计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2022/16(11)-2547-10
doi: 10.3778/j.issn.1673-9418.2104122

融合通道与空间注意力的编解码人群计数算法

余　鹰+,潘　诚,朱慧琳,钱　进,汤　洪
华东交通大学 软件学院,南昌 330013
+ 通信作者 E-mail: yuyingjx@163.com

摘　要:人群计数旨在准确地预测现实场景中人群的数量、分布和密度,然而现实场景普遍存在背景复杂、目标尺度多样和人群分布杂乱等问题,给人群计数任务带来极大的挑战。针对这些问题,提出了一种融合通道与空间注意力的编解码结构人群计数网络(CSANet)。该模型采用多层次编解码网络结构提取多尺度语义特征,并充分融合空间上下文信息,以此来解决复杂场景中行人尺度变化和分布杂乱的问题;为了降低复杂背景对计数性能的影响,在特征融合的过程中引入了通道与空间注意力,提高人群区域的特征权重,凸显感兴趣区域,同时降低弱相关背景区域的特征权重,抑制背景噪声干扰,最终提升人群密度图质量。为了验证算法的有效性,在多个经典人群计数数据集上进行了实验,实验结果表明,与现有的人群计数算法相比,CSANet具有良好的多尺度特征提取能力和背景噪声抑制能力,这使得密集场景下计数算法的准确性和鲁棒性均有较大提升。

关键词:人群计数;编解码网络;注意力;特征融合;深度学习
文献标志码:A　　**中图分类号:**TP391

Encoder-Decoder Network Fusing Channel and Spatial Attention for Crowd Counting

YU Ying+, PAN Cheng, ZHU Huilin, QIAN Jin, TANG Hong
College of Software, East China Jiaotong University, Nanchang 330013, China

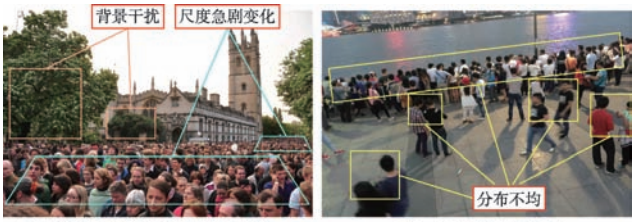
Abstract: The purpose of crowd counting is to accurately predict the number, distribution and density of crowds in real scenes. However, crowd counting often suffers from some problems such as complex background, diverse target scales, and cluttered crowd distribution, which strongly affects the precision of counting. To solve these problems, a channel and spatial attention-based encoder-decoder network for crowd counting (CSANet) is proposed. It uses a multi-level encoder-decoder network to extract multi-scale semantic features, and fully integrates spatial context information to solve the problem of pedestrian scale changes and messy distribution in complex scenes. To reduce the impact of complex background on counting performance, channel and spatial attention are introduced in the process of feature fusion to improve the quality of crowd density map by increasing the feature weights of crowd regions to highlight regions of interest, and decreasing the feature weights of weakly correlated background regions to suppress background noise interference. To verify the effectiveness of the proposed algorithm, experiments are conducted on several classical crowd counting datasets, and the experimental results show that CSANet performs well in multi-scale feature extraction and background noise suppression compared with existing crowd counting algorithms, which greatly improves the accuracy and robustness of counting algorithm in dense scenes.

Key words: crowd counting; encoder-decoder network; attention; feature fusion; deep learning

基金项目:国家自然科学基金(62163016,62066014);江西省自然科学基金(20212ACB202001,20202BABL202018)。
This work was supported by the National Natural Science Foundation of China (62163016, 62066014), and the Natural Science Foundation of Jiangxi Province (20212ACB202001, 20202BABL202018).
收稿日期:2021-05-07　　**修回日期:**2021-06-23

人群计数作为智能视频监控的重要组成部分，主要任务是分析统计场景中人群的数量、密度和分布，现已广泛应用在大型集会、旅游景点等人群密集的线下活动场景，在维护群众人身安全等方面发挥着巨大的作用。近年来，随着卷积神经网络[1-3]在计算机视觉领域的大放异彩，基于深度学习的人群计数算法取得了显著的进展，计数形式从简单的稀疏场景行人数量统计发展到了复杂密集场景的密度图计数，通过充分利用深度神经网络强大的特征表达能力，提升模型的计数精度。

随着计算机视觉和深度学习技术的快速发展，有关人群计数问题的研究已经取得了巨大的进展，优秀的模型和算法不断涌现，但是在人群密集场景中，要实现准确的计数依然存在诸多困难和挑战。如图1所示，该现实场景存在背景干扰、人群分布杂乱、行人尺度变化等问题，极大地影响了计数精度。在图1(a)中，远近景人群目标尺度差异较大，树与密集人群特征相似，容易对计数造成干扰；在图1(b)中，同样存在远近景目标尺度多样化问题，同时人群分布杂乱将对计数性能造成影响。



