·开发研究与工程应用·　　文章编号：1000-3428(2022)09-0314-07　　文献标志码：A　　中图分类号：TP391.41

基于背景抑制与上下文感知的人群计数网络

黄奕秋[1]，胡　晓[2]，杨佳信[1]，欧嘉敏[1]

（1.广州大学 电子与通信工程学院，广州 510006；2.广州大学 机械与电气工程学院，广州 510006）

摘　要：针对图像背景噪声、透视畸变等影响人群计数网络计数精度的问题，提出一种基于背景抑制与上下文感知的新网络。利用VGG-16网络提取图像特征，并分别将特征输入密度图生成模块和背景噪声抑制（BNS）模块中进行处理，生成密度特征图和空间注意力图。使用BNS模块优化密度特征图并生成初级密度图，以抑制图像中背景噪声干扰，提高人群区域的特征权重。为减少透视畸变对人群密度估计的影响，使用上下文感知增强网络优化初级密度图，并生成预测密度图。在ShanghaiTech、UCF-CC-50及UCF-QNRF 3个公开数据集上的实验结果表明，该网络相较于MCNN、SwitchCNN、CSRNet等网络的计算准确度较高，尤其在UCF-QNRF数据集上其平均绝对误差和均方误差分别为85.8、146.0，相较于其他网络最高分别下降69.0%和67.2%，能充分抑制图像背景噪声并有效减小透视畸变引起的误差，具有良好的泛化能力和较强的鲁棒性。

关键词：人群计数；深度学习；密度图；背景噪声；上下文感知

开放科学（资源服务）标志码（OSID）：

中文引用格式：黄奕秋，胡晓，杨佳信，等. 基于背景抑制与上下文感知的人群计数网络[J]. 计算机工程，2022，48(9)：314-320.

英文引用格式：HUANG Y Q，HU X，YANG J X，et al. Crowd counting network based on background suppression and context awareness[J]. Computer Engineering，2022，48(9)：314-320.

Crowd Counting Network Based on Background Suppression and Context Awareness

HUANG Yiqiu[1]，HU Xiao[2]，YANG Jiaxin[1]，OU Jiamin[1]

（1.School of Electronics and Communication Engineering，Guangzhou University，Guangzhou 510006，China；
2.School of Mechanical and Electrical Engineering，Guangzhou University，Guangzhou 510006，China）

【Abstract】To reduce the influence of background noise and perspective distortion in crowd counting tasks，a new network based on background suppression and context awareness is proposed. VGG-16 network is used to extract image features，which are input into Density Map Generation（DMG）and Background Noise Suppression（BNS）modules for processing to generate density feature and spatial attention maps. The BNS module is used to optimize a density feature map and generate a primary density map，to suppress noise information interference in the image and improve the characteristic weight of the crowd area. To reduce the influence of perspective distortion on counting density estimation，a Weight Enhancement-Context Aware Network（WE-CAN）is used to optimize the primary density map and generate the predicted density map. Experiment results on three public datasets，namely ShanghaiTech，UCF-CC-50 and UCF-QNRF show that the network has higher computational accuracy than Multi-Column Convolutional Neural Network（MCNN），Switching Convolutional Neural Network（SwitchCNN），Congested Scene Recognition Network（CSRNet）and other networks. Especially on UCF-QNRF，the Mean Absolute Error（MAE）of the proposed algorithm reach 85.8，and the Mean Square Error（MSE）reach 146.0. Compared with other algorithms，the highest decrease is 69.0% and 67.2%，respectively. The network proposed can also suppress background noise，reduce the error caused by perspective distortion，and has good accuracy and robustness.

【Key words】crowd counting；deep learning；density map；background noise；context awareness

DOI：10. 19678/j. issn. 1000-3428. 0062606

0　概述

近年来，随着城市居住人口的不断增加，人口密集程度不断上升，导致拥挤踩踏风险也随之增加。为实时检测人群数量变化，避免拥挤踩踏对公共安全、交通控制、智能交通等造成破坏，精确地监测人群的数量变

基金项目：国家自然科学基金（62076075）。

作者简介：黄奕秋（1997—），男，硕士研究生，主研方向为计算机视觉、人群计数；胡　晓，教授，博士；杨佳信、欧嘉敏，硕士研究生。

收稿日期：2021-09-06　　**修回日期：**2021-10-28　**E-mail：**798919380@qq.com

计算机工程 www.ecice06.com

化成为一个十分重要的课题。

基于计数方式的不同，人群计数算法可以分为传统的人群计数算法和基于深度学习的人群计数算法。传统的人群计数算法首先提取图像中行人特征，然后通过检测或回归的方法确定人群数量。其中，基于检测的方法[1]是通过检测图像中人的头部信息计算出人数，这种方法随着目标分布密集程度的增加，检测难度也随之增加。而基于回归的方法[2-3]则是利用图像的纹理特征与人数的映射关系估计出人群数量，能够减小人群拥挤情况下的估计误差。然而由于回归方法无法从图像中提取有助于完成人群计数任务的语义信息，导致当存在严重的遮挡问题和背景干扰时，算法性能难以提升。随着深度学习的发展，卷积神经网络（Convolution Neural Network，CNN）在许多计算机视觉任务中都取得了良好的效果[4-5]，在一定程度上突破了传统网络的限制，显著提升了计数性能。但图像存在的透视畸变问题和背景噪声不仅影响了网络的计算精度，而且约束了网络的泛化能力和鲁棒性。

