基于新型多尺度注意力机制的密集人群计数算法

万洪林*① 王晓敏① 彭振伟① 白智全③ 杨星海④ 孙建德②

①(山东师范大学物理与电子科学学院 济南 250358)

②(山东师范大学信息科学与工程学院 济南 250358)

③(山东大学信息科学与工程学院 青岛 266237)

④(青岛科技大学信息科学技术学院 青岛 266061)

摘 要：密集人群计数是计算机视觉领域的一个经典问题，仍然受制于尺度不均匀、噪声和遮挡等因素的影响。该文提出一种基于新型多尺度注意力机制的密集人群计数方法。深度网络包括主干网络、特征提取网络和特征融合网络。其中，特征提取网络包括特征支路和注意力支路，采用由并行卷积核函数组成的新型多尺度模块，能够更好地获取不同尺度下的人群特征，以适应密集人群分布的尺度不均匀特性；特征融合网络利用注意力融合模块对特征提取网络的输出特征进行增强，实现了注意力特征与图像特征的有效融合，提高了计数精度。在ShanghaiTech, UCF_CC_50, Mall和UCSD等公开数据集的实验表明，提出的方法在MAE和MSE两项指标上均优于现有方法。

关键词：人群计数；新型多尺度注意力；卷积神经网络；人工智能

中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2022)03-1129-08

DOI: 10.11999/JEIT210163

Dense Crowd Counting Algorithm Based on New Multi-scale Attention Mechanism

WAN Honglin① WANG Xiaomin① PENG Zhenwei① BAI Zhiquan③

YANG Xinghai④ SUN Jiande②

①(*School of Physics and Electronic Science, Shandong Normal University , Jinan* 250358, *China*)

②(*School of Information Science and Engineering, Shandong Normal University, Jinan* 250358, *China*)

③(*School of Information Science and Engineering, Shandong University, Qingdao* 266237, *China*)

④(*School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao* 266061, *China*)

Abstract: Dense crowd counting is a classic problem in the field of computer vision, and it is still subject to the influence of factors such as uneven scale, noise and occlusion. This paper proposes a dense crowd counting method based on a new multi-scale attention mechanism. Deep network includes backbone network, feature extraction network and feature fusion network. Among them, the feature extraction network includes feature branch and attention branch. It adopts a new multi-scale module composed of parallel convolution kernel functions, which can better obtain the characteristics of people at different scales to adapt to the uneven scale of dense population distribution features; The feature fusion network uses the attention fusion module to enhance the output features of the feature extraction network, realizes the effective fusion of attention features and image features, and improves counting accuracy. Experiments on public data sets such as ShanghaiTech, UCF_CC_50, Mall and UCSD show that the proposed method outperforms existing methods in both MAE and MSE indicators.

Key words: Crowd counting; New multi-scale attention; Convolutional neural network; Artificial intelligence

收稿日期：2021-02-25；改回日期：2021-10-23；网络出版：2021-11-11

*通信作者： 万洪林 visage1979@sdu.edu.cn

基金项目：国家自然科学基金(61971271)，山东省重点研发计划(2018GGX106008)

Foundation Items: The National Natural Science Foundation of China (61971271), The Key Research and Development of Shandong Province (2018GGX106008)

1 引言

随着庆祝活动、音乐会、体育赛事、公众游行等大型公共活动日益增多，人群大量集聚的场景不断出现，各种踩踏受伤事件也层出不穷，因此对密集人群进行有效监管非常必要。人群计数能够为大规模人群聚集的监管提供技术支持[1,2]。如果能够使用计算机视觉技术对相关场景的密集人群进行准确的人群密度估计，则会对减少危险事件的发生带来很大的帮助。目前，准确统计出在不同的场景下的人群总数仍然具有很大的难度，因此这一领域所面对的问题具有一定挑战性。早期人群计数大多采用传统的检测和回归方法，需要人为提取出图片中的低层次特征，并将其用特征框标记出来，标记框的数量即为图片中行人的数量，随着人群密度的不断提高，人与人之间的互相遮挡越来越严重，再加上密集人群分布不均、光照等因素的影响，这些问题都对密集人群计数提出了更高的挑战。近年来随着深度学习的快速发展，人群计数也更多地采用此方法。深度学习方法相对于传统的检测、回归方法，准确性和适用性要更好。通过卷积神经网络将卷积核与图像做卷积操作，通过一系列的卷积核，不断提取图像的特征，最后将提取到的高层特征进行分类生成密度图，再对密度图求和来统计人群的总体数量。但其只是将这些特征做了简单的操作，不能较好地利用这些特征。为此本文提出了一种新的基于新型多尺度注意力机制的密集人群计数方法。其网络结构分为主干网络、特征提取网络和特征融合网络。特征提取网络分为两个支路：特征支路和注意力支路。考虑到数据尺度特征的多样性，本文的两个支路都增加了新型多尺度模块，并在特征支路单独增加了Res结构，以便更好地获取不同尺度下的人群特征。注意力支路用于不断加强密集人群图像中的头部特征，从而使得头部区域的密度图相较而言更加明显。在特征融合网络中，通过注意力融合模块，将注意力特征与图像特征进行有效融合，进一步提高计数精度。在公开数据集(ShanghaiTech, UCF_CC_50, Mall, UCSD)上的实验获得了比其他方法更好的参数指标。

