基于多尺度增强网络的人群计数方法

徐 涛 段仪浓 杜佳浩 刘才华*

(中国民航大学计算机科学与技术学院 天津 300300)

(中国民航大学中国民航信息技术科研基地 天津 300300)

摘 要：人群计数研究普遍使用欧几里得损失函数，易造成图像局部相关性缺失，且现有研究方法未能充分提取人群图像中连续变化的尺度特征，影响了人群计数模型的性能。针对上述问题，该文提出一种基于多尺度增强网络的人群计数模型(MSEN)。首先，在多分支结构生成网络中引入区域性判别网络，将二者组合形成嵌入式GAN模块，以增强生成图像的局部相关性；之后，基于金字塔池化结构设计了尺度增强模块，将该模块连接在嵌入式GAN模块之后，进一步从不同区域提取不同尺度的局部特征，以最大程度地应对人群图像局部尺度连续变化的问题，从而增强整体模型的泛化能力。最后，在3个具有挑战性的人群计数公共数据集上进行了广泛的实验。实验结果表明，该文所述模型可有效提升人群计数问题的准确性和鲁棒性。

关键词：人群计数；图像局部相关性；多尺度特征；嵌入式GAN模块；尺度增强模块

中图分类号：TN911.73; TP391.4 **文献标识码**：A **文章编号**：1009-5896(2021)06-1764-08

DOI: 10.11999/JEIT200331

Crowd Counting Method Based on Multi-Scale Enhanced Network

XU Tao DUAN Yinong DU Jiahao LIU Caihua

(*College of Computer Science and Technology, Civil Aviation University of China, Tianjin* 300300, *China*)

(*Information Technology Base of Civil Aviation Administration of China, Civil Aviation University of China, Tianjin* 300300, *China*)

Abstract: The performance of the crowd counting methods is degraded due to the commonly used Euclidean loss ignoring the local correlation of images and the limited ability of the model to cope with multi-scale information. A crowd counting method based on Multi-Scale Enhanced Network(MSEN) is proposed to address the above problems. Firstly, an embedded GAN module with a multi-branch generator and a regional discriminator is designed to initially generate crowd density maps and optimize their local correlation. Then, a well-designed scale enhancement module is connected after the embedded GAN module to extract further local features of different scales from different regions, which will strengthen the generalization ability of the model. Extensive experimental results on three challenging public datasets demonstrate that the performance of the proposed method can effectively improve the accuracy and robustness of the prediction.

Key words: Crowd counting; Image local correlation; Multi-scale feature; Embedded GAN module; Scale-enhancement module

1 引言

人群计数旨在计算图像或视频场景中所包含的人数，是计算机视觉和智能监控领域的重要研究内容。随着城市人口的急剧增长，广场、火车站、机场航站楼等公共场所时常出现人群高度聚集的现象，存在着巨大的安全隐患。因此，准确预测场景中的人数可以有效地进行人流管控[1]和安防部署，对于社会公共安全具有重要的意义。此外，人群计数方法也可应用于细胞计数[2]、车辆计数[3]和动物迁徙观察[4]等领域，具有广泛的应用价值。由于人群规模和尺度在不同场景中存在着巨大的变化，因此人群计数仍是一项极具挑战性的研究。

早期人群计数研究方法大致可分为基于检测和基于回归两类。基于检测的方法采用目标检测器[5,6]

收稿日期：2020-04-28；改回日期：2020-10-12；网络出版：2020-10-16

*通信作者： 刘才华 chliu@cauc.edu.cn

基金项目：天津市自然科学基金(18JCYBJC85100)，中央高校基本科研业务基金项目中国民航大学专项(3122018C024)，中国民航大学科研启动项目(2017QD16X)

Foundation Items: The Natural Science Foundation of Tianjin (18JCYBJC85100), The Fundamental Research Funds for the Central Universities from the Civil Aviation University of China (3122018C024), The Scientific Research Startup Project of the Civil Aviation University of China (2017QD16X)

对图像中的行人进行逐一检测，统计检测结果以获取最终人数。这类方法在稀疏的人群场景下效果良好，但是在包含遮挡现象的拥挤场景中难以发挥作用。基于回归的方法则通过学习图像特征与相应人数之间的映射关系来实现。此类方法首先提取图像的低级特征(例如前景特征、边缘特征和纹理特征等)，之后利用不同的回归分析技术来建模特征到人数的映射关系。基于回归的方法避免直接通过目标检测进行计数，相对缓解了遮挡现象造成的影响。但是，这类方法的性能在很大程度上受限于低级特征提取的效果，难以在高度拥挤的人群场景较好地发挥作用，因此存在着一定的缺陷。

