·图形图像处理·　　文章编号：1000-3428(2022)05-0251-07　　文献标志码：A　　中图分类号：TP391

一种抗背景干扰的多尺度人群计数算法

郭爱心[1]，夏殷锋[2]，王大为[1]，芦　宾[1]

（1.山西师范大学 物理与信息工程学院，太原 030006；2.中国科学技术大学 自动化系，合肥 230026）

摘　要：人群计数技术以估计人群图片或视频中的人数为目标，可以有效预防人群踩踏事故的发生，广泛应用于安防预警、城市规划及大型集会管理等领域。然而，由于人群尺度变化、背景干扰、人群分布不均、遮挡和透视效应等因素的影响，单幅图片的人群计数仍是一项非常具有挑战性的任务。针对人群计数中多尺度变化和背景干扰问题，提出一种抗背景干扰的多尺度人群计数算法。以VGG16网络结构为基础，引入特征金字塔构建多尺度特征融合骨干网络解决人群多尺度变化问题，设计Double-Head-CC结构对融合后的特征图进行前景背景分割和密度图预测以抑制背景干扰。基于密度图的局部相关性和多任务学习，定义多重损失函数和多任务联合损失函数进行网络优化。在ShanghaiTech、UCF-QNRF和JHU-CROWD++数据集上进行训练和评测，实验结果表明，该算法能够很好地预测人群密度分布和人群数量，具有较高的准确性，且鲁棒性强、泛化性能良好。

关键词：人群计数；深度学习；特征金字塔；损失函数；密度图

开放科学（资源服务）标志码（OSID）：

中文引用格式：郭爱心，夏殷锋，王大为，等.一种抗背景干扰的多尺度人群计数算法[J].计算机工程，2022，48(5)：251-257.

英文引用格式：GUO A X，XIA Y F，WANG D W，et al. A multi-scale crowd counting algorithm with removing background interference[J].Computer Engineering，2022，48(5)：251-257.

A Multi-scale Crowd Counting Algorithm with Removing Background Interference

GUO Aixin[1]，XIA Yinfeng[2]，WANG Dawei[1]，LU Bin[1]

（1.College of Physics and Information Engineering，Shanxi Normal University，Taiyuan 030006，China；
2.Department of Automation，University of Science and Technology of China，Hefei 230026，China）

【Abstract】Crowd counting technology is aimed at estimating the number of people in crowd pictures or videos. The technology can effectively be applied to prevent stampede accidents and is widely used in security and early warning，urban planning，and management of large gatherings. However，due to crowd scale variation，background interference，uneven crowd distribution，occlusion，and perspective effect，it is still a very challenging task to count a single image. Aiming at the problem of multi-scale changes and background interference in crowd counting，a multi-scale crowd counting algorithm with removing background interference is proposed. Based on the VGG16 network structure，the feature pyramid is introduced to form the multi-scale feature fusion backbone network to solve the problem of the multi-scale changes. The Double-Head-CC structure is designed to perform foreground-background segmentation and density map prediction on the fused feature map to suppress the background interference. Based on the local correlation of the density map and multi-task learning，the multiple loss functions，and multi-task joint loss function are defined to optimize the network. The network model is trained and evaluated on the ShanghaiTech，UCF-QNRF，and JHU-CROWD++ datasets. Experimental results show that the algorithm can predict the population density distribution and number of the crowd well，with high accuracy，strong robustness，and good generalization performance.

【Key words】crowd counting；deep learning；feature pyramid；loss function；density map

DOI：10. 19678/j. issn. 1000-3428. 0061423

0　概述

随着城市化进程的不断推进和城镇人口规模的日益增大，大型集会中人群聚集拥挤现象带来的隐患已成为公共安全的重要课题。人群计数技术以估计人群图片或视频中的人数为目标，可以有效预防人群踩踏事故的发生，广泛应用于安防预警、城市规划及大型集会管理等领域。然而，由于人群尺度变