为减少图像中多尺度问题对计数任务的影响，ZHANG等[6]利用端到端的深度卷积神经网络实现不同场景的人群计数，此网络虽然提高了计数的精确性，但网络的跨场景计数能力较差。文献[7]提出多列卷积神经网络（Multi-Column CNN，MCNN），通过不同尺度的卷积核提取不同尺度大小的头部特征，提高网络跨场景计数的能力，但该网络的结构复杂，泛化能力有待提高。SAM等[8]提出切换卷积神经网络（Switching CNN，SwitchCNN），该网络虽然提高了泛化能力，但冗余结构较多，计算量较大。为解决MCNN这一类多列结构网络存在的网络冗余问题，LI等[9]提出基于扩展卷积的密集场景识别网络（Congested Scene Recognition Network，CSRNet），通过去除冗余的多列卷积层，采用7层空洞卷积构成一个单通道端到端的人群计数网络，在减少网络冗余的情况下提高了模型对透视畸变的抗干扰能力。与CSRNet类似，LIU等[10]设计了一个上下文感知网络（Context-Aware Network，CAN）以自适应预测人群密度所需的上下文信息，进而减小图像多尺度问题造成的干扰。值得注意的是，虽然CSRNet和CAN均减少了此问题造成的影响，但图像本身存在的背景噪声对人群计数的精度仍然造成影响，并未得到改善。

为抑制图像背景对人群计数网络性能的干扰，文献[11]提出注意力尺度网络（Attention Scaling Network，ASNet），并利用二进制掩码对背景和人群进行分割，但该网络并不能实现端到端的计算，网络计算量较大。ZHU等[12]通过对称双路径多尺度融合网络（Dual Path Multi-scale Fusion Network，SFANet），利用空间注意力图过滤密度图的噪声以提高计数准确度，但其生成的空间注意力图精度有待提高，无法对噪声进行充分抑制，网络的泛化能力和鲁棒性有待增强。

本文对SFANet进行改进，提出基于背景抑制与上下文感知的人群计数网络。通过构建背景噪声抑制（Background Noise Suppression，BNS）模块提高网络对图像背景噪声的抑制能力，并设计上下文感知增强网络结构，减少图像透视畸变对计数任务的干扰。

1　本文网络

1.1　网络结构

本文网络由VGG-16网络、密度图生成（Density Map Generation，DMG）模块、BNS模块以及上下文感知增强网络（Weight Enhancement-Context Aware Network，WE-CAN）4个部分组成，具体结构如图1所示。

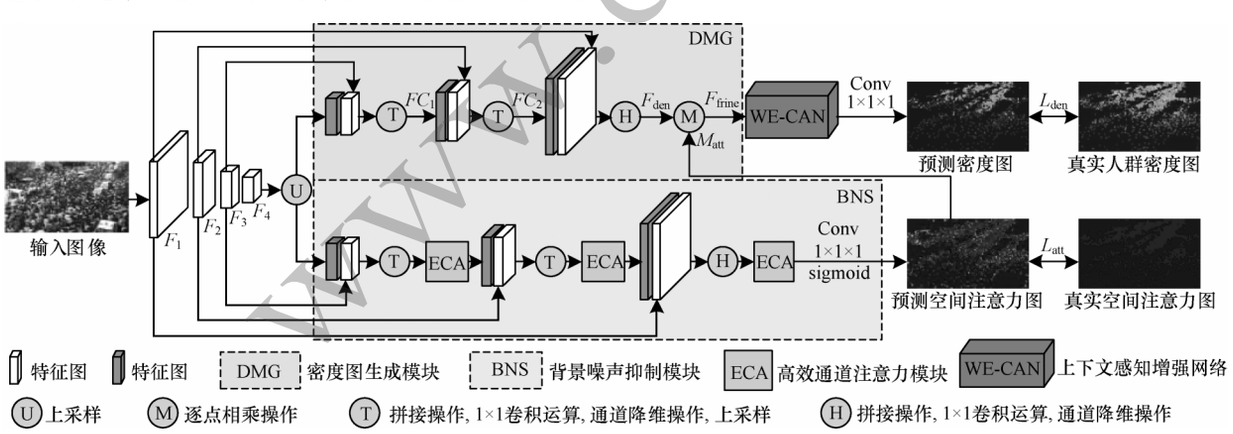


图1　本文网络结构

Fig.1　Structure of network in this paper

由图1可知，该网络首先利用特征提取网络VGG-16提取输入图像不同层次的特征并分别输入DMG模块和BNS模块中进行处理，生成密度特征图和空间注意力图。然后，将生成的密度特征图和空间注意力图相乘，得到能有效抑制噪声的初级密度图。最后，利用WE-CAN模块提取初级密度图上下文信息并进行优化处理，减少透视畸变引起的问题，进而获得高质量的预测密度图。