近年来，随着深度学习的发展，基于卷积神经网络(Convolutional Neural Network, CNN)的方法被广泛应用于人群计数研究。区别于直接预测人数，此类方法大多通过卷积神经网络预测一种人群密度图像，该图像的像素值反映原图中此位置的人群密度，对密度图像进行像素值累加即可获得原图对应的人数。Zhang等人[7]提出一种跨场景计数模型，由两个相关的学习目标(人群密度和人群数量)交替训练，并使用相似于目标场景的样本进行模型微调。该模型在训练和测试时还需用到一种人工制作的透视图像，因此在实际场景中可用性不高。为了解决多尺度问题，Zhang等人[8]提出多列卷积神经网络(Multi-column Convolutional Neural Network, MCNN)模型，构建具有不同卷积核大小的3分支网络以并行提取多尺度特征。但是，该模型仅能在几种特定尺度的场景下发挥作用，且各分支易学到相同模式的特征，造成参数冗余。受MCNN启发，Sam等人[9]提出Switch-CNN，利用所设计的分类网络将人群图像按照不同的密度分类，并根据密度等级为图像选择特定的回归网络。然而，该模型的表现很大程度上受限于分类网络的性能。Shen等人[10]提出对抗跨尺度一致性追求网络(Adversarial Cross-Scale Consistency Pursuit, ACSCP)，通过损失函数来约束局部图像块计数之和与整体图像计数相等，以此来模型增强跨尺度的一致性。此外，研究人员也尝试利用不同类型的卷积来提升模型性能。Li等人[11]提出拥挤场景识别网络(Congested Scene Recognition Network, CSR-Net)，利用扩张卷积在避免过度下采样的前提下增加网络的感知范围，从而更好地聚合多尺度信息。Cao等人[12]提出尺度聚合网络(Scale Aggregation Network, SANet)，利用反卷积来恢复下采样后图像的分辨率，从而生成高质量的密度图像。

目前，基于CNN的方法主要存在两个问题：(1)图像中的人群尺度呈现出区域性相似的特点，而人群计数研究普遍使用的是像素级欧几里得损失，其假设图像像素相互独立，难以捕捉人群尺度的局部相关性，因此影响了模型计数的准确性。(2)人群尺度在图像内与图像间都存在巨大的变化，而目前多数方法对于这种连续变化的多尺度信息处理能力有限，难以覆盖各种不同的尺度，因此所提取的特征无法准确描述人群信息，导致模型性能不佳。

针对以上问题，本文提出一种基于多尺度增强网络的人群计数模型(Multi-Scale Enhanced Network, MSEN)。首先，基于生成对抗网络(Generative Adversarial Networks, GAN)的思想设计了嵌入式GAN模块，其中生成网络由视觉几何组(Visual Geometry Group-16, VGG-16)[13]模型部分结构与多分支扩张卷积结构组合而成；引入块对抗生成网络(Patch-Generative Adversarial Networks, PatchGAN)[14]作为判别网络，利用其区域性判别机制引导生成网络，提升其所产生图像的局部相关性。其次，基于金字塔池化结构[15]设计了尺度增强模块，将该模块作用在嵌入式GAN模块所生成的图像上，进一步从不同区域提取不同尺度的局部特征，并生成最终的人群密度分布图像。所设计的GAN模块嵌入在整体模型中，其中的判别网络仅监督中间结果的生成模型，不参与最终人群密度分布图像的生成过程。在3个广泛使用的人群计数数据集上进行了实验，结果表明所述模型的性能优于其他对比方法。

2　多尺度增强网络

本文提出了基于多尺度增强网络的人群计数模型(MSEN)，该模型可视为一种嵌入式的GAN结构，其中嵌入式GAN模块学习人群特征并优化图像的局部相关性，尺度增强模块进一步提取局部多尺度特征并生成最终的人群密度图像。MSEN模型结构如图1所示，其包含3个部分：生成网络、判别网络以及尺度增强模块。生成网络与判别网络嵌入在整体模型中，构成嵌入式GAN模块。其中，生成网络由VGG-16模型部分结构与多分支扩张卷积结构组合而成学习人群不同尺度下的特征，判别网络仅监督中间结果的生成。此外，模型采用了跳跃连接设置，以保留输入图像的结构和上下文信息。

2.1　生成网络

受文献[11]启发，本文基于VGG-16模型构建生成网络的前端，该模型具有强大的特征提取能力与迁移学习能力，有利于复杂人群特征的提取。由于原VGG-16模型包含13个卷积层和5个池化层，

