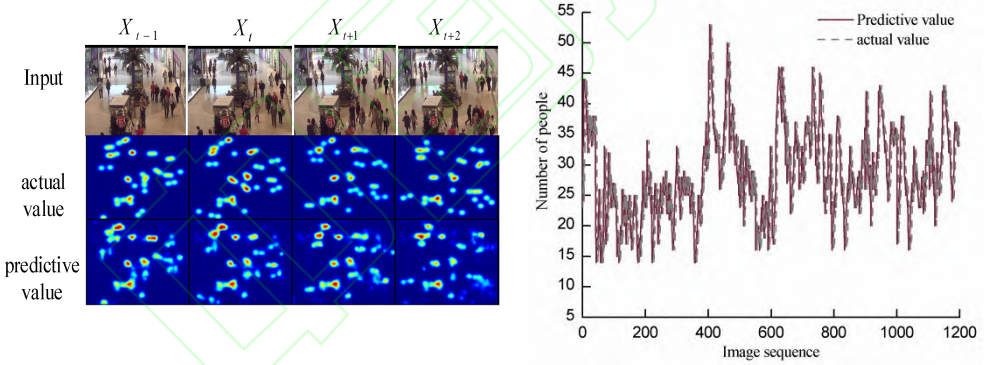

图 5 Mall 数据集序列图像实验及计数结果

Fig.5 Sequence image experiment and counting results of Mall data sets

表 2 Mall 数据集比较

Table 2 Comparison of Mall data sets

Method	f_{MAE}	f_{MSE}
ConvLSTM[3]	2.24	8.50
Bidirectional ConvLSTM[3]	2.10	7.60
Ridge Regression[20]	3.59	19.00
Methods in literature[21]	3.43	17.7
Paper network	1.95	7.50

由图 5 和表 2 能够看出，在 Mall 公开数据集中，本文方法在 f_{MAE} 和 f_{MSE} 较 ConvLSTM

分别降低了 12.9%和 11.8%，表明本文的方法对室内复杂的场景人群计数有较好的精度，说明该方法的准确性和鲁棒性都很好。

3.2.4 自建视频素材数据集

目前人们短距离和较长距离的出行离不开地铁和铁路，而现有数据集很少关注这两个方面，为满足智慧交通的要求，在重要节假日人群密集高峰期对入站和出站口的人群计数有很大的实际意义。本文建立的数据集包括来自兰州地铁入站及西站上车过程的监控视频，照片包括不同场景不同光照不同角度，一共选取 1200 帧图像组成本文所需的数据集，本文对数据集采用人工标注的方式，与 Mall 和 UCSD 数据集标记方法类似，手动标注每帧图片中人头的坐标点，训练网络分别选取 900 帧图片（各取 450 张），其余的 300 帧（每个场景各 150 帧）进行测试，本文自建视频素材数据集实验及结果测试如图 6 所示，与其他方法比较结果如表 3 所示。

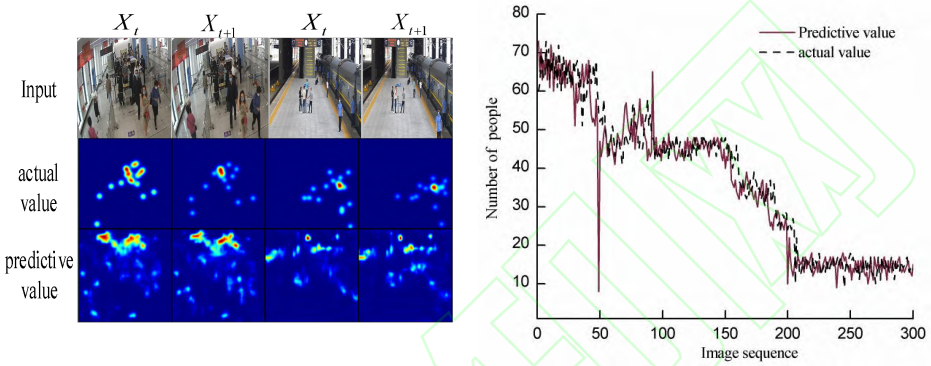

图 6 自建视频素材数据集序列图像实验及计数结果

Fig.6 Sequence image experiment and counting results of self-built video material data sets

