激光与光电子学进展 先进成像

Involution改进的卷积神经网络人群计数方法

李兆鑫，卢树华*，兰凌强，刘淇缘
中国人民公安大学信息网络安全学院，北京 102600

摘要 针对现有人群计数方法大多采用卷积操作提取特征，空间多样性特征信息提取和传递能力不足的问题，提出一种Involution改进的单列深层人群计数网络。该网络以VGG-16为基本框架，引入Involution算子替代卷积操作，并辅以残差链接提高对空间特征信息的感知和传递能力；采用膨胀卷积保持分辨率的同时扩大感受野，丰富深度语义特征；利用联合损失函数监督网络训练，提高计数准确性和全局信息相关性。所提方法在公开数据集ShangHaiTech、UCF-QNRF、UCF_CC_50上的性能较基线模型提升显著，并超越了诸多当前的先进算法。结果表明：所提人群计数方法具有较高的准确性和更好的鲁棒性。

关键词 人群计数；Involution算子；膨胀卷积；全局损失
中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP202259.1815004

Convolutional Neural Network Method for Crowd Counting Improved using Involution Operator

Li Zhaoxin, Lu Shuhua*, Lan Lingqiang, Liu Qiyuan
College of Information and Cyber Security, People's Public Security University of China, Beijing 102600, *China*

Abstract Most existing crowd counting methods use convolution operations to extract features. However, extracting and transmitting spatial diversity feature information are difficult. In this paper, we propose an Involution-improved single-column deep crowd-counting network to mitigate these problems. Using VGG-16 as the backbone, the proposed network uses an Involution operator combined with residual connection to replace the convolution operation, thereby enhancing the perception and transmission for spatial feature information. The dilated convolution was adopted to expand the receptive field while maintaining resolution to enrich deep semantic features. Additionally, we used the joint loss function to supervise the network training, improving counting accuracy and global information correlation. Compared with the baseline model, the performance of the proposed method across the ShangHaiTech, UCF-QNRF, and UCF_CC_50 datasets considerably is improved, demonstrating that our approach outperforms many current advanced algorithms. Furthermore, results show that the proposed crowd counting method has higher accuracy and better robustness than other methods.

Key words crowd counting; Involution operator; dilated convolution; global loss

1 引 言

人群计数在智能视频监控、城市空间规划以及交通流量控制等领域应用前景广阔，引起高度重视，成为计算机视觉领域研究的热点之一[1-3]。近年来，随着深度学习技术的快速发展，基于卷积神经网络(CNN)的人群计数方法取得显著进展[4-5]，此类方法一般通过将图片输入到卷积神经网络自动提取人群特征，经过处理得到人群密度图，再求和密度图的像素点得到人数。相比于传统的基于检测[6-7]或回归[8-9]的计数方法，基于卷积神经网络的人群计数方法准确性和鲁棒性显著提升。然而，现实场景中的人群计数仍然面临尺度变化、严重遮挡、空间分布不均和复杂背景干扰等问题[3,10-11]，导致准确率和鲁棒性受到较大限制。针对上述问题，现有方法大多通过加深网络深度来提取高层语义信息[12-14]、通过拓展网络宽度(多分支网络结构)来提高模型的多尺度特征感知能力[2-3,15]、通过引入空洞卷积来扩大感受野[16-19]、融合注意力机制来抑制背

收稿日期：2021-06-21；**修回日期**：2021-07-12；**录用日期**：2021-07-20
基金项目：公安学科基础理论研究专项(2021XKZX08)、中央高校基本科研业务经费重大项目(2021JKF102)
通信作者：*lushuhua@ppsuc.edu.cn

景干扰[2,11,17,20]等。这些方法广泛采用一般卷积算子提取特征，具有空间不变性和通道特异性，网络缺乏对空间特异性的感知能力，且存在通道冗余。

为提高模型的空间信息感知和传递能力，受文献[4,16,21-22]启发，本文提出一种 Involution 改进的 CNN 人群计数方法。其中，在网络前端，使用 VGG-16 的前 10 层提取浅层特征，然后引入 Involution 算子[21]替代卷积操作，并利用残差链接缓解梯度弥散和性能退化问题；在网络后端，使用膨胀卷积来生成密度图，以扩大感受野，提高密度图的质量；此外，利用联合损失函数来监督模型训练。在公开数据集上对所提方法进行了验证和测试，其准确率超过了诸多当前的先进算法，展现出较强的竞争力。

2 相关工作

根据发展阶段，人群计数方法大致分为两类，一类是传统方法，另一类是基于卷积神经网络的方法。

2.1 传统方法

早期人群计数工作中大多使用基于检测的方法，此类方法[6-7]使用检测器对一张图片中的行人整体或者行人身体部位进行检测，然后对检测到的数量进行累加求和，得出的结果作为该张图片总人数。还有部分学者使用基于回归的方法[8-9]进行人群计数，这类方法建立图像特征和人数的回归模型，通过提取特征估计场景中的人数。

