基于多尺度融合的深度人群计数算法

左静*，巴玉林
兰州交通大学自动化与电气工程学院，甘肃 兰州 730070

摘要 在人群计数统计时存在相机透视、人群重叠、人群遮挡等众多干扰因素，使人群计数的准确性不高。针对这一问题，提出一种多尺度融合的深度人群计数算法。首先，利用 VGG-16 网络的部分结构提取出人群底层特征信息；其次，以膨胀卷积理论为基础，构建多尺度特征提取模块，实现多尺度上下文特征信息的提取，降低模型参数量；最后通过将底层细节特征信息和高层语义特征信息融合的方式，提升模型计数性能和密度图质量。在三个公开数据集上对不同算法进行测试。实验结果表明，与其他人群计数算法相比，所提算法的平均绝对误差和方均误差均有不同程度的降低，说明所提算法具有较好的准确性、鲁棒性及良好的泛化性。

关键词 机器视觉；人群计数；密度图；卷积神经网络；膨胀卷积；特征融合
中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP57.241502

Population-Depth Counting Algorithm Based on Multiscale Fusion

Zuo Jing*, Ba Yulin
School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou, Gansu 730070, *China*

Abstract There are many interference factors such as camera perspective, crowd overlap, and crowd occlusion in crowd-counting statistics that decrease the accuracy of crowd counting. Aiming at addressing these problems, a population-depth counting algorithm based on multiscale fusion is proposed herein. First, the proposed algorithm uses the partial structure of the VGG-16 network to extract the underlying feature information of the crowd. Second, based on the dilated convolution theory, a multiscale feature extraction module is constructed to realize multiscale context feature information extraction and reduce the model parameter amount. Finally, the model counting performance and density-map quality are improved by fusing low-level detail feature information and high-level semantic feature information. Different algorithms are tested on three public datasets. The experimental results show that compared with other crowd counting algorithms, the average absolute error and mean square error of the proposed algorithm are reduced to varying degrees, indicating that the proposed algorithm exhibits good accuracy, robustness, and good generalization.

Key words machine vision; crowd counting; density map; convolutional neural network; dilated convolution; feature fusion
OCIS codes 150.1135; 100.4996

1 引 言

随着我国人口的快速增长和城市化进程的不断加快，体育赛事、交通出行、旅游观光等大规模人群聚集活动日益增多。人群的聚集往往会带来潜在危险，如踩踏事件等，因此，人群计数已成为公共安全领域的热门课题之一。然而由于人群分布不均、背景复杂、场景变化大、遮挡和透视变化等各种因素的影响，人群计数的准确性面临着巨大的挑战。

目前，人群计数方法主要分为基于检测的方法、基于回归的方法、基于卷积神经网络(CNN)的方法 3 种[1]。基于检测和回归的方法均从传统特征入手，如全局特征[2]、局部特征[3]及纹理特征[4]等，适

收稿日期：2020-04-20；**修回日期**：2020-05-21；**录用日期**：2020-06-01
基金项目：国家自然科学基金(61763025，61661027)、甘肃省自然科学基金(20JR5RA398)
* **E-mail**：1269132835@qq.com

用于人群稀疏场景。尽管研究人员对基于检测、回归的方法进行了大量的研究，但面对人群重叠、遮挡严重的高密度场景，这两种方法仍具有明显的局限性[5-7]。近年来，随着 CNN 在语义分割、目标识别等领域的快速发展，有学者将其应用于人群计数领域。Wang 等[8]首次将 CNN 应用于人群密度估计，但提出的模型没有考虑人群尺寸变化的问题，计数准确性和鲁棒性较差。为解决尺度变化问题，Zhang 等[9]提出了一种多列卷积网络(MCNN)，该网络通过三列不同的卷积网络结构提取场景中的多尺度特征信息，提升了人群计数的精度，但模型结构复杂，参数量巨大。Li 等[10]提出一种基于膨胀卷积的估计模型(CSRNet)，该模型通过具有连续不同膨胀率的卷积层来拟合人群的多尺度信息，减少了模型参数量，但大量细节信息被阉割。Cao 等[11]利用类似的 Inception 结构[12]，通过多个不同大小的卷积核提取多尺度特征，但融合后的特征参数量较大。Wang 等[13]提出一种多尺度膨胀卷积模型(MSCNN)，该模型结合了膨胀卷积和 Inception 结构的优势，但提取到的特征经 1×1 卷积融合后损失了底层特征，影响计数准确度。综上所述，多尺度特征提取和膨胀卷积虽然在一定程度上解决了人群计数问题中尺度变化和参数量较大的问题，但如何充分利用 CNN 学习到的多尺度人群特征信息，保证特征信息能够融合利用，提升密度图质量，仍然是目前存在的问题。