基金项目：国家自然科学基金（62004119）。

作者简介：郭爱心（1991—），女，助教、硕士，主研方向为计算机视觉、深度学习；夏殷锋，博士研究生；王大为、芦　宾，讲师、博士。

收稿日期：2021-04-25　　**修回日期：**2021-05-28　　**E-mail：**guoaixin@mail.ustc.edu.cn

化、背景干扰、人群分布不均、遮挡和透视效应等，单幅图片的人群计数仍是一项非常具有挑战性的任务。

根据人群特征提取的方式不同，现有的人群计数算法可分为基于传统手工特征的方法和基于卷积神经网络的算法[1]。基于卷积神经网络的人群计数算法能够自动提取特征，避免手工设计特征的局限性和复杂性，已成为人群计数的主流算法。文献[2]提出用多列卷积神经网络进行人群计数，不同的列使用不同大小的卷积核，分别处理大、中、小3种不同尺度的人，此后多列网络结构常用来解决尺度问题[3-5]。然而，多列结构使得网络臃肿并加重了计算资源的消耗，更多的研究者通过加深网络结构或者融合不同层次的特征来改进计数网络的性能。文献[6]选择利用去除全连接层的VGG网络作为前端网络，并引入空洞卷积来扩大感受野，生成高质量的人群密度图，提高了计数精度。文献[7]受目标检测领域特征金字塔网络[8]的启发，提出基于特征金字塔的全卷积网络，实现了不同层次特征图的融合，但在公开数据集上的实验结果有待提升。文献[9]设计一种编码解码结构人群计数网络，由编码器中的尺度聚合模块提取多尺度特征，再经过解码器生成高分辨率的人群密度图。文献[10]从复杂背景干扰的角度出发，将视觉注意机制应用于人群计数，通过生成注意力图指导网络进行密度图估计，但该模型的双列子网络的参数量冗余，并且不是端到端的可训练网络。此外，研究者还从多任务学习[11-12]、非监督学习[13-14]等角度进行了人群计数研究，但尺度问题和背景干扰仍是影响人群计数的关键因素。

针对上述问题，本文提出一种抗背景干扰的多尺度人群计数算法（Multi-Scale Crowd Counting algorithm with Removing Background Interference，MSCC-RBI）。该算法构建多尺度特征融合骨干网络来解决人群计数中的尺度问题，并通过设计Double-Head-CC（Double-Head for Crowd Counting）结构来抑制背景干扰，额外定义的多重损失函数可进一步提高预测密度图的质量，提升网络性能。最终在ShanghaiTech[2]、UCF-QNRF[15]和JHU-CROWD++[16]数据集上进行实验来验证算法的性能以及各个模块的有效性。

1 MSCC-RBI算法

MSCC-RBI算法设计由多尺度特征融合骨干网络、Double-Head-CC结构和多重损失函数三部分组成。

1.1 真实密度图的生成

目前公开的人群计数相关数据集基本上只是标记了图片中人头的位置，并不是人群密度图，因此需要先将人头位置转化为真实密度图。真实密度图可以用2D高斯核滤波器与人头位置函数进行卷积得到。设图片中人头的坐标为x_i，对应的位置函数为$\delta(x-x_i)$，若图片中共标记了N_1个人头，则该图片对应的人群密度图$F(x)$如式（1）所示：

$$F(x)=\sum_{i=1}^{N_1}\delta(x-x_i)\times G_\sigma(x) \tag{1}$$

其中：$G_\sigma(x)$为2D高斯核滤波器。本文将该方法生成的人群密度图作为真实密度图，即网络训练的标签。

1.2 多尺度特征融合骨干网络

鉴于VGG16[17]优异的性能以及规范的网络结构，本文选择VGG16作为基础网络，引入额外的特征金字塔结构以解决人群计数任务中行人尺度变化问题。在此基础上，构建多尺度特征融合骨干网络，其网络结构如图1所示。其中：MP表示最大池化操作，本文使用2×2的最大池化，池化后特征图尺寸是池化前的1/2；UP表示上采样操作，本文上采样操作使得特征图尺寸变为原来的2倍；C3、C4、C5、P3、P4、P5为特征图；⊕符号表示逐像素相加。VGG16基础网络由5个卷积模块和4个最大池化层构成，待计数图片输入到基础网络后，经过一系列卷积和池化，会产生不同分辨率的特征图。低层的特征图分辨率大，包含边缘、轮廓等丰富的细节信息，高层特征图分辨率小，包含更高级的语义特征。不同等级特征中存在语义和细节信息的不平衡，使得单一层次的特征难以解决行人尺度剧烈变化问题，故本文提取了VGG16基础网络第三、四、五卷积块输出的特征图C3、C4、C5，对这3个层次的特征图进行特征融合来丰富不同尺度人群的特征表达，在多个尺度上构建具有丰富语义信息的特征。本文采用特征金字塔结构[8]，通过自顶向下的上采样和横向连接对不同层次的特征进行融合。首先对基础网络产生的不同层次的特征图进行3×3的卷积操作，统一特征图的通道数，然后对当前层次的特征图进行上采样，使其大小变为原来的2倍，将上采样后和上一层次卷积后的特征图进行逐像素相加，后续再进行3×3的卷积来降低特征上采样导致的混叠效应，最终得到融合后的特征图。