计算机工程 www.ecice06.com

1.2 密度图生成模块

DMG模块的作用是生成高分辨率的初级密度图，其结构如图1中DMG模块所示。由于低层特征包含更多的纹理信息，高层特征包含头部等位置信息[13]，因此DMG模块采用特征金字塔的结构[14]能使输入的低层特征图和高层特征图的信息有效互补，从而得到高分辨率的初级密度图。

在DMG模块中，首先提取VGG-16网络中4个池化层的中间特征图，分别记为$\boldsymbol{F}_1$、$\boldsymbol{F}_2$、$\boldsymbol{F}_3$、$\boldsymbol{F}_4$；然后依次对$\boldsymbol{F}_1$、$\boldsymbol{F}_2$、$\boldsymbol{F}_3$、$\boldsymbol{F}_4$进行特征融合。DMG模块首先利用特征融合策略对$\boldsymbol{F}_4$进行两倍上采样，并与$\boldsymbol{F}_3$级联，利用T操作处理得到通道数为256的中间特征图$\boldsymbol{FC}_1$；然后，使用T操作将$\boldsymbol{FC}_1$和$\boldsymbol{F}_2$进行特征融合并进行降维操作，生成通道数为128的特征图$\boldsymbol{FC}_2$；最后，利用H操作对$\boldsymbol{FC}_2$与$\boldsymbol{F}_1$级联后的特征进行融合，在经过特征通道降维后得到通道数为32的高分辨率的密度特征图$\boldsymbol{F}_{\mathrm{den}}$，其大小为原始输入的1/2。

生成高分辨率的密度特征图$\boldsymbol{F}_{\mathrm{den}}$后，为了减少图像背景对计数结果的干扰，将$\boldsymbol{F}_{\mathrm{den}}$与BNS模块生成的空间注意力图$\boldsymbol{M}_{\mathrm{att}}$进行相乘操作，实现对噪声的抑制，最后生成初级密度图$\boldsymbol{F}_{\mathrm{refine}}$。此过程的表达式如式(1)所示：

$$\boldsymbol{F}_{\mathrm{refine}}=\boldsymbol{F}_{\mathrm{den}}\otimes\boldsymbol{M}_{\mathrm{att}} \tag{1}$$

其中：$\otimes$表示逐点相乘操作。

1.3 背景噪声抑制模块

BNS模块旨在生成对图像背景噪声起抑制作用的空间注意力图$\boldsymbol{M}_{\mathrm{att}}$，其结构如图1中BNS模块所示。由式(1)可知，空间注意力图$\boldsymbol{M}_{\mathrm{att}}$对噪声的抑制能力越强，其与$\boldsymbol{F}_{\mathrm{den}}$相乘后生成初级密度图$\boldsymbol{F}_{\mathrm{refine}}$的精确度越高。因此，为增强空间注意力图抑制噪声的能力，本文设计了基于通道域注意力ECA模块[15]和多层次信息融合的BNS模块。由于低层特征包含细节边缘的信息，且高层特征图具有区分头部区域与噪声的有效语义信息，因此BNS模块使用ECA模块充分提取不同层次的有效信息，提高头部特征信息表达能力，从而生成高质量的空间注意力图。通道域注意力模块ECA的结构如图2所示，其中，K为卷积核，GAP为全局平均池化。

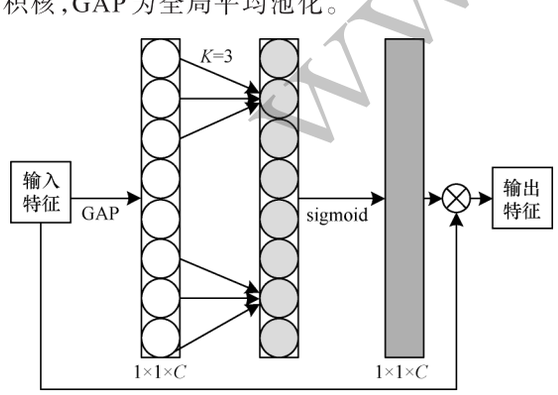


图2 ECA模块结构

Fig.2 Structure of ECA module

由图2可知，ECA模块的引入不仅能增强低层特征图中细节边缘信息的表达，而且能提高高层特征图中用于区分头部和背景的语义信息权重。具体而言，ECA模块首先利用卷积核大小为K的一维卷积获取特征图中每个通道以及K个相邻通道之间的交互信息，从而计算出不同通道的权重，此过程用$\mathrm{C1D}_K(\cdot)$表示。然后，将计算得到的权重与原始特征图相乘，以增强不同特征中有效信息的表达[15]，此过程如式(2)所示：

$$\boldsymbol{y}=\mathrm{sigmoid}\left(\mathrm{C1D}_K\left(\mathrm{GAP}(\boldsymbol{x})\right)\right)\otimes\boldsymbol{x} \tag{2}$$