(a) 背景干扰和多尺度变化
(a) Background interference and diverse scales

(b) 人群分布杂乱
(b) Cluttered crowd distribution

图1 人群计数的挑战

Fig.1 Challenge of crowd counting

为了解决行人尺度变化问题，一些学者试图通过引入多阵列卷积结构来感受不同尺度的行人特征[4-5]，以提高模型预测精度。尽管这些方法增强了算法对多尺度特征的感知能力，但同时也带来了无效的冗余分支结构和大量训练时间。对于背景噪声干扰，Liu等人[6]试图使用注意力机制去抑制背景区域。通过级联方式，预先训练注意力图生成器，检测前景人群区域，抑制弱相关复杂背景信息，然后使用人群密度估计器进行人群计数。此时，场景图片已经聚焦在前景人群区域，可以有效减少背景噪声的干扰。这类方法对注意力生成器要求极高，容易造成前景和背景的误判，也不能自适应地在线调整背景区域范围，可能在计数之前引入误差，增加了计数任务的复杂性。

针对上述问题，本文提出了一种融合通道与空间注意力的编解码结构人群计数网络(channel and spatial attention-based encoder-decoder network for crowd counting，CSANet)，以解决计数任务中存在的目标尺度变化、人群分布杂乱以及背景噪声干扰等问题。在编码阶段，通过不同深度层次的卷积提取人群的不同尺度特征；在解码阶段，使用卷积和上采样操作逐步恢复空间语义信息，并将多尺度语义信息与空间上下文信息充分融合，然后注入通道和空间注意力，使网络关注点聚焦在感兴趣前景人群区域，进一步降低弱相关背景干扰，以此提高密度图的生成质量。本文的主要贡献如下：

(1)提出了一种融合通道与空间注意力的编解码结构计数网络，通过将多尺度信息与空间上下文信息进行融合以提高图像特征的鲁棒性，最终提升计数精度。

(2)将多维度注意力机制引入人群计数，使得端到端的计数网络能够自适应地聚焦前景人群区域，降低弱相关背景区域的干扰，提升生成密度图质量。

1 相关工作

人群计数任务所遇到的挑战主要为场景拥挤、人群尺度变化多样和人群分布杂乱等。为了降低其带来的计数精度下降问题，主要研究路线大致可分为传统方法和基于深度学习的方法。传统方法使用经过预训练的分类器人工提取目标底层特征[7-8]，然后判别出行人从而实现计数；基于深度学习的方法利用卷积神经网络自动学习人群特征并生成场景密度图，密度图中不仅包含行人数量信息，还有丰富的空间位置信息。

1.1 传统方法

传统方法可分为基于检测和基于回归两类。基于检测的方法[9]首先通过滑动窗口提取图像特征，然后使用已经训练好的分类器来识别行人。此类方法在人群稀疏的场景中计数效果良好，但是在复杂的人群密集场景中，由于行人之间的严重遮挡和背景杂乱干扰，导致无法提取完整的个体特征，计数性能较差。为了克服密集场景中行人特征不完整等问题，研究者设计出判别身体部分特征的检测器[10]，但是算法仍然难以胜任高密度场景的计数需求。基于此，提出了另一种自适应的回归预测方法[11]，直接从场景中提取特征，然后学习图像特征至人群数量的映射关系。

总之，传统方法大都依赖人工提取的特征。由于现实环境复杂，人群变化等因素普遍存在，导致人工提取的特征判别性不强，从而计数模型应用时预测效果较差。

1.2　基于深度学习的方法

近些年，深度学习技术在图像分类[12]、目标检测[13-14]、语义分割[15]等视觉任务上的应用表现抢眼。相对于使用传统技术，使用深度学习技术可以使算法的性能得到显著提升，并且其更擅长处理复杂场景问题。因此，基于卷积神经网络（convolutional neural networks，CNN）的人群计数方法的研究陆续开展[16-18]，并取得了卓有成效的进展。其主要过程是通过卷积神经网络提取特征，再利用全卷积形式生成包含人群数量和空间位置信息的人群分布密度图。

为了处理多尺度问题，已有模型大多采用多阵列卷积神经网络架构[4-5]，通过不同的感受野去提取行人多尺度特征。Sindagi等人[19]提出了一种上下文金字塔网络（contextual pyramid CNN，CP-CNN），通过融合全局和局部上下文信息，来提高生成密度图的质量和人数预测的精度；Sam等人[20]提出Switch-CNN（switching convolutional neural network）模型，通过训练密度分类器，将图像划分为局部图像块，用分类器自适应地输出对应等级；Cao等人[21]提出了一种基于编解码结构的尺度聚焦网络（scale aggregation network，SANet），利用多尺度聚焦模块来提取行人多尺度特征。此类方法的计数性能相比传统方法虽然有了很大突破，但是其网络结构冗余，参数量过大，导致模型训练困难。为了简化网络复杂度和提高训练效率，单列网络架构重新获得关注。Li等人提出单列计数网络CSRNet（network for congested scene recognition）[22]，通过空洞卷积扩大感受野，以捕获多尺度特征同时降低网络模型的参数量。为了解决背景噪声干扰问题，Liu等人[6]提出了一种用于人群计数的可形变卷积网络（attention-injective deformable convolutional network for crowd understanding，AD-CrowdNet），该网络融合了注意力机制，让模型只关注人群区域，从而忽略背景噪声的干扰。此外，亦有研究通过将图像语义分割技术应用于人群计数领域，以去除背景噪声。总之，如何增强特征的尺度适应性和降低背景噪声干扰仍然是人群计数领域目前重点关注的问题。