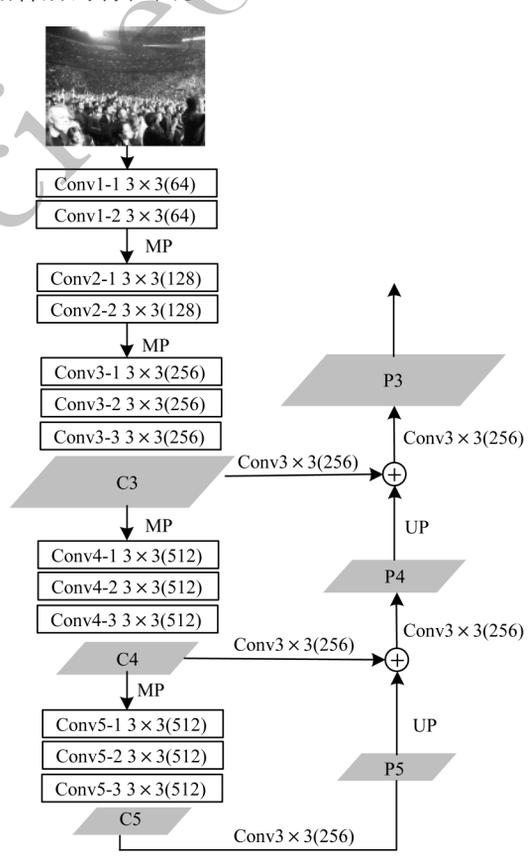


图1 多尺度特征融合骨干网络

Fig.1 Multi-scale feature fusion backbone network

特征图P3是金字塔结构最终输出的特征图，其中丰富的多尺度行人特征表示可以有效提高中小尺度行人计数精度。多尺度特征融合骨干网络的输出特征图P3将作为Double-Head-CC结构的输入。

1.3　Double-Head-CC结构

根据文献[18]，对提取的特征进行与任务相关的预测的网络部分称为头部(head)网络。文献[19]受COCO2018目标检测冠军团队算法启发提出了Double-Head结构，它将目标检测中检测框的分类和回归任务分别在全连接和卷积这两种不同的head上实现，取得了比单一head更好的结果。而现有人群计数算法通常基于单一head且更加关注特征提取过程，常采用简单的卷积层来回归密度图，这种简单的head设计容易受背景噪声因素的干扰，从而导致预测密度图的背景区域出现亮像素，影响计数精度。为此，本文引入前景和背景的分类任务，将人群计数问题转化为多任务学习问题，并设计2个head构成适用于人群计数的Double-Head-CC结构，进行掩膜的生成和密度图的回归。如图2所示，Double-Head-CC结构由DRH(Density Regression Head)和MCH(Mask Classification Head)两部分组成，MCH为DRH提供了与前景背景区域相关的掩膜，可以有效抑制背景噪声的干扰。DRH由3个卷积层和2个ReLU层组成，输入为多尺度特征融合骨干网络的输出特征图P3，输出为2个通道的特征图，分别代表前景和背景对应的初始密度图。MCH由2个卷积层、1个ReLU层和1个Softmax层组成，输出为前景和背景的掩膜，MCH对应的任务实质上是对某一像素点是前景还是背景进行分类，对DRH输出的特征图和MCH输出的掩膜进行逐像素相乘，则可以得到前景密度图和背景密度图，然后两者相加则可以得到预测密度图。

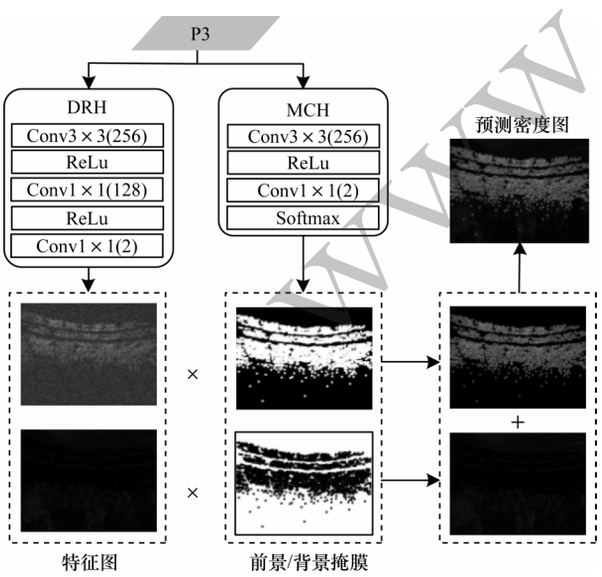


图2　Double-Head-CC结构

Fig.2　Double-Head-CC structure

从图2可以看出，Double-Head-CC结构中DRH对应的任务为密度图回归，MCH对应的任务为前景背景分类，这两个任务都是有监督学习。密度图回归任务的标签为真实密度图$F(x)$，而前景背景分类任务目前缺乏精确的标签，考虑到分类任务是作为密度图回归任务的辅助，对背景和人群的粗略像素进行分类即可，故本文基于真实密度图的阈值生成前景背景分类的标签$S(x)$，其规则如式(2)所示，真实密度图中不为0的区域标记为前景，为0的区域标记为背景，由此得到$S(x)$。

$$S(x)=\begin{cases}1, F(x)>0\\0, F(x)=0\end{cases} \tag{2}$$

1.4　多重损失函数

由1.3节已知网络的输出和标签，还需设计适当的损失函数，才能有效地进行网络训练。为此，本文提出了多重损失函数，并引入了交叉熵损失函数。多重损失函数用来优化密度图回归任务，交叉熵损失函数用来优化前景背景分类任务，多重损失函数和交叉熵损失函数构成多任务联合损失函数。

