第 58 卷 第 8 期/2021 年 4 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于单列深度时空卷积神经网络的人群计数

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 突发性人群聚集会给人们的人身安全带来隐患，因此，对高风险区域进行有效的人群计数具有重要意义。针对多列神经网络结构臃肿、冗余信息多及耗时长等问题，提出了一种基于单列深度时空卷积神经网络的人群计数模型，并对模型进行改进，以满足视频图像计数的需要。首先，在全卷积神经网络(FCN)中加入空洞卷积和跳级连接特征融合，以提高网络提取特征的能力。然后，为了减少视频监控产生的角度畸变对计数结果的影响，在长短期记忆(LSTM)网络结构中加入空间变换模块；为了提高网络计数结果的精确性，用残差连接方式连接改进的 FCN 和关联时序的 LSTM 网络。最后，在 UCSD、Mall 和自建人群数据集上分别进行测试，结果表明，相比其他模型，本模型的人群计数准确性和鲁棒性更好。

关键词 图像处理；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 **doi**：10.3788/LOP202158.0810011

Crowd Counting Based on Single-Column Deep Spatiotemporal Convolutional Neural Network

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

School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou, Gansu 730070, China

Abstract Sudden mass gatherings are detrimental to people's safety. Therefore, it is paramount to conduct effective crowd counting in high-risk areas. Aiming at the problems of multicolumn neural network structure is bloated, redundant information and time consuming, we proposed a crowd counting model based on a single-column deep spatiotemporal convolutional neural network and modified it for video image counting. First, a fully convolutional network (FCN) is added to the feature fusion of dilated convolution and level jump connection to improve the ability of the network to extract features. Then, to reduce the influence of the angle distortion generated by the video surveillance on the counting results, a spatial transformation module is added to the long short-term memory (LSTM) network structure. Further, the residual connection method is used to connect and improve the FCN and associated timing LSTM network to improve the accuracy of the network counting results. Finally, tests are performed on UCSD, Mall, and self-built population data sets. Results show that the crowd counting accuracy and robustness of the model are better compared with other models.

Key words image processing; neural networks; crowd counting; deep spatiotemporal network; dilated convolution; spatial transformation

OCIS codes 100.4996; 100.3008; 100.2000

1 引 言

监控设备的快速发展使图像和视频的数据量不断增长，对视频内容分析的需求也越来越大。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，使其在计算机视觉任务中得到了广泛的研究

收稿日期：2020-07-28；**修回日期**：2020-08-22；**录用日期**：2020-09-14

* **E-mail**：xuyan@mail.lzjtu.cn

与应用。随着科技的创新和城市交通的部署，地铁已成为人们出行的首选交通工具。在出行高峰期，不可避免地会引发人为事故。为了满足智慧交通的要求，提高社会的安全性，在地铁监控系统中引入人群计数具有重要意义，但以往对人群计数的研究大多集中在单张图像上[1-3]，难以满足视频监控的需求。

已有的人群计数方法大致可分为基于检测的方法、基于回归的方法、基于密度估计的方法以及基于深度学习的方法四类。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]，用检测器对单独的个人或身体部位进行逐次检测，累计后得到统计结果[6]。这种方法计数精度较高，但在密集和复杂场景下无法检测出小物体和被遮挡的头、身体，且耗时较长。基于回归的方法以人群为整体估计人群密度，能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]，简化了复杂的个体检测任务。但训练需要大量的标记数据，算法复杂度较高，且不能定位人群的空间位置关系。基于密度估计的方法不仅解决了检测和特征回归的局限性，且能估计图像中任意区域的目标数量，将计数问题转为估计图像密度的问题，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中得到了广泛应用，基于深度学习的方法根据网络特性差异可分为基于卷积神经网络（CNN）、基于尺度感知模型、基于上下文感知模型和基于多任务模型的方法。Zhang 等[10]提出用全卷积网络（FCN）结合长短期记忆（LSTM）网络进行视频车辆计数的方法。Xiong 等[3]从视频图像在时间上存在关联的角度，提出了一种融合 CNN 与 LSTM 网络的 ConvLSTM 人群密度估计模型，可充分利用视频序列的时间信息。Li 等[11]在网络中加入空洞卷积，形成能识别高密度场景的空洞卷积网络（CSRNet）模型，可提取高层次语义特征，准确估计密集场景的人群分布情况。该模型虽然参数量较少，但容易丢失细节信息[12]。