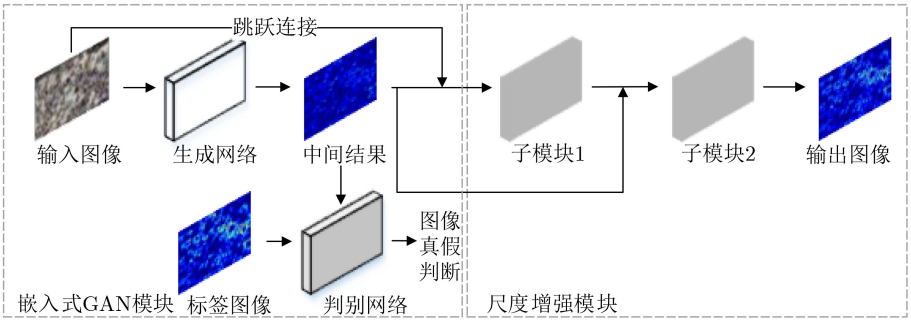


图 1 MSEN模型结构示意图

因此其网络深层的特征图尺寸非常小，不利于小尺度目标的建模。为了避免过度采样造成的小尺度目标信息丢失问题，本文首先删除了原VGG-16模型的全连接层，之后利用其前10个卷积层和3个池化层来构建生成网络前端。为了聚合更丰富的多尺度信息，本文设计了多分支结构以构建生成网络的后端。多分支结构基于扩张卷积所设计，可在不增加参数量的前提下扩大网络的感知范围，有利于应对图像间人群规模和尺度的变化。后端网络由3个分支组成，每个分支包含具有不同扩张因子的扩张卷积，扩张因子依次为1, 2, 4。扩张因子为1的分支用于捕获小尺度目标的特征，其余分支则扩大感知范围以捕获大尺度目标的特征。如第1节所述，彼此独立的分支难以学习到不同模式的特征，易导致参数冗余。因此，本文将各分支网络的特征图在每层进行拼接，并使用1×1卷积进行跨通道特征聚合，加强各分支间的信息交互，从而充分利用各分支提取特征的互补性，使输出的特征图更具表达能力和尺度多样性。生成网络的具体结构如图2所示，图2中方框内的参数表示为“卷积层-卷积核大小-通道数量-扩张因子”。

2.2 判别网络

区域性判别网络最早应用于图像转换任务，受其启发，本文采用PatchGAN[14]来构建嵌入式GAN模块中的判别网络，其具体结构表示如下：$C(4, 64, 2)$-$C(4, 128, 2)$-$C(4, 256, 2)$-$C(4, 512, 1)$-$C(4, 1, 1)$，其中C表示卷积层，括号内的参数依次为卷积核大小、通道数量以及卷积步长。除最后一层外，每个卷积层之后添加了批量标准化层(Batch Normalization, BN)与LeakyReLU激活函数。区别于常规判别网络，本文所采用的网络为全卷积网络，其输出为一个$N\times N$矩阵，而非标量值。矩阵中的每一个元素映射于原图的一个局部图像块，反映该图像块的真实性。针对此矩阵计算误差，可使网络更加关注于图像的局部区域，有利于引导生成网络得到局部相关性更高的人群密度图像。

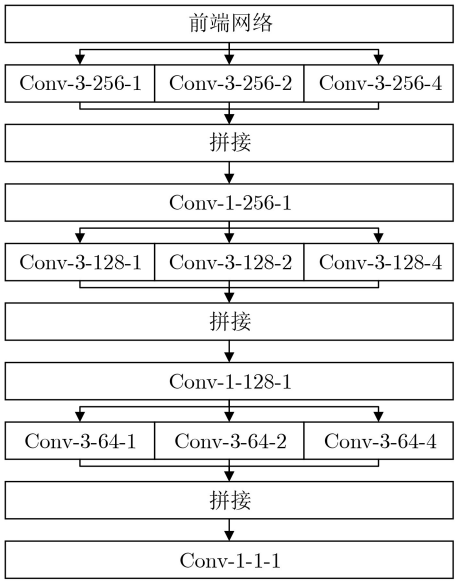


图 2 生成网络结构示意图

2.3 尺度增强模块

上文所述的嵌入式GAN模块学习人群特征并优化了密度图像的局部相关性。在此基础上，本文设计了尺度增强模块，以进一步从不同区域提取不同尺度的局部特征，从而增强模型的泛化能力。

尺度增强模块由两个具有相同结构的子模块串联构成，子模块则基于金字塔池化结构所设计。如图3所示，对于上一层网络的输入，子模块首先通过两个大小为3×3的前端卷积层进行特征提取，之后将前端卷积层的输出按4个级别进行平均池化。由于人群图像中的场景为包含众多目标的复杂场景，且人群规模和尺度呈现连续变化的特点，而传统金字塔池化结构中的全局平均池化不足以反映不同目标各自的尺度特征，因此本文将4个级别的池化尺寸依次设置为2×2, 3×3, 6×6, 8×8。上述操作将特征图按比例划分为多个大小不同的子区域，并对每个子区域进行平均池化，由此来反映每个子区域的局部特征。之后，将各自比例的池化结果通过大小为1×1的卷积层进行降维，并使用双线性插值操作上采样到原始特征图的尺寸，随后与原始特征