2 相关工作

早期行人计数主要采用的是基于检测的方法，但这类方法不够准确，效率较低。随着深度学习技术的飞速发展，人们更加倾向使用卷积神经网络(Convolutional Neural Networks, CNN)实现对于密集人群的计数。Shang等人[3]提出了一种使用CNN的端到端计数估计方法，将整个图片作为输入，最后直接输出人群总数。针对图像密度、视角信息差异大的问题，Zhang等人[4]提出了MCNN方法，即通过使用多个卷积核大小不同的网络来捕捉不同尺度的目标特征信息，以增强模型的稳定性。通过估计具有任意人群密度和任意视角的图像，从而生成图像或视频中真实的人群密度图。与MCNN类似，Onoro-Rubio等人[5]提出了一种尺度感知计数模型Hydra，通过尺度放缩的思想考虑了视角差异带来的影响，即使没有任何明确的场景信息，这一模型也能估计各种各样的拥挤场景中的密度。Marsden等人[6]受到尺度感知模型的启发，提出了一种基于Resnet-18[7]架构的网络，可同时实现人群计数、暴力行为检测和人群密度等级分类的工作。Li等人[8]首先提出MCNN的劣势：训练时间长以及无效分支结构，然后提出使用空洞卷积以获得更大的感受域并提取更深层次的特征。Sam等人[9]提出了选择卷积神经网络（Switch-CNN）来提升人群计数的精确度，首先由几个卷积核大小不同的CNN作为密度图预测的回归器，然后再由一个选择分类器来为每一张输入图像选取最优回归器，将其得到结果作为最终结果。此外，Wang等人[10]提出了一种数据收集器和贴标机，它可以生成合成人群场景，不需要任何人力就可以对图片进行注释。在此基础上，作者还构建了一个大规模、多样化的合成数据集。Li等人[11]提出了一种针对可自由移动人体在单个摄像机场景下估计深度密度图的方法。Wang等人[12]设计了一个包含结构特征编码器、空间上下文学习解码器以及密度回归模块在内的网络结构。这一网络从频道维度和空间维度两个维度来获取空间上下文信息，以此来提高网络性能。Chen等人[13]提出了相关区域预测方法，即统计密度图中的像素之和代表输入图像中落入相应局部区域的数量。这一方法丢弃了详细的空间信息，使网络更加关注计数而不是对具体每个人进行定位，从而相应地提高计数的准确性。多列结构在一定程度上解决了人群计数存在的尺度变化问题，但不同CNN学习到的多尺度人群特征如何在保证信息不丢失的情况下充分融合利用，提高输出密度图质量，仍是多列结构没有解决的难题。为此孟月波等人[14]提出了一种编码-解码结构的多尺度卷积神经网络来进行人群计数，提升了密度图的输出质量。编码器采集更加丰富的尺度信息，解码器对编码器的输出进行上采样，实现了高层语义信息和前端低层特征信息的融合。左静等人[15]提出了一种多尺度融合的深度人群计数算法，以膨胀卷积理论为基础，构建多尺度特征提取模块，以此实现上下文特

征信息提取。最后经过特征融合得到更高质量的密度图。Zou等人[16]提出了自适应容量多尺度卷积神经网络(ACM-CNN)，它可以为输入的不同部分分配不同的容量。该模型以输入图像的重要区域为中心，在满足人群密集度的前提下，优化其容量分配。尽管取得了很大进展，但由于密集人群计数场景下人群分布不均、光照、遮挡等因素带来的影响，上述方法仍然存在改进空间。