其中：$\boldsymbol{x}$表示输入特征；$\boldsymbol{y}$表示输出结果；sigmoid表示sigmoid激活函数；在网络中K值设为3。

为进一步突出空间注意力图中前景和后景差异[12]，BNS模块在输出空间注意力图前使用1×1×1卷积对特征图进行降维，并使用sigmoid函数将空间注意力图像素点的值映射为(0,1)，从而区分图像中头部区域与背景区域，此过程的表达式如式(3)所示：

$$\boldsymbol{M}_{\mathrm{att}}=\mathrm{sigmoid}\left(\boldsymbol{W}*\boldsymbol{f}_{\mathrm{att}}+\boldsymbol{b}\right) \tag{3}$$

其中：$\boldsymbol{M}_{\mathrm{att}}$表示最终生成的空间注意力图；$\boldsymbol{f}_{\mathrm{att}}$表示经过3个通道域注意力模块优化后的中间特征图；*表示卷积操作；$\boldsymbol{W}$和$\boldsymbol{b}$分别表示一维卷积的权重和偏置。

1.4 上下文感知增强网络

为增强特征图中上下文信息的表达，减少图像透视畸变造成的影响，本文在CAN[10]模块的基础上结合特征权重增强模块(Feature Weight Enhance Module，WE)进行改进，提出WE-CAN模块。其中WE模块用于增强特征图中多尺度上下文信息的表达，其结构如图3所示。由图3可知，WE模块的功能是提取DMG模块所生成初级密度图$\boldsymbol{F}_{\mathrm{refine}}$中的多尺度特征并增强其中有效信息的表达。WE模块首先对$\boldsymbol{F}_{\mathrm{refine}}$进行平均池化[16-17]，并使用一个1×1卷积和2个3×3卷积以不同的感受野提取$\boldsymbol{F}_{\mathrm{refine}}$的多尺度信息。为有效提取不同尺度的头部特征，本文首先根据文献[16]中提出的空洞空间卷积池化金字塔中对卷积核的设置，将1×1卷积的空洞率设为1，2个3×3卷积的空洞率分别设为6和12。然后，对提取到的头部特征与池化后的$\boldsymbol{F}_{\mathrm{refine}}$进行级联后降维输出，记为$\boldsymbol{F}_A$。其次，利用通道域注意力模块(Squeeze and Excitation Module，SE)[18]强化$\boldsymbol{F}_{\mathrm{refine}}$中有效的多尺度信息表达，生成具有显著信息的特征，记为$\boldsymbol{F}_S$。最后，对$\boldsymbol{F}_A$和$\boldsymbol{F}_S$进行相乘操作，生成多尺度特征密度图$\boldsymbol{F}_m$，该过程可以表述为：

$$\boldsymbol{F}_m=\boldsymbol{W}*\left(\boldsymbol{F}_A\otimes\boldsymbol{F}_S\right)+\boldsymbol{b} \tag{4}$$