现有的人群计数方法关于视频序列帧中人群数量的时间相关性模型研究较少。针对摄像机视角畸变导致的物体识别难度大问题，本文提出了一种基于单列深度时空 CNN 的人群计数模型。模型前端的密度生成模块采用在单列 FCN 融入空洞卷积和跳级连接的方式，可提取图像的细节特征，提高生成密度图的质量；模型后端的时空变换计数模块将 LSTM 网络结构和空间变换（ST）模块相结合，改善了摄像机视角畸变导致的物体识别难度大问题。

2　人群计数模型

2.1　基于单列深度时空计数的 CNN

在监控场景不同的人流分布、光照变化、人群遮挡以及拍摄角度畸变情况下，有效计算人群数量是一个极大的挑战。尽管多列网络已经取得了很大的进展，但其不同分支中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，无法提升最终生成的密度图质量。针对该问题，改进了基于单列深度时空计数的 CNN，密度图生成主干网络 FCN 的主要作用是将传统网络进行卷积化[13]，然后加入空洞卷积和跳跃连接，以提高网络提取特征的能力；时空变换部分将 LSTM 网络和 ST 模块进行结合，以实现图像时序关联计数。模型的结构如图 1 所示，其中，X_t 为网络训练的图像，t 为图像的序列，FC 为全连接层。

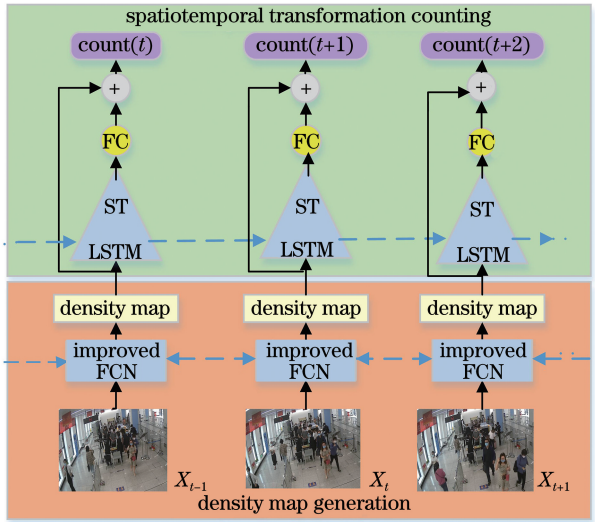

图 1　基于单列深时空计数的 CNN

Fig. 1　CNN based on single-column deep spatiotemporal counting

整个网络主要分为两个模块，一个是密度图生成模块，该模块主要由改进的 FCN 组成；另一个模块为基于 LSTM 网络实现的时空变换计数模块。网络通过残差学习将改进的 FCN 和 LSTM 网络相结合，从而根据时序估计人群数量。改进的 FCN 主要将像素级特征映射到人群密度中，在 LSTM 网络中插入 ST 模块。ST 模块能进行端到端的训练，可合并到 CNN 并使用标准的反向传播算法进行训练，用参数矩阵确定分区的位置以及分区的调整角度和旋转角度。

2.1.1　密度图生成模块

密度图生成模块的主干网络为 FCN[14]，为了在深层网络中增加感受野、降低计算量，需通过池化或卷积方式进行降采样，但会降低图像的分辨率，导致信息丢失。为了解决该问题，在原始 FCN 中增加空洞卷积[15]，以提高人群计数的精度。空洞卷积由 Chen 等[16]提出，首先，在非零滤波器抽头之间插入小孔滤除上采样；然后，将特征响应双线性插值变回原始图像的尺寸，在不增加参数量的情况下有效扩大了感受野；其次，将几个空洞卷积层的输出值与第二个最大池化层进行结合，以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，将第一个反卷积上采样得到的 1/8 分辨率热图与模型正向卷积操作得到的 1/8 分辨率特征图进行融合操作（Fuse operation），以减少上采样过程中的信息丢失，获取更多的特征。网络模型正向卷积的前两层卷积池化获得的特征比较低级，不利于对后续人群密度特征进行抽象描述，因此直接对 1/8 分辨率图像进行上采样，得到与输入图像分辨率相同的图像。用网络最后一个 1×1 卷积核作为回归器，将特征映射到人群密度中，生成人群密度图。该网络的结构如图 2 所示，其中，Conv 为卷积层，maxpool 为最大池化操作，dilated Conv 为空洞卷积，deConv 为反卷积。