针对上述方法的不足和存在的问题，本文提出一种膨胀卷积方法与特征融合方法结合的人群计数模型。模型前端采用 VGG-16 的部分结构提取出图像中的一般特征信息；模型中端以膨胀卷积为基础，采用类似 Inception 的结构对具有不同膨胀率的卷积核进行组合，构建多尺度特征提取模块，以全面提取图像中的多尺度上下文信息，降低模型参数量；模型后端将网络提取的底层细节信息与高层语义信息相融合，提升模型对人群图像中小尺度目标的感知能力和输出密度图的质量。

2 人群计数算法

基于 CNN 的人群计数算法主要包括 5 个步骤：数据集的准备(训练集和测试集)；样本的标注，生成真实密度图标签；模型结构的搭建；网络模型的训练，通过训练集中的标签数据，不断对比优化调整模型参数，直到达到预定标准；采用测试集测试模型结构的有效性。

2.1 人群密度图

基于 CNN 的人群计数算法以密度图作为标签进行网络训练，通过对回归的人群密度图进行积分求和来获得总人数。对于密集场景下的人群计数，由于存在身体重叠、遮挡的情况，现有数据集会将图像中所有的人头位置标记出来，通过基于高斯核的计算方法求得人群密度图。假设 x_i 为图像中的一个人头标注点，将人头表示为 $\delta(x-x_i)$，因此一张拥有 N 个人头标注点的图像就可以表示为

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

通过与高斯核 G_σ 进行卷积，得到的连续密度函数表达式为

$$F(x)=\sum_{i=1}^{N}\delta(x-x_i)*G_\sigma(x)。 \tag{2}$$

由于存在透视失真的情况，每一个像素点 x_i 所表示的人头的面积存在较大差别。针对这个问题，Zhang 等[9]采用几何自适应高斯核生成密度图。假设人群分布是均匀的，对于给定的人头坐标 x_i，其到 k 个邻居的距离可以表示为 $\{d_1^i, d_2^i,\cdots,d_k^i\}$，对应的平均距离为 $\bar{d}_i=\frac{1}{k}\sum_{j=1}^{k}d_j^i$。则人群密度函数为

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

式中：β 为高斯参数。根据实验结果可知，$\beta=0.3$ 时效果最好。图 1 为经过自适应高斯核处理的人群密度结果图。

2.2 网络结构

为解决人群图像的计数问题，提出基于膨胀卷积与特征融合的卷积神经网络模型。所提模型分为 3 部分，即特征提取、多尺度特征提取模块(MSB)及浅层深层特征融合，整体网络结构如图 2 所示，网络输入为一幅人群图像，输出为人群密度图像。为适应人群图像中的人物尺度变化，提升高密度人群图像中小目标的计数性能，所提模型采用膨胀卷积结构，并结合多尺度特征提取和浅层深层特征融合方法，通过一个端到端的网络结构方式，以达到直接从人群图像到人群密度图的效果。

2.3 多尺度特征提取模块

在提取特征图时，基于 CNN 的人群密度估计为降低网络参数量、扩大感受野，在网络结构设计时往往会增加池化层对图像进行下采样操作，虽然池

化操作能够有效地降低网络参数量、扩大感受野，但同时也降低了特征图的分辨率，使得部分人群信息被淹没，影响最终人群密度图的质量和人群计数准确度。膨胀卷积[14-15]是针对语义分割问题中池化操作会降低图像分辨率、丢失信息而提出的一种卷积思路。在传统的卷积层中引入一个称为“膨胀率”的新参数。该参数定义了卷积核处理数据时各值的间距，通过设置不同的膨胀率来达到在不增加参数量的前提下，扩大感受野，保证输出特征图尺寸不变的目的。