2.2 基于卷积神经网络的方法

与传统人群计数方法不同，卷积神经网络实现了端到端的训练，在准确率和鲁棒性方面均取得了较大提升，根据网络结构，可以分为单列和多列（多分支）卷积神经网络的方法。

CSRNet[16]是单列卷积神经网络方法的代表之一，前端采用 VGG-16 前 10 层提取特征，后端采用膨胀卷积层来扩大感受野，同时保持分辨率，提升密度图的质量，该网络结构简单，但是对尺度变化信息感知具有一定局限。为了提高特征尺度连续性和信息传递能力，Dai 等[12]提出了一种单列深层的计数网络（DSNet），该网络主要由 3 个密集膨胀卷积块构成。在每个卷积块的内部都包含若干个膨胀率不同的膨胀卷积层，彼此之间通过残差链接相连。此外，他们引入了多尺度密度水平一致性损失以克服特征相似性问题，提高模型性能。Zhang 等[23]提出一种尺度自适应网络（SaCNN），以解决尺度变化和透视失真问题。Wang 等[20]提出一种空间上下文学习网络（SCLNet）处理拥挤场景的人群计数问题，该网络在人群计数和定位方面有较高竞争力。单列卷积神经网络方法简单，但对大尺度变化问题应对有限。

针对图片中尺度变化这一挑战问题，Zhang 等[24]提出了 3 分支计数网络（MCNN），该网络从人群图片中提取不同感受野特征，为后续设计多尺度感知网络提供了良好的借鉴。为提取不同层次的互补尺度信息，Zeng 等[13]提出了一种多分支人群计数网络（DSPNet），前端采用卷积神经网络提取浅层特征，而网络后端采用“最大比率组合”策略。此外，DSPNet 进行基于 RGB 图像的推理，易于模型学习特征，减少了上下文信息的丢失。受多任务学习启发，Zhu 等[2]提出了一种多任务多分支人群计数网络（AMCNN），该网络包含分级密度估计器（HDE）和辅助计数分类器（AUCC）。HDE 采用分层策略以由粗到细的方式挖掘语义特征，以解决尺度变化和视角失真的问题；AUCC 则被用来实现计数分类任务，是对密度估计的补充。Miao 等[25]提出了一种端到端的深度人群计数网络（ST-CNN），该网络将 2D 卷积神经网络（C2D）和三维卷积神经网络（C3D）统一在同一个框架下学习时空特征，此外在生成的密度图上执行合并方案，同时利用时空信息进行人群计数任务。尽管上述多列/多分支网络可提取多尺度信息，但是结构较为复杂，且存在相似尺度信息冗余。

3 人群计数方法

3.1 网络架构

引入 Involution 算子，提出一种单列深层人群计数网络，结构如图 1 所示，其中 VGG-16 代表特征提取器，INV 代表 Involution 核，半椭圆线代表残差链接，特征图后的数字表示通道数。网络前端采用 VGG-16 前 10 层作为浅层特征提取器；在网络中部采用 3 个 Involution 算子和残差链接以提高对空间多样性特征信息的提取和传递能力；在后端使用膨胀卷积扩大感受野，并利用联合损失函数监督模型训练；最后使用 1×1 的卷积核对特征图进行降维，输出密度图，完成计数。

3.2 Involution 算子

3.2.1 基本性质

现有主流人群计数算法广泛采用卷积操作，卷积具有两个性质：空间无关性和通道特异性。卷积的两个性质在平衡参数和性能之间有着重要意义，但是空间不变性也一定程度上减弱了卷积核对空间信息的学习能力。为了尽量减少参数，一般卷积核的取值都较小，限制了其捕捉大范围上下文信息的能力。此外，卷积核在通道维度上存在冗余，会增加运算成本。

Involution 与卷积不同，在通道维度上共享参数，在空间维度上则不相同，即具有空间特异性和通道无关性。Involution 核与卷积核亦不相同，Involution 核是基于单个像素的而不是其与相邻像素的关系。具体地说，Involution 核在空间范围上是不同的，但在通道上是共享的，这样设计使得 Involution 核可以在更大范围上捕捉上下文信息，而且共享通道也一定程度上减少了核冗余问题。