其中：$\boldsymbol{F}_m$表示WE模块生成的多尺度特征密度图。

计算机工程 www.ecice06.com

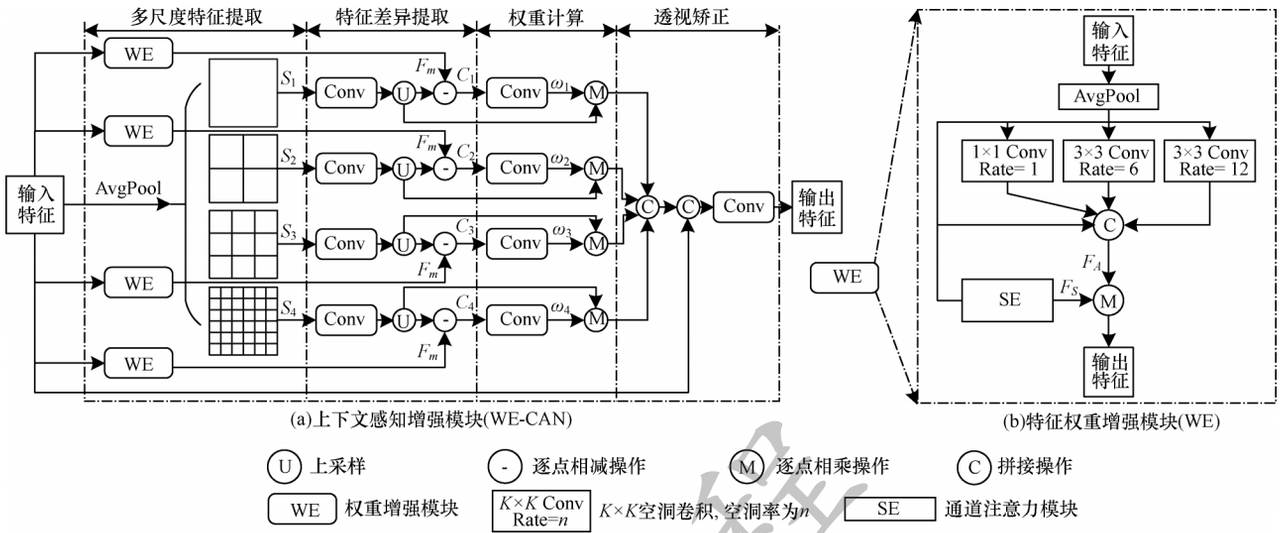


图3　WE-CAN模块结构

Fig.3　Structure of WE-CAN module

结合WE模块对CAN[10]模块进行改进，本文提出的WE-CAN模块能提取特征中更详细的空间信息和全局信息，增强网络自适应优化多级上下文信息的能力，其结构如图3(a)所示。由图3(a)可知，WE-CAN模块分为4个步骤：首先使用4个不同核大小的平均池化层对初级密度图$\boldsymbol{F}_{\text{refine}}$进行多尺度特征提取，生成4个不同尺寸的感受野，以感知上下文特征$\boldsymbol{S}_j$，比例分别为1、2、3、6[10]；其次，将$\boldsymbol{S}_j$与WE模块生成的$\boldsymbol{F}_m$相减，从而提取出目标特征与相邻特征的特征差异$\boldsymbol{C}_j$，实现WE-CAN模块中的特征差异提取，其定量表述为式(5)；然后，WE-CAN模块将计算出的$\boldsymbol{C}_j$输入到权重计算网络中，利用一维卷积计算不同尺度在输入特征图中的尺度权重$\boldsymbol{\omega}_j$，此过程可表述为式(6)；最后，利用尺度权重$\boldsymbol{\omega}_j$对初级密度图$\boldsymbol{F}_{\text{refine}}$进行透视矫正并降维输出，使网络生成高质量的预测密度图$\boldsymbol{D}^{\text{pre}}$，该过程可表述为式(7)：

$$\boldsymbol{C}_j=\boldsymbol{S}_j-\boldsymbol{F}_m \tag{5}$$

$$\boldsymbol{\omega}_j=\text{sigmoid}\left(\boldsymbol{W}*\boldsymbol{C}_j+\boldsymbol{b}\right) \tag{6}$$

$$\boldsymbol{D}^{\text{pre}}=\boldsymbol{W}*\left(\left[\boldsymbol{F}_{\text{refine}}\left|\frac{\sum_{j=1}^{4}\boldsymbol{\omega}_j\otimes\boldsymbol{S}_j}{\sum_{j=1}^{4}\boldsymbol{\omega}_j}\right.\right]\right)+\boldsymbol{b} \tag{7}$$

其中：$[\bullet|\bullet]$表示级联操作。

2　实验与结果分析

2.1　数据集

ShanghaiTech、UCF-CC-50以及UCF-QNRF数据集是目前人群计数研究中的主流数据集详情如下：

1)ShanghaiTech数据集是由ZHANG[8]等提出的大型开源数据集，该数据集由Part A和Part B 2个部分组成。Part A包含从互联网中随机收集的482张图片，其中300张用于训练，182张用于测试。Part B则包含716张拍摄于上海繁华街道的图片，其中400张用于训练，316张用于测试。

2)UCF-CC-50数据集[19]是一个包含拥挤人群图片的开源数据集，一共包括50张不同分辨率的人群图片。该数据集图片人数变化大，人群范围跨度大，从94人到4 543人，平均每张图片1 280人。为更好地验证模型的准确性，在实验过程中使用与文献[20]相同的5折交叉验证法：首先在不重复的前提下将数据集平均划分为5份，每份10张图像。然后每次选择不同的一份作为测试集，将余下的4份作为训练集进行训练，分别获得5组测试结果，最后取5组测试结果的平均值作为模型的性能指标，并与其他网络进行比较。

3)UCF-QNRF数据集[20]是一个复杂且拥挤的人群数据集，共包含1 251 642处人体标注。该数据集共有1 535张图像，训练集由1 201张图像组成，其余334张作为测试集。UCF-QNRF数据集不仅包含多种多样的环境背景，而且数据集中图像的分辨率差异较大，具有一定的挑战性。

本文采用的3个数据集对比结果如表1所示。

表1　本文数据集信息

Table 1　Information of datasets in this paper

数据集	分辨率/像素	图片数量	最多人数	最少人数	平均人数
ShanghaiTech Part A	不统一	482	3 139	33	501
ShanghaiTech Part B	768×1 024	716	578	9	123
UCF_CC_50	不统一	50	4 543	94	1 279
UCF_QNRF	不统一	1 535	12 865	49	815

2.2　数据处理

2.2.1　真实人群密度图生成

由于现有的人群计数数据集只标注了图像中人头的位置，因此需要将人头位置转化为真实人群密度图以提供更多的监督信息。本文按照文献[8]中生成密度图的方法，对所有的目标均采用内核大小

计算机工程 www.ecr.com

相同的高斯滤波器进行处理。首先使用$\delta(x-x_i)$表示图像在像素点x_i上的头部标注，然后利用标准差为σ_i的高斯核滤波器$\boldsymbol{G}_{\sigma_i}$对每个头部标注进行高斯模糊处理，进而得到相应的真实人群密度图$\boldsymbol{D}^{\mathrm{GT}}$，该过程可定量描述为：

$$\sigma_i=\beta\overline{d}_i \tag{8}$$

$$\boldsymbol{D}^{\mathrm{GT}}(x)=\sum_{i=1}^{C}\delta(x-x_i)G_{\sigma_i}(x) \tag{9}$$