2　CSANet模型

本文提出的融合通道与空间注意力的人群计数模型CSANet的网络结构如图2所示。整体采用了易于端到端训练的编解码架构。其中，编码器使用VGG16[1]网络的前13层作为主干，构建特征提取网络，提取多个不同深度层次的语义特征，来辨识场景中的多尺度人群；解码器在逐步恢复空间信息的同时，将多尺度信息与空间上下文信息充分融合，以增强网络的表征能力。并且融入通道与空间注意力模块，聚焦前景人群区域，抑制弱相关背景特征，以生成高质量、高分辨率的密度图进行人群计数。

2.1　编解码器Encoder-Decoder

编解码器包含两部分，其中编码器可以提取不同尺度行人特征。为了提取多层次更具有表征能力的深度特征，且易于网络的搭建和训练，本部分选取了经过预训练的VGG16网络前13层作为编码器的

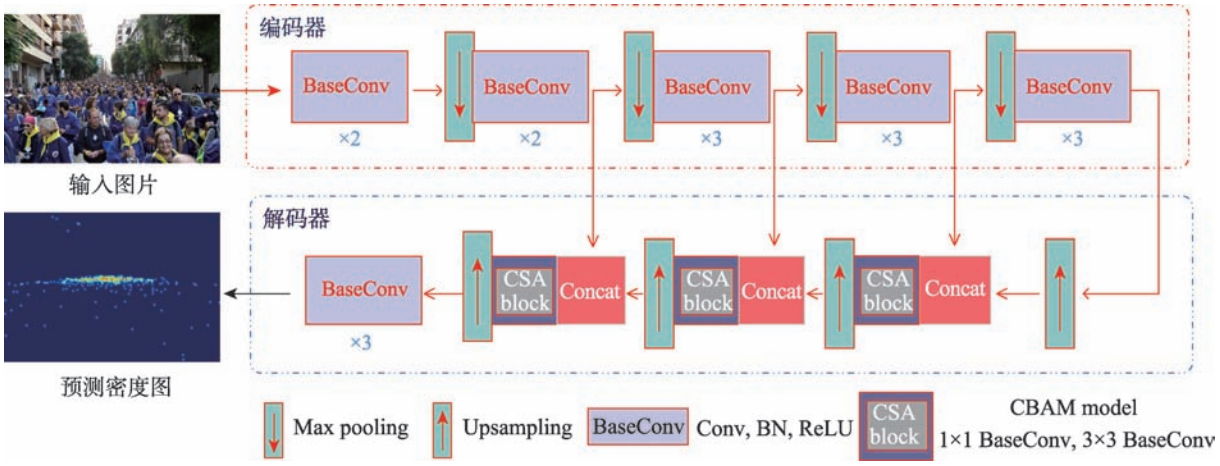


图2　CSANet网络结构

Fig.2　Architecture of CSANet

主干网络。在训练的过程中,保留4个具有代表性的不同层次深度语义特征Conv2_2、Conv3_3、Conv4_3、Conv5_3,其尺寸分别为原始输入图片分辨率的1/2、1/4、1/8、1/16,这些不同深度提取的特征可以捕获不同尺度的行人信息。随着网络深度递增,特征图分辨率逐渐减小,维度逐步增加。解码器主要用于逐步恢复图像空间特征信息与聚焦前景人群区域。通过解码恢复的多层次深度特征与编码器各阶段输出的对应层特征进行融合,最大程度上减少卷积和下采样等操作造成的特征损失,并进一步整合空间上下文信息。在融合之后,对特征添加通道与空间注意力,以此来凸出前景人群区域,抑制弱相关背景区域特征的权重。解码器对不同阶段特征图进行融合主要是对两个特征图进行通道拼接,特征融合之后新的特征图分辨率大小不变,通道为两者之和。其网络参数配置如表1所示。

表1 网络参数

Table 1 Network parameters

编码器	解码器
Conv1_1(3-64-1)	Upsampling
Conv1_2(3-64-1)	Concat
Max pooling	CBAM module
Conv2_1(3-128-1)	Conv6_1(1-256-1)
Conv2_2(3-128-1)	Conv6_2(3-256-1)
Max pooling	Upsampling
Conv3_1(3-256-1)	Concat
Conv3_2(3-256-1)	CBAM module
Conv3_3(3-256-1)	Conv7_1(1-128-1)
Max pooling	Conv7_2(3-128-1)
Conv4_1(3-512-1)	Upsampling
Conv4_2(3-512-1)	Concat
Conv4_3(3-512-1)	CBAM module
Max pooling	Conv8_1(1-64-1)
Conv5_1(3-512-1)	Conv8_2(3-64-1)
Conv5_2(3-512-1)	Upsampling
Conv5_3(3-512-1)	Conv9_1(3-32-1)
	Conv9_2(3-32-1)
	Conv10_1(1-1-1)