3　网络结构

针对当前密集人群计数存在的问题，本文提出了基于新型多尺度注意力机制的密集人群计数算法。其基本思想，一是通过双通道特征提取网络取代传统的单通道网络结构，将人头定位与密度图结合，实现更丰富的特征提取；二是引入新型多尺度模块，增强网络对不同尺度特征的适应性；三是引入空间注意力机制，进一步丰富特征形态，从而为高质量的密度图生成奠定基础。

本文提出的网络结构分为3部分，即主干网络、特征提取网络和特征融合模块(如图1所示)。

3.1　主干网络

主干网络主要用于图像特征的提取，本文采用的骨干网络为VGG-16，其中有4层特征作为主干网络的输出特征，分别是conv2_2，conv3_3，conv4_3和conv5_3(如图2所示)。

3.2　特征提取网络

本文提出的网络模型中，特征提取网络采用新型注意力机制。它包括两个支路：特征支路与注意力支路。特征支路用来提取图像中的人群分布特征；注意力支路则用于准确估计人头位置，对得到的人群密度图进行修正，得到较高质量的人群密度估计图。

特征支路包括基础特征提取模块(如图2所示)、新型多尺度模块和辅助结构。基础特征提取模块主要用于将低分辨率特征恢复为高分辨率特征，为密集人群计数的密度图估计提供更丰富的空间分布信息。注意力支路包括基础注意力模块和新型多尺度模块。在本文中，基础注意力模块的结构与基础特征模块相同，作用是将低分辨率特征恢复为高分辨率特征，有利于人头位置的精准定位。

针对特征提取网络，本文提出了新型多尺度模块，用于改善两个支路的输出特征，提高计算效率。随着神经网络深度的不断增加，网络参数体量越来越大，而其中大量参数的权值趋于零，冗余度高，浪费计算资源。解决此问题的一种方法就是引入稀疏滤波器。由此Szegedy提出了inception结构。经典的inception是由1×1，3×3，5×5卷积层和一个池化层(pooling)组成的并行结构(如图3所示)。卷积核的大小直接决定了对不同目标的感知能力。本文考虑到密集人群图像中人的大小的变化范围，为提取图像中的大尺度人群特征，我们将inception结构中的池化层，替换为7×7卷积层(如图4所示)。

同时考虑到为了提高网络计算效率，我们进一步将上述5×5卷积层，替换为2个级联的3×3卷积层，将7×7卷积层替换为3个级联的3×3卷积层。替换前后其感受野范围不会改变[17]。

由此我们提出了新型多尺度模块(如图5所示)。

新型多尺度模块增强了特征支路中人群密度特征的集中度，进一步扩大了感受野，使得每一层输出的特征图上的像素点在输入图片上映射的区域增大。同时新型多尺度模块也能够在注意力支路中增强人头位置信息。

3.3　特征融合模块

特征融合模块的作用是将注意力支路的输出特征作用于特征支路的输出特征，通过相乘的方式实现两者融合，得到更高质量的人群密度图。其中起关键作用的是注意力融合模块，其结构如图6所示。

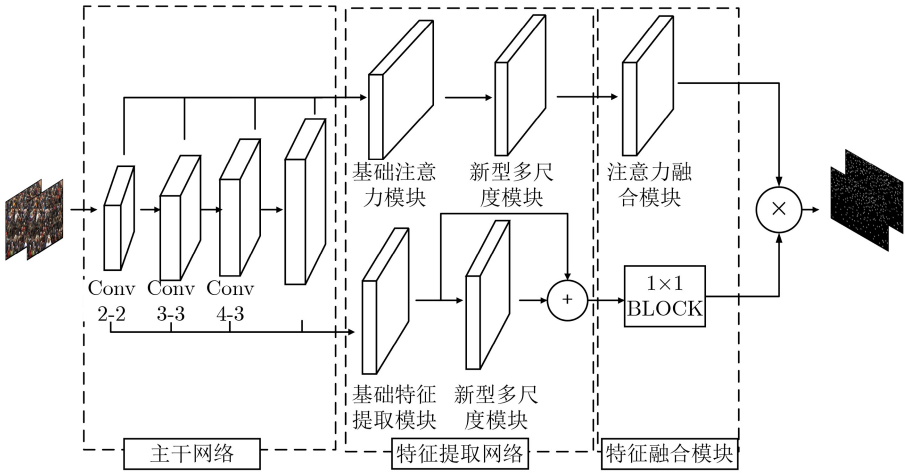

图 1 本文提出的网络结构

在高层次特征中，丰富而抽象的特征信息，对网络的特征辨识能力提出了更高的要求。在注意力融合模块中，通过矩阵变换及其组合，特征维度或元素位置发生变化，即通道信息发生改变，从而实现了特征重组。这些重组后的特征能够进一步丰富密集人群密度图的特征描述，提高网络辨识能力。本文中注意力定义为