其中：标准差σ_i由K最近邻算法计算得出；$\overline{d}_i$表示x_i和k个相邻点之间的距离，k设为3；β为超参数，根据文献[7]将其设为0.3。

2.2.2 真实空间注意力图生成

为使网络生成预测空间注意力图，实现对预测密度图的权重分配，本文根据文献[12]设定阈值的方式对真实人群密度图$\boldsymbol{D}^{\mathrm{GT}}$进行二值化处理，将人群区域和背景区域进行分割，从而获得真实空间注意力图$\boldsymbol{A}^{\mathrm{GT}}$，该过程可定量描述为：

$$\forall x\in\boldsymbol{D}^{\mathrm{GT}},\boldsymbol{A}_i^{\mathrm{GT}}(x)=\begin{cases}0,x<t\\1,x\geqslant t\end{cases} \tag{10}$$

其中：x表示$\boldsymbol{D}^{\mathrm{GT}}$中不同位置像素点的值；$t$表示阈值，在本文实验中将其设为0.001。

2.3 训练方法

本文基于Pytorch框架，在Ubuntu18.04系统和RTX 2080Ti GPU条件下进行实验。为提高训练速度和计数精度，本文导入VGG-16预训练模型并将其作为前端特征提取器，其余网络参数初始值由均值为0、标准差为0.01的高斯分布随机生成。同时，采用Adam优化器训练，学习率设为1×10^{-4}，权重衰减设为5×10^{-3}，迭代次数设为800次，批量化大小设为8。

2.4 多任务联合损失函数

为获得精确的人群估计结果，本文根据文献[12]的方法，将多任务联合损失函数分别用于训练网络生成有效的预测空间注意力图和精准的预测密度图。此外，基于现有的研究，本文利用欧氏几何距离测量预测人群密度图和真实人群密度图之间的差值并将其作为损失函数，以用于训练网络模型生成预测密度图。损失函数的表达式如式(11)所示：

$$L_{\mathrm{den}}=\frac{1}{N}\left\|F(X_i;\Theta)-\boldsymbol{D}_i^{\mathrm{GT}}\right\|_2^2 \tag{11}$$

其中：$\boldsymbol{F}(\boldsymbol{X}_i;\boldsymbol{\Theta})$表示预测的人群密度图；$\boldsymbol{\Theta}$表示网络学习参量；$x_i$和$\boldsymbol{D}^{\mathrm{GT}}$分别表示输入图像和真实人群密度图。

为得到有效的空间注意力图，本文在训练过程中采用二进制交叉熵损失(Binary Cross Entropy Loss，BCELoss)作为损失函数，计算人群空间注意力图与真实人群空间注意力图之间的差值。损失函数的表达式如式(12)所示：

$$L_{\mathrm{att}}=-\frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{A}_i^{\mathrm{GT}}\log_a(P_i)+(1-\boldsymbol{A}_i^{\mathrm{GT}})\log_a(1-P_i)\right\|_1 \tag{12}$$

其中：$\boldsymbol{A}_i^{\mathrm{GT}}$表示真实人群空间注意力图；$P_i$是经过sigmoid函数处理后预测空间注意力图中每个像素的概率值。

在训练过程中使用多任务结合损失函数实现端到端的训练，多任务联合损失函数表达式如式(13)所示：

$$L=L_{\mathrm{den}}+\alpha L_{\mathrm{att}} \tag{13}$$

其中：α为超参数，设其为0.1。

2.5 评价指标

现有的研究主要使用平均绝对误差(Mean Absolute Error，MAE)和均方误差(Mean Square Error，MSE)对模型进行评估，MAE反映了网络的准确性，而MSE体现了网络的泛化能力。其计算过程分别如式(14)和式(15)所示：

$$R_{\mathrm{MAE}}=\frac{1}{N}\sum_{i=1}^{N}\left|C_i-C_i^{\mathrm{GT}}\right| \tag{14}$$

$$R_{\mathrm{MSE}}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_i-C_i^{\mathrm{GT}})^2} \tag{15}$$

其中：N表示测试集中测试图像的数量；C_i和C_i^{GT}分别表示网络预测的人群总数和真实人群密度图中的人群总数。

2.6 结果分析

为验证本文网络的有效性，在人群计数任务数据集ShanghaiTech、UCF-CC-50以及UCF-QNRF数据集上分别进行实验，并与现有的其他网络包括MCNN[7]，SwitchCNN[8]，CSRNet[9]，多尺度聚合网络(Multi-Scale Aggregation Network，SANet)[21]，CAN[10]，关系注意力神经网络(Relational Attention Network，RANet)[22]，SFANet[12]，泛密度神经网络(Pan-Density Neural Network，PaDNet)[23]，密度感知卷积神经网络(Density-Aware CNN，DensityCNN)[24]以及多尺度感知人群计数神经网络(Scale-Aware Crowd Counting Network，SACCN)[25]进行对比分析，实验结果如表2所示。

