计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(4):143−151 [doi: 10.15888/j.cnki.csa.009474]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多尺度特征融合的人群密度检测①

余梦飞, 杨海波, 卢　鑫, 贾军营

(沈阳工业大学 信息科学与工程学院, 沈阳 110870)
通信作者: 杨海波, E-mail: yanghb@sut.edu.cn

摘　要: 基于深度学习的人群密度检测算法取得了巨大进步, 但该算法在实际复杂场景中的检测准确性和鲁棒性还有很大的提升空间. 复杂场景下目标尺度不一致和背景信息干扰等因素使得人群密度检测成为一项具有挑战性的任务. 针对该问题, 提出了一种基于多尺度特征融合的人群密度检测网络. 该网络首先利用不同分辨率图像并行交互提取人群粗细粒度特征, 并引入多层次特征融合机制, 以充分利用多层尺度信息. 其次采用空间和通道注意力机制突出人群特征权重, 聚焦感兴趣的人群, 降低背景信息干扰, 生成高质量密度图. 实验结果表明, 在多个典型的公共数据集上与具有代表性的人群密度检测方法相比, 多尺度特征融合的人群密度检测网络具有良好的准确性和鲁棒性.

关键词: 人群计数; 特征融合; 多尺度; 注意力; 人群密度

引用格式: 余梦飞,杨海波,卢鑫,贾军营.基于多尺度特征融合的人群密度检测.计算机系统应用,2024,33(4):143–151. http://www.c-s-a.org.cn/1003-3254/9474.html

Crowd Density Detection Based on Multi-scale Feature Fusion

YU Meng-Fei, YANG Hai-Bo, LU Xin, JIA Jun-Ying

(School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China)

Abstract: The crowd density detection algorithm based on deep learning has made great progress, while there is still a lot of room for improvement in the detection accuracy and robustness of the algorithm in actual complex scenes. Factors such as inconsistent object scales and background information interference in complex scenes make crowd density detection a challenging task. Aiming at this problem, this study proposes a crowd density detection network based on multi-scale feature fusion. The network first uses images of different resolutions to interactively extract coarse and fine-grained features of the crowd and introduces a multi-level feature fusion mechanism to make full use of multi-level scale information. Secondly, the study utilizes the spatial and channel attention mechanism to highlight the weight of crowd characteristics, focus on interested crowds, reduce background information interference, and generate high-quality density maps. Experimental results show that the crowd density detection network with multi-scale feature fusion has better accuracy and robustness than representative crowd density detection methods on multiple typical public datasets.

Key words: population count; feature fusion; multi-scale; attention; population density

人群密度检测技术主要目的是实时统计并掌握某一场景下的人群密度信息. 现已在公共安全、城市规划、安全监测等领域得到广泛应用[1]. 如在旅游景点、大型广场等人群密集场景下, 通过时刻统计人群密度避免踩踏事件的发生. 在商场、大型超市等场景下, 利用人群密度信息进行店铺或商品的合理排布.

① 基金项目: 辽宁省自然科学基金 (1645773678079)
收稿时间: 2023-10-07; 修改时间: 2023-11-09, 2023-12-04; 采用时间: 2023-12-15; csa 在线出版时间: 2024-03-01
CNKI 网络首发时间: 2024-03-07

目前, 采集数据时拍摄角度不一致、距离差异以及场景信息复杂等因素造成了人群目标尺度不一致和背景干扰的问题[2], 该问题对人群密度检测结果造成严重的影响. 针对这些问题, 大多数研究者采用增加网络深度和广度的方法提升网络特征提取能力, 但是这类方法准确率还有较大提升空间[3].

为了有效解决上述问题, 本文提出了一种基于多尺度特征融合的人群密度检测网络 (crowd density detection network based on multiscale feature fusion, CDBM). CDBM 使用多分辨率分支结构利用目标图像多分辨率特征捕获不同尺度特征信息和不同粒度信息, 丰富了目标人群的特征信息; 通过多阵列空洞卷积提升检测网络对目标人群尺度不一致的适应能力. 另一方面为了凸显目标人群区域, 降低背景信息对人群密度检测网络的影响, 在多尺度特征提取中融合双注意力机制应对目标人群多尺度特征变化和复杂背景信息干扰问题. 在多个公共数据集上的实验表明, 该网络具有较优的密度检测性能. 本文的创新性如下.

(1) 提出了一个多尺度特征融合的人群密度检测网络 (CDBM). 该网络采用并行交互结构提取多层次人群目标信息, 从有限的网络中获取更丰富的尺度信息, 提升网络的全局尺度特征的处理能力.

(2) 设计了一个多分辨率特征提取模块 (multi resolution feature extraction module, MRFM) 利用交互融合结构提取丰富的多尺度信息, 提升了网络对多尺度信息的感知能力.