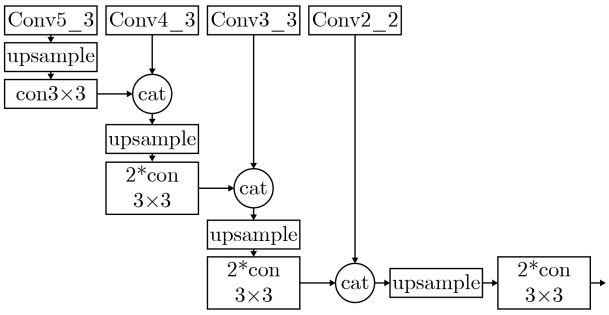

图 2 基础特征提取模块，在本文亦被采用为基础注意力模块

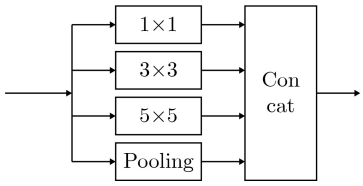

图 3 传统Inception结构

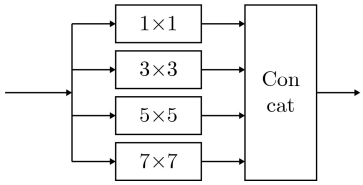

图 4 改进Inception结构

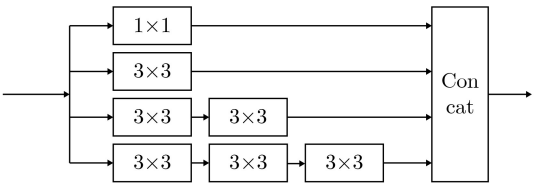

图 5 新型多尺度模块

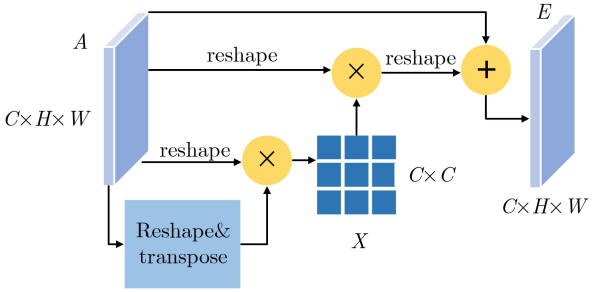

图 6 注意力融合模块

$$X \in R^{C\times C}: x_{ji} = \frac{\exp(A_i \cdot A_j)}{\sum_{i=1}^{C}\exp(A_i \cdot A_j)} \tag{1}$$

其中，C为通道数；A表示密度图特征值；x_{ji}度量第i个对第j个通道的影响。输出特征为

$$E \in R^{C\times H\times W}: E_j = \beta\sum_{i=1}^{C}(x_{ji}A_i) + A_j \tag{2}$$

其中，β表示尺度系数，初始化为0，并逐渐地学习分配到更大的权重。每个通道的结果特征为$\boldsymbol{E}$，是所有通道特征和原始特征的加权和。

3.4 损失函数

本文选取欧氏距离作为网络模型的损失函数，将网络输出的人群密度估计图回归到 ground truth 的密度图。损失函数定义为

$$L = \frac{1}{2N}\sum_{i=1}^{N}||F(X_i) - D(X_i)||^2 \tag{3}$$

其中，X_i是输入图像；$F(X_i)$是估计密度图；N是训练图像数量；$D(X_i)$是图像X_i的 ground truth。

本文引入了注意力损失函数，其中A_i^{GT}是注意力真值；P_i是由激活函数激活的预测注意力图中每个像素的概率。

$$L_{\mathrm{att}} = -\frac{1}{N}\sum_{i=1}^{N}(A_i^{\mathrm{GT}}\lg(P_i) + (1-A_i^{\mathrm{GT}})\lg(1-P_i)) \tag{4}$$