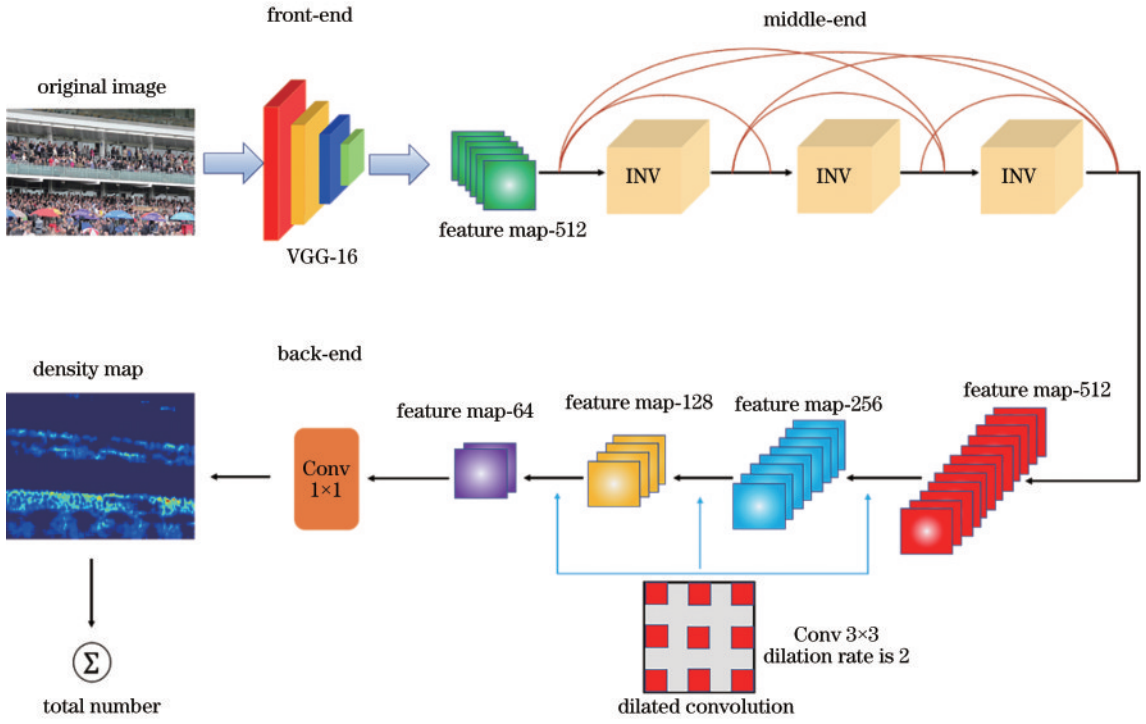

图 1　人群计数网络结构

Fig. 1　Crowd counting network structure

3.2.2　Involution 核的生成和工作原理

对于一组输入和输出的特征图，可以分别将其表示为 $\boldsymbol{F}\in\mathbb{R}^{C_{\text{in}}\times H\times W}$ 和 $\boldsymbol{F}_{\text{out}}\in\mathbb{R}^{C_{\text{out}}\times H\times W}$，其中 C_{in} 表示输入特征图通道数，C_{out} 表示输出特征图通道数，H 和 W 分别表示特征图的高度和宽度。由于卷积的通道特异性，将作用于这组特征图的 C_{out} 组卷积核表示为 $\boldsymbol{C}\in\mathbb{R}^{C_{\text{in}}\times C_{\text{out}}\times K\times K}$，$K$ 为卷积核的大小，每组卷积核对输入特征图进行处理后，会生成相应的输出 $\boldsymbol{F}'_c\in\mathbb{R}^{H\times W}$，$c=1,2,\cdots,C_{\text{in}}$。把这些输出整合，可以得到该组卷积核的输出，然后将 C_{out} 组卷积核分别作用于输入特征图，得到输入特征图 $\boldsymbol{F}_{\text{out}}\in\mathbb{R}^{C_{\text{out}}\times H\times W}$。

Involution 在设计上则与卷积相反，对于一组输入特征图，将通道数 C 分为 G 组，在同一组上保持通道无关性，而在同一组内的不同空间位置上使用不同的卷积核，即空间特异性，可将从某个像素生成 Involution 核的过程表示为

$$\boldsymbol{I}_{i,j}=\phi(\boldsymbol{F}_{i,j})=\boldsymbol{W}_1\sigma(\boldsymbol{W}_0\boldsymbol{F}_{i,j}),\tag{1}$$

式中：$\phi(\cdot)$ 为 Involution 核的生成函数，由两个线性变化 $\boldsymbol{W}_0\in\mathbb{R}^{\frac{C}{r}\times C}$ 和 $\boldsymbol{W}_1\in\mathbb{R}^{(K\times K\times G)\times\frac{C}{r}}$ 构成，r 是通道缩减比率；$\sigma(\cdot)$ 是中间的批归一化（BN）和 ReLU 函数。对于输入特征图 $\boldsymbol{F}$ 某一像素上的特征向量 $\boldsymbol{F}_{i,j}\in\mathbb{R}^{1\times 1\times C}$，通过 $\phi\left(\boldsymbol{W}_0-\boldsymbol{F}_{\text{BN}}-\boldsymbol{F}_{\text{ReLU}}-\boldsymbol{W}_1\right)$ 得到特征向量 $\boldsymbol{F}'_{i,j}\in\mathbb{R}^{K\times K\times G}$，$K$ 为 Involution 核的大小；再将其通过 reshape 展开成核的形状，即可得到此像素点上的 Involution 核；然后将输入特征图上这个坐标点邻域的特征向量与其进行乘加运算，得到最终输出的特征图 $\boldsymbol{F}_{\text{out}}$。图 2 展示了这一过程（此处取 $G=1$）。相较于一般卷积，Involution 可以在更大的空间范围总结上下文信息，而且可以在不同空间位置自适应地分配权重，从而加强了模型对空间域中信息的特征提取能力，一定程度上提升了密度图的质量。