(3) 设计了一个多尺度注意力模块 (multi scale attention module, MSAM) 以提取多层次的感兴趣人群特征信息. 该模块利用注意力机制和多列空洞卷积提升了对人群特征信息的敏感度和提取多层次特征的能力.

1 相关工作

随着计算机视觉领域的不断发展, 人们对人群密度检测也越来越关注, 针对不同的场景等问题, 研究者们提出了许多人群密度检测方法. 现有的人群密度检测算法根据特征提取方式的不同可分为基于传统的方法和基于深度学习的方法.

1.1 基于传统的方法

传统的关于人群密度检测的方法基本上都是采用先进行特征提取, 然后进行特征处理预测人群数目. 这些传统的人群密度检测的方法主要分为 3 种, 基于聚类的方法、基于检测的方法和基于回归的方法[4,5]. 其中，基于检测的方法主要是对单个行人整体和头部局部特征提取进行训练实现密度检测任务[6], 但仅适用于非常稀疏的场景. 基于回归的方法主要是通过学习图像特征到人群密度或密度图之间的映射关系获取人群密度信息[7], 但在密集场景下, 人群密度检测效果还有较大提升空间. 基于传统的人群密度检测方法在特征提取上具有明显的局限性, 无法获取丰富的目标人群特征信息, 对于密集场景下, 人群密度检测不能满足实际需求.

1.2 基于深度学习的方法

Li 等人[8]提出单通道检测人群密度网络 (CSRNet), 通过使用空洞卷积神经网络, 在保持分辨率的同时扩大感知域. Liu 等人[9]提出 DSSINet, 使用不同尺寸的图片作为输入来得到不同尺度的特征, 又通过条件随机场来融合不同层之间的特征, 通过信息传递的方式对多尺度特征进行细化. Xu 等人[10]提出 AutoScale 通过在网络输入调节图像, 使得网络获取不同尺度的特征图, 丰富了网络特征信息, 获取人群更多细节特征. Zand 等人[11]提出一种多列多尺度的人群计数网络 (MPS) 利用多列不同网络深度进行获取人群的多尺度特征信息. Zhu 等人[12]提出 SFANet, 利用双路径方式将人群区域特征图与多尺度特征进行融合生成高质量密度图, 从而提高检测准确性. Oh 等人[13]提出了 DUBNet, 利用单列可扩展的网络提高了生成人群密度图质量, 采用点估计提升网络的人群计数性能. Cao 等人[14]提出了一种尺度聚合网络 (SANet), 利用不同大小卷积核提取图像特征, 并利用反卷积进行恢复图像分辨率生成高质量密度图. Jiang 等人[15]提出了编码器-解码器网络 (TEDNet), 通过利用多分支的多尺度编码获取图像的多尺度特征信息. Lian 等人[16]提出一种双路径检测网络 (DPDNet), 通过利用双回归头进行图像密度图的生成, 进一步提升了网络对小尺寸特征的检测能力. Wang 等人[17]提出一种分布匹配进行人群计数的网络 (DM-Count), 通过利用最优传输来进行衡量归一化预测密度图和真实密度图的相似性, 来进行提升网络性能.

上述方法在人群密度检测方面具有比较好的效果, 可以减小遮挡、畸变等问题带来的检测误差, 但是目前人群密度检测技术应用在各类场景中, 遇到的主要的问题是复杂场景的干扰和人群尺度不一致, 这些方法在复杂背景干扰下对不同尺度的人群进行密度检测

上还有提升的空间. 因此本文主要是针对人群尺度不一和背景干扰的问题进行研究改进.

2 网络架构

针对当前复杂场景下人群密度检测出现的问题, 本文提出了基于多尺度特征融合的人群密度检测网络. 该网络主要是由多分辨率特征提取模块、多尺度注意力模块和分割注意力模块组成, 如图 1 所示. CDBM 利用多分辨率特征提取模块获取人群目标图像丰富的多层语义信息; 多尺度注意力模块内部使用多个多尺度模块提取不同层次目标特征, 通过注意力机制聚焦目标行人特征区域, 使得提取的目标特征更加精细化, 有效避免背景信息干扰; 提取出的目标特征信息送入分割注意力模块得到最终的高质量密度图.