在Conv$X_Y(K\text{-}C\text{-}S)$中,X_Y代表卷积所在层的深度,K表示卷积核大小,C为卷积核个数,S为步长。最后输出的密度图分辨率大小与原始输入图片的相等。Upsampling使用双线性插值将分辨率扩大至输入特征的2倍,Concat为特征融合操作,将输入的2组特征图进行通道拼接,CBAM module为通道与空间特征注意力模块。

2.2 通道与空间注意力模块

背景噪声干扰问题给人群计数任务带来了严峻的挑战,复杂背景可能极大降低模型的预测精度。视觉注意力机制的作用已经在大量的工作中被证实,它在关键特征提取以及模型性能增强等方面有着良好的效果。如果将注意力机制应用于人群计数,将有助于模型更加关注感兴趣的人群区域,从而抑制弱相关背景信息的影响。Woo等人[23]提出的CBAM(convolutional block attention module)注意力模型可以在通道和空间两个维度上添加注意力,相较于单通道域或单空间域注意力,更适合人群计数任务。因为人群计数模型生成的特征图不仅包含人群数量信息,还包含空间位置信息。对于一个给定的中间特征图,CBAM模块会沿着通道和空间两个独立的维度依次推断注意力图,然后将注意力图与输入特征图相乘以进行自适应特征优化来提高感兴趣区域的权重。添加CBAM注意力模块时,一般将其添加到网络每个卷积层之后或结合残差添加。

为了增强模型在多层次特征融合之后对人群区域的聚焦能力,CSANet网络在解码器部分添加了CBAM注意力模块,融合方式如图3所示。编码器和解码器提取的特征图在对应层次进行通道叠加,以充分整合空间上下文信息,再使用通道与空间注意力模块,对其前景行人区域进行关注,并对背景区域特征权重进行抑制。具体过程为:首先将编码阶段提取的多尺度特征F_e与对应层解码恢复的特征F_d做特征叠加操作,得到特征累加之后的特征图F',如式(1)所示:

$$F'=F_e\oplus F_d \tag{1}$$

其中,$\oplus$为特征通道叠加操作,F'为多层信息融合之后的特征图,并作为注意力模块的输入,然后依次利用通道和空间注意力模块微调输入特征F',得到最终经过加权之后的特征图F_{Att}。通道注意力模块学习通道上的权重信息,再按通道元素相乘,作为后一阶段的输入;空间注意力模块学习空间权重,与输入特征空间相乘,如式(2)和式(3)所示:

$$M_c(F)=\sigma((MLP(F^c_{avg}))+(MLP(F^c_{max}))) \tag{2}$$

$$M_s(F)=\sigma(f^{7\times7}([F^s_{avg};F^s_{max}])) \tag{3}$$

σ为Sigmoid函数,输入特征图$F\in\mathbf{R}^{C\times H\times W}$,通道注意力为$M_c\in\mathbf{R}^{C\times1\times1}$,$F^c_{avg}\in\mathbf{R}^{C\times1\times1}$和$F^c_{max}\in\mathbf{R}^{C\times1\times1}$为每个单独通道上的平均池化和最大池化,$MLP$为多层感知机,这里仅使用了一个隐藏层,其神经元个数为

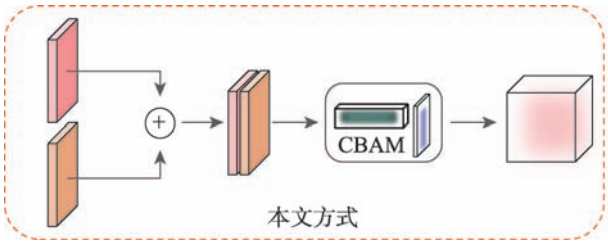


图3　注意力融合方式

Fig.3　Fusing attention method

$\mathbf{R}^{C/r\times1\times1}$，$r$ 为参数缩减率，$r=16$；空间注意力为 $M_s\in\mathbf{R}^{1\times H\times W}$，$F^s_{avg}\in\mathbf{R}^{1\times H\times W}$ 和 $F^s_{max}\in\mathbf{R}^{1\times H\times W}$ 为所有通道上的全局平均池化和最大池化，做通道相加操作，$f^{7\times7}$ 为 7×7 卷积。

2.3　损失函数

在训练过程中，使用欧式距离评估真实密度图与预测密度图之间的差异，其定义如式(4)所示：

$$L_{\text{density}}=\frac{1}{N}\sum_{i=1}^{N}\|Z(X_i;\theta)-Z_i^{\text{GT}}\|^2 \tag{4}$$

N 是一次训练图片的总数量，X_i 为第 i 张训练图片，$Z(X_i;\theta)$ 为第 i 张图片的预测密度图，其中 $i\in[1,N]$，θ 为网络模型参数，Z_i^{GT} 为第 i 张训练图片的真实密度图。