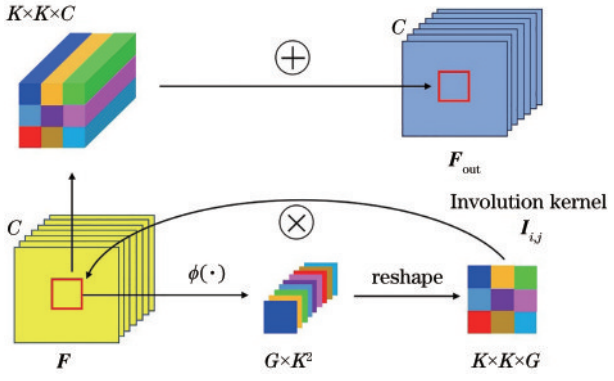

图 2　Involution 核的生成和工作原理

Fig. 2　Generation and working principle of Involution kernel

3.3　损失函数

现有的人群计数算法[16-17,20,24,26]大多采用欧氏距离计算密度图和标签密度图之间各像素点的误差。这一欧氏距离损失函数只考虑像素误差，忽略了密度图和标签密度图之间的全局上下文信息相关性，也忽略了计数性能。因此，受文献[27]的启发，引入欧氏距离 L_e 和全局损失 L_c 两种损失函数，通过设置不同权重进

行联合优化，用来衡量密度图和标签密度图之间的全局上下文信息相关性，损失函数表达式分别为

$$L_e(\Theta)\frac{1}{2N}\sum_{i=1}^{N}\left\| F(\boldsymbol{X}_i,\Theta)-\boldsymbol{F}_i\right\|_2^2, \tag{2}$$

$$L_c(\Theta)=\frac{1}{N}\sum_{i=1}^{N}\left| P^{\text{avg}}\left[F(\boldsymbol{X}_i,\Theta)\right]-P^{\text{avg}}\left(\boldsymbol{F}_i\right)\right|, \tag{3}$$

式中：N是训练集图像的数量；$\boldsymbol{X}_i$和$\boldsymbol{F}_i$分别代表第i张图片和第i张图片对应的标签密度图；Θ表示模型的参数；$F(\boldsymbol{X}_i,\Theta)$代表模型使用参数$\Theta$得到的密度图；$P^{\text{avg}}(\cdot)$代表平均池化操作，输出大小设为1。总损失函数为两部分损失的加权和，表达式为

$$L(\Theta)=L_e(\Theta)+\alpha L_c(\Theta)。 \tag{4}$$

为了确定系数α，分别进行了多次实验，最终确定系数取值，如表1所示。

表1 不同数据集中系数α的取值

Table 1 Value of coefficient α in different datasets

Dataset	α
SHHA	1000
SHHB	100
UCF-QNRF	1000
UCF_CC_50	100

4 实验与结果分析

4.1 标签密度图的生成

所提算法采用监督学习的方法进行训练，用到的标签是包含了每一张输入图像真实人群分布的真值密度图，也叫标签密度图，标签密度图质量的好坏一定程度上影响模型的性能。基于Zhang等[24]的工作，本文采用自适应高斯核的方法生成标签密度图。该方法可以根据当前的尺度自适应地改变高斯核的大小，以在高度拥挤的场景下能得到效果较好的人群密度图，提高模型的精度。假设图片中一个人头的像素点表示为$\delta(\boldsymbol{x}-\boldsymbol{x}_m)$，可以将有$M$个人的标签密度图表示为标记图像和高斯核$\boldsymbol{G}_{\sigma_m}$的卷积：

$$F(\boldsymbol{x})=H(\boldsymbol{x})=\sum_{m=1}^{M}\delta(\boldsymbol{x}-\boldsymbol{x}_m)*\boldsymbol{G}_{\sigma_m},\ \sigma_m=\beta\bar{d}_m, \tag{5}$$

式中：σ_m表示高斯核的大小，其值由β和$\bar{d}_m$共同决定；β为一常数，由实验确定；$\bar{d}_m$表示某图像中某一个头部和其周围k个头部之间的平均距离，将这一距离作为决定高斯核大小的依据。取$\beta=0.3$，$k=3$。