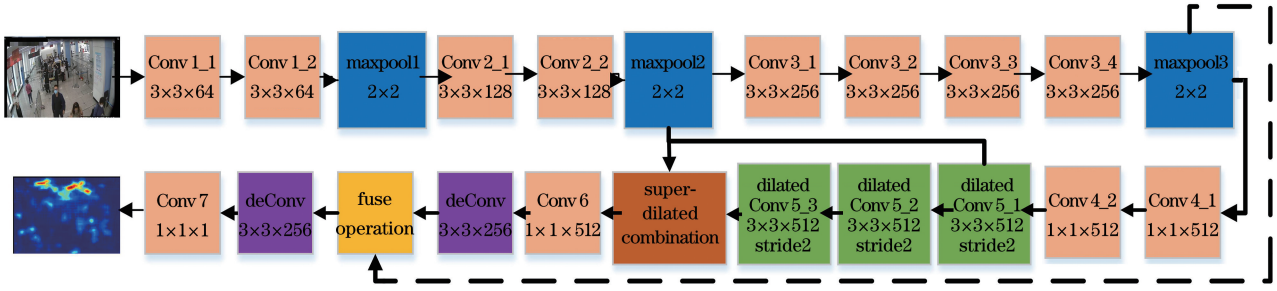

图 2　改进的 FCN 结构

Fig. 2　Structure of the improved FCN

2.1.2　时空变换计数模块

基于文献[10]中视频监控车辆的计数方法，对原始网络进行改进，以实现对视频人群的计数。摄像机的视角会导致成像中人的尺度和旋转角度发生变化，从而影响人群计数的准确性。为了解决该问题，在最后一层 LSTM 网络中插入一个 ST 模块。在第 i 次迭代中，用 ST 模块确定要细化的图像区域，可表示为

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

式中，a_i 为当前迭代的内存单元，g_i 为隐状态，X_{LSTM} 为获取密度图过去信息和变化信息的函数。ST 模块由 Jaderberg 等[17]提出，能在人群密度图中动态定位一个注意力区域，并对其进行尺度变换和旋转，变换矩阵可表示为

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

式中，θ^i 为变换参数，可利用隐状态 g_i 计算。根据变换矩阵 $\boldsymbol{T}_i$ 从完整密度图 $\boldsymbol{M}_{i-1}$ 中提取的区域密度图 $\boldsymbol{d}_i$ 可表示为

$$\boldsymbol{d}_i = X_{\mathrm{ST}}(\boldsymbol{M}_{i-1}, \boldsymbol{T}_i), \tag{3}$$

式中，X_{ST} 为 ST 模块，区域密度图 $\boldsymbol{d}_i$ 可通过双线性插值调整到给定的尺寸 $w \times h$。将经过 ST 模块的区域密度图 $\boldsymbol{d}_i$ 与相同大小的真实密度图进行对比，计算区域损失函数 L_{ST}，通过该损失函数优化 ST 模块的参数，网络结构如图 3 所示。其中，$\boldsymbol{M}_t$ 为输出密度图，GT 为真值密度图，SL 为改进后插入网络的整体模块。

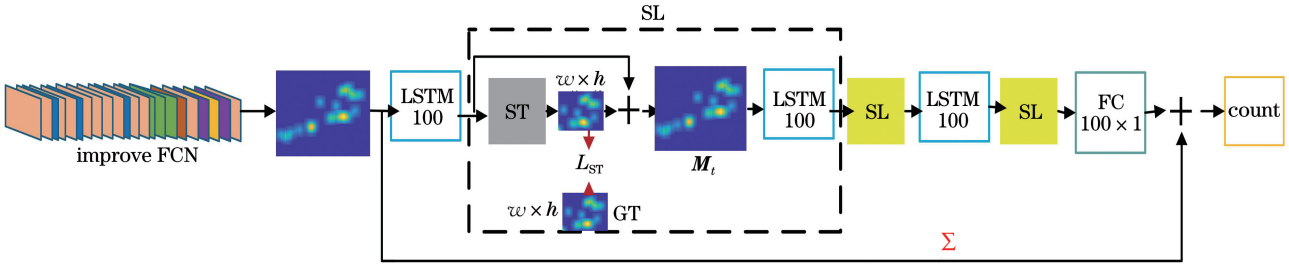

图 3　ST 计数网络的结构

Fig. 3　Structure of the ST counting network

2.2 网络模型训练

2.2.1 更新函数

LSTM 网络在 t 时刻的更新方程可表示为

$$\begin{aligned}\boldsymbol{e}_t&=\sigma_e(W_{xe}\boldsymbol{x}_t+W_{ke}\boldsymbol{k}_{t-1}+w_{ce}\boldsymbol{c}_{t-1}+\boldsymbol{b}_e)\\\boldsymbol{f}_t&=\sigma_f(W_{xf}\boldsymbol{x}_t+W_{kf}\boldsymbol{k}_{t-1}+w_{cf}\boldsymbol{c}_{t-1}+\boldsymbol{b}_f)\\\boldsymbol{o}_t&=\sigma_o(W_{xo}\boldsymbol{x}_t+W_{ko}\boldsymbol{k}_{t-1}+w_{co}\boldsymbol{c}_t+\boldsymbol{b}_o)\\\boldsymbol{c}_t&=\boldsymbol{f}_t\boldsymbol{c}_{t-1}+\boldsymbol{e}_t\tanh(W_{xc}\boldsymbol{x}_t+W_{kc}\boldsymbol{k}_{t-1}+\boldsymbol{b}_c)\\\boldsymbol{k}_t&=\sigma_t\tanh(\boldsymbol{c}_t)\end{aligned},\tag{4}$$

