基于多重注意力引导的人群计数算法*

杨倩倩,何　晴,彭思凡,殷保群
(中国科学技术大学 信息科学技术学院,安徽 合肥 230026)

摘　要:针对实际场景中存在的人群非均匀分布问题,提出了一种基于多重注意力引导的人群计数算法。首先,基于轻量级金字塔切分注意力机制构建了自顶向下的特征融合路径,旨在促进高层语义信息和低层空间细节的融合,生成高级语义和空间细节兼备的高质量特征图;然后,提取并融合多尺度上下文信息,以此生成关注于不同密度分布模式的注意力权重图;最后,通过注意力权重图指导密度回归网络识别不同分布状态下的行人目标,增强模型对密度变化的适应性,生成高质量人群密度图。在ShanghaiTech、UCF_QNRF 和 JHU-CROWD++三个数据集上进行了大量的实验来说明所提算法的先进性。

关键词:人群计数;密度估计;注意力机制;特征金字塔
中图分类号:TP309　　文献标识码:A　　DOI:10.20044/j.csdg.2097-1788.2022.01.017
引用格式:杨倩倩,何晴,彭思凡,等. 基于多重注意力引导的人群计数算法[J]. 网络安全与数据治理,2022,41(1):108-116.

Multi-attention convolutional network for crowd counting

Yang Qianqian, He Qing, Peng Sifan, Yin Baoqun
(School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract: Aiming at the problem of non-uniform crowd distribution in practical scenes, this paper proposes a crowd counting algorithm based on multi-attention mechanism. A top-down feature fusion path is constructed based on the lightweight pyramid split attention mechanism, which aims to promote the fusion of high-level semantic features and low-level spatial details, resulting in high-quality feature maps with both semantics and spatial details. Then multi-scale context information is extracted and fused to generate attention weight maps that focus on different density distribution patterns. At last, the density regression network is guided by the attention weight maps to identify pedestrian targets in different distributions, enhancing the model's adaptability to density variation, so as to generate high-quality crowd density maps. Abundant experiments on three datasets including ShanghaiTech, UCF_QNRF and JHU-CROWD++ were conducted to prove the effectiveness of the proposed network.

Key words: crowd counting; density map estimation; attention mechanism; feature pyramid network

0 引言

由于人群所在的位置和行动轨迹具有主观性强、自由度高的特点,监控视频采集的图像包含大量杂乱分布的人群,不同局部区域的人群密度差异巨大,增大了人群计数算法的估计难度。如图1所示,在同一人群场景中,多个局部区域表现为人口极度聚集,而部分区域人口稀疏甚至是孤立的个体,难以预测的行人位置将导致密度图中不同位置的密度值之间存在巨大差异,对算法感知不同密度分布模式的能力提出了更高的要求。为解决上述问题,本文提出基于多重注意力引导的人群计数算法,将特征金字塔机制和注意力机制相结合,促进语义信息和空间细节的融合,并通过注意力图引导模型生成对应不同分布状态的密度图。

1 概述

为解决计数问题中的非均匀分布问题,研究者主要设计了基于分类或检测等辅助任务和注意力机制的计数方法,让模型提取到各种分布模式下的

*基金项目:中国人工智能学会-华为 MindSpore 学术奖励基金(6213000091)

图 1　人群非均匀分布图像示例

全局和局部语义信息来识别分布在不同密集程度下的行人。CP-CNN[1]在多列密度估计框架的基础上搭建了两个额外的人群密度分类网络，分别学习将输入图像或局部图像块按照密度差异分类为 5 个类别，以此来学习全局和局部区域的上下文信息，并将提取的上下文语义信息融合到密度图估计任务中，从而生成高质量的人群密度图。然而，该算法对全局或局部区域人数分类只能提取到粗糙的语义特征，复杂场景下需要细粒度的密度语义信息。Liu 等人[2]认为在不知道每个行人具体位置的情况下，计数模型在低密度区域容易产生过估计现象。为了解决此问题，作者提出的 Decidenet 将检测和回归方法结合来处理不同密度区域的人群，能够自适应地根据实际密度情况决定合适的计数方法，在低密度区域采用检测方法估计人数，在拥挤区域则采用回归方法捕捉人群信息。然而，采用的 Faster-RCNN[3]等检测网络往往十分复杂，难以将其和密度图回归网络以端到端的方式训练，且训练过程繁琐。Jiang 等人[4]指出在不同的密度模式的图像区域中，数据驱动的计数算法容易出现过估计或欠估计。作者提出了一种基于注意力机制的人群计数算法，该网络包括密度注意力网络和注意力尺度网络，后者给前者提供与不同密度区域相关的注意力信息，并采用尺度因子帮助缩小不同区域的估计误差。然而，该模型由 2 个 VGG-16[5]组成，网络参数量过大并且结构较为复杂，同时需要对子网络分别训练，这些弊端导致该模型训练困难且不利于计数性能的提升。