表 3 自建视频素材数据集比较

Table 3 Comparison of self-built video material data sets

Method	f_{MAE}	f_{MSE}
ConvLSTM[3]	4.51	5.91
MCNN[22]	3.81	4.92
Our without ST	4.32	5.21
Paper network	3.51	5.10

由图 6 和表 3 可看出，当视频图片中存在遮挡的物体时，本文方法的计数精度有所下降，分析该原因可能是网络对密集遮挡人群的图片训练较少，网络没能学习到足够的特征，f_{MAE} 相比最佳方法 MCNN 只降低了 8.1%，而 f_{MSE} 也只降低了 7.1%，在今后的研究中还需改进。

3.3 模型对比实验分析

（1）密度生成模块引入空洞卷积的网络性能对比

为验证本文提出的在单列深度全卷积神经网络中加入空洞卷积对视频图像人群计数的有效性，实验设置性能验证对比实验，在基础网络即无空洞卷积无时空变换模块(No dilated No ST)和加入空洞卷积无时空变换模块网络(dilated+No ST)进行比较，测试集分别在 UCSD、Mall 和本文自建视频素材 3 个数据集上进行，测试帧的选择与 3.2 节提到的三个数据集的测试帧都相同。测试结果如下表 4 所示。

表 4 各数据集验证性实验结果

Table 4 Verification experiment results of each data sets

datasets	No dilated No ST		dilated+NoST	
	f_{MAE}	f_{MSE}	f_{MAE}	f_{MSE}
UCSD	1.71	4.25	1.52	4.13
Mall	2.89	9.01	2.13	8.51
Self-built video material	4.74	6.65	4.32	5.21

由表中对比内容看出，在网络前端加入空洞卷积的改进方式的性能相比原基础 FCN 无改进的网络(No dilated No ST)分别在 3 个数据集上 f_{MAE} 和 f_{MSE} 在准确性和鲁棒性上都有显著的提升。因空洞卷积在不增加参数个数的情况下能有效扩大感受野，说明本文引入的空洞卷积提高了密度生成模块在提取人群特征上的能力。

（2）时空变换计数模块网络性能对比

为验证本文提出的在 LSTM 网络中加入空间变换模块对视频图像人群计数准确度的提高，实验将无空洞卷积无时空变换模块(No dilated No ST)的网络、无空洞卷积与加入时空变换模块的网络(No dilated+ST)和本文的加入空洞卷积并结合时空变换模块的网络(dilated+ST)进行比较，在 UCSD、Mall 和本文自建视频素材 3 个数据集上测试，测试数据集图片设置与实验（1）相同。测试结果如下表 5 所示。

表 5 各数据集验证性实验结果

Table 5 Verification experiment results of each data sets

datasets	No dilated No ST		No dilated+ ST		dilated+ST	
	f_{MAE}	f_{MSE}	f_{MAE}	f_{MSE}	f_{MAE}	f_{MSE}
UCSD	1.71	4.25	1.41	3.52	1.05	1.59
Mall	2.89	9.01	2.01	8.43	1.95	7.50
Self-built video material	4.74	6.65	4.33	5.23	3.51	5.10

由表中能够看出，将无空洞卷积与加入时空变换模块的网络(No dilated+ST)与基础网络(No dilated No ST)相比，在 f_{MAE} 和 f_{MSE} 的性能上有明显的提高，说明 ST 模块的加入对网络计数结果的准确性有贡献；对比实验无空洞卷积与加入时空变换模块的网络(No dilated+ST)结果与本文所提的空洞卷积与时空变换模块两者相结合的联合网络(dilated+ST)相比，仅有 ST 模块的网络(No dilated+ST)计数性能效果比本文的联合网络效果差；但将此实验结果与表 4 中的仅有空洞无 ST 模块的网络(dilated+No ST)比较，仅有 ST 模块的网络比仅有空洞卷积的网络性能高，说明 ST 模块比引入空洞卷积的贡献大，根据最终实验结果显示两者的结合网络实现的效果是最好的。