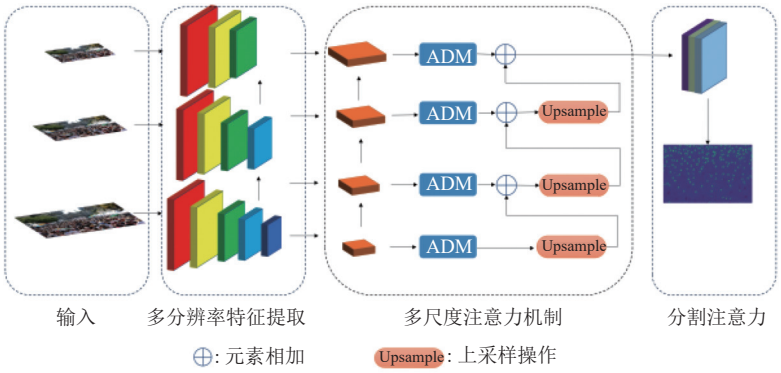


图 1 多尺度特征融合的人群密度检测网络架构

2.1 多分辨率特征提取模块

多分辨率特征提取模块 (MRFM) 采用不同分辨率图像分列提取图像空间和尺度特征[18], 获取更丰富的细节特征应对人群目标尺度变化问题. 以往的多阵列网络各列工作基本是相互独立的, 造成提取的尺度信息具有局限性, 所以本文提出了多分辨率特征提取模块, 用于交互融合提取图像特征, 通过利用高分辨率图像提取精确的局部信息和低分辨率图像学习丰富的上下文信息, 将多分辨率提取的特征信息进行结合, 丰富了人群的特征信息, 而且进一步增强网络的鲁棒性, 当某一分辨率图像受一些问题影响时, 其他分辨率图像仍然包含有用信息, 有利于提高网络在复杂场景下的性能. 多分辨率特征提取模块架构如图 2 所示.

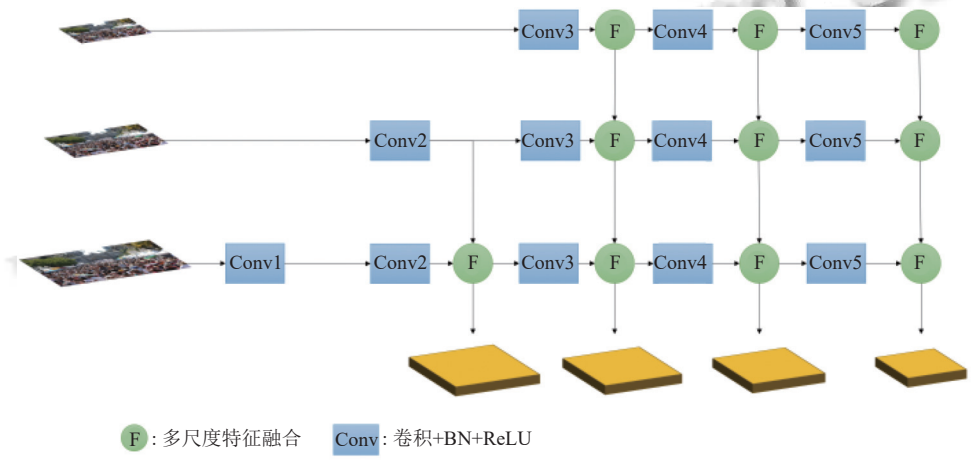


图 2 多分辨率特征提取模块

从图 2 中可知, 由于高分辨率和低分辨率具有不同的行人细节信息和全局内容, 所以 MRFM 通过调整输入图像的分辨率, 分别对不同分辨率图像提取特征信息. 由于不同分辨率图像携带特征信息量的不同, 需要使用不同深度网络进行特征提取, 便于充分提取网络特征信息和减少网络计算量. 经实验得出, MRFM 特征输入分辨率为原图大小的 (2, 1, 0.5) 倍的效果最佳. MRFM 采用 VGG-16 网络前 13 层卷积层作为骨干网

络提取特征信息, 将骨干网络分为 5 个层次分别为 Conv1–Conv5. 首先构建 3 个图像 (A_1, A_2, A_3), 其中 $A_2 \in R^{H\times W}$ 是原始图像, $A_1 \in R^{2H\times 2W}$ 是原始图像利用双线性插值操作获取的高分辨率图像、$A_3 \in R^{1/2H\times 1/2W}$ 是原始图像通过下采样操作获取的低分辨率图像. A_1、A_2、A_3 分别传入对应的子网络中. A_k 在子网络卷积块提取到特征, 本文通过将不同子网络中相同分辨率的特征输入特征融合模块进行特征融合增强交互, 提高尺度变化的鲁棒性, 传入 k–1 个子网络的下一层, 并作为多尺度注意力模块的输入. 最后通过生成的注意力图和特征图融合生成高质量人群密度图, 降低背景干扰和人群特征尺度变化带来的误差. 多分辨率特征提取网络在训练过程中采用同时训练的方式使得所有分支能够同时共享信息, 综合整体网络获取全局特征, 同时训练能够直接优化最终任务损失, 有助于端到端的优化, 有效获取丰富的人群特征信息.