整个网络使用以下损失函数进行训练，其中α在实验中设置为0.1。

$$L_{\mathrm{all}} = L + \alpha L_{\mathrm{att}} \tag{5}$$

4 实验结果

本文实验的硬件配置为： CPU Xeon-E5，GPU Quadro P5000 / 16GB和128GB内存；软件环境是Ubuntu 16.04和Pytorch 1.0。

4.1 评估指标

现有的传统人群计数方法均采用平均绝对误差(Mean Average Error, MAE)和均方误差(Mean Square Error, MSE)两种误差来评估模型的性能。本文亦采用MAE以及MSE两项指标来评价密集人群技术网络的性能，其定义为

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|a_i - a_i^{\mathrm{GT}}|$$

$$\mathrm{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|a_i - a_i^{\mathrm{GT}}|^2} \tag{6}$$

其中，N 代表图片的数量；a_i代表本文网络得到的

估计密度图；a_i^{GT}表示对应图像的ground truth。这两项指标可以直观地反映每个算法对人群数量估计的准确程度，是各算法均采用的公平标准的评价指标。MAE和MSE越低，说明网络学习的误差越小。上述两项指标在人群计数中被普遍采用。

4.2 数据集

本文实验数据集包括ShanghaiTech[4]，UCF_CC_50[18]，Mall[19]和UCSD[20]。

上海数据集包含了1198个图像以及330165个注释头文件，它分为A和B两部分。其中A部分由300张训练图片和182张测试图片组成，B部分由400张训练图片和316张测试图片组成。上海数据集是具有不同场景和不同密度级别的数据集，非常具有挑战性，也是最有代表性的数据集。

UCF_CC_50是包括各种密度和视角的不同场景。为了捕捉场景类型的多样性，作者收集了不同的图像，像音乐会、抗议、体育馆和马拉松等场景。它包含了50个不同分辨率的图像，每个图像平均有1280人，整个数据集中共标记了63075个人。个数从94到4543不等，图像之间存在很大差异。

Mall数据集是一个具有不同光照条件以及人群密度的数据集，是使用安装在购物中心的监视摄像机收集的数据集。除了具有各种密度水平外，它还具有不同的活动模式。另外，数据集中的场景还具有严重的透视畸变，导致对象的大小和外观大的变化，该数据集还呈现了由场景对象引起的严重遮挡的挑战。数据集中的视频序列由2000帧大小为320×240的帧组成，其中标记为行人的6000个实例。前800帧用于训练，剩余的1200帧用于评估。

UCSD数据集是为人数统计创建的第1批数据集。数据集是从人行道的摄像机收集的。该数据集由来自视频的2000帧大小为238×158 以及每5个帧中每个行人的地面实况(Ground Truth)注释组成。此数据集共包含49885个行人实例，我们将601到1400作为训练集，剩余的1200张图片用来测试。

4.3 实验结果

我们评估了本文模型在主要人群数据集ShanghaiTech，UCF_CC_50，Mall和UCSD上的计数性能。并与其他人群计数模型的MAE和MSE指标进行比较。表1—表4列出了在4个数据集上不同模型的实验结果，可以看出本文的模型要优于其他方法。

4.3.1 ShanghaiTech

本文网络估计的人群密度图与ground truth的对比如图7所示。

4.3.2 UCF_CC_50

考虑到UCF_CC_50图像数量少，该数据集发布者定义了一种交叉验证协议，以此实现样本容量的扩增。我们也采用了相同的5次交叉验证策略，即将整个数据集样本均分为5份，每次训练取其中4份样本作为训练集，剩余的1份作为测试集，一共进行5次训练和测试。最后计算5次实验的MAE和MSE的均值作为测试结果。表2为本文方法对UCF_CC_50数据集的实验结果与其他方法的对比(batch size=8)，可以看出本文方法取得了更优的实验结果。

表 1　ShanghaiTech数据集实验结果

方法	Part A		Part B	
	MAE	MSE	MAE	MSE
MCNN[4]	110.2	173.2	26.4	41.3
EDMNet[14]	76.5	100.2	15.4	26.3
MSFNet[15]	63.4	97.2	9.6	14.3
Switching-CNN[9]	90.4	135.0	21.6	33.4
CSRNet[8]	68.2	115.0	10.6	16.0
SCAR[21]	66.3	114.1	9.5	15.2
MRA-CNN[22]	74.2	112.5	11.9	21.3
ACSPNet[23]	85.2	137.1	15.4	23.1
ACM-CNN[16]	72.2	103.5	17.5	22.7
SFANet[24]	59.8	99.3	26.0	30.5
FPNet[33]	108.6	126.3	26.0	30.5
本文方法	**57.1**	**91.9**	**6.87**	**9.8**