式中，$\boldsymbol{e}_t$ 为输入门状态值，$\boldsymbol{f}_t$ 为遗忘门状态值，$\boldsymbol{o}_t$ 为输出门状态值，$\boldsymbol{c}_t$ 为激活向量，$\boldsymbol{k}_t$ 为隐藏输出向量，$\boldsymbol{k}_{t-1}$ 为 $t-1$ 时刻隐藏层的状态值，$\boldsymbol{c}_{t-1}$ 为 $t-1$ 时刻的记忆单元状态值，W 为门控网络的权重参数，$\boldsymbol{b}$ 为偏置参数。$\boldsymbol{e}_t$、$\boldsymbol{f}_t$ 和 $\boldsymbol{o}_t$ 可通过门控网络的权重参数、偏置参数和输入的 $\boldsymbol{x}_t$ 和 $\boldsymbol{k}_{t-1}$ 求得，一维向量 $\boldsymbol{x}_t$ 为改进 FCN 最后输出的密度映射。LSTM 网络隐藏层的激活函数为 Sigmoid 函数[18]，记忆单元状态值和隐藏层状态值的激活函数用双曲正切函数更新。

2.2.2 密度图

通过人群计数获得视频图像的人群密度图，用积分值表示图像中的人数。若 $\boldsymbol{x}_z$ 为图像 $\boldsymbol{x}$ 中第 z 个人头标记点的位置，函数 $\delta(\boldsymbol{x}-\boldsymbol{x}_z)$ 表示该人头的密度图，则图像 $\boldsymbol{x}$ 的密度图可表示为

$$H(\boldsymbol{x})=\sum_{z=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_z),\tag{5}$$

式中，N 为人头标记点的个数。将高斯核滤波器 G_δ 与(5)式进行卷积，得到密度图

$$M(\boldsymbol{x})=H(\boldsymbol{x})*G_\delta(\boldsymbol{x})。\tag{6}$$

根据透视原理可知，用二维彩色图像反映三维人群图像场景的过程中，每个像素点代表的人头所占面积会发生变化，(5)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，用自适应高斯滤波器 $\boldsymbol{G}_{\delta_z(\boldsymbol{x})}$ 与(5)式进行卷积，得到的最终密度图为

$$M(\boldsymbol{x})=\sum_{z=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_z)*\boldsymbol{G}_{\delta_z(\boldsymbol{x})},\ \delta_z(\boldsymbol{x})=\phi\bar{r}_z,\tag{7}$$

式中，$\bar{r}_z$ 为标记点 $\boldsymbol{x}_z$ 与其邻近 k_0 个人头的平均距离，ϕ 为可调参数，大量实验表明，$\phi=0.3$ 时，得到的密度图质量最好。

2.2.3 损失函数

用欧氏距离计算估计的人群密度与地面真值的误差，密度图估计的损失函数可表示为

$$L_{\text{Density}}=\frac{1}{2F}\sum_{i=1}^{F}\sum_{p=1}^{p}\left\|\hat{M}_i(p;\theta_{\text{FCN}})-M_i(p)\right\|_2^2,\tag{8}$$

$$L_{\text{ST}}=\frac{1}{2F}\sum_{i=1}^{F}\left\|\hat{M}_i^{\text{ST}}(w\times h)-M_i(w\times h)\right\|_2^2,\tag{9}$$

式中，F 为训练单批样本的数量(Batch size)，$\hat{M}_i(p;\theta_{\text{FCN}})$ 为 FCN 估计的第 i 帧图像中像素 p 对应的人群密度，$M_i(p)$ 为第 i 帧图像中像素 p 对应的人群密度真值，θ_{FCN} 为 FCN 中需学习的参数，$\hat{M}_i^{\text{ST}}(w\times h)$ 为第 i 帧经过 ST 模块后尺寸为 $w\times h$ 的密度图，$M_i(w\times h)$ 为其对应的真值密度图。人群计数任务中通过基本计数对整帧图像的密度图进行整合，残差计数通过 LSTM 网络学习获得，将两者相加得到最后估计的人群数量，可表示为

$$\hat{C}_i=\sum_{p=1}^{p}\hat{M}_i(p)+G(\hat{M}_i;\gamma,\eta),\tag{10}$$

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

$$L_{\text{count}}=\frac{1}{2F}\sum_{i=1}^{F}(\hat{C}_i-C_i)^2,\tag{11}$$

式中，C_i 为第 i 帧中人群数量的真值，$\hat{C}_i$ 为估算的第 i 帧图像中的人群数量。综上所述，网络总损失函数可表示为

$$L_{\text{Global}}=\lambda(L_{\text{Density}}+\beta L_{\text{ST}})+L_{\text{count}},\tag{12}$$