4.2 评价标准

为了定量评价算法的准确性和稳定性，采用平均绝对误差(MAE)和均方误差(MSE)作为衡量预测结果的指标。算法的准确性由MAE评价，算法的稳定性由MSE评价。MAE和MSE的定义分别为

$$E_{\text{MA}}=\frac{1}{N}\sum_{i=1}^{N}\left| C_i^{\text{ET}}-C_i^{\text{GT}}\right|, \tag{6}$$

$$E_{\text{MS}}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left| C_i^{\text{ET}}-C_i^{\text{GT}}\right|^2}, \tag{7}$$

式中：C_i^{ET}为算法估计出的第i张图片的行人个数；C_i^{GT}为第i张图片中真实行人个数，由标签密度图像素点求和得到。

4.3 实验配置

实验是在Ubuntu系统下进行的，训练使用的GPU为Nvidia Tesla P100。迭代次数为1000，Batchsize设为1。使用Adam优化器对学习率进行动态调整，学习率初始值设置为0.00001，在400轮迭代之后每两个迭代学习率调整一次，衰减率为0.995。在对图像进行预处理时修改了图像的大小，在保持原本图像比例的条件下，将图像的宽和高限制为不超过1024，且保证能被16整除。

4.4 数据集

ShangHaiTech：ShangHaiTech(SHH)数据集分为PartA(SHHA)和PartB(SHHB)两个部分[24]，共1198张图片，330165个注释头。其中A部分由482张图片构成，训练集300张，测试集182张。B部分共有716张图片，其中训练集400张，测试集316张。

UCF-QNRF：UCF-QNRF由1535张分辨率为2013×2902的图片组成[28]，其包含有1251642个标有注释点的人，单张图片最大人数可达12865。实验中使用其中1201张图片作为训练集对模型进行训练，使用334张图片作为测试集对模型进行测试。

UCF_CC_50：UCF_CC_50数据集[29]是一个十分具有挑战性的数据集，虽然照片总数只有50，但是分辨率各不相同，并且标记总人数高达63075。该数据集场景丰富多样，音乐会、体育场等都包含其中。在实验中随机抽取其中40张图片进行训练，10张图片进行验证。由于图片数量的限制，还采用五折交叉验证对所提模型进行评估。

4.5 结果分析

为验证所提方法的有效性，在公开的大规模人群计数数据集ShanghaiTech、UCF-QNRF及UCF_CC_50上进行了训练和测试，并与当前诸多先进算法进行了对比，结果如表2所示。从表2可以看出，所提方法在3个公开数据集上的MAE和MSE均表现出较强的竞争力，其中在SHHB稀疏场景和UCF_CC_50密集场景数据集上，均取得最优性能，表明所提方法泛化性较好。与部分经典的单列CNN算法如CSRNet[16]、DSPNet[13]、SCLNet[20]、TEDNet[30]等相比，所提方法的两个评价指标表现优异；与部分多列多分支计数算法如MCNN[24]、CMTL[31]、Switching CNN[26]、AMCNN[2]、PCCNet[10]等相比，所提方法在准确性和鲁棒性方面亦有显著提高。这可归因于Involution算子能使模型更灵活地提取空间多样性特征，且自注意力属性能够抑制复杂背景干扰。需要指出的是，在解决空间尺度变化问题时，现有方法

表 2　不同方法的结果对比

Table 2　Comparison results of the different methods

Method	SHHA		SHHB		UCF-QNRF		UCF_CC_50	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN[24]	110.2	173.2	26.4	41.3	243.5	364.7	467.0	498.5
Switching CNN[26]	90.4	135.0	21.6	33.4	228.0	445.0	318.1	439.2
CMTL[31]	101.3	152.4	20.0	31.1	252	514	322.8	397.9
CSRNet[16]	68.2	115.0	10.6	16.0	120.3	208.5	266.1	397.5
SANet[32]	67.0	104.5	8.4	13.6	—	—	—	—
ACSPNet[18]	85.2	137.1	15.4	23.1	—	—	—	—
PCCNet[10]	73.5	124.0	11.0	19.0	149.0	247.0	240.0	315.5
AMCNN[2]	76.1	110.7	15.3	27.4	—	—	—	—
LSC-CNN[33]	66.4	117.0	8.1	12.7	120.5	218.2	255.6	302.7
TEDNet[30]	64.2	109.1	8.2	12.8	113.0	188.0	249.4	354.5
SCLNet[20]	67.9	102.9	9.1	14.1	109.6	182.5	258.9	326.2
DSPNet[13]	68.2	107.8	8.9	14.0	107.5	182.7	243.3	307.6
MSCANet[34]	66.5	109.4	—	—	104.1	183.8	242.8	329.8
Method in Ref. [15]	61.9	100.5	7.4	11.7	104.8	182.3	212.3	289.6
AMS-Net[11]	63.8	108.5	7.3	11.8	86.5	167.2	236.5	319.2
Proposed method	61.1	101.3	7.0	11.3	102.5	181.7	202.0	288.7