针对以上问题，本文提出一种基于多重注意力引导的人群密度估计算法(Multi-Attention Convolutional Network for Crowd Counting，MACN)来完成密度图回归任务，采用多重注意力机制指导网络提取多种分布模式下的人群特征，改善不同密集程度下的密度图质量，从而提升人群计数算法的整体性能，主要包括特征提取模块(Feature Extraction Module，FEM)、特征融合模块(Feature Fusion Module，FFM)、上下文注意力模块 (Context Attention Module，CAM) 和密度图生成模块(Density Map Generation Module，DGM)。

2 密度估计网络结构设计

针对复杂场景中存在的人群非均匀分布问题，设计了一个端到端的计数网络来实现密度图估计，采用注意力机制获取像素级别的语义信息来指导密度图回归任务，感知不同分布模式下的行人，具体结构如图 2 所示。

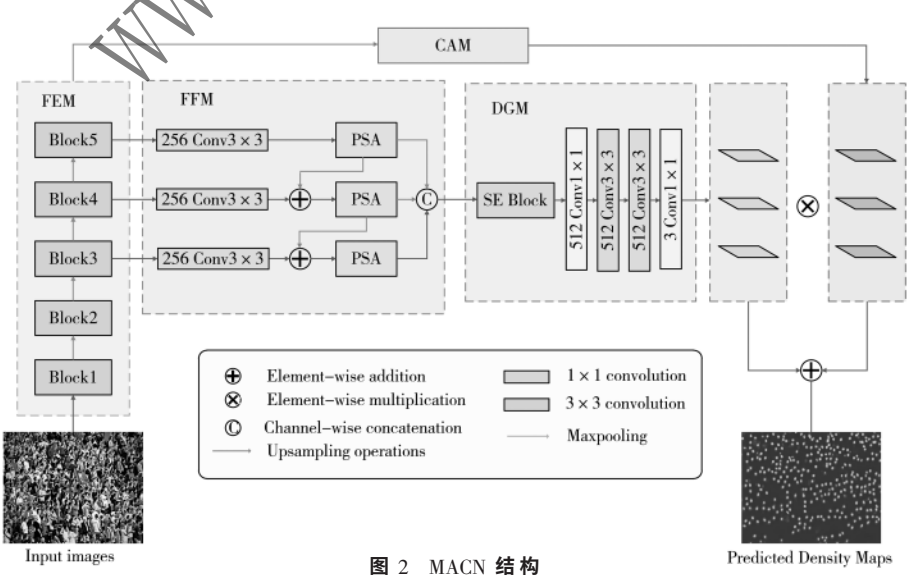

图 2　MACN 结构

2.1 特征提取模块

VGG-16 具有强大的特征提取能力和优秀的迁移学习能力,采用小卷积核代替大卷积核,能够在保证特征表达能力的同时不消耗过多计算资源,所以采用 VGG-16 的前 13 个卷积层作为特征提取模块,根据池化层位置进一步分 5 个子模块,其具体配置如表 1 所示。

表 1 FEM 网络结构

层	类别	数目	卷积核尺寸
1	卷积层	64	3×3×3
2	卷积层	64	64×3×3
	最大池化层		
3	卷积层	128	64×3×3
4	卷积层	128	128×3×3
	最大池化层		
5	卷积层	256	128×3×3
6	卷积层	256	256×3×3
7	卷积层	256	256×3×3
	最大池化层		
8	卷积层	512	256×3×3
9	卷积层	512	512×3×3
10	卷继承	512	512×3×3
	最大池化层		
11	卷积层	512	512×3×3
12	卷积层	512	512×3×3
13	卷积层	512	512×3×3

2.2 特征融合模块

Lin 等人[6]在目标检测系统 Faster R-CNN 中引入了特征金字塔结构(Feature Pyramid Network, FPN),它是一种包含自下而上路径、横向连接和自上而下路径的特征融合机制,能够充分融合来自网络中间不同层次的特征。Zhang 等人[7]提出一种轻量化金字塔切分注意力模块(Pyramid Split Attention, PSA),将其嵌入到 ResNet[8]中来提取更细粒度的空间信息并建立通道之间的长距离依赖,有效提升了分类网络的性能。受到上述工作启发,本文将金字塔切分注意力模块引入特征金字塔机制,搭建了自上而下的特征融合路径。

PSA 的网络结构如图 3 所示,共包含四个主要步骤。假设输入特征 X 的尺寸为 $C\times H\times W$,首先在通道维度上将输入向量均分为 S 组,采用分组卷积的方法让 S 个不同尺寸的卷积核并行处理每组向量,并采用级联的方式对结果进行融合得到 F。然后采用通道注意力模块为每组特征学习注意力权重向量,权重的计算方式将在后面详细介绍。第三步是采用 Softmax 函数注意力向量进行调整得到 Z',从而自适应地选择不同的空间尺寸。最后,将调整后的注意力向量 Z' 和 F 做逐像素乘积得到输出特征。