3　训练方法

本章将详细阐述端到端人群计数模型 CSANet 的训练环境，包括真实密度图的生成方式、数据增强方法以及实验参数和硬件配置。

3.1　真实密度图

由于当下主流人群计数数据集通常只提供人头中心点的坐标位置信息，而模型对于单个像素点的预测效率低下，普遍做法是将坐标点进行区间扩散，以提升模型的学习效率。本文使用几何自适应高斯核生成密度图，作为预测学习的标签，具体如式(5)所示：

$$F(x)=\sum_{i=1}^{N}\delta(x-x_i)*G_{\sigma^i}(x),\ \sigma^i=\beta\bar{d} \tag{5}$$

其中，x 为当前图像中的每个像素点，x_i 为第 i 个人头中心点坐标，$G(x)$ 为高斯核滤波器，$\bar{d}$ 为人头坐标点 x_i 与其最近的 K 个人头的平均距离。参照文献[22]的参数设置，将 β 设为 0.3。

3.2　数据增强

由于人群数据集图片数量有限，而标注图片代价过高，为了获得更多的图片用于训练，本文在数据输入网络之前对数据集中的图片进行了一系列数据增强操作。具体为对每张图片随机裁剪出分辨率大小为 400×400 的局部图像块，如图 4 所示。对于边长不足 400 的图片，对其进行双线性插值，使得边长增大到 400。再对裁剪出的局部图像块随机进行镜像翻转，调整对比度和灰度来扩大数据量，以获得更丰富的训练数据。

图4　随机裁剪示例

Fig.4　Example of random cropping

3.3　实验设置

实验所使用的操作系统为 Windows 10，深度学习框架为 PyTorch 1.6.0，使用两块显存为 11 GB 的 NVIDIA-1080Ti 显卡。

编码器部分使用基于 ImageNet[24] 预训练的 VGG16 网络的前 13 层参数对网络进行初始化，其他参数则利用均值为 0，方差为 0.01 的高斯函数进行随机初始化。模型训练过程中，使用学习率为 1E－4 的 Adam 优化器进行模型优化，训练迭代次数收敛即停止。对于 UCF-QNRF 数据集，其平均尺寸为 2 013×2 902，分辨率过大，训练效率低，因此在进行数据增强之前，本文使用双线性插值方法将其大小统一调整至 1 024×768。

4　实验与分析

为了验证算法的有效性和性能，在 4 个经典人群计数数据集上进行了实验。与已有计数算法相比，CSANet 性能更优，而且训练过程更加简单、灵活。本章首先介绍计数模型的评价指标，然后简单描述用于实验的 4 个数据集的基本情况，并比较分析了各个算法的实验结果。

4.1　评价指标

平均绝对误差(mean absolute error，MAE)和均方根误差(root mean square error，RMSE)是人群计数算法常用的评价指标；MAE 和 RMSE 均可以表示预测人数与真实人数的差异程度，但是 MAE 通常用来

评估模型的准确性，而RMSE通常用来度量被评估模型的鲁棒性。MAE和RMSE的值越小，表示模型性能越好，其计算方法如式(6)和式(7)所示：

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|C_i-C_i^{GT}| \tag{6}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_i-C_i^{GT})^2} \tag{7}$$

其中，N为数据集图像总数；C_i为第i张图片的预测人数；C_i^{GT}为第i张图片的真实人数。

4.2 数据集与实验分析

4.2.1 ShanghaiTech数据集

ShanghaiTech[5]是一个大型的人群计数数据集，共标注了1 198幅图像，人头总数为330 165个。按照数据来源和场景稀疏程度划分，可分为Part_A和Part_B这两部分，其中Part_A随机采集自互联网，人群分布较为密集，共有300幅图像作为训练集，182幅图像作为测试集；而Part_B采集自上海市的部分监控视频，人群分布较为稀疏，有400幅图像作为训练集，316幅图像作为测试集。该数据集的实验结果如表2所示。

表2 不同计数方法在ShanghaiTech数据集上的性能比较

Table 2 Performance comparison of different methods on ShanghaiTech dataset

方法	Part_A		Part_B	
	MAE	RMSE	MAE	RMSE
MCNN[5]	110.2	173.2	26.4	41.3
CP-CNN[19]	73.6	106.4	20.1	30.1
CSRNet[22]	68.2	115.0	10.6	16.0
ASD[25]	65.6	98.0	8.5	13.7
PACNN[26]	66.3	106.4	8.9	13.5
SFANet[27]	59.8	99.3	**6.9**	**10.9**
DUBNet[28]	64.6	106.8	7.8	12.2
CSANet	**59.4**	**97.1**	7.1	11.2

与已有算法相比，CSANet在Part_A上的性能指标MAE与RMSE均达到了最优值，而在Part_B上，性能仅次于SFANet。总损失变化趋势如图5所示，训练之初由于随机程度较高，损失较大，但是随着模型不断迭代训练，损失呈现明显的下降趋势并趋于稳定；Part_A部分在整体可控范围内波动，Part_B部分在400次迭代之后基本达到了稳定状态。