大多采用多列或多分支网络，尽管准确率较高，但是网络结构较为复杂，所提方法与其他方法相比，在计数性能和效率方面均表现优越，且网络结构更简捷，验证了所提方法的先进性。图 3 为所提方法生成的密度图实例，结果表明所提方法在密集人群和复杂背景干扰下，预测结果与真实值较为一致。

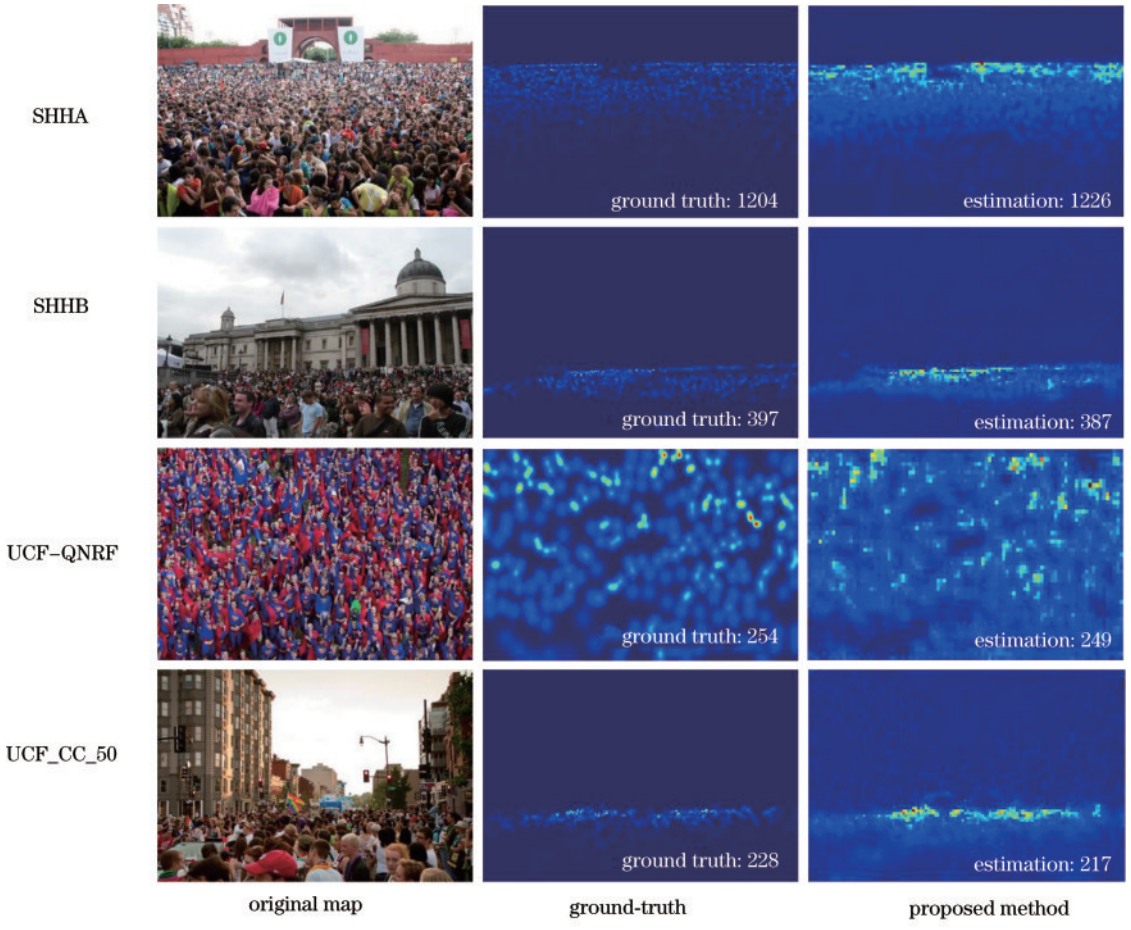

图 3　密度图生成实例

Fig. 3　Examples of density map generation

4.6　消融实验

为了验证网络每个组分的有效性，在SHHA数据集上进行了消融实验，结果如表3所示，其中VGG代表网络前端所采用的前10层VGG-16网络，INV代表Involution层。可以看出，在其他条件不变的情况下，通过逐步增加Involution算子数，调整膨胀率，引入残差链接和全局损失函数，模型的性能不断提高，MAE和MSE分别降至61.1和101.3，较基线模型分别减小了11.6%和4.6%。此外，为了衡量所提方法的时间和空间复杂度，亦对单张图片的测试时间和模型的参数量分别进行了计算，结果如表3所示，可以看出，引入Involution算子后会增加模型参数量，但是能够显著缩短测试时间。为了直观展现出所提方法各部分的有效性，对消融实验进行了可视化，结果如图4所示，随着各组分的增加，密度图效果也逐渐提升，计数误差逐步下降。