将 FEM 模块的 5 个子模块的输出特征按照从下到上的顺序表示为 C_1, C_2, C_3, C_4, C_5,考虑到浅层特征具有较多噪声,故只采用最上面三层特征作为 FFM 的输入信号。本节采用上述 PSA 模块代替 FPN 中的标准卷积层来消除上采样带来的混叠效应,并提取更细致的空间信息。首先,采用 1×1 卷积层将三层特征的通道维度都调整为 256,得到 $C_3^{'}$, $C_4^{'}$, $C_5^{'}$。然后,将第五层的特征 $C_5^{'}$ 输入到 PSA 模块产生

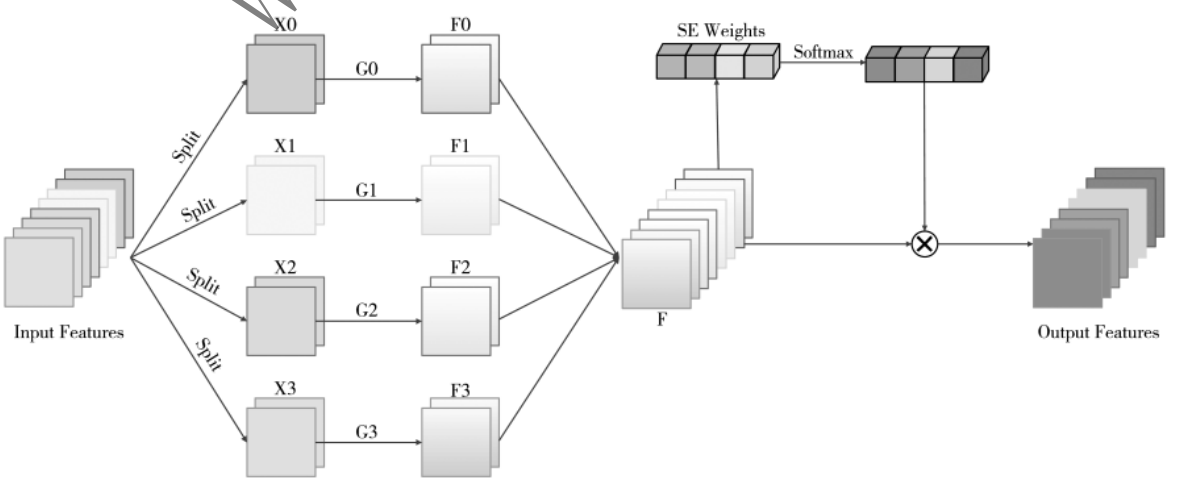

图 3 PSA 结构

该层的输出特征 P_5。采用双线性插值方法将其上采样得到与 C_4' 相同尺寸的特征 P_5'，对 C_5' 和 P_5' 执行逐像素相加并输入到 PSA 模块生成第四层的输出特征 P_4。以此类推，获得第三层的输出特征 P_3。最后，采用级联的方式，将三层的输出特征融合生成同时具有高级语义和空间细节的融合特征。FFM 中三层输出特征的计算方式如式(1)所示：

$$P_i=\begin{cases} f(\mathrm{conv}_1(C_i)) & i=5 \\ f(\mathrm{conv}_1(C_i))+\mathrm{up}(P_{i+1}) & i=3,4 \end{cases} \tag{1}$$

其中 $f(\cdot)$ 表示 PSA 模块的映射关系，$\mathrm{conv}(\cdot)$ 表示 1×1 卷积运算，$\mathrm{up}(\cdot)$ 表示双线性插值函数。

2.3 上下文注意力模块

注意力机制是一种让模型关注重点目标区域的信息筛选机制，能够帮助网络提取到与人群目标更相关的有利信息。上下文注意力模块首先提取全局和局部上下文信息并进行融合，基于该融合语义信息生成能感知不同密度分布模式的注意力图，给不同密度级别的密度图分配对应的权重，让模型关注到不同分布状态的行人，从而克服图像中人群杂乱分布导致的密度差异，其具体结构如图 4 所示。