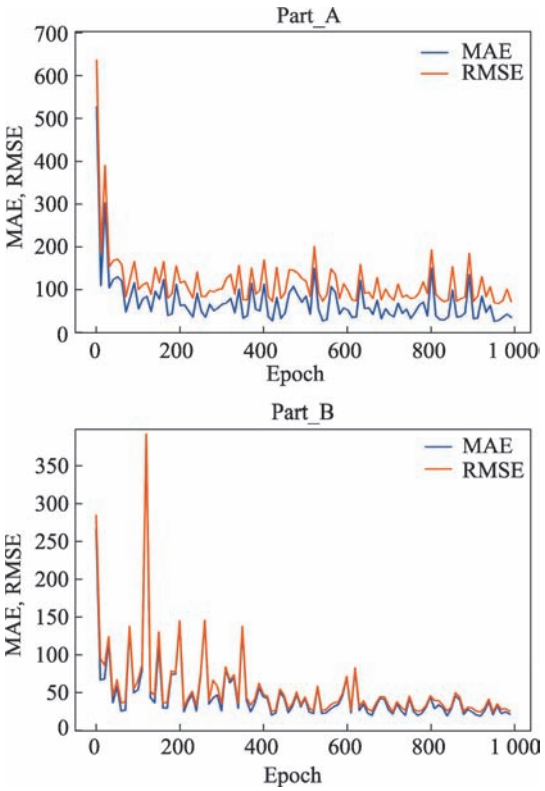


图5 ShanghaiTech数据集训练过程

Fig.5 Training process on ShanghaiTech dataset

4.2.2 UCF_QNRF数据集

UCF_QNRF[29]是一个挑战性极大的数据集，场景丰富且人群分布杂乱，共标注了1 535幅图像，其中训练集有1 201幅图像，测试集有334幅图像，标注总人数达到了1 251 642。

表3显示了各种人群计数算法在UCF_QNRF数据集上的实验结果。由表3可见，CSANet网络的两个性能指标MAE和RMSE均为最优，证明CSANet模型在跨场景计数时具有较好的性能。CSANet的训练损失曲线如图6所示，前500次迭代的波动较大，500次后逐渐趋于稳定。

表3 不同计数方法在UCF_QNRF数据集上的性能比较

Table 3 Performance comparison of different methods on UCF_QNRF dataset

方法	MAE	RMSE
MCNN[5]	277.0	426.0
CMTL[30]	252.0	514.0
IA-DCCN[31]	125.0	186.0
RANet[32]	111.0	190.0
DUBNet[28]	105.6	180.5
CSANet	**104.8**	**175.4**

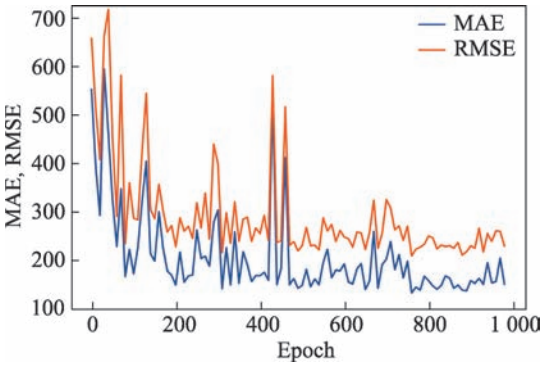


图6　UCF_QNRF数据集训练过程

Fig.6　Training process on UCF_QNRF dataset

4.2.3　UCF_CC_50数据集

UCF_CC_50数据集[33]中的图像全部采集自互联网，其场景包括音乐会、游行示威等人群高度密集的场合，总共有50幅不同分辨率、不同视角拍摄的极度密集图像，共标注人头数量为63 974个，每幅图像标注人数从最低94人到最高4 543人不等，平均每张图片标注的人头数为1 280个，其数量远超其他人群计数数据集。数据集使用5折标准交叉验证训练，实验结果如表4所示。由表4可见，即使是在极端密集的场景中，CSANet网络的计数准确性和鲁棒性依然优于已有模型。

4.2.4　实验结果可视化分析

为了更好地说明模型的预测效果，本小节展示了CSANet网络在不同数据集上预测的部分密度图，如图7所示。其中，第1行图片选自ShanghaiTech Part_A测试集，代表了高度拥挤和严重背景干扰场景的预测效果；第2行图片选自ShanghaiTech Part_B测试集，表示了在正常街道中，人群分布不均时的预测效果；第3行为UCF_QNRF测试集图片，来自一个游行集会场景。由绝大多数场景的可视化表现可知，CSANet模型生成的人群分布密度图非常接近真实的人群分布密度图，说明CSANet具有良好的多尺度特征提取能力和背景噪声抑制能力。

表4　不同计数方法在UCF_CC_50数据集上的性能比较

Table 4　Performance comparison of different methods on UCF_CC_50 dataset

方法	MAE	RMSE
MCNN[5]	377.6	509.1
Switch-CNN[20]	318.1	439.2
CP-CNN[19]	295.8	320.9
CSRNet[22]	266.1	397.5
ASD[25]	196.2	270.9
PACNN[26]	267.9	357.8
CSANet	**191.3**	**262.8**