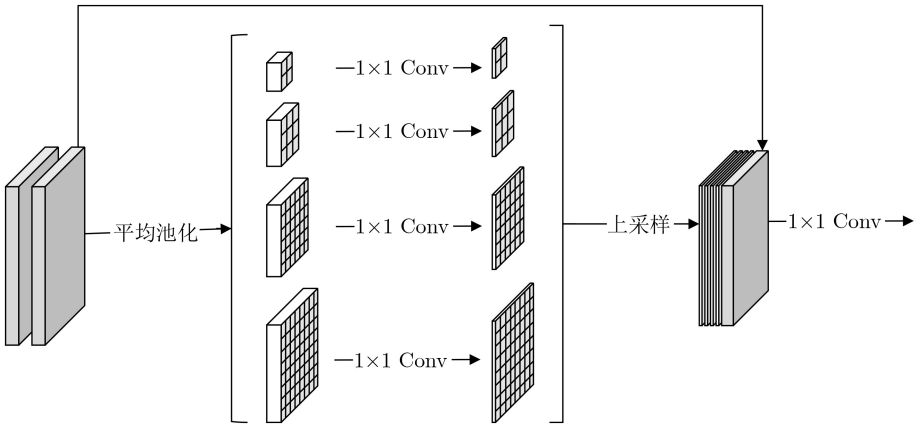


图 3 尺度增强子模块结构示意图

图拼接。最后，本文使用一个大小为3×3的后端卷积层对拼接后的特征图进行跨通道聚合，从而产生子模块的最终输出。

本文将原始图像经跳跃连接后输入第1个子模块，将第1个子模块的输出与嵌入式GAN模块的输出拼接后输入第2个子模块。通过上述操作，尺度增强模块可进一步从不同区域提取不同尺度的局部特征，以应对图像内人群尺度连续变化的特点，实现整体模型泛化能力的增强。

2.4 损失函数

人群计数研究普遍使用的欧几里得损失假设像素相互独立，忽略了图像的局部相关性。因此本文使用3项损失函数联合优化所述模型，分别为L_1损失、对抗性损失与欧几里得损失。L_1损失与对抗性损失约束嵌入式GAN模块产生的初步预测图像并优化其局部相关性，得出欧几里得损失约束模型的最终预测图像。L_1损失定义如式(1)

$$L_1 = \frac{1}{n}\sum_{i=1}^{n}||G(x_i) - y_i||_1 \tag{1}$$

其中，n为训练样本的数量，x_i为输入图像，y_i为对应的标签图像，G为生成网络，$G(x_i)$为生成网络根据输入图像产生的模型中间预测结果。对抗性损失定义如式(2)

$$\min_G \max_D L_{\mathrm{A}}(G,D) = E_{y\sim P_{\mathrm{data}(y)}}[\lg D(y)] + E_{x\sim P_{\mathrm{data}(x)}}[\lg(1 - D(G(x)))] \tag{2}$$

其中，x为输入图像，y为对应的标签图像，G为生成网络，D为判别网络，$G(x)$为生成网络根据输入图像产生的模型中间预测结果。欧几里得损失定义如式(3)

$$L_{\mathrm{E}} = \frac{1}{n}\sum_{i=1}^{n}||m_i - y_i||_2^2 \tag{3}$$

其中，n为训练样本的数量，m_i为模型最终预测的密度图像，y_i为对应的标签图像。3项损失函数经加权组合形成模型最终的目标函数，定义如式(4)

$$L = \alpha L_{\mathrm{A}} + \beta L_1 + L_{\mathrm{E}} \tag{4}$$

其中，α与β为平衡3项损失的权重，二者的取值依据将于后续对比实验部分说明。

2.5 训练步骤

由于本文设计的多尺度增强网络为一种嵌入式的GAN结构，因此整体模型无法遵循传统GAN模型的训练步骤。受文献[14]启发，本文采用一种新的交替训练步骤来优化所述模型，在该训练步骤中，生成网络将进行两次参数更新，具体步骤如下：