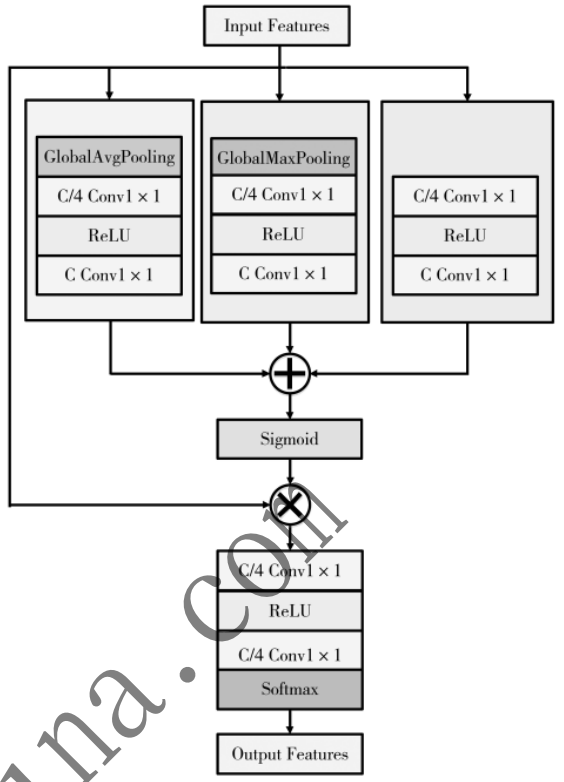

图 4 CAM 结构

Dai 等人[9]指出只利用全局上下文信息将使得模型更关注于分布于全局的目标，并弱化小尺寸目标的图像信号，局部上下文信息的使用将有助于模型关注更多的空间细节信息，从而生成更高质量的特征图。所以 CAM 模块分别设计了全局和局部上下文提取模块，并对提取的多尺寸上下文语义进行融合。考虑到平均池化和最大池化具有不同的特性，平均池化对特征图上的每个像素点都产生反馈信号，而最大池化更加关注于区域内相应最大的位置，所以采用全局平均池化(Global Average Pooling，GAP)和最大池化(Global Max Pooling，GMP)两种池化机制来提取全局上下文信息。输入特征首先分别经过 GAP 和 GMP 来聚合全局空间信息，然后采用两个 1×1 卷积捕获各通道间全局信息的依赖关系。同时设计了局部上下文信息提取模块，不采用全局池化操作，通过小尺寸的卷积操作建立局部连接来捕捉局部上下文特征。最后对提取的上下文信息进行融合得到多尺寸上下文语义特征，并经过两个 1×1 卷积进一步增大特征的感受野并调整通道数量，生成最终的注意力权重。

2.4 密度图生成模块

SE Block[10]能够对输入特征通道之间的相关性进行建模，通过网络学习的方式自动获得每个特征通道的重要程度。本节采用 SE Block 来对特征融合模块得到的多级语义融合特征进行通道维度的权重调整，其具体结构如图 5 所示。

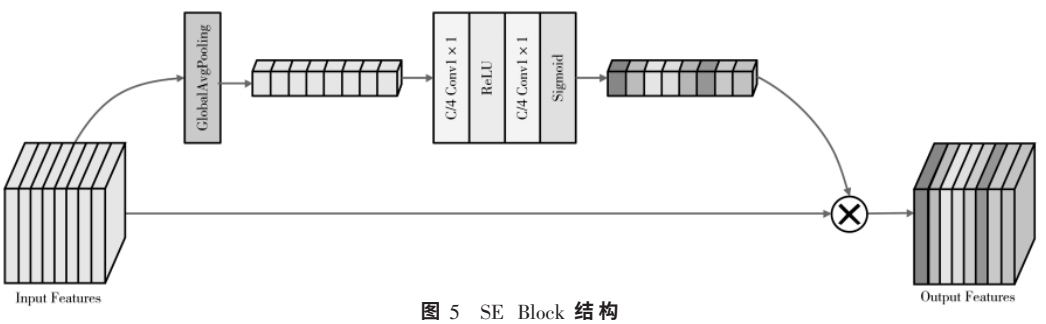

图 5 SE Block 结构

SE Block 首先通过全局平均池化聚合空间信息,为每个通道分别生成一个全局描述符。假设输入特征 f 的尺寸为 $C\times W\times H$,于是第 c 个通道的描述符的计算过程如式(2)所示:

$$z_c=\frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H}f_c(i,j) \tag{2}$$

其中 f_c 表示第 c 个通道的输入特征。然后将全部描述符输入到两个全连接层进行非线性变换,并采用 Sigmoid 激活函数实现特征的重新标定,于是得到所有通道的注意力权重:

$$s=\mathrm{Sigmoid}(W_2\cdot(\mathrm{ReLU}(W_1\cdot z))) \tag{3}$$

其中 W_1 和 W_2 分别表示两全连接层的权重,z 表示 C 个通道描述符组成的整体描述符。于是调整后的特征如式(4)所示:

$$f'=s\cdot f \tag{4}$$

通道注意力模块对多级语义融合特征进行了通道权重的调整,加强了对密度估计更重要的通道,生成了更具辨别力的特征。为了提取更深层的语义信息,首先采用 1×1 卷积层将通道维数降为 512 通道,再采用 2 个 3×3 卷积层进一步扩大特征的感受野,最后采用 1×1 卷积层生成 3 个对应不同密度级别的密度图。最后,将 CAM 生成的感知不同密度模式的注意力图和 DGM 生成的不同区域的密度图进行逐像素相乘,相加融合生成最终的高质量密度图,具体计算过程如式(5)所示:

$$\mathrm{Density\_Map}=\sum_{i=1}^{3}A_i\odot D_i \tag{5}$$

其中,A_i 和 D_i 分别表示 CAM 和 DGM 生成的第 i 个通道的特征,$\odot$ 表示逐像素相乘。

2.5 损失函数

本文采用欧式距离衡量估计密度图和真值密度图之间的差距,定义如式(6)所示:

$$L^D(\theta)=\frac{1}{N}\sum_{i=1}^{N}||F_i-F(X_i;\theta)||^2 \tag{6}$$

其中,θ 是 MACN 中待优化网络参数的集合,N 是训练样本数量,X_i 表示第 i 张输入图像,F_i 和 $F(X_i;\theta)$ 分别表示真值密度图和估计密度图。此外,采用交叉熵损失监督 CAM 模块生成的注意力图,计算方式如式(7)所示:

$$L^A(\theta)=-\frac{1}{N}\sum_{j=1}^{3}\sum_{i=1}^{N}y_i^j\log(A^j(X_i;\theta)) \tag{7}$$

其中,y^j 和 $A^j(X_i;\theta)$ 分别表示第 j 个密度等级注意力权重的真值标签和估计值。y^j 的获得方式将在 3.1 节详细介绍。因此,总体损失函数如式(8)所示:

$$L(\theta)=L^D(\theta)+\alpha L^A(\theta) \tag{8}$$

其中 α 表示超参数,在实验中被设置为 1。

3 算法实现

3.1 地面真值生成

假设 x_i 位置存在一个人头标记,可将其表示为一个单位冲击函数 $\delta(x-x_i)$,可将包含 M 个人头标记的输入图像表示为:

$$H(x)=\sum_{i=1}^{M}\delta(x-x_i) \tag{9}$$

然后采用归一化的高斯核对每个人头标记做平滑处理,即将 $H(x)$ 与高斯核做卷积来生成真值密度图,计算方式如式(10)所示:

$$F(x)=\sum_{i=1}^{M}\delta(x-x_i)*G_{\mu,\sigma}(x) \tag{10}$$

其中 μ 和 σ 分别表示高斯核的均值和方差,$G_{\mu,\sigma}(x)$ 表示归一化的二维高斯核。本文采用 15×15 的固定尺寸高斯核生成真值密度图。

此外,为了监督 CAM 模块生成的注意力图,本文依据上述生成的真值密度图得到注意力真值标签,计算过程如式(11)所示:

$$\begin{cases}y^1(x)=\begin{cases}1 & 0\leqslant F(x)\leqslant \mathrm{avg}\\ 0 & \text{其他}\end{cases}\\ y^2(x)=\begin{cases}1 & \mathrm{avg}\leqslant F(x)\leqslant \mathrm{avg}+\beta\\ 0 & \text{其他}\end{cases}\\ y^3(x)=\begin{cases}1 & \mathrm{avg}+\beta\leqslant F(x)\\ 0 & \text{其他}\end{cases}\end{cases} \tag{11}$$

其中 $y^i(x)$ 表示第 i 个密度级别的注意力标签,avg 表示整张真值密度图所有像素的平均值,β 是超参数,被设置为 0.2。

3.2 数据增强处理

本节采用多种方法对数据进行增强,防止训练过程出现过拟合现象。首先,从原始图片中裁剪出 9 张 1/4 原图尺寸的图像块,并将每个图像块随机裁剪到 256×256 的固定尺寸。然后,以 0.5 的概率对所有图像块做水平翻转,以 0.3 的概率在[0.5,1.5]范围内对图像块做伽马变换,以 0.1 的概率将彩色图片转换成灰度图像,以 0.25 的概率将彩色图像的 RGB 通道切换。最后,以 0.25 的概率像图像加入均值为 0、标准差为 5 的高斯噪声。

3.3 训练过程

本文以端到端的方式训练 MACN。特征提取模块采用在 ImagNet 上预训练的 VGG-16 的参数进行初始化，剩余其他所有卷积层的权重都采用均值为 0，标准差为 0.01 的高斯分布初始化。采用 Adam 优化器来优化模型参数，初始学习率和衰减率分别被设置为 1×10^{-5} 和 0.995，批尺寸设置为 8。跟随 Gao 等人[11]的做法，训练时将密度图乘以一个较大的放大倍数(100)来实现更快的收敛速度和更准确的结果，更好地平衡损失权重。本文所有工作皆是基于 PyTorch 深度学习框架，采用 11 GB 显存的 Nvidia-2080Ti GPU 显卡实现训练和测试的加速。

4 实验分析

参考其他优秀人群计数方法[12-13]，本文采用平均绝对误差(Mean Absolute Error，MAE)和均方误差(Mean Squared Error，MSE)来评估算法性能，计算公式如式(12)、式(13)所示：

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}\left|z_i-\hat{z}_i\right| \tag{12}$$

$$\mathrm{MSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(z_i-\hat{z}_i)^2} \tag{13}$$