表2 不同网络在公开数据集上的MAE与MSE结果比较

Table 2 Comparison on MAE and MSE of different networks on public datasets

网络	Shanghai-Tech Part A		Shanghai-Tech Part B		UCF-CC-50		UCF-QNRF	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN[7]	110.2	173.2	26.4	41.3	377.6	509.1	277.0	422.0
SwitchCNN[8]	90.4	135.0	21.6	33.4	318.1	439.2	228.0	445.0
CSRNet[9]	68.2	115.0	10.6	16.0	266.1	397.5	—	—
CAN[10]	62.3	100.0	7.8	12.2	212.2	243.7	107.0	183.0
SFANet[12]	59.8	99.3	6.9	10.9	219.6	316.2	100.8	174.5
SANet[21]	67.0	104.5	8.4	13.6	—	—	—	—
RANet[22]	59.4	102.0	7.9	12.9	239.8	319.4	111.0	190.0
PaDNet[23]	59.2	98.1	8.1	12.2	185.8	278.3	96.5	170.2
DensityCNN[24]	63.1	106.3	9.1	16.3	244.6	341.8	101.8	186.5
SACCN[25]	59.2	98.0	6.8	10.5	178.0	258.5	96.1	167.8
本文网络	56.6	97.6	6.3	10.2	160.6	224.7	85.8	146.0

www.ecice06.com 计算机工程

由表2可知，对于ShanghaiTech数据集，本文网络模型在密集程度较高的Part A子集上的MAE和MSE分别为56.6和97.6，均达到了先进水平。同时在人群密度较低的Part B子集中，本文网络也取得了所有对比网络中最佳的MAE和MSE，分别为6.3和10.2，相比SFANet分别下降了8.7%和6.4%。在样本量较少，但视角丰富的UCF-CC-50数据集中，本文网络取得了对比网络中最低的MAE和MSE，分别为160.6和224.7，与SFANet相比分别下降了26.9%和28.9%。在场景丰富、图像视角多样的UCF-QNRF数据集中，本文网络的MAE和MSE分别为85.8和146.0，取得了对比网络中的最佳结果，与SFANet相比MAE下降了14.9%，MSE下降了16.3%。

表2的实验数据说明本文网络在拥挤人群和稀疏人群场景中均具有较高的准确性，而且在图像畸变较为严重、背景环境复杂多样的拥挤人群场景中具有良好性能，体现了该网络良好的泛化能力和较强的鲁棒性。

2.7　消融性实验

为验证BNS模块和WE-CAN模块的有效性，在ShanghaiTech数据集上进行实验，分析不同模块对网络性能的影响。

消融实验结果如表3所示。首先以SFANet作为基线网络，并在其基础上增加BNS模块进行实验，从而验证BNS模块的有效性。此外，为验证WE-CAN模块的有效性，在基线网络添加了BNS模块的基础上分别添加CAN模块和WE-CAN模块进行对比实验。由表3可知，SFANet增加BNS模块后在Part A数据集上的MAE下降1.5%，在Part B数据集的MAE下降2.9%，MSE下降4.6%。当SFANet依次增加BNS模块和CAN模块后，Part A数据集的MAE相比SFANet下降3.5%，Part B数据集的MAE下降4.3%，MSE下降4.6%。当网络增加了WE-CAN模块后，在Part A数据集的MAE下降5.3%，Part B数据集的MAE和MSE分别下降了8.7%和6.4%。上述消融性实验验证了本文设计的BNS模块和WE-CAN模块在人群计数任务中的有效性和合理性。

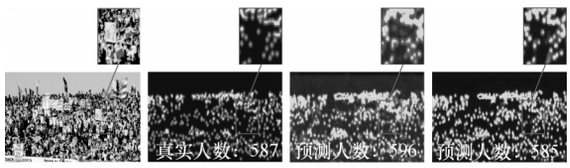
表3　消融实验结果

Table 3　Results of ablation experiment

网络	ShanghaiTech Part A		ShanghaiTech Part B	
	MAE	MSE	MAE	MSE
SFANet网络	59.8	99.3	6.9	10.9
SFANet+BNS网络	58.9	100.3	6.7	10.4
SFANet+BNS+CAN网络	57.7	100.4	6.6	10.4
SFANet+BNS+WE-CAN网络	56.6	97.6	6.3	10.2

2.8　结果可视化

将本文网络和SFANet生成的密度图进行对比，结果如图4所示（彩色效果见《计算机工程》官网HTML版）。



(a)原始图像　(b)真实密度图　(c)SFANet网络　(d)本文网络

图4　本文网络与SFANet的预测密度图比较

Fig.4　Comparison of predict density map between SFANet and network in this paper

由图4可知，当拥挤人群存在旗帜等遮挡物的时候，SFANet无法很好区分人群和背景区域，容易将旗帜上的图案判定为人体特征，导致最终预测结果误差较大。而本文网络能有效抑制噪声，使生成的预测密度图能准确突出人群区域，提高人群计数任务的计算精度。