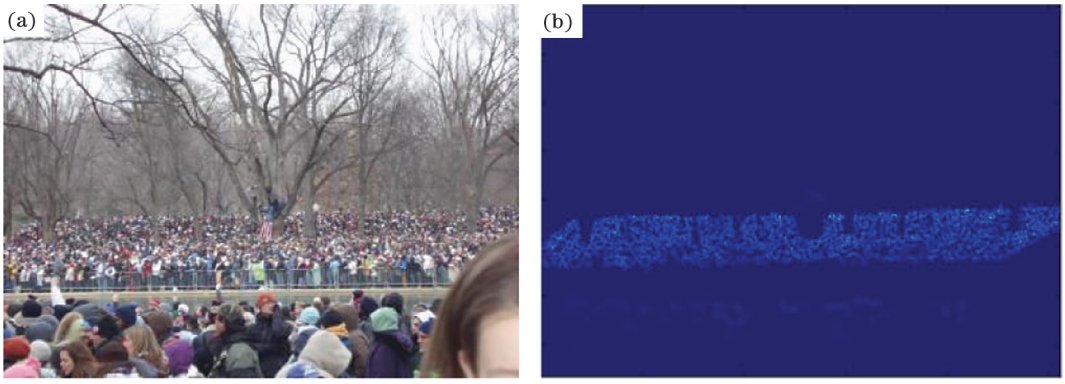

图 1 人群密度图。(a) 原图；(b) 几何自适应高斯核

Fig. 1 Crowding density map. (a) Original image; (b) geometric adaptive Gaussian kernel

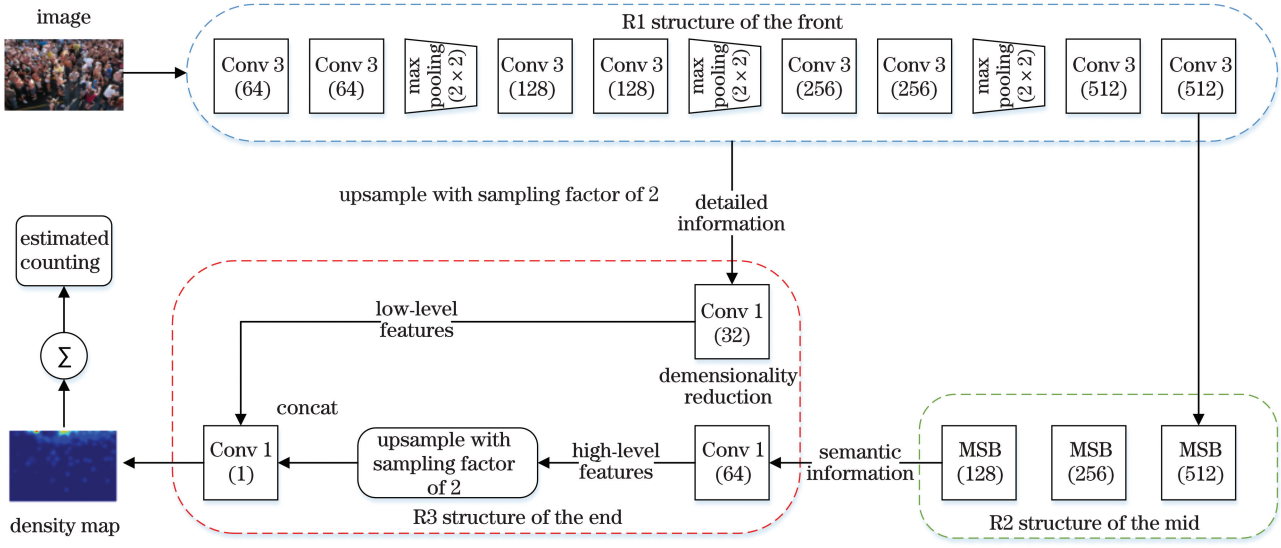

图 2 总体结构图

Fig. 2 Diagram of overall structure

图 3 为膨胀卷积的原理图。当膨胀率为 1 时，感受野尺寸为 3×3，保持不变；当膨胀率为 2 时，在保持卷积核参数不变的情况下，感受野尺寸扩大为 5×5；同理，当膨胀率为 3 时，感受野尺寸扩大为 7×7。

人群计数问题最大的挑战在于人群尺度的多变性。由于相机透视效果的影响，人群图像中人的尺度存在较大差异。靠近镜头位置的人会以较大尺度呈现，反之则以较小的尺度呈现。传统人群计数模型通常采用多列不同卷积网络去拟合不同人群尺度。虽然该类模型解决了人群尺度变化问题，但由于存在多列卷积网络，需要分别训练各列网络，使得训练复杂，无法实现端到端的训练效果，且随着卷积核尺度的增大，其模型参数量也大量增加。鉴于膨胀卷积具有特征提取能力优越、参数量小的优点，本文采用类似于 Inception 模块的设计，在每个模块中并排堆叠几列具有不同膨胀率的膨胀卷积层，通过不同的膨胀率去等效不同的感受野，进而提取图像的多尺度信息。图 4 为构建的 MSB。