其中，N 代表测试图片数目，z_i 与 $\hat{z}_i$ 分别代表真值与算法的预测值。

4.1 ShanghaiTech 数据集实验

ShanghaiTech 数据集[13]共包括1 198 张人群图片，根据密度差异被分为 Part_A 和 Part_B 两个部分。Part_A 由互联网上获取的 482 张人群图片组成，人群分布较为密集，训练集和测试集分别包含 300 张和 182 张图片；Part_B 中的 716 张图片均拍摄于上海街头，人群分布较为稀疏，训练集和测试集分别包含 400 和 316 张图片。

表 2 给出了在本数据集上 MACN 算法与其他代表性算法的实验结果。由分析可知，MACN 在两个子数据集上同时实现了最好的 MAE 和 MSE 指标，相比于第二名分别将 MAE 指标优化了 4.6%和 9.2%，在人群密集的 Part_A 和人群较为稀疏的 Part_B 上都能实现较明显的性能提升，说明 MACN 算法对人群密度变化具有较好适应性。图 6 展示了测试集的部分估计密度图和真实密度图的可视化样例，第一行、第二行表示 Part_A 上的估计结果，第三、四行表示 Part_B 上的估计结果；从左到右每一列分别表示人群图片、真值密度图和估计密度图。

表 2 ShanghaiTech 数据集结果

算法	Part_A		Part_B	
	MAE	MSE	MAE	MSE
CP-CNN[1]	73.6	106.4	20.1	30.1
MCNN[13]	110.2	173.2	26.4	41.3
PaDNet[14]	59.2	98.1	8.1	12.2
PACNN+[15]	62.4	102.0	7.6	11.8
BL[16]	62.8	101.8	7.7	12.7
ADCrowdNet[17]	70.9	115.2	7.7	12.9
HA-CCN[18]	62.9	94.9	8.1	13.4
DUBNet[19]	64.6	106.8	7.7	12.5
MACN(本文)	56.5	94.5	6.9	10.9

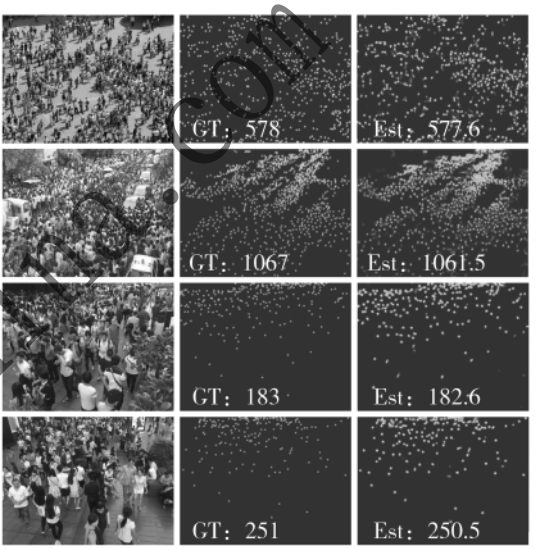

图 6 ShanghaiTech 上 MACN 算法生成密度图可视化样例

4.2 UCF_QNRF 数据集实验

Ideers 等人[20]提出了 UCF_QNRF 数据集，图片主要来源于朝圣素材和网页搜索，共包含 1 535 张图片，比 ShanghaiTech 数据集具有更大的规模，且图像中人群总数浮动较大，人群场景更为复杂。

表 3 列出了 MACN 算法与其他优秀算法在 UCF-QNRF 上的实验结果。和 MAE 排名第二的 PaDNet 相比，MACN 将 MAE 指标在数值上减小了 2.7%，这表明该算法在极度密集的场景中仍能对人群数量进行准确估计。图 7 展示了测试集的部分估计密度图和真实密度图的可视化样例，可以看到 MACN 算法能够有效应对图像中存在的人群非均匀分布问题，生成接近真实人群分布的估计结果。

4.3 JHU-CROWD++数据集实验

Sindagi 等人[27]提出的 JHU-CROWD++数据集是

表 3 UCF_QNRF 数据集结果

算法	MAE	MSE
MCNN[13]	277	426
PaDNet[14]	96.5	170.2
DUBNet[19]	105.6	180.5
CSRNet[21]	120.3	208.5
TEDNet[22]	113	188
PCCNet[23]	132	191
RAZ-Net[24]	116	195
RANet[25]	111	190
HYGNN[26]	100.8	185.3
MACN(本文)	93.9	170.8

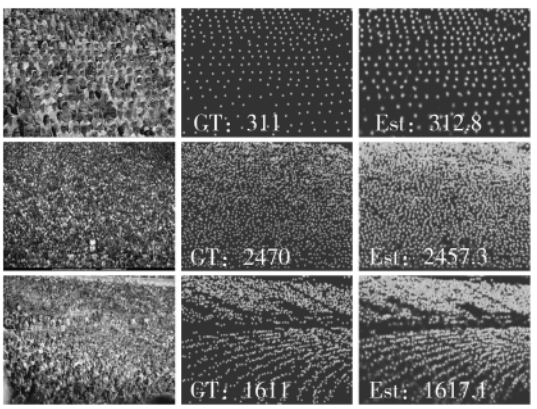

图 7 UCF_QNRF 上 MACN 算法生成密度图可视化样例

一个新型的大规模无约束数据集,由不同场景与环境下的 4 372 张人群图像组成,由训练集(2 272 张)、验证集(500 张)、测试集(1 600 张)三个部分组成,人头标注数量高达 151 万。为了进一步研究人群计数算法在各种密度情况下的性能,测试集与验证集的人群图像被细分为高密度、中密度和低密度三个类别。同时考虑到恶劣天气的影响,此数据集还包含单独由复杂天气人群图像组成的子集。下面将在总体、低密度、中密度、高密度以及天气环境这 5 种情况下分析人群计数算法的准确率和鲁棒性,如表 4 所示。