4.3　消融实验

为了验证CSANet网络中各模块的有效性，在ShanghaiTech数据集上做了相关的消融实验，结果如表5所示。

主干网络为CSANet网络中设计的编解码部分，由表5可见，其计数精度优于绝大多数经典计数网络，表现出了骨干网络强大的特征提取能力。在融

图7　结果可视化

Fig.7　Result visualization

表5 ShanghaiTech数据集消融实验

Table 5 Ablation study on ShanghaiTech dataset

模块	Part_A		Part_B	
	MAE	RMSE	MAE	RMSE
主干网络	60.9	99.2	7.7	12.2
主干网络+注意力	**58.5**	**96.1**	**7.1**	**11.2**

入通道与空间注意力模块之后，CSANet网络的计数效果显著提升。本节还对消融实验的结果进行了可视化，如图8所示。由图8可见，对于图中红色框中的背景区域部分，主干网络已经能够获得比较准确的密度图，但是经过注意力前景增强和背景抑制之后可以看出，密度图的前景部分更加显著，背景误差也相对减少。

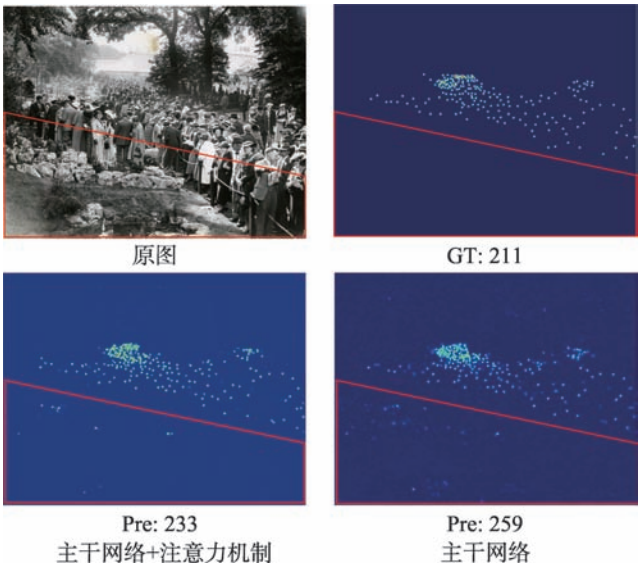


图8 消融实验结果可视化

Fig.8 Visualization of ablation study results

5 结束语

本文提出了一种融合通道与空间注意力的编解码人群计数网络CSANet。该模型能够以端到端的形式进行训练，整体采用了编解码结构以提取多尺度特征和充分融合空间上下文信息，并加以通道与空间注意力模块来提升前景行人区域的权重，并抑制弱相关背景特征，以此生成高质量的密度图。经过实验分析，证明CSANet网络具有良好的准确性与鲁棒性。未来的工作中，将考虑如何采用可形变卷积等方面，更加准确地聚焦人群区域，以进一步提高人群计数的精度。

参考文献：

[1] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2016-08-22)[2021-01-06]. https://arxiv.org/abs/1608.06197.

[2] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 27-30, 2016. Washington: IEEE Computer Society, 2016: 770-778.

[3] HAN K, WANG Y H, TIAN Q, et al. GhostNet: more features from cheap operations[C]//Proceedings of the 2020 IEEE Conference Computer Vision and Pattern Recognition, Seattle, Jun 13-19, 2020. Piscataway: IEEE, 2020: 1580-1589.

[4] ZHANG C, LI H S, WANG X G, et al. Cross-scene crowd counting via deep convolutional neural networks[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, Jun 7-12, 2015. Washington: IEEE Computer Society, 2015: 833-841.

[5] ZHANG Y Y, ZHOU D, CHEN S Q, et al. Single-image crowd counting via multi-column convolutional neural network [C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 27-30, 2016. Washington: IEEE Computer Society, 2016: 589-597.

[6] LIU N, LONG Y C, ZOU C Q, et al. ADCrowdNet: an attention-injective deformable convolutional network for crowd understanding[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-20, 2019. Piscataway: IEEE, 2019: 3225-3234.

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

[8] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, Jun 20-26, 2005. Washington: IEEE Computer Society, 2005: 886-893.

[9] DOLLAR P, WOJEK C, SCHIELE B, et al. Pedestrian detection: an evaluation of the state of the art[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 34(4): 743-761.

[10] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis

and Machine Intelligence, 2009, 32(9): 1627-1645.

[11] CHEN K, GONG S G, XIANG T, et al. Cumulative attribute space for age and crowd density estimation[C]//Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, Jun 23-28, 2013. Washington: IEEE Computer Society, 2013: 2467-2474.

[12] HOWARD A, SANDLER M, CHU G, et al. Searching for MobileNetV3[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 1314-1324.

[13] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[EB/OL]. (2016-01-06) [2021-01-06]. https://arxiv.org/abs/1506.01497.

[14] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2021-01-06]. https://arxiv.org/abs/2004.10934.

[15] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//LNCS 11211: Proceedings of the 15th European Conference on Computer Vision, Munich, Sep 8-14, 2018. Cham: Springer, 2018: 833-851.

[16] 余鹰, 朱慧琳, 钱进, 等. 基于深度学习的人群计数研究综述[J]. 计算机研究与发展, 2021, 58(12): 2724-2747.
YU Y, ZHU H L, QIAN J, el al. Survey on deep learning based crowd counting[J]. Journal of Computer Research and Development, 2021, 58(12): 2724-2747.