MSB 以原始的 3×3 卷积核为基础，堆叠 3 列具有不同膨胀率的膨胀卷积核，以适应不同尺度的人群特征信息。通过膨胀率为 1 的卷积核提取人群

图像较小尺度信息；膨胀率为 2 的卷积核提取中等尺度人群信息；膨胀率为 3 的卷积核提取较大尺度人群特征信息。最后对提取的特征信息进行像素级的融合，从而揭示图像的上下文信息。

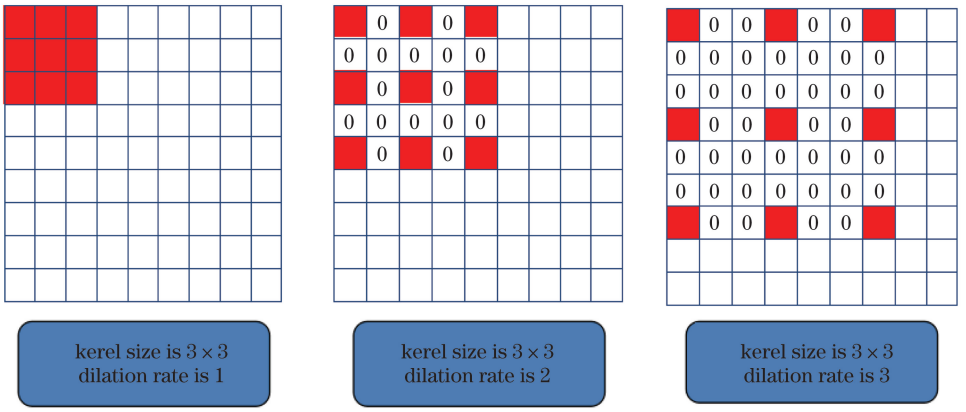

图 3 膨胀卷积原理图

Fig. 3 Principle diagram of dilated convolution

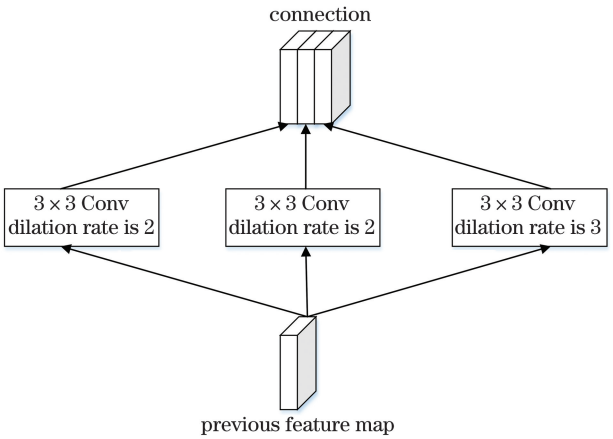

图 4 MSB 结构

Fig. 4 Structure of MSB

2.4 特征融合

除尺度变化问题之外，人群计数还面临着另外一个问题，即小尺度目标的计数问题。人群图像，尤其是高密度图像，往往会存在着许多尺度很小的目标。在 CNN 中，连续多次下采样和逐层抽象化的结构特性使得生成的特征图呈现出鲜明的金字塔结构[16]。越靠近高层的特征，越接近语义信息，越靠近底层的特征，越接近细节特征。其中高层特征表现为目标整体感知信息，而底层特征表现为目标局部细节信息，如边缘、角点等。在应对人群计数中的小目标检测问题时，小目标检测对特征细节要求较高。但由于特征图的层层传递，高层特征往往包含更多的语义信息，细节信息保留不足，导致模型对小尺度的人头目标感知较差。为进一步提升模型计数性能和尺度感知能力，在模型结构设计时，将高层语义信息与底层细节信息相融合，并通过融合后的整体特征图实现密度图的回归。一方面可以使得融合后的特征图包含更丰富的信息；另一方面可以有效提升密度图的生成质量，因为特征融合后生成的密度图保留了更多的细节特征，变得更加清晰。