对于密度图回归任务，多数人群计数算法都是采用欧几里得损失函数进行优化的，欧几里得损失函数$L(\Theta)$定义如式(3)所示：

$$L(\Theta)=\frac{1}{N}\sum_{i=1}^{N}\left\| M(X_i;\Theta)-F_i(X_i)\right\|^2 \tag{3}$$

其中：Θ为待优化的网络参数；N为训练样本的数量；X_i为第i个训练样本；$M(X_i;\Theta)$、$F_i(X_i)$分别为第i个训练样本的网络预测密度图和真实密度图。由式(3)可知，欧几里得损失函数是逐像素进行计算的，即认为预测密度图和真实密度图中的每个像素是独立的，这种损失函数的计算忽略了密度图的局部相关性，无法反映预测密度图和真实密度图之间的结构性差别，进而影响预测密度图的生成质量和人群计数精度。针对欧几里得损失函数的局限性，本文提出了多重损失函数。考虑到池化层是基于局部相关性思想提出的，故本文通过2×2的平均池化操作对密度图局部区域的像素值求平均，池化后的密度图实际上已包含局部相关性信息，考虑到不同空间尺度上预测密度图和真实密度图应尽量一致，多次进行平均池化(AP)操作，构成密度图金字塔，然后在每个层次求其欧几里得损失，构成多重损失函数，在多个尺度上进行优化。如图3所示，本文将Double-Head-CC结构生成的预测密度图进行3次2×2的平均池化操作，得到1/2密度图、1/4密度图和1/8密度图，真实密度图也同步进行池化作为标签。

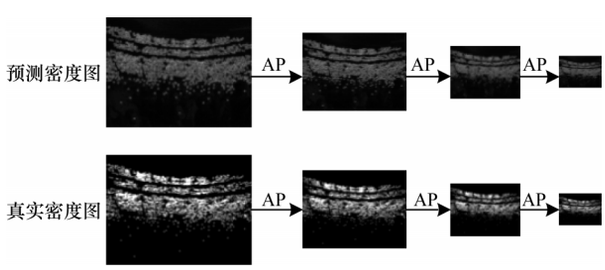


图3　多重损失函数示意图

Fig.3　Schematic diagram of multiple loss function

对于每一阶段的密度图，其损失函数按照欧几里得损失进行计算，其定义如式(4)所示：

$$L_j(\Theta)=\frac{1}{N}\sum_{i=1}^{N}\left\| M_j(X_i;\Theta)-F_{ij}(X_i)\right\|^2 \tag{4}$$

其中：$L_j(\Theta)$为第j次平均池化后的损失函数；$M_j(X_i;\Theta)$、$F_{ij}(X_i)$分别为第j次平均池化后第i个训练样本的网络预测密度图和真实密度图。得到每一阶段密度图的损失函数后，多重损失函数$L_{\mathrm{ml}}(\Theta)$的定义如式(5)所示：

$$L_{\mathrm{ml}}(\Theta)=L_0(\Theta)+2L_1(\Theta)+4L_2(\Theta)+8L_3(\Theta) \tag{5}$$

由多重损失函数的定义可知，它考虑了密度图像素点之间的局部相关性，使得密度图回归任务损失函数的设计更为合理。

对于前景背景分类任务，本文采取交叉熵损失函数进行优化。交叉熵损失函数$L_{\mathrm{ce}}(\Theta)$的定义如式(6)所示：

$$L_{\mathrm{ce}}(\Theta)=-\frac{1}{N}\sum_{i=1}^{N}y_i\log_a(C_{\mathrm{seg}}(X_i;\Theta)) \tag{6}$$

其中：$C_{\mathrm{seg}}(X_i;\Theta)$表示前景背景预测为真实类别的概率；$y_i$表示真实类别。

在密度图回归任务和前景背景分类任务的损失函数基础上，本文定义了多任务联合损失函数$L_{\mathrm{mtjl}}(\Theta)$作为最终网络训练的损失函数，其定义如式(7)所示：

$$L_{\mathrm{mtjl}}(\Theta)=L_{\mathrm{ml}}(\Theta)+\gamma L_{\mathrm{ce}}(\Theta) \tag{7}$$