步骤 1　加载训练数据集，进行数据预处理；

步骤 2　初始化模型训练参数，输入训练数据；

步骤 3　提升式(2)的梯度，以更新判别网络的参数；

步骤 4　降低式(1)与式(2)之和的梯度，以更新生成网络的参数；

步骤 5　降低式(3)的梯度，以分别更新生成网络与尺度增强模块的参数；

步骤 6　重复步骤3～5，直至训练结束。

3　实验与分析

3.1 实验数据集

为了验证所述模型的有效性，本文采用人群计数研究常用的3个数据集进行实验，分别为ShanghaiTech[8], UCF_CC_50[16], UCF-QNRF[17]。ShanghaiTech数据集包含1198张人群图像，其中共标注了330165个目标行人。该数据集分为Part_A和Part_B两个部分。Part_A部分共包含482张搜集于互联网的人群图像，具体分为300张训练图像和182张测试图像。Part_B部分共包含716张拍摄于上海某步行街的人群图像，具体分为400张训练图像和316张测试图像，相对而言，Part_B部分图

像场景中的人群较为稀疏。UCF_CC_50数据集共包含50张搜集于互联网的人群图像，其中共标注了63075个目标行人。其中图像包含人数平均为1280人，单张图像包含人数94～4543不等。该数据集所含数据量较少，因此本文遵循文献[15]所提出的5折交叉验证方法来进行实验。UCF-QNRF数据集共包含1535张人群图像，其中共标注了1.25×10^6个目标行人。该数据集具体分为1201张训练图像和334张测试图像，单幅图像包含人数49～12865不等。上述3个数据集的基本信息如表1所示。

3.2 评价指标

本文采用人群计数研究普遍使用的两个评价指标来评估模型的性能，分别为平均绝对误差(Mean Absolute Error, MAE)和均方误差(Mean Square Error, MSE)。MAE反映模型预测准确性，MSE反映模型预测鲁棒性，二者数值越低，表明模型性能越好。

3.3 实验设置

本文所采用的实验环境为：Intel Xeon(R) Sliver 4110 2.10 GHz CPU, Quadro P5000 GP(16G显存)。使用的操作系统为Ubuntu 16.04，采用的深度学习框架为Pytorch框架。本文采用在ImageNet数据集上预训练的VGG-16模型参数来初始化生成网络的前端，其余各网络的参数使用均值为0，标准差为0.01的高斯分布随机初始化。模型通过Adam算法进行优化，学习率固定为0.0000001，总迭代次数为30000次。

对于ShanghaiTech Part_A, UCF_CC_50和UCF-QNRF数据集，本文采用几何自适应高斯核为其制作标签密度图像；而对于ShanghaiTech Part_B数据集，因其图像中的人群较为稀疏，本文采用固定高斯核为其制作标签密度图像。此外，对于ShanghaiTech与UCF_CC_50数据集，本文以原始图像尺寸进行训练，设置批处理大小为1，并通过随机水平翻转来进行数据增强。由于UCF-QNRF数据集中均为高分辨率图像(例如9000×6000)，本文遵循文献[18]提出的训练方法，将原始图像裁剪为16张不重叠的且尺寸为224×224的子图像，并设置批处理大小为16进行训练。

表 1 数据集基本信息对比

数据集	图像数量	分辨率	最小人数	最大人数	密度水平
ShanghaiTech Part_A	482	不等	33	3139	高
ShanghaiTech Part_B	716	768×1024	9	578	低
UCF_CC_50	50	不等	94	4543	极高
UCF-QNRF	1535	不等	49	12865	极高

3.4 实验结果分析

ShanghaiTech数据集的实验结果如表2所示，本文将所述模型与7种近年人群计数研究的主流方法进行了比较。对于Part_A部分，所述模型获得了最低的MAE值，相比方法TEDnet降低了1.1%，所述模型的MSE值也接近于该指标表现最好的方法ACSCP。对于Part_B部分，所述模型分别获得了最低的MAE值与MSE值，其中MAE指标与方法TEDnet持平，MSE指标相比TEDnet降低了3.9%。在ShanghaiTech数据集两个部分的实验结果表明，所述模型在拥挤和稀疏的人群场景中均可表现出良好的性能。

UCF_CC_50数据集的实验结果如表3所示，本文同样将所述模型与7种近年来人群计数研究的主流方法进行了比较。所述模型在MAE指标与MSE指标上均获得了最低值，其中MAE指标相比方法TEDnet降低了9.1%，MSE指标降低了12.4%。该数据集所含样本数量较小，仅为50张图像。实验结果表明，所述模型对于小样本数据也可表现出良好的适应性。

UCF-QNRF数据集是2018年公布的最新数据集之一，目前使用该数据集进行评估的方法相对较少，本文将所述模型与4种主流方法进行了比较，结果如表4所示。所述模型获得了具有竞争力的MAE