式中，λ 为人群密度估计损失函数的权重，β 为局部密度图的损失权重。通过调整 λ、β 提升模型的整体性能，用批处理 Adam 优化算法和反向传播算法优化损失函数。

3 实验设计及结果分析

3.1 实验环境及参数设置

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对数据集的检测计数情况，实验中的密度图生成和 ST 计数两个任务是联合训练的，以减少不必要的参数，同时模型能得到更好的训练结果。操作系统为 Windows10，CPU 为锐龙 3700x，GPU 为 rtx2060，实验框架为 CUDA10＋anaconda3＋python3.6＋pytorch，训练网络模型过程中，批大小为 30，LSTM 网络的时间步长为 10，网络的初始学习率(Learning rate)为 10^{-4}，迭代次数为 10^5，(12)式中的参数 $\lambda=0.1$，$\beta=0.001$。

3.2　数据集训练

3.2.1　评价指标

用平均绝对误差(MAE)和均方误差(MSE)评估模型在人群密度估计和计数准确性上的性能，可表示为

$$f_{\mathrm{MAE}}=\frac{1}{F}\sum_{i=1}^{F}|x_i-\hat{x}_i|, \tag{13}$$

$$f_{\mathrm{MSE}}=\sqrt{\frac{1}{F}\sum_{i=1}^{F}(x_i-\hat{x}_i)^2}, \tag{14}$$

式中，x_i 和 $\hat{x}_i$ 分别为第 i 帧图像中的实际人数和模型估计出的人数。

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，分辨率为 238 pixel×158 pixel，每秒传输帧数(FPS)为 10 frame，每帧图像有 11～46 不等的人数。实验参数与文献[19]相同，将 601～1400 帧作为训练帧，数据集中剩余的 1200 帧作为测试帧。表 1 为不同模型在 UCSD 数据集上的准确性，图 4 为用本模型得到每帧图像的预测值和真实值分布。可以看出，相比 Bidirectional ConvLSTM 模型，本模型在 UCSD 数据集上的 f_{MAE} 降低了 7.08%，相比 ConvLSTM 模型，本模型的 f_{MSE} 降低了 11.17，原因可能是数据集中人群数量较少且遮挡少。

表 1　不同模型在 UCSD 数据集上的性能指标

Table 1　Performance indexes of different models on the UCSD data set

Models	f_{MAE}	f_{MSE}
ConvLSTM[3]	1.30	1.79
Bidirectional ConvLSTM[3]	1.13	1.43
Ref. [19]	2.24	7.97
Ref. [20]	2.25	7.82
Ref. [10]	1.54	3.02
Ref. [21]	2.07	6.86
Ours	1.05	1.59

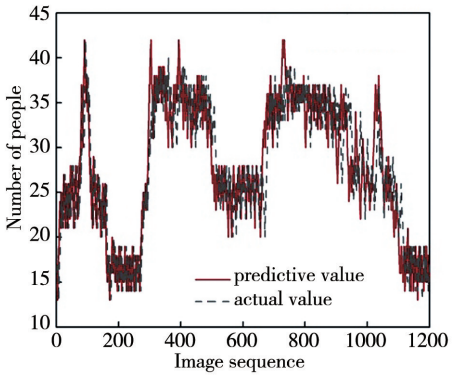

图 4　本模型在 UCSD 数据集上的计数结果

Fig. 4　Counting results of our model on the UCSD data set

3.2.3　Mall 数据集

Mall 数据集由安装在购物中心的监控摄像机拍摄的图像组成[20]，目标总人数在 6000 左右，图像像素为 640 pixel×480 pixel，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、透视畸变和遮挡严重等情况。将前 800 帧图像作为训练帧，其余 1200 帧图像作测试帧。本模型在 Mall 数据集中的计数结果如图 5 所示，不同模型在 Mall 数据集上的性能如表 2 所示。