其中：γ为密度图回归任务和前景背景分类任务之间的平衡系数，本文选取$\gamma=1$。

2 网络优化与评价标准

多尺度特征融合骨干网络和Double-Head-CC结构组成了MSCC-RBI网络模型，多任务联合损失函数作为目标函数进行网络优化。

2.1 数据增强

人群图片中通常包含大量的人群，数据标注困难且成本较高，故目前标注的人群数据集中样本数量有限，为了得到更多的训练样本和更好的训练结果，本文进行了数据增强。对每张图片随机截取大小为原图1/4的9张图片，并将得到的图片进行水平翻转。根据文献[20]，考虑到光照变化，以0.3的概率采用参数为[0.5,1.5]的伽马变换对数据集中的图片进行处理，以0.1的概率随机地将包含灰度图的数据集中的彩色图片转换为灰度图。本文进一步以0.25的概率对数据集中彩色图片的RGB通道进行随机交换，以0.25的概率对数据集中的图片增加平均值为0、标准差为5的高斯噪声。通过裁减、水平翻转、伽马变换、通道变换、高斯噪声等方法得到了增强后的训练数据。

2.2 网络训练

本文是基于PyTorch深度学习框架进行网络设计和训练的。在进行网络参数初始化时，使用预训练的VGG16和均值为0、标准差为0.01的高斯分布进行初始化。网络优化算法选取Adam算法，初始学习率设置为10^{-5}，学习率衰减参数设置为0.995。

2.3 评价标准

平均绝对误差(Mean Absolute Error，MAE)和均方误差(Mean Squared Error，MSE)是人群计数中常用的算法评价标准，其定义如式(8)和式(9)所示：

$$M_{\mathrm{MAE}}=\frac{1}{N_t}\sum_{i=1}^{N_t}\left|C_i^{\mathrm{pred}}-C_i^{\mathrm{gt}}\right| \tag{8}$$

$$M_{\mathrm{MSE}}=\sqrt{\frac{1}{N_t}\sum_{i=1}^{N_t}\left|C_i^{\mathrm{pred}}-C_i^{\mathrm{gt}}\right|^2} \tag{9}$$

其中：N_t为测试图片的数量；C_i^{pred}为网络模型预测的第i张图片中的人数；C_i^{gt}为第i张图片中的实际人数。平均绝对误差(MAE)评价的是算法的准确性，而均方误差(MSE)评价的是算法的鲁棒性。

3 实验结果与分析

本文在ShanghaiTech[2]、UCF-QNRF[15]和JHU-CROWD++[16]数据集上训练并评测了MSCC-RBI算法，并通过消融实验验证了MSCC-RBI算法设计的合理性和有效性。

3.1 ShanghaiTech数据集实验

ShanghaiTech数据集共标记了1 198张图片共计330 165个人头位置，分为Part_A与Part_B两个部分。Part_A中的图片来源于互联网，人群分布较为密集，图片分辨率差异大，训练集包含300张图片，测试集包含182张图片；Part_B中的图片在上海街头拍摄获得，人群分布密度较低、人群尺度变化大且场景多样，训练集包含400张图片，测试集包含316张图片。

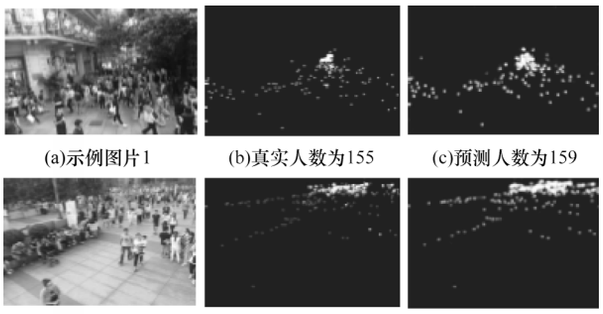
表1为本文所提MSCC-RBI算法与7种当前主流的具有代表性的人群计数算法在ShanghaiTech数据集上的比较结果，其中粗体为结果最优。由表1可知，相比于其他7种算法，MSCC-RBI算法在Part_A部分的MAE最优，使得MAE下降了1.7%；在Part_B部分，MSCC-RBI算法的MAE和MSE均为最优，MAE和MSE分别下降了8.3%和15.6%，体现了MSCC-RBI算法的优越性。

表1 ShanghaiTech数据集上的不同算法性能比较结果

Table 1 Performance comparison of different algorithms on the ShanghaiTech dataset