为更直观检测 ST 模块对计数任务的影响，对网络的准确性进行了实验，如下图 7 所示。

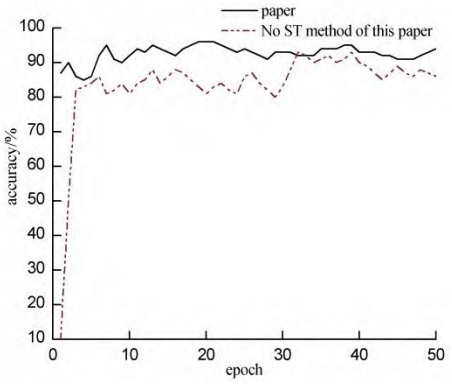

图 7 改进模型准确度

Fig.7 Improved model accuracy

根据 Mall 数据集人群密度变化大及疏密分布较平衡的特点，选用 Mall 数据集的 801-1200 帧作为测试集对网络改进的准确度进行结果验证，得到 50 次迭代后模型准确率，从图 7 可看出，改进前后的网络准确率都在 70%以上，没有加入 ST 模块的网络准确率都在 80%～90%之间，而本文改进网络准确率都在 90%以上。

（3）改进模型前后训练损失值对比

下图 8 为网络收敛效果对比图。从图中能够看出，改进后的网络收敛性得到了很大的提高。本文网络模型在训练前期，损失曲线波动起伏，这是由于前期 FCN 卷积层未完成更多参数的学习，特征中加入了冗余的细节信息，模型训练会受到误导。随着迭代次数的增加，原始网络更容易发生过拟合现象，而改进模型则表现更稳定。

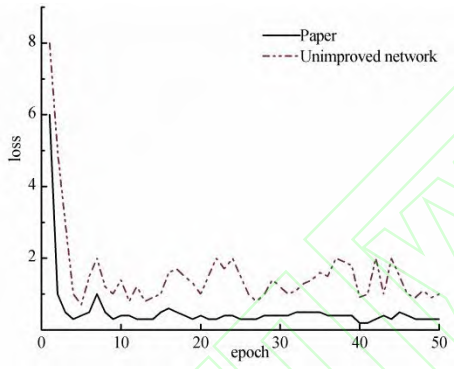

图 8 训练损失曲线

Fig.8 Training loss curve

由以上对比内容看出，空洞卷积的引入、ST 模块与 LSTM 网络的结合皆对人群计数任务有贡献。其中能够明显看出 ST 模块对网络性能提升最明显，其次是空洞卷积效果，将两者与整体网络连接方式结合时效果是最佳的。此次对比实验结果表明，本文提出的方法在视频图像中人群计数的准确性上效果显著。

4 结　论

本文提出基于单列深度时空计数的卷积网络方法实现视频图像中的人群计数，整体网络结构主要分为密度生成和时空变换计数两个模块；针对单列网络提取图片特征不全丢失信息导致人群计数不精确的问题，采用增加空洞卷积和跳跃连接融合的方式，大大提高了网络提取特征的能力；同时由于摄像机视角的问题造成成像中人尺度和旋转角度的巨大变化难以识别的问题，采用空间变换模块(ST)插入在 LSTM 网络中，通过在数据集上测试检测速度和检测精确度都达到了较高的水准，证明本文改进的网络具有现实应用的价值。本模型在对于存在遮挡和密集的人群的计数上，还存在明显的不足，需要在今后的研究实验中加以改进。

参考文献

[1] V. A. Sindagi, V. M. Patel. A survey of recent advances in CNN-based single image crowd counting and density estimation[J]. Pattern Recognition Letters, 2017, 3-16

[2] Y. M. Marsden, K. Mc Guinness, S. Little, et al. Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification[J]. ar Xiv preprint ar Xiv:1705.10698, 2017, 1-18

[3] F. Xiong , X. Shi , D.-Y. Yeung , Spatiotemporal modeling for crowd counting in videos[C]// 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 5161-5169 .

[4] Y. Cong, H. Gong, S. C. Zhu. Flow mosaicking: Real-time pedestrian counting without scene-specific learning[C]. CVPR, 2009, 1093-1100

[5] B. Li, J. Zhang, Z. Zhang. A people counting method based on head detection and tracking[C]. International Conference on Smart Computing, IEEE, 2015, 136-141

[6] Fan C C.Research on Algorithms of Density Estimation and Crowd Counting Based on Convolutional Neural Network[D].He Fei:Anhui University,2019:3-20.
范超超.基于卷积神经网络的密度估计及人群计数的算法研究[D].合肥:安徽大学,2019:3-20.