此外，本文网络能生成较精确的预测空间注意力图和预测密度图，且能进行可视化，可视化结果如图5所示（彩色效果见《计算机工程》官网HTML版）。



(a)原始图像

(b)真实空间注意力图

(c)预测空间注意力图

(d)真实密度图

(e)预测密度图

图5　本文网络的实验效果展示

Fig.5　Experimental effect display of network in this paper

由图5左数第1列、第2列图可知，针对存在背景干扰较严重的拥挤人群场景的图像，本文网络能将图像中存在的汽车、树木等遮挡物识别为背景元素，减少遮挡物对预测结果的干扰。由图5左数第3列图可知，针对目标尺度变化较大的图像，本文网络生成的预测空间注意力图和预测密度图均能较好地对图像中的多尺度目标进行检测。由图5左数第4列图可知，针对稀疏人群场景的图像，本文网络也能生成精确的预测空间注意力图和预测密度图，体现了该网络良好的跨场景计数能力。

综上所述，本文网络能生成高质量的预测密度图，在有效抑制图像噪声干扰的同时，也能有效减少透视畸变引起的估计误差，展示了该网络在人群计数领域中优越的性能。

计算机工程 www.ecsec.com

3 结束语

本文提出基于背景抑制与上下文感知的人群计数网络，通过引入DMG模块有效提取不同层次特征的有效信息，生成高分辨率的特征图，并利用BNS模块充分抑制图像噪声干扰，提高模型的人群计数精度。此外，采用WE-CAN模块解决图像存在的透视畸变问题，在对网络模型进行训练时采用多任务的联合损失函数，以提高模型的计数精度。实验结果表明，本文网络在UCF-QNRF数据集上的平均绝对误差和均方误差分别为85.8、146.0，相较于MCNN、SwitchCNN、CSRNet等网络最高分别下降69.0%和67.2%，具有良好的泛化能力和较强的鲁棒性。下一步将引入雾霾、暴雨等复杂天气变化的数据样本和存在低光照、过曝等光照变化的数据样本对模型进行训练，提高网络在复杂环境下的计算精度。

参考文献

[1] IDREES H, SOOMRO K, SHAH M. Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(10): 1986-1998.

[2] OJALA T, PIETIKAINEN M, MAENPAA T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987.

[3] CHAN A B, VASCONCELOS N. Bayesian poisson regression for crowd counting[C]//Proceedings of the 12th International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2009: 545-551.

[4] 冯兴杰，张乐，曾云泽. 基于多注意力CNN的问题相似度计算模型[J]. 计算机工程，2019，45(9)：284-290.
FENG X J, ZHANG L, ZENG Y Z. Question similarity calculation model based on multi-attention CNN[J]. Computer Engineering, 2019, 45(9): 284-290. (in Chinese)

[5] CHOLLET F. Xception: deep learning with depthwise separable convolutions[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 1800-1807.

[6] ZHANG C, LI H S, WANG X G, et al. Cross-scene crowd counting via deep convolutional neural networks[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 833-841.

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

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

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

[10] 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.

[11] JIANG X H, ZHANG L, XU M L, et al. Attention scaling for crowd counting[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 4705-4714.

[12] ZHU L, ZHAO Z J, LU C, et al. Dual path multi-scale fusion networks with attention for crowd counting[EB/OL]. [2021-09-06]. https://arxiv.org/abs/1902.01115.

[13] 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)

[14] LIN T Y, DOLLÁR 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.

[15] WANG Q L, WU B G, ZHU P F, et al. ECA-net: efficient channel attention for deep convolutional neural networks[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 11531-11539.

[16] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[17] 翟强，王陆洋，殷保群，等. 基于尺度自适应卷积神经网络的人群计数算法[J]. 计算机工程，2020，46(2)：250-254，261.
ZHAI Q, WANG L Y, YIN B Q, et al. Crowd counting algorithm based on scale adaptive convolutional neural network[J]. Computer Engineering, 2020, 46(2): 250-254, 261. (in Chinese)

[18] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 7132-7141.

[19] IDREES H, SALEEMI I, SEIBERT C, et al. Multi-source multi-scale counting in extremely dense crowd images[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2013: 2547-2554.

[20] 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: 544-559.

[21] CAO X K, WANG Z P, ZHAO Y Y, et al. Scale aggregation network for accurate and efficient crowd counting[C]//Proceedings of 2018 European Conference on Computer Vision. Berlin, Germany: Springer, 2018: 757-773.

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

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

[24] JIANG X H, ZHANG L, ZHANG T Z, et al. Density-aware multi-task learning for crowd counting[J]. IEEE Transactions on Multimedia, 2021, 23: 443-453.

[25] YI Q S, LIU Y X, JIANG A W, et al. Scale-aware network with regional and semantic attentions for crowd counting under cluttered background[EB/OL]. [2021-08-01]. https://arxiv.org/abs/2101.01479.

编辑 赖玉玲

计算机工程 www.ecice06.com