算法	Part_A		Part_B	
	MAE	MSE	MAE	MSE
MCNN[2]	110.2	173.2	26.4	41.3
RANet[3]	59.4	102.0	7.9	12.9
CSRNet[6]	68.2	115.0	10.6	16.0
SANet[9]	67.0	104.5	8.4	13.6
ADCrowdNet[10]	63.2	**98.9**	7.7	12.9
CFF[11]	65.2	109.4	7.2	12.2
TEDnet[21]	64.2	109.1	8.2	12.8
MSCC-RBI	**58.4**	99.9	**6.6**	**10.3**

图4为MSCC-RBI算法在ShanghaiTech数据集上的结果示例，示例图片包含背景干扰和多尺度行人。由图4可知，预测密度图和真实密度图的分布高度相似，预测人数接近真实人数。



(a)示例图片1　(b)真实人数为155　(c)预测人数为159

(d)示例图片2　(e)真实人数为315　(f)预测人数为295

图4　ShanghaiTech数据集上真值和预测结果对比示例

Fig.4　Comparison examples of true and predicted results on ShanghaiTech dataset

3.2　UCF-QNRF数据集实验

UCF-QNRF数据集由IDREES等[15]收集并公开，共标记了1 535张图片共计1 251 642个人头位置，其中1 201张为训练样本，334张为测试样本。UCF-QNRF数据集中图片场景和拍摄角度多样，且分辨率都较高，在进行网络训练时，为节约内存，本文将图片较长的一边统一为1 024像素。

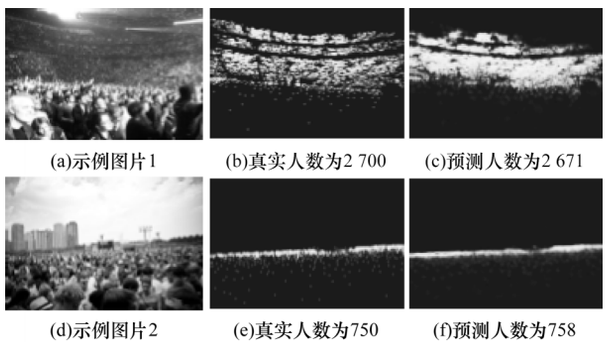
表2为本文所提MSCC-RBI算法与7种当前主流的具有代表性的人群计数算法在UCF-QNRF数据集上的比较结果，其中粗体为结果最优。由表2可知，MSCC-RBI算法的MAE最优且下降了2.5%，人群计数准确性最高，说明本文所提MSCC-RBI算法具有较高的准确性和鲁棒性。

表2　UCF-QNRF数据集上的不同算法性能比较

Table 2　Performance comparison of different algorithms on UCF-QNRF dataset

算法	UCF-QNRF	
	MAE	MSE
MCNN[2]	277.0	426.0
RANet[3]	111.0	190.0
TEDnet[21]	113.0	188.0
CAN[22]	107.0	183.0
S-DCNet[23]	104.4	176.1
SFCN[24]	102.0	171.4
DSSI-Net[25]	99.1	**159.2**
MBTTBF-SCFB[26]	97.5	165.2
MSCC-RBI	**95.1**	172.5

MSCC-RBI算法在UCF-QNRF数据集的结果示例如图5所示。虽然示例图片1中的背景灯光点和示例图片2中的树叶在形态和尺度上与人群高度相似，但预测结果与真值仍非常接近。



(a)示例图片1　(b)真实人数为2 700　(c)预测人数为2 671

(d)示例图片2　(e)真实人数为750　(f)预测人数为758

图5　UCF-QNRF数据集上真值和预测结果对比示例

Fig.5　Comparison examples of true and predicted results on UCF-QNRF dataset

3.3　JHU-CROWD ++数据集实验

JHU-CROWD++数据集是由约翰霍普金斯大学视觉和图像理解实验室于2020年公布的大规模人群计数数据集，该数据集包含了不同密度、不同光照条件以及恶劣天气（雨、雪、雾等）下的4 372张人群图片，共计1 515 005个人头标注，其中训练样本2 272个，验证样本500个，测试样本1 600个。

表3为本文所提MSCC-RBI算法与7种当前主流的具有代表性的人群计数算法在JHU-CROWD++验证集上的比较结果，表4为测试集上的比较结果，其中粗体为结果最优。

表3　JHU-CROWD++验证集上的不同算法性能比较

Table 3　Performance comparison of different algorithms on JHU-CROWD++ validation set