[17] GAO G S, GAO J Y, LIU Q J, et al. CNN-based density estimation and crowd counting: a survey[EB/OL]. (2020-03-28)[2021-01-06]. https://arxiv.org/abs/2003.12783.

[18] YU Y, ZHU H L, WANG L W, et al. Dense crowd counting based on adaptive scene division[J]. International Journal of Machine Learning and Cybernetics, 2021, 12(4): 931-942.

[19] SINDAGI V A, PATEL V M. Generating high-quality crowd density maps using contextual pyramid CNNs[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Oct 22-29, 2017. Washington: IEEE Computer Society, 2017: 1861-1870.

[20] 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, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 4031-4039.

[21]CAO X K, WANG Z P, ZHAO Y Y, et al. Scale aggregation network for accurate and efficient crowd counting[C]//LNCS 11209: Proceedings of the 15th European Conference on Computer Vision, Munich, Sep 8-14, 2018. Cham: Springer, 2018: 734-750.

[22] LI Y H, ZHANG X F, CHEN D M. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. Washington: IEEE Computer Society, 2018: 1091-1100.

[23] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//LNCS 11211: Proceedings of the 15th European Conference on Computer Vision, Munich, Sep 8-14, 2018. Cham: Springer, 2018: 3-19.

[24] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, Jun 20-25, 2009. Washington: IEEE Computer Society, 2009: 248-255.

[25] WU X J, ZHENG Y B, YE H, et al. Adaptive scenario discovery for crowd counting[C]//Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, May 12-17, 2019. Piscataway: IEEE, 2019: 2382-2386.

[26] SHI M J, YANG Z H, XU C, et al. Revisiting perspective information for efficient crowd counting[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-20, 2019. Piscataway: IEEE, 2019: 7279-7288.

[27] ZHU L, ZHAO Z J, LU C, et al. Dual path multi-scale fusion networks with attention for crowd counting[EB/OL]. (2019-02-04)[2021-01-06]. https://arxiv.org/abs/1902.01115.

[28] OH M, OLSEN P, RAMAMURTHY K N. Crowd counting with decomposed uncertainty[C]//Proceedings of the 2020 AAAI Conference on Artificial Intelligepnce, New York, Feb 7-12, 2020. Menlo Park: AAAI Press, 2020: 11799-11806.

[29] IDREES H, TAYYAB M, ATHREY K, et al. Composition loss for counting, density map estimation and localization in dense crowds[C]//LNCS 11206: Proceedings of the 15th European Conference on Computer Vision, Munich, Sep 8-14, 2018. Cham: Springer, 2018: 544-559.

[30] SINDAGI V A, PATEL V M. CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting[C]//Proceedings of the 14th IEEE International

Conference on Advanced Video and Signal Based Surveillance, Lecce, Aug 29-Sep 1, 2017. Washington: IEEE Computer Society, 2017: 1-6.

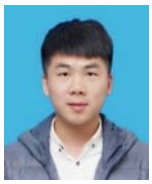
[31] SINDAGI V A, PATEL V M. Inverse attention guided deep crowd counting network[C]//Proceedings of the 16th IEEE International Conference on Advanced Video and Signal Based Surveillance, Taipei, China, Sep 18-21, 2019. Piscataway: IEEE, 2019: 1-8.

[32] ZHANG A R, SHEN J Y, XIAO Z H, et al. Relational attention network for crowd counting[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27 -Nov 2, 2019. Piscataway: IEEE, 2019: 6788-6797.

[33] IDREES H, SALEEMI I, SEIBERT C, et al. Multi-source multi-scale counting in extremely dense crowd images[C]//Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, Jun 23-28, 2013. Washington: IEEE Computer Society, 2013: 2547-2554.

余鹰(1979—),女,博士,副教授,硕士生导师,CCF会员,主要研究方向为机器学习、计算机视觉、粒计算等。

YU Ying, born in 1979, Ph.D., associate professor, M.S. supervisor, member of CCF. Her research interests include machine learning, computer vision, granular computing, etc.

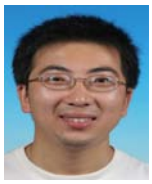


潘诚(1995—),男,硕士研究生,主要研究方向为机器学习、计算机视觉。

PAN Cheng, born in 1995, M.S. candidate. His research interests include machine learning and computer vision.

朱慧琳(1996—),女,硕士研究生,主要研究方向为计算机视觉。

ZHU Huilin, born in 1996, M.S. candidate. Her research interest is computer vision.



钱进(1975—),男,博士,教授,硕士生导师,CCF会员,主要研究方向为粒计算、大数据挖掘、机器学习。

QIAN Jin, born in 1975, Ph.D., professor, M.S. supervisor, member of CCF. His research interests include granular computing, big data mining and machine learning.

汤洪(1998—),男,硕士研究生,主要研究方向为深度学习、计算机视觉。

TANG Hong, born in 1998, M.S. candidate. His research interests include deep learning and computer vision.