2.5 损失函数

参考目前多数人群计数算法，将欧氏距离作为网络的损失函数，以计算估计密度图与真实密度图之间的差异。损失函数的定义为

$$L(\boldsymbol{\theta})=\frac{1}{2N}\sum_{i=1}^{N}\left\|\boldsymbol{F}(\boldsymbol{X}_i;\boldsymbol{\theta})-\boldsymbol{F}_i\right\|^2, \tag{4}$$

式中：$\boldsymbol{\theta}$ 为网络中参数集合；$\boldsymbol{X}_i$ 为第 i 张输入图像；$\boldsymbol{F}(\boldsymbol{X}_i;\boldsymbol{\theta})$ 和 $\boldsymbol{F}_i$ 分别为第 i 张图像对应的预测密度图和真实密度图。

3 实 验

分别在三个公开数据集 ShanghaiTech、UCF_CC_50、WorldExpo′10 上对所提算法进行训练和测试。在训练阶段，将测试集图像与真实密度图标签输入到模型中，通过反向误差传播的方式不断修正特征参数，以训练模型。在测试阶段，通过一定的评价指标对模型的准确性和鲁棒性进行评估。

3.1 评价指标

实验采用平均绝对误差（MAE）和方均误差（MSE）作为模型准确性和鲁棒性的评价指标，表达式分别为

$$E_{\mathrm{MA}}=\frac{1}{N}\sum_{i=1}^{N}\left|Z_i-Z_i^{\mathrm{T}}\right|, \tag{5}$$

$$E_{\mathrm{MS}}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|Z_i-Z_i^{\mathrm{T}}\right|^2}, \tag{6}$$

式中：Z_i 和 Z_i^{T} 分别为第 i 张测试图像中的预测人数和实际人数。MAE 值越小，表示模型的准确率越高，MSE 值越小，表示模型的鲁棒性越好。

3.2 数据集

1）ShanghaiTech 数据集

ShanghaiTech 数据集[9]作为一种适应性最广的人群数据集，由 Part_A 和 Part_B 两部分组成，总共拥有 1198 张图像，标记了 330165 个人。Part_A 部分的 482 张图像是从网上下载得到的，图像中的人群密度很高，其中 300 张图像作为训练集，182 张图像作为测试集。Part_B 部分的 716 张图像来源于上海街道的监控图像，人群密度相对稀疏，其中 400 张图像作为训练集，316 张图像作为测试集。

2）UCF_CC_50 数据集

UCF_CC_50 数据集[17]是人群计数领域最具挑战性的数据集，该数据集有 50 张图像，共标记 63075 个人，包括各种密度和不同视角的不同场景。在整个数据集中，人群数量差异巨大，人数从 94 到 4543 不等，每幅图像的平均人数为 1280。

3）WorldExpo′10 数据集

WorldExpo′10 数据集[18]是目前包含图像最多的人群数据集，该数据集提供了 3980 幅图像，全部来源于上海世博会期间 108 个摄像机拍摄的 1132 个视频序列。数据集标记总人数为 199923 人，从 1 人到 253 人不等，其中 3380 幅图像作为训练集。测试集由 S1、S2、S3、S4、S5 等 5 种不同的视频序列组成，每个视频序列拥有 120 幅图像，构成 5 种不同场景。

3.3 训练过程

数据预处理：数据集中的图像分辨率被设置为 224×224，为防止出现过拟合现象，采用数据增强方法对数据集进行扩充。将训练集中的每张图像随机裁剪成 9 个图像块，图像块的尺寸为原图的 1/4。最后对训练集和测试集中的图像进行打乱排序操作，增加模型的泛化能力。

参数配置：模型的训练参数如表 1 所示。

表 1 模型参数

Table 1 Parameters of model

Parameter	Content	Parameter	Content
Learning rate	0.001	Momentum	0.9
Optimizer	Adma	Normalization	0.001
Weight-decay	0.05	Batch-size	1

3.4 模型评估

为验证模型结构的有效性，对所提算法与目前在人群计数领域取得较好成绩的几种算法进行比较。

1）ShanghaiTech 数据集评估

表 2 为各种人群计数算法在 ShanghaiTech 数据集上的处理结果。从表 2 可知：在 Part_A 部分，所提算法取得了最小的 MAE 值和 MSE 值；在 Part_B 部分，所提算法的 MAE 和 MSE 值仅次于 DADNet[19]。图 5 展示了所提算法在 ShanghaiTech 数据集上生成的人群密度图。