表 2　UCF_CC_50实验结果

方法	MAE	MSE
MCNN[13]	377.6	509.1
MSFNet[15]	257.2	380.8
Switching-CNN[9]	318.1	439.2
CSRNet[8]	266.1	397.5
ic-CNN[25]	260.9	365.5
SCAR[21]	259.0	374.0
MRA-CNN[22]	240.8	352.6
ACSPNet[16]	275.2	383.7
ACM-CNN[16]	291.6	337.0
SDA-MCNN[26]	306.6	313.2
SFANet[24]	219.6	316.2
FPNet[33]	463.0	501.6
本文方法	**175.2**	**233.6**

表 3　Mall实验结果

方法	MAE	MSE
EDMNet[14]	1.80	5.36
R-FCN[27]	6.02	5.46
Faster R-FCN[28]	5.91	6.60
BidirectionalConvLSTM[29]	2.10	7.6
DigCrowd[30]	3.21	16.4
ACM-CNN[16]	2.3	3.1
本文方法	**1.57**	**2.03**

在表2中可以看到与之前最好的方法相比，本文方法的平均绝对误差(MAE)结果是175.2比最好的方法要低44.4，均方误差(MSE)也有明显的降低。

4.3.3 Mall数据集

表3为本文方法对Mall数据集实验的实验结果(batch size=8)，可以看到与之前最好的方法相比，MAE结果是1.57，比之前最好的方法要好0.23，MSE结果是2.03，比之前最好的方法要好1.07。

4.3.4 UCSD数据集

网络在多次降采样后输出特征过于模糊，影响了计数精度。因此，本文通过双线性插值将UCSD的分辨率扩大为960×640，其ground truth也进行相同比例的插值。提升分辨率能够提高人群密度估计的精度，适于注意力机制的作用发挥。由表4看出，本文方法取得了更优的实验结果(batch size=8)。

4.4 消融实验

在实验最后本文进行了消融实验，以确认本文包含的各个网络结构带来的影响。本文将1×1，3×3，5×5，7×7的基础特征提取网络简称为D，将新型多尺度密度图估计模块称为ND，将1×1，3×3，5×5，7×7的多尺度注意力模块称为M，将新型多尺度注意力模块称为NM，将注意力融合模块称为C。消融实验对最具代表性的同时也是具备相当难度ShanghaiTech-PartA数据集进行。实验结果(表5)证明了本文网络的不同部分对结果的改善程度。从表5可以看出，以Backbone + D + M为原型，在增加了注意力融合模块C后，网络Backbone+ D+M+C的MAE和MSE分别减少了0.8和3.9。将D替换为ND、将M替换为NM后，网络Backbone + ND + NM+C的MAE和MSE继续分别减少了0.7和0.8。这充分证明了新型多尺度模块和注意力模块对网络性能的改进作用。

表 4 UCSD实验结果

方法	MAE	MSE
MCNN[13]	1.07	1.35
Switching-CNN[9]	1.62	2.10
BidirectionalConvLSTM[29]	1.13	1.43
ACSCP[31]	1.04	1.35
CSRNet[8]	1.16	1.47
SaNet[32]	1.02	1.29
ACSPNet[23]	1.02	1.28
ACM-CNN[16]	1.01	1.29
SFANet[24]	0.82	1.07
FPNet[33]	1.67	3.91
本文方法	**0.97**	**1.27**