[7] Q. Wen, C. Jia, Y. Yu. People Number Estimation in the Crowded Scenes Using Texture Analysis Based on Gabor Filter[J]. Journal of Computational Information Systems, 2011, 7(11)

[8] T. Liu and D. Tao, On the Robustness and Generalization of Cauchy Regression[C]. 2014 4th IEEE International Conference on Information Science and Technology, Shenzhen, 2014, pp. 100-105, doi: 10.1109/ICIST.2014.6920341.

[9] Lempitsky V S , Zisserman A . Learning To Count Objects in Images[C]// Advances in Neural Information Processing Systems 23: Conference on Neural Information Processing Systems A Meeting Held December. Curran Associates Inc. 2010,10:1007.

[10] S. Zhang, G. Wu, J. P. Costeira,et al. FCN-rLSTM: Deep Spatio-Temporal Neural Networks for Vehicle Counting in City Cameras[C],2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 3687-3696, doi: 10.1109/ICCV.2017.396.

[11] LI Y,ZHANG X,CHEN D.Csrnet:Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.Salt Lake:IEEE Press,2018:1091-1100.

[12] Zuo J, Ba Y L. A depth population counting algorithm based on multi-scale fusion[J/OL]. Laser&OptoelectronicsProgress:1-12[2020-08-17].http://kns.cnki.net/kcms/detail/31.1690.TN.20200623.1440.014.html.
左静,巴玉林.一种多尺度融合的深度人群计数算法[J/OL].激光与光电子学进展:1-12[2020-08-17].http://kns.cnki.net/kcms/detail/31.1690.TN.20200623.1440.014.html.

[13] Wu Z H, Gao Y M, Li J,et al. Fully Convolutional Network Method of Semantic Segmentation of Class Imbalance Remote Sensing Images[J]. Acta Optica Sinica, 2019,39(04):401-412.
吴止锾,高永明,李磊,薛俊诗.类别非均衡遥感图像语义分割的全卷积网络方法[J].光学学报,2019,39(04):401-412.

[14] Marsden M , Mcguinness K , Little S , et al. Fully Convolutional Crowd Counting On Highly Congested Scenes[J]. 2016,10:5220.

[15] Gao L, Song W D, Tan H,et al. Multi-scale expansion convolutional neural network resources satellite image cloud recognition[J]. Acta Optica Sinica, 2019,39(01):0104002.
高琳,宋伟东,谭海,刘阳.多尺度膨胀卷积神经网络资源三号卫星影像云识别[J].光学学报,2019,39(01):299-307.

[16] 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, 2018, 40(4):834-848.

[17] Max Jaderberg, Karen Simonyan, Andrew Zisserman,et al.Spatial transformer networks, in Advances in neural information processing systems[J]. 2015, pp. 2017-2025.

[18] Greff K , Srivastava R K , Jan Koutník, et al. LSTM: A Search Space Odyssey[J]. IEEE Transactions on Neural Networks & Learning Systems, 2016, 28(10):2222-2232.

[19] A. B. Chan, Zhang-Sheng John Liang and N. Vasconcelos, Privacy preserving crowd monitoring: Counting people without people models or tracking[C],2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, 2008, pp. 1-7, doi: 10.1109/CVPR.2008.4587569.

[20] Chen K , Loy C C , Gong S , et al. Feature Mining for Localised Crowd Counting[C]// British Machine Vision Conference. 2012,10:5244.

[21] K. Chen, S. Gong, T. Xiang ,et al.Cumulative Attribute Space for Age and Crowd Density Estimation[C],2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, 2013, pp. 2467-2474, doi: 10.1109/CVPR.2013.319.

[22] Y. Zhang, D. Zhou, S. Chen, et al. Single-Image Crowd Counting via Multi-Column Convolutional Neural Network[C],2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 589-597, doi: 10.1109/CVPR.2016.70.

网络首发：

标题：单列深度时空卷积神经网络的人群计数研究
作者：鱼春燕,徐岩,缑丽莎,南哲锋
收稿日期：2020-07-28
录用日期：2020-09-14
DOI：10.3788/lop58.081011

引用格式：
鱼春燕,徐岩,缑丽莎,南哲锋. 单列深度时空卷积神经网络的人群计数研究[J].激光与光电子学进展,2021,58(08):081011.