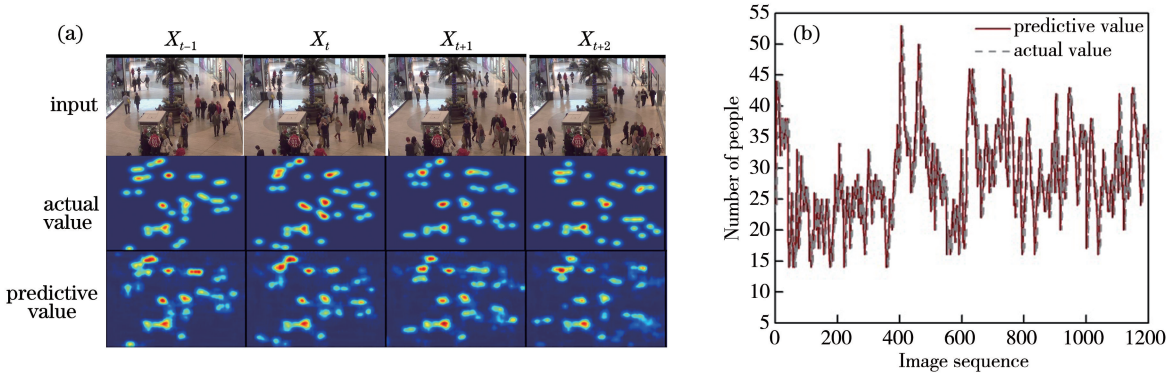

图 5　本模型在 Mall 数据集上的实验结果。(a)密度图；(b)计数结果

Fig. 5　Experimental results of our model on the Mall data set. (a) Density map; (b) counting result

从图 5 和表 2 可以发现，相比 ConvLSTM 模型，本模型在 Mall 公开数据集中的 f_{MAE} 和 f_{MSE} 分别降低了 12.95%和 11.76%，这表明本模型对室内复杂的场景人群计数精度较高，且鲁棒性较好。

3.2.4　自建视频数据集

目前人们的短距离和较长距离出行离不开地铁和铁路，但现有数据集很少关注这两个方面。为满足智慧交通的要求，需在重要节假日人群密集高峰

表 2　不同模型在 Mall 数据集上的性能指标

Table 2　Performance indexes of different models on the Mall data set

Model	f_{MAE}	f_{MSE}
ConvLSTM[3]	2.24	8.50
Bidirectional ConvLSTM[3]	2.10	7.60
Ref. [20]	3.59	19.00
Ref. [21]	3.43	17.70
Ours	1.95	7.50

期对入站口和出站口的人群进行计数。实验建立的数据集包括兰州地铁入站口及西站上车过程监控视频中的图像，包括不同场景、不同光照、不同角度共 1200 帧图像。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，手动标注每帧图像中人头的坐标点，训练网络分别选取 900 帧图像（每个场景各 450 帧），其余 300 帧（每个场景各 150 帧）图像用于测试。本模型在自建数据集中的测试结果如图 6 所示，不同模型的对比结果如表 3 所示。可以发现，本模型的 f_{MAE} 比多列卷积神经网络（MCNN）降低了 7.87%，而 f_{MSE} 比没有添加 ST 模块时下降了 2.11%。当视频图像中存在遮挡的物体时，本模型的计数精度有所下降，原因是网络对密集遮挡人群的图像训练较少，没有学习到足够多的特征，因此，还需进一步改进。

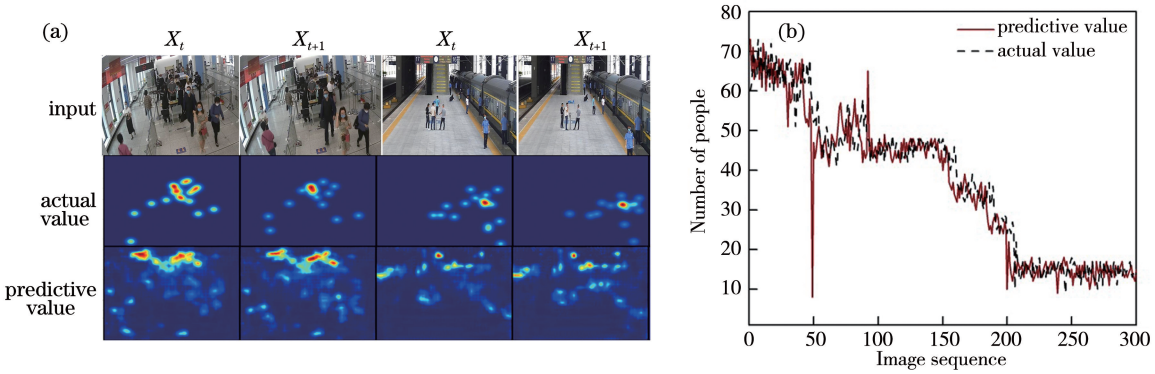

图 6　本模型在自建数据集上的实验结果。(a)密度图；(b)计数结果

Fig. 6　Experimental results of our model on the self-built data set. (a) Density map; (b) counting result