表3　在SHHA数据集上的消融实验研究

Table 3　Ablation study on the SHHA dataset

Component	MAE	MSE	Time /ms	Parameters /10^6
VGG (dilation rate is 2) (baseline)	69.1	106.2	166.54	11.54
VGG+1INV(dilation rate is 2)	66.9	105.2	145.38	11.81
VGG+2INV(dilation rate is 2)	64.9	101.4	133.85	12.08
VGG+3INV(dilation rate is 1)	67.0	108.5	132.20	12.35
VGG+3INV(dilation rate is 2)	63.6	103.6	127.53	12.35
VGG+3INV(dilation rate is 2)+residual connection	62.9	106.5	133.63	12.35
VGG+3INV(dilation rate is 2)+residual connection+Loss	61.1	101.3	136.43	12.35

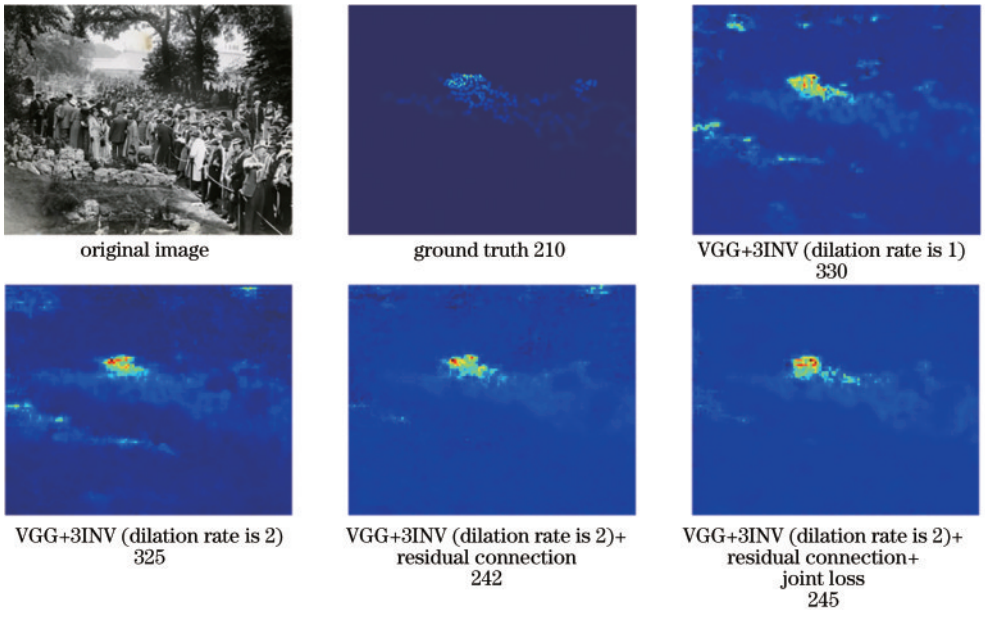

图4　消融实验实例

Fig. 4　Example of the ablation experiment

5　结　　论

提出一种基于Involution改进的人群计数方法，引入Involution算子替代普通的卷积操作，使网络可以在较大的空间范围中捕获多样性信息，加强了空间特征提取能力；在每个Involution层之间，采用残差链接的方式提高信息传递能力。此外，引入了全局损失与欧氏损失联合优化策略监督模型训练，提高了预测密度图和标签密度图之间的上下文信息相关性。在3个公开数据集上进行了验证和测试，结果表明，与其他主流算法相比，所提方法有较好的准确性和鲁棒性，是一种具有较强竞争力的人群计数算法。

参　考　文　献

[1] Zhu F S, Yan H, Chen X Y, et al. A multi-scale and multi-level feature aggregation network for crowd counting[J]. Neurocomputing, 2021, 423: 46-56.

[2] Zhu M, Wang X Q, Tang J, et al. Attentive multi-stage convolutional neural network for crowd counting[J]. Pattern Recognition Letters, 2020, 135: 279-285.

[3] 余鹰, 朱慧琳, 钱进, 等. 基于深度学习的人群计数研究综述[J]. 计算机研究与发展, 2021, 58(12): 2724-2747.
Yu Y, Zhu H L, Qian J, et al. Survey on deep learning

based crowd counting[J]. Journal of Computer Research and Development, 2021, 58(12): 2724-2747.

[4] 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.

[5] 蒋妮, 周海洋, 余飞鸿. 基于计算机视觉的目标计数方法综述[J]. 激光与光电子学进展, 2021, 58(14): 1400002.
Jiang N, Zhou H Y, Yu F H. Review of computer vision based object counting methods[J]. Laser & Optoelectronics Progress, 2021, 58(14): 1400002.

[6] Dollár 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, 2012, 34(4): 743-761.

[7] Topkaya I S, Erdogan H, Porikli F. Counting people by clustering person detector outputs[C]//2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance, August 26-29, 2014, Seoul, Korea (South). New York: IEEE Press, 2014: 313-318.

[8] Chan A B, Liang Z S J, Vasconcelos N, et al. 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: 10139874.