对比 MACN 算法和其他先进算法在 JHU-CROWD++测试集上的性能。在整体性能上,MACN 算法的 MAE 指标比第二名优化了 6.8%;在低密度、中密度、高密度、恶劣天气等子类上 MACN 分别将 MAE 指标比第二名在数值上优化了 0.2、1.9、16.6 和 7.9。JHU-CROWD++测试集的部分估计密度图和真实密度图的可视化样例如图 8 所示。分析表明,MACN 算法在总体与所有子类上都实现了最准确的计数结果并能够保持较高的鲁棒性,验证了 MACN 算法的泛化性和准确性。

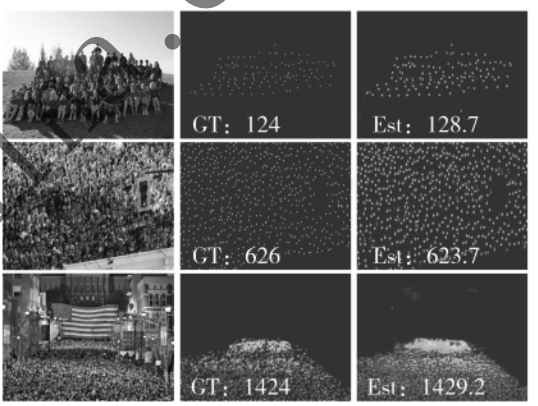

图 8 JHU-CROWD++上 MACN 算法生成密度图可视化样例

4.4 消融实验

为验证各子模块的有效性,本小节在 Shanghai-

表 4 JHU-CROWD++数据集结果

算法	低密度		中密度		高密度		天气环境		总体	
	MAE	MSE	MAE	MSE	MAE	MSE	MA	MSE	MAE	MSE
MCNN[13]	97.1	192.3	121.4	191.3	618.6	1 166.7	330.6	852.1	188.9	483.4
BL[16]	10.1	32.7	34.2	54.5	352.0	768.7	140.1	675.7	75.0	299.9
CSRNet[21]	27.1	64.9	43.9	71.2	356.2	784.4	141.4	640.1	85.9	309.2
CMTL[28]	58.5	136.4	81.7	144.7	635.3	1225.3	261.6	816.0	157.8	490.4
SANet[29]	17.3	37.9	46.8	69.1	397.9	817.7	154.2	685.7	91.1	320.4
SFCN[30]	16.5	55.7	38.1	59.8	341.8	758.8	122.8	606.3	77.5	297.6
DSSINEet[31]	53.6	112.8	70.3	108.6	525.5	1 047.4	229.1	760.3	133.5	416.5
MBTTBF[32]	19.2	58.5	41.6	66.0	352.2	760.4	138.7	631.6	81.8	299.1
LSC-CNN[33]	10.6	31.8	34.9	55.6	601.9	1 172.2	178.0	744.3	112.7	454.4
MACN(本文)	9.9	19.9	32.3	10.9	325.2	754.5	114.9	638.1	69.9	295.3

Tech Part_A 上对不同结构进行实验。表 5 展示了五种不同结构的实验结果，其中 W/O FFM+CAM+SE 表示将 FFM 模块、CAM 模块和 DGM 中的 SE Block 移除；W/O FFM 表示移除 FFM 模块；W/O CAM 表示移除 CAM 模块；W/O SE 表示移除 SE Block；MACN 表示完整结构。由实验结果可知，同时采用 FFM、CAM、SE Block 能够获得最优的 MAE 和 MSE 指标，大幅提升模型性能。

表 5 参数选择实验

算法	MAE	MSE
W/O FFM+CAM+SE	67.6	105.1
W/O FFM	60.1	103.6
W/O CAM	61.3	102.4
W/O SE	58.5	100.3
MACN(本文)	56.5	94.5

5 结论

本文提出了一种基于多重注意力引导的密度估计网络 MACN，来解决人群计数任务中的人群杂分布问题。首先，该算法基于轻量化的切分注意力模块，构建了一条自上而下的特征融合路径以促进不同语义级别特征之间的流动，旨在增强特征的表达能力。然后，建立上下文注意力模块来提取并融合多尺度上下文信息，以生成对应不同密度级别的注意力图，从而引导密度估计网络识别不同分布模式下的行人目标。最后，凭借大量实验结果充分说明 MACN 有助于提升各种场景下的计数性能。

参考文献

[1] SINDAGI V A, PATEL V M. Generating high-quality crowd density maps using contextual pyramid cnns[C]// Proceedings of the IEEE International Conference on Computer Vision, F, 2017.