表 3　不同模型在自建数据集上的性能指标

Table 3　Performance indexes of different models on self-built data set

Model	f_{MAE}	f_{MSE}
ConvLSTM[3]	4.51	5.91
MCNN[22]	3.81	4.92
Ours without ST	4.32	5.21
Ours	3.51	5.10

3.3　模型对比实验及分析

1）引入空洞卷积前后网络的性能

为验证在单列深度 FCN 中加入空洞卷积对视频图像人群计数精度的提升，将该网络与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（dilated＋No ST）网络进行对比，测试实验分别在 UCSD、Mall 和自建数据集上进行，测试帧的选择与 3.2 节一致，测试结果如表 4 所示。可以发现，在网络前端加入空洞卷积后，相比原始 FCN 在 3 个数据集上的 f_{MAE} 和 f_{MSE} 均有显著提升，原因是空洞卷积在不增加参数量的情况下能有效扩大感受野，验证了引入空洞卷积可提高密度图生成模块的特征提取能力。

表 4　不同数据集的验证性实验结果 1

Table 4　Confirmation experiment results1 of different data sets

Data set	No dilated No ST		dilated＋No ST	
	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	4.74	6.65	4.32	5.21

2）引入时空变换计数模块后网络的性能

为验证在 LSTM 网络中加入 ST 模块对视频图像人群计数精度的提升，将无空洞卷积、无时空变换模块、无空洞卷积与加入时空变换模块（No dilated＋ST）的网络和加入空洞卷积并结合时空变换模块的网络（dilated＋ST）进行对比，在 UCSD、Mall 和自建数据集上的测试结果如表 5 所示。可

以发现，相比基础网络，引入无空洞卷积与 ST 模块的网络在 f_{MAE} 和 f_{MSE} 性能上有明显提高，验证了加入 ST 模块可提升网络计数结果的准确性。对比加入无空洞卷积、加入 ST 模块的网络与加入两者的联合网络发现，仅有 ST 模块的网络计数性能比联合网络效果差，但比仅有空洞卷积的网络性能高，这表明 ST 模块对精度提升的贡献比空洞卷积大。

表 5　不同数据集的验证性实验结果 2

Table 5　Confirmation experiment results 2 of different data sets

Data set	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	4.74	6.65	4.33	5.23	3.51	5.10

为更直观检测 ST 模块对计数任务的影响，根据 Mall 数据集人群密度变化大及疏密分布比较平衡的特点，选用 Mall 数据集 801～1200 帧作为测试集验证改进网络的准确性，得到 50 次迭代后模型的准确率如图 7 所示，可以发现，改进前后的网络准确率都在 70%以上，没有加入 ST 模块的网络准确率为 80%～90%，改进后网络的准确率均在 90%以上。

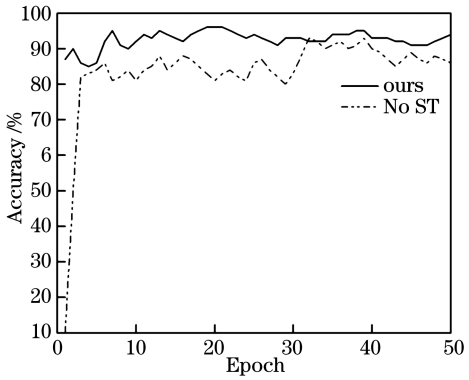

图 7　不同模型的准确率

Fig. 7　Accuracies of different models

3）改进模型前后的训练损失值

图 8 为改进网络前后损失函数的收敛情况，可以发现，改进后的网络收敛性得到了很大提高。本网络模型在训练前期损失曲线波动较大，原因是前期 FCN 的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，原始网络模型更容易发生过拟合现象，而改进模型的表现更稳定。

综上所述，引入空洞卷积、结合 ST 模块与 LSTM 网络均可提升人群计数任务的精度，且 ST 模块对网络性能的提升更明显，其次为空洞卷积，将两者与整体网络用连接方式结合时效果最佳。