表 2 ShanghaiTech数据集实验结果

方法	Part_A		Part_B	
	MAE	MSE	MAE	MSE
MCNN[8]	110.2	173.2	26.4	41.3
Switch-CNN[9]	90.4	135.0	21.6	33.4
ACSCP[10]	75.7	**102.7**	17.2	27.4
CSRNet[11]	68.2	115.0	10.6	16.0
SANet[12]	67.0	104.5	8.4	13.6
PACNN[19]	66.3	106.4	8.9	13.5
TEDnet[20]	64.2	109.1	**8.2**	12.8
MSEN (本文)	**63.5**	106.2	**8.2**	**12.3**

表 3 UCF_CC_50数据集实验结果

方法	MAE	MSE
MCNN[8]	377.6	509.1
Switch-CNN[9]	318.1	439.2
ACSCP[10]	291.0	404.6
CSRNet[11]	266.1	397.5
PACNN[19]	267.9	357.8
SANet[12]	258.4	334.9
TEDnet[20]	249.4	354.5
MSEN (本文)	**226.7**	**310.6**

值，同时获得了最低的MSE值。相比方法TEDnet，所述模型的MAE指标降低了15.2%，MSE指标也与之接近。该数据集具有样本数量多，场景复杂等特点，在此情况下所述模型的预测准确性有待提高。同时，所述模型的预测鲁棒性较好，表明其具有良好的泛化能力。

3.5 消融实验

为了进一步验证所述模型各部分结构的有效性，本文基于ShanghaiTech Part_A数据集设计了模型结构对比实验，具体关注模型结构的3个因素：是否采用嵌入式GAN结构、尺度增强子模块的数量、是否采用跳跃连接设置。为了平衡模型性能与资源开销，将尺度增强子模块的最大数量限制为2。具体而言，本文基于排列组合原理构造了10种不同结构的模型，并将各模型的具体描述与对应结果展示于表5，其中尺度增强子模块记作E，跳跃连接记作S：

(1) 仅包含生成网络，记作G。

(2) 在模型(1)的基础上增加了判别网络，构成生成对抗网络，记作GAN。

(3～6) 模型结构均为非嵌入式GAN结构(分别对应于(7～10)的嵌入式GAN结构)，记作GAN*。在此类模型中，本文将原生成网络与尺度增强模块组合，将组合后的整体结构作为独立生成网络，并使用判别网络直接监督模型的最终输出。

表 4　UCF-QNRF数据集实验结果

方法	MAE	MSE
MCNN[8](CL)	277.0	426.0
Switch-CNN[9] (CL)	228.0	445.0
CL[18]	132.0	191.0
TEDnet[20]	**113.0**	188.0
MSEN (本文)	114.1	**159.5**

(7) 嵌入式GAN结构，之后连接1个尺度增强子模块。

(8) 在模型(7)的基础上增加了跳跃连接设置。

(9) 嵌入式GAN结构，之后连接2个尺度增强子模块。

(10)在模型(9)的基础上增加了跳跃连接设置，即为本文所提多尺度增强网络模型(MSEN)。

由表5可知，模型(2)的性能优于模型(1)，表明引入区域性判别网络可优化图像局部相关性并提升模型计数准确性；模型(4), (8)的性能分别优于模型(3), (7)，表明采用跳跃连接设置有助于重建输入图像的结构和全局上下文信息；模型(9)的性能优于模型(7)，表明采用两个尺度增强子模块更有利于据合图像各区域的多尺度局部特征；在具有相同配置的前提下，采用嵌入式GAN结构的模型的性能均优于对应的非嵌入式GAN结构模型，且模型(5), (6)在所有模型中性能最差，原因或为原生成网络与尺度增强模块组合构成的独立生成网络的结构较为复杂，参数量过大，导致整体模型在训练时难以收敛，由此也证明了采用嵌入式GAN结构的有效性。

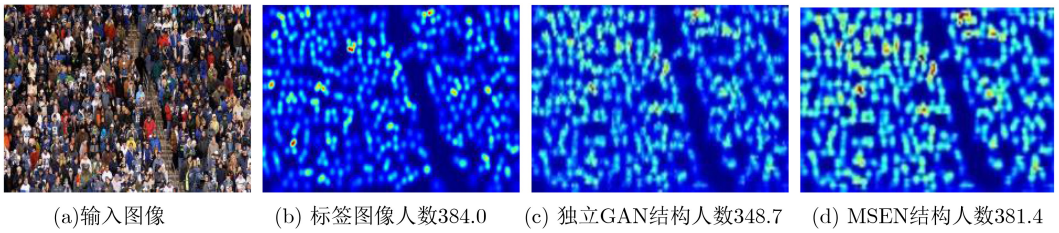
此外，为了进一步证明在嵌入式GAN模块之后连接尺度增强模块的有效性，本文将模型(2)与模型(10)预测图像的结果对比展示于图4，二者的结构分别为GAN结构与本文所述的MSEN结构，区别为模型是否包含尺度增强模块。可以看出，由模型(10)，即本文所述MSEN结构预测的图像可以更好地反映人群分布的热点情况，且根据预测图像计算出的人数更加接近标签图像实际包含的人数，因此进一步证明了尺度增强模块的有效性。