表 2 ShanghaiTech 数据集上的计数结果比较

Table 2 Comparison of counting results on ShanghaiTech dataset

Algorithm	Part_A		Part_B	
	MAE	MSE	MAE	MSE
Algorithm in Ref.[18]	181.8	277.7	32.0	49.8
MCNN[9]	110.2	173.2	26.4	41.3
SCNN[20]	90.4	135	21.6	33.4
MSCNN[21]	83.8	127.4	17.7	30.2
CSRNet[10]	68.2	115.0	10.6	16.0
DADNet[19]	64.2	99.9	8.8	13.5
Proposed algorithm	63.4	97.2	9.6	14.3

2）UCF_CC_50 数据集评估

表 3 为不同算法在 UCF_CC_50 数据集上的实验结果。由表 3 可知，相较于其他的人群计数算法，所提算法在性能方面均有不同程度的提升。与目前在人群计数领域取得良好效果的 CSRNet 算法相比，所提算法在 MAE 值和 MSE 值上分别降低了 8.9 和 16.7。

表 3 UCF_CC_50 数据集上的计数结果比较

Table 3 Comparison of counting results on UCF_CC_50 dataset

Algorithm	MAE	MSE
Algorithm in Ref.[18]	467.0	498.5
MCNN[9]	377.6	509.1
Algorithm in Ref.[22]	338.6	424.5
SCNN[20]	318.1	439.2
DADNet[19]	285.5	389.7
CSRNet[10]	266.1	397.5
Proposed algorithm	257.2	380.8

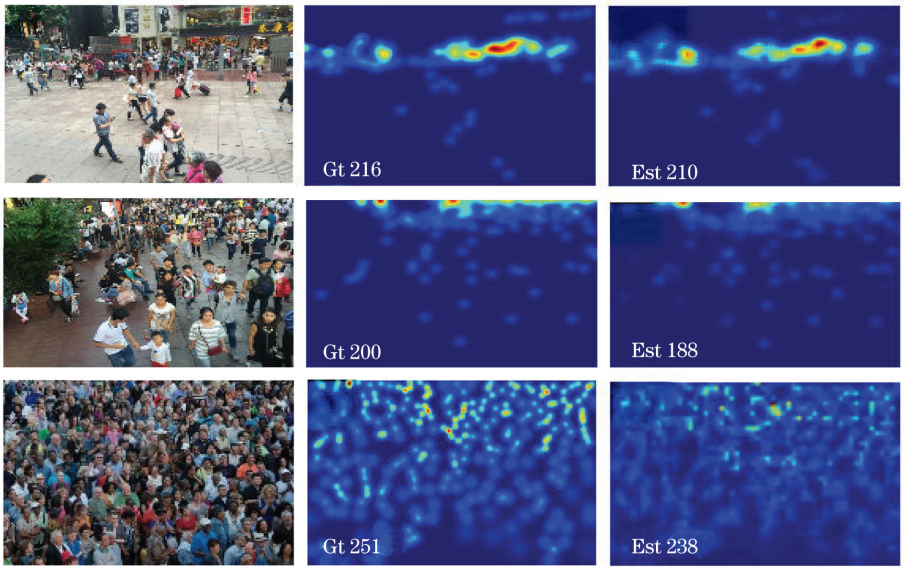

图 5 ShanghaiTech 数据集实验结果。(a) 原图；(b) 真值图；(c) 密度估计图

Fig. 5 Experimental results on ShanghaiTech dataset. (a) Original images; (b) ground-truth images; (c) estimated crowd density maps

3) WorldExpo′10 数据集评估

表 4 为不同算法在 WorldExpo′10 数据集上的计数结果比较，分别对 5 种测试场景序列进行统一的结果比较，并求出比较结果的综合平均值。根据比较结果可知，虽然所提算法在场景 S2、S3、S5 中的准确度低于相应文献中提出的方法，但在场景 S1、S4 及最终的综合平均值方面取得了良好的表现，使得准确度有所提升。