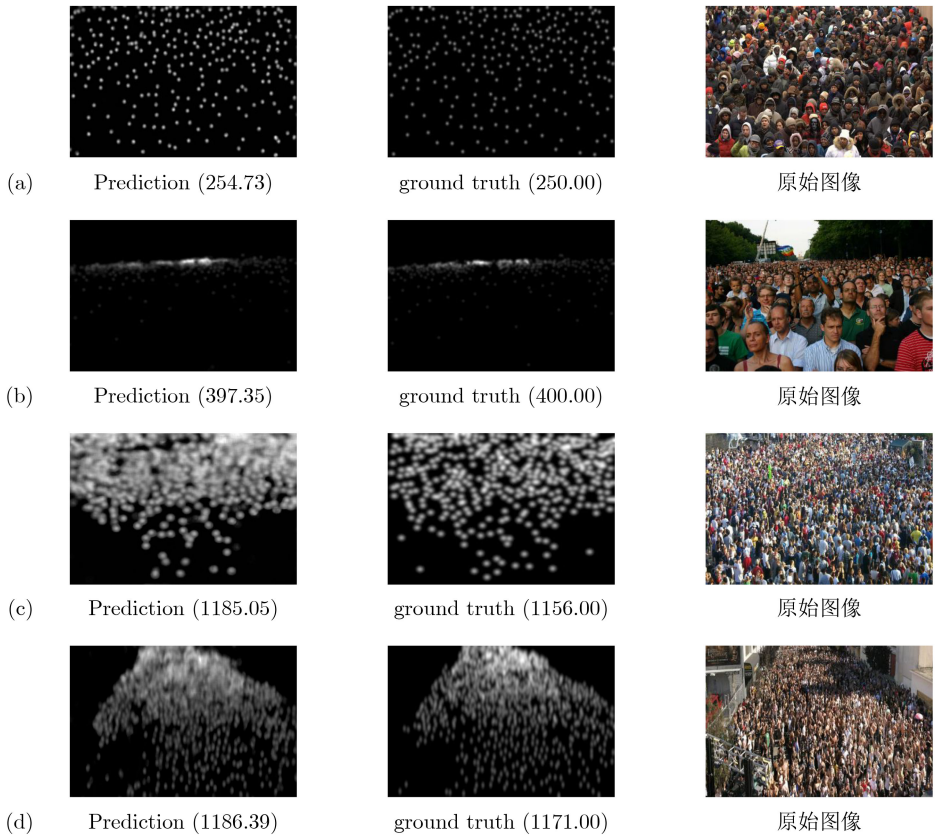

图 7 密度估计图、ground truth以及原始图像

表 5 消融实验结果

方法	MAE	MSE
Backbone + D +M	58.6	96.6
Backbone+D+M+C	57.8	92.7
Backbone + ND +NM+C	**57.1**	**91.9**

5 结论

本文提出了一种新型卷积神经网络结构，用于密集人群计数。该网络利用新型多尺度注意力模块对语义丰富的深层特征进行处理，以获得更加丰富的语义信息；利用注意力机制对深层多尺度特征进行处理以抑制非头部区域，使头部区域的信号更加明显。同时，本文引入的新型多尺度模块能够使深层特征的空间集中度变高，扩大感受野，得到更高质量的人群密度图。在深层的特征中，通过注意力融合模块提高特征辨别度，以此来提高网络的性能。实验结果证明了本文方法的有效性。

参 考 文 献

[1] ARTETA C, LEMPITSKY V, and ZISSERMAN A. Counting in the wild[C]. Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 2016: 483–498.

[2] ARTETA C, LEMPITSKY V, NOBLE J A, *et al*. Interactive object counting[C]. Proceedings of the 13th European Conference on Computer Vision, Zurich, Switzerland, 2014: 504–518.

[3] SHANG Chong, AI Haizhou, and BAI Bo. End-to-end crowd counting via joint learning local and global count[C]. Proceedings of 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, USA, 2016: 1215–1219.

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

[5] OÑORO-RUBIO D and LÓPEZ-SASTRE R J. Towards perspective-free object counting with deep learning[C]. Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 2016: 615–629.

[6] 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]. Proceedings of the 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, Lecce, Italy, 2017: 123–126.

[7] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778.

[8] LI Yuhong, ZHANG Xiaofan, and CHEN Deming. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes[C]. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 1091–1100.

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

[10] WANG Qi, GAO Junyu, LIN Wei, *et al*. Learning from synthetic data for crowd counting in the wild[C]. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 8190–8199.

[11] LI Zhengqi, DEKEL T, COLE F, *et al*. Learning the depths of moving people by watching frozen people[C]. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 4516–4525.

[12] WANG Shunzhou, LU Yao, ZHOU Tianfei, *et al*. SCLNet: Spatial context learning network for congested crowd counting[J]. *Neurocomputing*, 2020, 404: 227–239. doi: 10.1016/j.neucom.2020.04.139.

[13] CHEN Xinya, BIN Yanrui, GAO Changxin, *et al*. Relevant region prediction for crowd counting[J]. *Neurocomputing*, 2020, 407: 399–408. doi: 10.1016/j.neucom.2020.04.117.

[14] 孟月波, 纪拓, 刘光辉, 等. 编码-解码多尺度卷积神经网络人群计数方法[J]. 西安交通大学学报, 2020, 54(5): 149–157.
MENG Yuebo, JI Tuo, LIU Guanghui, *et al*. Encoding-decoding multi-scale convolutional neural network for crowd counting[J]. *Journal of Xi'an Jiaotong University*, 2020, 54(5): 149–157.