3.6 损失函数权重选择实验

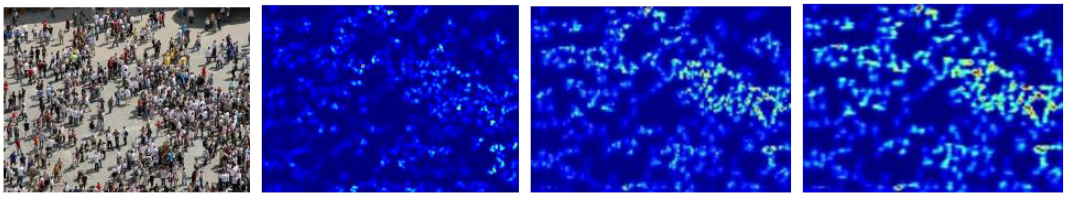
为了对损失函数中的权重取值依据进行说明，本文对了不同参数权重下模型的性能。从简化模型

表 5　不同结构的模型及其对应的实验结果

模型序号	结构概述	嵌入式	尺度增强子模块数量	跳跃连接	MAE
(1)	G	–	–	–	67.5
(2)	GAN	–	–	–	65.6
(3)	GAN*GAN*(E×1)	–	1	–	65.3
(4)	GAN*GAN*(E×1+S)	–	1	√	65.2
(5)	GAN*GAN*(E×2)	–	2	–	66.5
(6)	GAN*GAN*(E×2+S)	–	2	√	66.4
(7)	嵌入式GAN+ E×1	√	1	–	65.0
(8)	嵌入式GAN+ E×1+S	√	1	√	64.7
(9)	嵌入式GAN+ E×2	√	2	–	64.1
(10)	嵌入式GAN+ E×2+S (MSEN)	√	2	√	**63.5**



(a)输入图像 (b) 标签图像人数384.0 (c) 独立GAN结构人数348.7 (d) MSEN结构人数381.4



(e) 输入图像 (f) 标签图像人数579.0 (g) 独立GAN结构人数522.0 (h) MSEN结构人数575.9

图 4 独立GAN结构与MSEN结构的预测图像与计算人数示例

训练流程的角度出发，本文首先对比了各项损失函数回传梯度的大小，并将权重α设置为3，之后选取6个代表性数值作为权重β的基准数值，通过对比实验来确定其最终取值。实验结果如图5所示，随着β取值大小的增加，模型的MAE指标持续降低。当β=1时，损失函数中L_1与L_E的权重相等，同时模型获得最低的MAE指标。当β取值继续增加，即L_1与L_E之间的权重差距逐渐增大时，MAE指标迅速增加，即模型性能开始下降。因此，当β取值为1时，模型性能达到最佳。

4 结论

为了解决人群计数研究忽略图像局部相关性以及模型对多尺度特征提取能力有限的问题，本文提出了一种基于多尺度增强网络的人群计数模型(MSEN)，将所设计的多分支生成网络与引入的区域性判别网络组合构成嵌入式GAN模块，在其之后连接基于金字塔池化结构所设计的尺度增强模块，使用3项损失函数对整体模型进行联合训练，使所述模型在提升预测图像局部相关性的同时提升了多尺度特征的提取能力，从而提升模型最终的计数准确性与鲁棒性。本文在3个人群计数公共数据集上进行了广泛的实验，实验结果证明了所述模型的有效性。

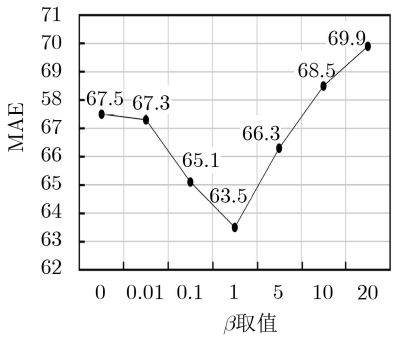


图 5 不同β取值与对应的模型MAE值

参 考 文 献

[1] 陈朋, 汤一平, 王丽冉, 等. 多层次特征融合的人群密度估计[J]. 中国图象图形学报, 2018, 23(8): 1181–1192. doi: 10.11834/jig.180017.

[2] XIE Weidi, NOBLE J A, and ZISSERMAN A. Microscopy cell counting and detection with fully convolutional regression networks[J]. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 2018, 6(3): 283–292. doi: 10.1080/21681163.2016.1149104.