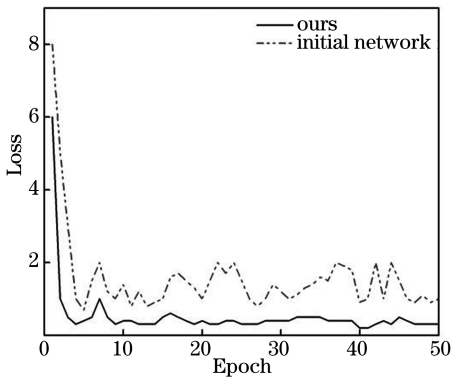

图 8　改进前后网络的训练损失曲线

Fig. 8　Training loss curves of the network before and after the improvement

4　结　　论

提出了一种基于单列深度时空计数 CNN 的人群计数模型，整体网络结构分为密度图生成和时空变换计数两个模块；针对单列网络提取的图像特征不完整导致人群计数不精确的问题，通过引入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；针对摄像机视角畸变导致的难以识别问题，将 ST 模块插入 LSTM 网络中。在 3 个数据集上的测试结果表明，相比其他模型，本模型的测试检测速度和检测精度均较好，验证了本模型的实际应用价值。但本模型对于存在遮挡和密集人群的计数，还存在明显的不足，今后还需通过实验进行改进。

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] Marsden M, McGuinness K, Little S, et al. ResnetCrowd: a residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification[C]//2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), August 29-September 1, 2017, Lecce, Italy. New York: IEEE Press, 2017: 1-7.

[3] Xiong F, Shi X J, Yeung D Y. Spatiotemporal modeling for crowd counting in videos[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 5161-5169.

[4] Yang C, Gong H F, Zhu S C, et al. Flow mosaicking: real-time pedestrian counting without scene-specific learning[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE Press, 2009: 1093-1100.

[5] Li B, Zhang J, Zhang Z, et al. A people counting method based on head detection and tracking[C]//2014 International Conference on Smart Computing, November 3-5, 2014, Hong Kong, China. New York: IEEE Press, 2014: 136-141.

[6] Fan C C. Research on density estimation and crowd counting algorithms based on convolutional neural network[D]. Hefei: 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] Wen Q, Jia C, Yu Y, et al. People number estimation in the crowded scenes using texture analysis based on gabor filter[J]. Journal of Computational Information Systems, 2011, 7(11): 3754-3763.

[8] Liu T L, Tao D C. On the robustness and generalization of Cauchy regression[C]//2014 4th IEEE International Conference on Information Science and Technology, April 26-28, 2014, Shenzhen, China. New York: IEEE Press, 2014: 100-105.

[9] Lempitsky V, Zisserman A. Learning to count objects in images[C]//Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, December 6-9, 2010, British Columbia, Canada. New York: Curran Associates Inc. 2010: 1324-1332.

[10] Zhang S H, Wu G H, Costeira J P, et al. FCN-rLSTM: deep spatio-temporal neural networks for vehicle counting in city cameras[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice. New York: IEEE Press, 2017: 3687-3696.

[11] Li Y H, Zhang X F, Chen D M. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 1091-1100.

[12] Zuo J, Ba Y L. Population-depth counting algorithm based on multiscale fusion[J]. Laser & Optoelectronics Progress, 2020, 57(24): 241502.
左静, 巴玉林. 基于多尺度融合的深度人群计数算法[J]. 激光与光电子学进展, 2020, 57(24): 241502.

[13] Wu Z H, Gao Y M, Li L, et al. Fully convolutional network method of semantic segmentation of class imbalance remote sensing images[J]. Acta Optica Sinica, 2019, 39(4): 0428004.
吴止锾, 高永明, 李磊, 等. 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(4): 0428004.

[14] Marsden M, McGuinness K, Little S, et al. Fully convolutional crowd counting on highly congested scenes[EB/OL]. [2020-07-05]. https://arxiv.org/abs/1612.00220.

[15] Gao L, Song W D, Tan H, et al. Cloud detection based on multi-scale dilation convolutional neural network for ZY-3 satellite remote sensing imagery[J]. Acta Optica Sinica, 2019, 39(1): 0104002.
高琳, 宋伟东, 谭海, 等. 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(1): 0104002.

[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] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[EB/OL]. [2020-07-02]. https://arxiv.org/abs/1506.02025.

[18] Greff K, Srivastava R K, Koutník J, et al. LSTM: a search space odyssey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(10): 2222-2232.

[19] Chan A B, Liang Z S John, Vasconcelos N. Privacy preserving crowd monitoring: counting people without people models or tracking[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2008, Anchorage, AK, USA. New York: IEEE Press, 2008: 1-7.

[20] Chen K, Loy C C, Gong S G, et al. Feature mining for localised crowd counting[C]//Procedings of the

British Machine Vision Conference 2012, September 3-7, 2012, Surrey, UK. Guildford: BMVA Press, 2012: 1-11.

[21] Chen K, Gong S G, Xiang T, et al. Cumulative attribute space for age and crowd density estimation [C]//2013 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. New York: IEEE Press, 2013: 2467-2474.

[22] Zhang Y Y, Zhou D S, Chen S Q, et al. Single-image crowd counting via multi-column convolutional neural network[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 589-597.