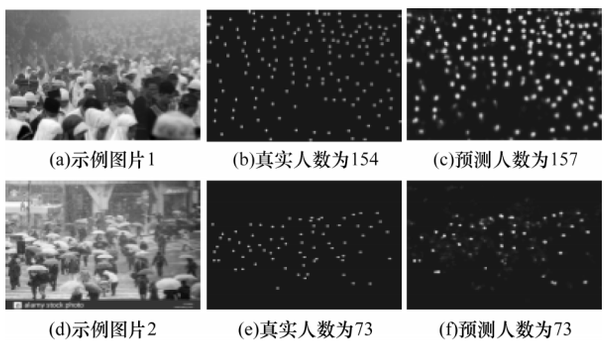
算法	JHU-CROWD++	
	MAE	MSE
MCNN[2]	160.6	377.7
CSRNet[6]	72.2	249.9
SANet[9]	82.1	272.6
CG-DRCN-Res101[16]	57.6	244.4
SFCN[24]	62.9	247.5
DSSI-Net[25]	116.6	317.4
MBTTBF[26]	73.8	256.8
MSCC-RBI	**51.9**	**220.3**

表4　JHU-CROWD++测试集上的不同算法性能比较

Table 4　Performance comparison of different algorithms on JHU-CROWD++ test set

算法	JHU-CROWD++	
	MAE	MSE
MCNN[2]	188.9	483.4
CSRNet[6]	85.9	309.2
SANet[9]	91.1	320.4
CG-DRCN-Res101[16]	71.0	278.6
SFCN[24]	77.5	297.6
DSSI-Net[25]	133.5	416.5
MBTTBF[26]	81.8	299.1
MSCC-RBI	**63.1**	**271.5**

由表3、表4可知，MSCC-RBI算法在验证集和测试集上都取得了最优的结果。本文在JHU-CROWD++数据集上选取了雾天和雨天两张恶劣天气下的图片进行示例，由图6可知，MSCC-RBI算法对恶劣天气造成的前景背景对比模糊的场景也有很高的适用性。



(a)示例图片1 (b)真实人数为154 (c)预测人数为157

(d)示例图片2 (e)真实人数为73 (f)预测人数为73

图6 JHU-CROWD++数据集上真值和预测结果对比示例

Fig.6 Comparison examples of true and predicted results on JHU-CROWD++ dataset

3.4 消融实验

为验证和分析MSCC-RBI算法设计的合理性和有效性，本文在ShanghaiTech数据集的Part_A部分进行了消融实验。本文在多尺度特征融合骨干网络的基础上增加密度图回归模块DRH组成基线，分别增加Double-Head-CC结构模块（不重复增加DRH模块）和多重损失函数模块进行实验，消融实验的结果对比如图7所示。

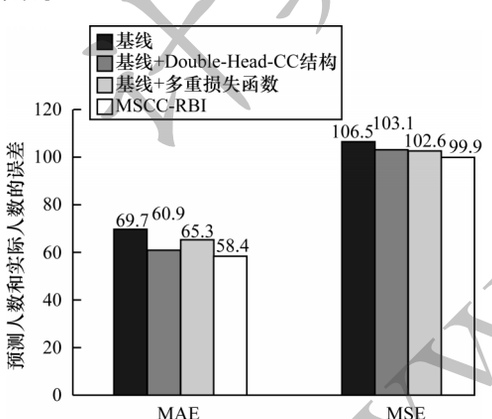


图7 消融实验结果对比

Fig.7 Comparison of ablation experiment results

由图7可知，在基线上增加Double-Head-CC结构可使MAE下降12.6%，MSE下降3.2%，表明Double-Head-CC结构对人群计数的精度和算法的鲁棒性有很大的提升作用。在基线上增加多重损失函数模块可使MAE下降6.3%，MSE下降3.7%。MSCC-RBI算法在基线的基础上，同时增加了Double-Head-CC结构模块和多重损失函数模块，使得MAE下降了16.2%，MSE下降了6.2%，表明Double-Head-CC结构和多重损失函数对模型的改进是同向的。

上述消融实验的结果验证了本文所提Double-Head-CC结构模块、多重损失函数模块和MSCC-RBI算法设计的合理性和有效性。

3.5 模型参数与计数实时性

为进一步分析算法模型的参数规模和人群计数实时性，本文将输入图片的大小设置为1 024×768像素，在GeForce RTX 2080 GPU上进行了测试，结果如表5所示。以基线为参照，MSCC-RBI算法模型的参数量相对于基线仅增加了0.3 MB，每秒浮点运算次数（Floating-point Operations Per Second，FLOPS）增加了14.62G，其中Double-Head-CC结构的参数量为0.921 MB，FLOPS为45.274G，由此可知，Double-Head-CC结构的设计不会引入过多的参数量和FLOPS，整体网络模型参数规模较小。而在计数实时性方面，MSCC-RBI算法模型的推理时间为49.94 ms，每秒帧数（Frames Per Second，FPS）为20.02，与基线相差不大，能够实现快速人群计数。

表5 模型参数与推理效能

Table 5 Model parameter and inference efficiency

算法模型	参数量/MB	运算次数/10^9	推理时间/ms	帧率/(frame·s^{-1})
基线	19.47	355.12	47.79	20.92
MSCC-RBI	19.77	369.74	49.94	20.02