[3] 左静, 窦祥胜. 视频车辆分类与计数的模型与应用[J]. 运筹与管理, 2020, 29(1): 124–130.

[4] CUI Kai, HU Cheng, WANG Rui, *et al*. Deep-learning-based extraction of the animal migration patterns from weather radar images[J]. *Science China Information Sciences*, 2020, 63(4): 140304. doi: 10.1007/s11432-019-2800-0.

[5] 孙彦景, 石韫开, 云霄, 等. 基于多层卷积特征的自适应决策融合目标跟踪算法[J]. 电子与信息学报, 2019, 41(10): 2464–2470. doi: 10.11999/JEIT180971.

SUN Yanjing, SHI Yunkai, YUN Xiao, *et al*. Adaptive strategy fusion target tracking based on multi-layer convolutional features[J]. *Journal of Electronics & Information Technology*, 2019, 41(10): 2464–2470. doi: 10.11999/JEIT180971.

[6] 蒲磊, 冯新喜, 侯志强, 等. 基于空间可靠性约束的鲁棒视觉跟踪算法[J]. 电子与信息学报, 2019, 41(7): 1650–1657. doi: 10.11999/JEIT180780.

PU Lei, FENG Xinxi, HOU Zhiqiang, *et al*. Robust visual tracking based on spatial reliability constraint[J]. *Journal of Electronics & Information Technology*, 2019, 41(7): 1650–1657. doi: 10.11999/JEIT180780.

[7] ZHANG Cong, LI Hongshen, WANG Xiaogang, *et al*. Cross-scene crowd counting via deep convolutional neural networks[C]. The IEEE Conference on Computer Vision and

Pattern Recognition, Boston, USA, 2015: 833–841. doi: 10.1109/CVPR.2015.7298684.

[8] ZHANG Yingying, ZHOU Desen, CHEN Siqin, *et al*. Single-image crowd counting via multi-column convolutional neural network[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 589–597. doi: 10.1109/CVPR.2016.70.

[9] SAM D B, SURYA S, and BABU R V. Switching convolutional neural network for crowd counting[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 4031–4039.

[10] SHEN Zan, XU Yi, NI Bingbing, *et al*. Crowd counting via adversarial cross-scale consistency pursuit[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 5245–5254. doi: 10.1109/CVPR.2018.00550.

[11] LI Yuhong, ZHANG Xiaofan, and CHEN Deming. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 1091–1110. doi: 10.1109/CVPR.2018.00120.

[12] CAO Xinkun, WANG Zhipeng, ZHAO Yanyun, *et al*. Scale aggregation network for accurate and efficient crowd counting[C]. The 15th European Conference on Computer Vision, Munich, Germany, 2018: 757–773.

[13] SIMONYAN K and ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]. The International Conference on Learning Representations, San Diego, USA, 2015: 1–14.

[14] ISOLA P, ZHU Junyan, ZHOU Tinghui, *et al*. Image-to-image translation with conditional adversarial networks[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 5967–5976.

[15] ZHAO Hengshuang, SHI Jianping, QI Xiaojuan, *et al*. Pyramid scene parsing network[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6230–6239. doi: 10.1109/CVPR.2017.660.

[16] IDREES H, SALEEMI I, SEIBERT C, *et al*. Multi-source multi-scale counting in extremely dense crowd images[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013: 2547–2554. doi: 10.1109/CVPR.2013.329

[17] IDREES H, TAYYAB M, ATHREY K, *et al*. Composition loss for counting, density map estimation and localization in dense crowds[C]. The 15th European Conference on Computer Vision, Munich, Germany, 2018: 544–559. doi: 10.1007/978-3-030-01216-8_33.

[18] QU Yanyun, CHEN Yizi, HUANG Jingying, *et al*. Enhanced pix2pix dehazing network[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 8152–8160. doi: 10.1109/CVPR.2019.00835.

[19] SHI Miaojing, YANG Zhaohui, XU Chao, *et al*. Revisiting perspective information for efficient crowd counting[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 7271–7280.

[20] JIANG Xiaolong, XIAO Zehao, ZHANG Baochang, *et al*. Crowd counting and density estimation by trellis encoder-decoder networks[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 6126–6135. doi: 10.1109/CVPR.2019.00629.

徐　涛：男，1962年生，教授、博士生导师，研究方向为智能信息处理、图像处理.

段仪浓：男，1994年生，硕士生，研究方向为计算机视觉与模式识别.

杜佳浩：男，1994年生，硕士生，研究方向为计算机视觉与模式识别.

刘才华：女，1987年生，讲师、博士，研究方向为机器学习与计算机视觉.

责任编辑：余　蓉