表 4 WorldExpo′10 数据集上的准确度比较

Table 4 Accuracy comparing on WorldExpo′10 dataset

%

Algorithm	Accuracy					Average accuracy
	S1	S2	S3	S4	S5	
Algorithm in Ref. [18]	9.8	14.1	14.3	22.2	3.7	12.8
MCNN[9]	3.4	20.6	12.9	13.0	8.1	11.6
SCNN[20]	4.4	15.7	10.0	11.0	5.9	9.4
CP-CNN[23]	2.9	14.7	10.5	10.4	5.8	8.86
CSRNet[10]	2.9	11.5	8.6	16.6	3.4	8.6
RRSC[24]	2.9	15.0	7.2	14.7	2.6	8.5
Proposed algorithm	2.6	15.3	9.8	9.4	4.7	8.36

3.5 验证实验与分析

1) 模型结构性验证

为验证所提算法提出的 MSB、特征融合方式两项工作的有效性，将在 ShanghaiTech 数据集上进行验证性实验。实验结果如表 5 所示。由表 5 可以看出：MSB、特征融合方式对人群计数性能均有贡献；MSB 对模型性能影响最为显著，特征融合方式次之。实验结果不仅验证了人群多尺度信息提取对计数准确性的显著影响，同时表明特征融合方法能够有效提升模型性能。

表 5 模型性能对比

Table 5 Comparison of model performance

Algorithm	Part _A		Part_B	
	MAE	MSE	MAE	MSE
No MSB	123.2	194.7	26.5	42.4
No feature fusion	70.5	119.4	11.3	18.7
Proposed algorithm	63.4	97.2	9.6	14.3

图 6 为特征融合对网络收敛效果的影响。从图 6(a)可以发现，相较于未特征融合模型，特征融合模型能够收敛到更好的局部最优点，测试损失更低。从图 6(b)可以发现，特征融合模型在迭代前期，训练误差波动起伏。这是由于在训练前期，特征融合模型的卷积层有更多参数尚未完成学习，模型整体特征中掺杂了无用的细节信息，使得模型训练受到误导。但随着迭代次数的增加，模型的误差整体具有同一趋势，表明特征融合模型已经能有效学习不同阶段的特征图参数。

2) 泛化性能验证

在人群计数领域，模型的泛化性能是检验模型

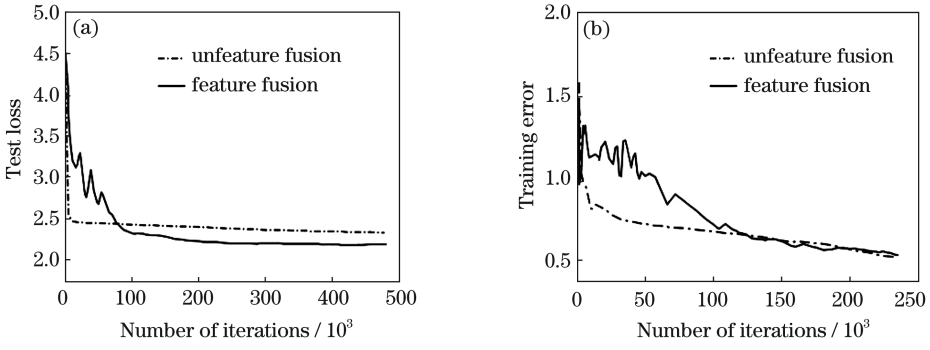

图 6 特征融合对测试损失和训练误差的影响。(a) 测试损失；(b) 训练误差

Fig. 6 Effect of feature fusion on training error and test loss. (a) Test loss; (b) training error

算法性能的重要指标。为测试模型的泛化能力，通过迁移学习的方式在 ShanghaiTech 和 UCF_CC_50 数据集上对所提算法进行跨数据集测试。本次测试分为两组：第一组，在源域 ShanghaiTech 数据集 Part_A 部分训练网络模型，在目标域 UCF_CC_50 数据集上测试人群计数的结果和误差；第二组，在源域 UCF_CC_50 数据集上训练网络模型，在目标域 ShanghaiTech 数据集 Part_A 部分测试人群计数的结果和误差。两组迁移学习的实验结果如表 6 所示。

表 6 迁移学习的实验结果

Table 6 Experimental results of transfer learning

Group	MAE	MSE
Group 1	290.3	458.7
Group 2	115.7	169.5

对比表 2、3、6 的实验结果可知：在 UCF_CC_50 数据集测试过程中，第一组的实验结果在准确度方面相差不大(290.3 vs 257.2)，表明经过 Part_A 部分训练的模型能够较好地适应 UCF_CC_50 数据集的任务；在 ShanghaiTech 数据集 Part_A 部分的测试过程中，虽然 UCF_CC_50 数据集训练样本有限，人群密度变化极为激烈，导致第二组的实验结果出现一定程度的波动(115.7 vs 63.4)，但相较于其他人群计数算法，所提算法仍有较好的处理结果。