4 结束语

本文提出一种抗背景干扰的多尺度人群计数算法MSCC-RBI。通过构建多尺度特征融合骨干网络融合不同层次的特征，设计Double-Head-CC结构抑制背景干扰并生成密度图，并定义了多重损失函数和多任务联合损失函数进行网络优化。在ShanghaiTech、UCF-QNRF和JHU-CROWD++数据集上的实验结果表明，MSCC-RBI算法具有较高的准确性、较强的鲁棒性和良好的泛化能力。下一步将从提高密度图质量和引入难分负样本等角度出发，增强算法对背景信息的鲁棒性。

参考文献

[1] 王陆洋. 基于卷积神经网络的图像人群计数研究[D]. 合肥：中国科学技术大学，2020.
WANG L Y. Image crowd counting based on convolutional neural network[D]. Hefei: University of Science and Technology of China, 2020. (in Chinese)

[2] ZHANG Y Y, ZHOU D S, CHEN S Q, et al. Single-image crowd counting via multi-column convolutional neural network[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 589-597.

[3] ZHANG A R, SHEN J Y, XIAO Z H, et al. Relational attention network for crowd counting[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2019: 6787-6796.

[4] CHENG Z Q, LI J X, DAI Q, et al. Improving the learning of multi-column convolutional neural network for crowd counting[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York, USA: ACM Press, 2019: 1897-1906.

[5] 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. New York, USA: ACM Press, 2019: 1823-1832.

[6] LI Y H, ZHANG X F, CHEN D M. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 1091-1100.

[7] 马皓，殷保群，彭思凡. 基于特征金字塔网络的人群计数算法[J]. 计算机工程，2019，45(7)：203-207.
MA H, YIN B Q, PENG S F. Crowd counting algorithm based on feature pyramid network[J]. Computer Engineering, 2019, 45(7): 203-207. (in Chinese)

[8] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 936-944.

[9] CAO X K, WANG Z P, ZHAO Y Y, et al. Scale aggregation network for accurate and efficient crowd counting[C]// Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2018: 734-750.

[10] LIU N, LONG Y C, ZOU C Q, et al. ADCrowdNet: an attention-injective deformable convolutional network for crowd understanding[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2019: 3220-3229.

[11] SHI Z L, METTES P, SNOEK C. Counting with focus for free[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2019: 4199-4208.

[12] TIAN Y K, LEI Y M, ZHANG J P, et al. PaDNet: pan-density crowd counting[J]. IEEE Transactions on Image Processing, 2020, 29(5): 2714-2727.

[13] SAM D B, SAJJAN N N, MAURYA H, et al. Almost unsupervised learning for dense crowd counting[C]// Proceedings of AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2019: 8868-8875.

[14] BAI S, HE Z Q, QIAO Y, et al. Adaptive dilated network with self-correction supervision for counting[C]// Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 4593-4602.

[15] IDREES H, TAYYAB M, ATHREY K, et al. Composition loss for counting, density map estimation and localization in dense crowds[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2018: 532-546.

[16] SINDAGI V, YASARL R, PATEL V M. JHU-CROWD++: large-scale crowd counting dataset and a benchmark method [EB/OL]. [2021-03-20]. https://arxiv.org/abs/2004.03597.

[17] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2021-03-20]. http://arxiv.org/abs/1409.1556.pdf.

[18] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. [2021-03-20]. https://arxiv.org/pdf/2004.10934.pdf.

[19] WU Y, CHEN Y P, YUAN L, et al. Rethinking classification and localization for object detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 10186-10195.

[20] DAI F, LIU H, MA Y K, et al. Dense scale network for crowd counting[EB/OL]. [2021-03-20]. https://arxiv.org/pdf/1906.09707.pdf.

[21] JIANG X L, XIAO Z H, ZHANG B C, et al. Crowd counting and density estimation by trellis encoder-decoder networks[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2019: 6126-6135.

[22] LIU W Z, SALZMANN M, FUA P. Context-aware crowd counting[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2019: 5094-5103.

[23] XIONG H P, LU H, LIU C X, et al. From open set to closed set: counting objects by spatial divide-and-conquer[C]// Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2019: 8361-8370.

[24] WANG Q, GAO J Y, LIN W, et al. Learning from synthetic data for crowd counting in the wild[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2019: 8190-8199.

[25] LIU L B, QIU Z L, LI G B, et al. Crowd counting with deep structured scale integration network[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2019: 1774-1783.

[26] SINDAGI V A, PATEL V M. Multi-level bottom-top and top-bottom feature fusion for crowd counting[C]// Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2019: 1002-1012.

编辑　索书志