[9] Chan A B, Vasconcelos N. Bayesian Poisson regression for crowd counting[C]//2009 IEEE 12th International Conference on Computer Vision, September 29-October 2, 2009, Kyoto, Japan. New York: IEEE Press, 2009: 545-551.

[10] Gao J Y, Wang Q, Li X L. PCC net: perspective crowd counting via spatial convolutional network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(10): 3486-3498.

[11] Zhang B, Wang N Y, Zhao Z, et al. Crowd counting based on attention-guided multi-scale fusion networks[J]. Neurocomputing, 2021, 451: 12-24.

[12] Dai F, Liu H, Ma Y, et al. Dense scale network for crowd counting[EB/OL]. (2019-06-24) [2021-06-17]. https://arxiv.org/abs/1612.00220.

[13] Zeng X, Wu Y P, Hu S Z, et al. DSPNet: deep scale purifier network for dense crowd counting[J]. Expert Systems With Applications, 2020, 141: 112977.

[14] 鱼春燕, 徐岩, 缑丽莎, 等. 基于单列深度时空卷积神经网络的人群计数[J]. 激光与光电子学进展, 2021, 58(8): 0810011.
Yu C Y, Xu Y, Gou L S, et al. Crowd counting based on single-column deep spatiotemporal convolutional neural network[J]. Laser & Optoelectronics Progress, 2021, 58(8): 0810011.

[15] Wang Y J, Zhang W, Huang D X, et al. Multi-scale supervised network for crowd counting[J]. IET Image Processing, 2020, 14(17): 4701-4707.

[16] 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.

[17] Guo D, Li K, Zha Z J, et al. DADNet: dilated-attention-deformable ConvNet for crowd counting[C]//Proceedings of the 27th ACM International Conference on Multimedia, October 21-25, 2019, Nice, France. New York: IEEE Press, 2019: 1823-1832.

[18] Ma J J, Dai Y P, Tan Y P. Atrous convolutions spatial pyramid network for crowd counting and density estimation[J]. Neurocomputing, 2019, 350: 91-101.

[19] 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.
Zuo J, Ba Y L. Population-depth counting algorithm based on multiscale fusion[J]. Laser & Optoelectronics Progress, 2020, 57(24): 241502.

[20] Wang S Z, Lu Y, Zhou T F, et al. SCLNet: spatial context learning network for congested crowd counting [J]. Neurocomputing, 2020, 404: 227-239.

[21] Li D, Hu J, Wang C, et al. Involution: Inverting the inherence of convolution for visual recognition[EB/OL]. (2021-04-11)[2021-06-17]. https://arxiv.org/abs/2103.06255.

[22] Li J Q, Yan H. Crowd counting method based on cross column features fusion[J]. Computer Science, 2021, 48 (6): 118-124.

[23] Zhang L, Shi M J, Chen Q B. Crowd counting via scale-adaptive convolutional neural network[C]//2018 IEEE Winter Conference on Applications of Computer Vision, March 12-15, 2018, Lake Tahoe, NV, USA. New York: IEEE Press, 2018: 1113-1121.

[24] 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, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 589-597.

[25] Miao Y Q, Han J G, Gao Y S, et al. ST-CNN: spatial-temporal convolutional neural network for crowd counting in videos[J]. Pattern Recognition Letters, 2019, 125: 113-118.

[26] Sam D B, Surya S, Babu R V. Switching convolutional neural network for crowd counting[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 4031-4039.

[27] Cheng J, Chen Z, Zhang X Y, et al. Exploit the potential of multi-column architecture for crowd counting [EB/OL]. (2020-07-28)[2021-06-17]. https://arxiv.org/abs/2007.05779.

[28] Idrees H, Tayyab M, Athrey K, et al. Composition loss for counting, density map estimation and localization in dense crowds[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11206: 544-559.

[29] Idrees H, Saleemi I, Seibert C, et al. Multi-source multi-scale counting in extremely dense crowd images[C]// 2013 IEEE Conference on Computer Vision and Pattern

Recognition, June 23-28, 2013, Portland, OR, USA. New York: IEEE Press, 2013: 2547-2554.

[30] Jiang X L, Xiao Z H, Zhang B C, et al. Crowd counting and density estimation by trellis encoder-decoder networks [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 6126-6135.

[31] Sindagi V A, Patel V M. CNN-Based cascaded multi-task learning of high-level prior and density estimation for crowd counting[C]//2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, August 29-September 1, 2017, Lecce, Italy. New York: IEEE Press, 2017: 17287241.

[32] Cao X K, Wang Z P, Zhao Y Y, et al. Scale aggregation network for accurate and efficient crowd counting[M]// Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11209: 757-773.

[33] Sam D B, Peri S V, Sundararaman M N, et al. Locate, size, and count: accurately resolving people in dense crowds via detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(8): 2739-2751.

[34] Zhang Y N, Zhao H L, Duan Z D, et al. Congested crowd counting via adaptive multi-scale context learning [J]. Sensors, 2021, 21(11): 3777.