[15] 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): 307–315.
ZUO Jing and BA Yulin. Population-depth counting algorithm based on multiscale fusion[J]. *Laser & Optoelectronics Progress*, 2020, 57(24): 307–315.

[16] ZOU Zhikang, CHENG Yu, QU Xiaoye, *et al*. Attend to count: Crowd counting with adaptive capacity multi-scale CNNs[J]. *Neurocomputing*, 2019, 367: 75–83. doi: 10.1016/j.neucom.2019.08.009.

[17] SZEGEDY C, LIU Wei, JIA Yangqing, *et al*. Going deeper with convolutions[C]. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 1–9.

[18] IDREES H, SALEEMI I, SEIBERT C, *et al*. Multi-source multi-scale counting in extremely dense crowd images[C]. Proceedings of 2013 IEEE Conference on Computer Vision

and Pattern Recognition, Portland, USA, 2013: 2547–2554.

[19] CHEN Ke, LOY C C, GONG Shaogang, *et al*. Feature mining for localised crowd counting[C]. Proceedings of the British Machine Vision Conference, Surrey, UK, 2012: 3–5.

[20] CHAN A B, LIANG Z S J, and VASCONCELOS N. Privacy preserving crowd monitoring: Counting people without people models or tracking[C]. Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, USA, 2008: 1–7.

[21] GAO Junyu, WANG Qi, and YUAN Yuan. SCAR: Spatial-/channel-wise attention regression networks for crowd counting[J]. *Neurocomputing*, 2019, 363: 1–8. doi: 10.1016/j.neucom.2019.08.018.

[22] ZHANG Youmei, ZHOU Chunluan, CHANG Faliang, *et al*. Multi-resolution attention convolutional neural network for crowd counting[J]. *Neurocomputing*, 2018, 329: 144–152.

[23] MA Junjie, DAI Yaping, and TAN Y P. Atrous convolutions spatial pyramid network for crowd counting and density estimation[J]. *Neurocomputing*, 2019, 350: 91–101. doi: 10.1016/j.neucom.2019.03.065.

[24] ZHU Liang, ZHAO Zhijian, LU Chao, *et al*. Dual path multi-scale fusion networks with attention for crowd counting[J]. arXiv: 1902.01115, 2019.

[25] RANJAN V, LE H, and HOAI M. Iterative crowd counting[C]. Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, 2018: 278–293.

[26] YANG Biao, ZHAN Weiqin, WANG Nan, *et al*. Counting crowds using a scale-distribution-aware network and adaptive human-shaped kernel[J]. *Neurocomputing*, 2020, 390: 207–216. doi: 10.1016/j.neucom.2019.02.071.

[27] DAI Jifeng, LI Yi, HE Kaiming, *et al*. R-FCN: Object detection via region-based fully convolutional networks[C]. Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 2016: 379–387.

[28] REN Shaoqiang, HE Kaiming, GIRSHICK R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1137–1149. doi: 10.1109/TPAMI.2016.2577031.

[29] XIONG Feng, SHI Xingjian, and YEUNG D Y. Spatiotemporal modeling for crowd counting in videos[C]. Proceedings of 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 5161–5169.

[30] XU Mingliang, GE Zhaoyang, JIANG Xiaoheng, *et al*. Depth Information Guided Crowd Counting for complex crowd scenes[J]. *Pattern Recognition Letters*, 2019, 125: 563–569. doi: 10.1016/j.patrec.2019.02.026.

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

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

[33] 邓远志, 胡钢. 基于特征金字塔的人群密度估计方法[J]. 测控技术, 2020, 39(6): 108–114.
DENG Yuanzhi and HU Gang. Crowd density evaluation method based on feature pyramid[J]. *Measurement & Control Technology*, 2020, 39(6): 108–114.

万洪林：男，1979年生，副教授，博士，主要研究方向为计算机视觉、人工智能.

王晓敏：女，1998年生，硕士生，主要研究方向为图像处理、人群计数.

彭振伟：男，1995年生，硕士，主要研究方向为图像处理、人群计数.

白智全：男，1978年生，教授，博士生导师，主要研究方向为协作通信技术、无线光通信技术.

孙建德：男，1978年生，教授、博士生导师，主要研究方向为多媒体信息处理、分析、理解及其应用.

责任编辑：陈 倩