[2] JIANG L, GAO C, MENG D, et al. DecideNet: counting varying density crowds through attention guided detection and density estimation[J]. Computer Vision and Pattern Recognition(cs.CV), 2018: 5197-5206.

[3] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

[4] JIANG X, ZHANG L, XU M, et al. Attention scaling for crowd counting[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2020.

[5] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv: 14091556, 2014,

[6] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[J]. IEEE Computer Society, 2017: 936-944.

[7] ZHANG H, ZU K, LU J, et al. EPSANet: an efficient pyramid squeeze attention block on convolutional neural network[J]. arXiv preprint arXiv: 2105.14447, 2021.

[8] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[J]. arXiv preprint arXiv: 1512.03385, 2016.

[9] DAI Y, GIESEKE F, OEHMCKE S, et al. Attentional feature fusion[J]. arXiv preprint arXiv: 2009.14082, 2020.

[10] Hu Jie, Shen Li, ALBANIE S, et al. Squeeze and excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(8): 2011-2023.

[11] GAO J, LIN W, ZHAO B, et al. C^3 Framework: an open-source pytorch code for crowd counting[J]. arXiv preprint arXiv: 1907.02724, 2019.

[12] SAM D B, SURYA S, BABU R V. Switching convolutional neural network for crowd counting[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017.

[13] ZHANG Y, ZHOU D, CHEN S, et al. Single-image crowd counting via multi-column convolutional neural network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[14] TIAN Y, LEI Y, ZHANG J, et al. PaDNet: pan-density crowd counting[J]. arXiv preprint arXiv: 1811.22805, 2018.

[15] SHI M, YANG Z, XU C, et al. Revisiting perspective information for efficient crowd counting[J]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2019: 7271-7280.

[16] MA Z, WEI X, HONG X, et al. Bayesian loss for crowd count estimation with point supervision[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision(ICCV), 2020.

[17] LIU N, LONG Y, ZOU C, et al. Adcrowdnet: an attention-

injective deformable convolutional network for crowd understanding[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[18] SINDAGI V A, PATEL V M. HA-CCN: hierarchical attention-based crowd counting network[J]. IEEE Transactions on Image Processing, 2019, 29: 323-335.

[19] OH M H, OLSEN P, RAMAMURTHY K N. Crowd counting with decomposed uncertainty[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11799-11806.

[20] IDREES H, TAYYAB M, ATHREY K, et al. Composition loss for counting, density map estimation and localization in dense crowds[C]//Proceedings of the European Conference on Computer Vision(ECCV), 2018.

[21] LI Y, ZHANG X, CHEN D. Csrnet: dilated convolutional neural networks for understanding the highly congested scenes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[22] JIANG X, XIAO Z, ZHANG B, et al. Crowd counting and density estimation by trellis encoder-decoder networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[23] GAO J, WANG Q, LI X. PCC net: perspective crowd counting via spatial convolutional network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, PP(99): 1-1.

[24] LIU C, WENG X, MU Y. Recurrent attentive zooming for joint crowd counting and precise localization[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2019.

[25] ZHANG A, SHEN J, XIAO Z, et al. Relational attention network for crowd counting[C]//Proceedings of the IEEE International Conference on Computer Vision, 2019.

[26] LUO A, YANG F, LI X, et al. Hybrid graph neural networks for crowd counting[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

[27] SINDAGI V A, YASARLA R, PATEL V M. JHU-CROWD++: large-scale crowd counting dataset and a benchmark method [J]. arXiv preprint arXiv: 200403597, 2020.

[28] 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(AVSS). IEEE, 2017.

[29] CAO X, WANG Z, ZHAO Y, et al. Scale aggregation network for accurate and efficient crowd counting[C]//Proceedings of the European Conference on Computer Vision(ECCV), 2018.

[30] WANG Q, GAO J, LIN W, et al. Learning from synthetic data for crowd counting in the wild[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[31] LIU L, QIU Z, LI G, et al. Crowd counting with deep structured scale integration network[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[32] SINDAGI V A, PATEL V M. Multi-level bottom-top and top-bottom feature fusion for crowd counting[C]//Proceedings of the IEEE International Conference on Computer Vision, 2019.

[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, 2020, 43(8): 2739-2751.

(**收稿日期**: 2022-04-01)

作者简介:

杨倩倩(1997-),**女**,**硕士研究生**,**主要研究方向**:**深度学习、人群计数**。

何晴(1998-),**女**,**硕士研究生**,**主要研究方向**:**深度学习、人群计数**。

彭思凡(1995-),**男**,**博士研究生**,**主要研究方向**:**深度学习、人群计数**。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部