4 结 论

结合膨胀卷积与特征融合的优点，提出一种基于膨胀卷积与特征融合的多尺度人群计数模型。以膨胀卷积为基础构建多尺度特征提取模块，用于解决密集场景中人群尺度变化巨大、模型参数量大的问题。通过将网络底层中的细节信息与高层的语义信息相融合，网络终端的特征表达能力得到增强，提升了模型对小尺度目标信息的提取能力。最后，通过对比不同人群计数算法在三个公开数据集上的处理结果，可知，所提算法应对密集场景下的人群计数时有着良好的准确性和鲁棒性。

由于所提算法在设计过程中采用欧氏距离作为损失函数，在密度估计过程中，其会使生成的预测密度图变得模糊，鉴于对抗损失函数在图像语义分割领域取得的良好效果，后续在人群计数研究过程中可以将其作为损失函数的选取点。

参 考 文 献

[1] Luo H L. Study on crowd counting and density estimation based on depth convolution neural network[D]. Chongqing: Chongqing University, 2018: 10-11.
罗红玲. 基于深度卷积神经网络的人群计数与密度估计研究[D]. 重庆：重庆大学, 2018: 10-11.

[2] 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), June 20-25, 2005, San Diego, CA, USA. New York: IEEE Press, 2005: 886-893.

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

[4] Li M, Zhang Z X, Huang K Q, et al. Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection [C] // 2008 19th International Conference on Pattern Recognition, December 8-11, 2008, Tampa, FL, USA. New York: IEEE Press, 2008.

[5] Paragios N, Ramesh V. AMRF-based approach for

real-time subway monitoring[C]//Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, December 8-14, 2001, Kauai, HI, USA. New York: IEEE Press, 2001.

[6] Johnson J, Alahi A, Li F F. Perceptual losses for real-time style transfer and super-resolution[M]// Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9906: 694-711.

[7] Chen K, Loy C C, Gong S G, et al. Feature mining for localised crowd counting[C]//Procedings British Machine Vision Conference, September 3-7, 2012, Surrey. Durham: British Machine Vision Association Press, 2012: 21.

[8] 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-MM′15, October 13-30, 2015, Brisbane, Australia. New York: ACM Press, 2015: 1299-1302.

[9] 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.

[10] 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, USA. New York: IEEE Press, 2018: 1091-1100.

[11] 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.

[12] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015.

[13] WangY J, Hu S Y, Wang G D, et al. Multi-scale dilated convolution of convolutional neural network for crowd counting[J]. Multimedia Tools and Applications, 2020, 79(1/2): 1057-1073.

[14] 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.

[15] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[EB/OL]. (2016-04-30)[2020-04-29]. https://arxiv.org/abs/1511.07122.

[16] Lin Z J, Luo Z, Zhao L, et al. Multi-scale convolution target detection algorithm with feature pyramid[J]. Journal of Zhejiang University (Engineering Science), 2019, 53(3): 533-540.
林志洁, 罗壮, 赵磊, 等. 特征金字塔多尺度全卷积目标检测算法[J]. 浙江大学学报(工学版), 2019, 53(3): 533-540.

[17] 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.

[18] Zhang C, Li H S, Wang X G, et al. Cross-scene crowd counting via deep convolutional neural networks[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 833-841.

[19] 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, Nice, France. New York: ACM Press, 2019: 1823-1832.

[20] 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 (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 4031-4039.

[21] Zeng L K, Xu X M, Cai B L, et al. Multi-scale convolutional neural networks for crowd counting[C]//2017 IEEE International Conference on Image Processing (ICIP), September 17-20, 2017, Beijing, China. New York: IEEE Press, 2017: 465-469.

[22] Marsden M, McGuinness K, Little S, et al. Fully convolutional crowd counting on highly congested scenes[EB/OL]. (2017-01-17)[2020-04-29].

https://arxiv.org/abs/1612.00220.

[23] Sindagi V A, Patel V M. Generating high-quality crowd density maps using contextual pyramid CNNs [C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 1879-1888.

[24] Wan J, Luo W H, Wu B Y, et al. Residual regression with semantic prior for crowd counting[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: 4031-4040.