由于多分辨率特征提取模块中每个子网络输出的相同分辨率特征图是由不同感受野中提取出的, 具有一定的互补性. 为了多尺度模块能提取更为丰富的图像信息, 本文设计了多尺度特征融合模块. 如图 3 所示, 第 k 个子网络的输入特征为 F_k, 将 F_k 通过 1×1 卷积进行调节网络通道信息, 再与 F_{k-1} 在通道上进行连接得到特征 F'_{k-1}, 然后 F'_{k-1} 再通过一个 3×3 卷积进行融合得到 S_{k-1} 作为第 k–1 个子网络下一层的输入. 其融合融合过程为:

$$S_{k-1} = \mathrm{Conv}(\mathrm{cat}(\mathrm{Conv}(F_k), F_{k-1})) \tag{1}$$

其中, cat(·)为特征融合操作, 将输入的特征图进行通道连接.

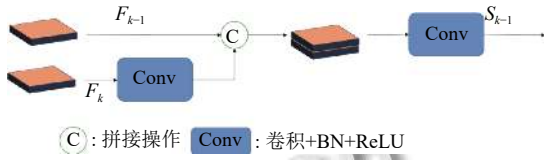


图 3 多尺度特征融合模块

2.2 多尺度注意力模块

由于卷积神经网络的卷积核大小决定了卷积网络的感受野的大小, 不同大小的感受野可以获取图像不同尺度的语义信息, 利用不同大小卷积核可以获取更为丰富的图像特征. 因此本文设计了多尺度注意力模块 (MSAM) 采用多列空洞卷积和注意力机制降低人群特征尺度变化带来的误差以及背景对特征提取带来的干扰. 但是由于多列网络结构各分支都是相互独立的, 而且只能提取特定的尺度信息, 不能有效地应对尺度连续变化的问题. 所以本文通过特征融合将各个子网络分支进行交互融合, 可以有效聚合人群尺度的上下文信息, 提高特征信息的交互能力. 本文还采用了空间通道注意力机制能够充分地抑制背景信息的干扰, 增强对人群特征感兴趣区域的权重. MSAM 网络结构如图 4 所示.

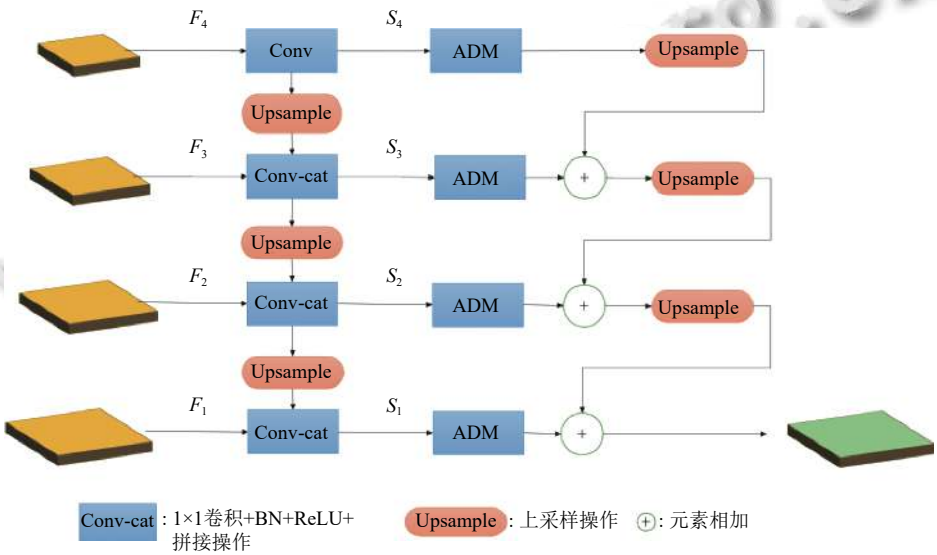


图 4 多尺度注意力模块

由图 4 可知, MSAM 通过将 MRFM 提取出的特征图 (F_1、F_2、F_3、F_4) 作为输入, 为了充分地提取图像特征的多层次语义信息, MSAM 对输入的特征图从上往下 ($F_4 \rightarrow F_3 \rightarrow F_2 \rightarrow F_1$) 进行语义交互融合, 上一层

特征图通过 1×1 卷积调整特征图通道数, 通过上采样操作调整特征图的高和宽并与下层特征图进行拼接操作, 再通过过一个 3×3 卷积进行语义信息融合传输给下一层特征图和多尺度模块 (ADM) 进行多尺度特征提取. 如图 5 所示, ADM 是由 4 个具有不同感受野的分支和空间通道注意力模块 (CBAM) 组成. 每个分支都具有 3 个不同空洞率的 3×3 的空洞卷积, 其空洞率分别为$d\in\{(1,2,3),(3,4,5),(5,6,7),(7,8,9)\}$, 利用这种空洞率的设置方式可以减少空洞卷积的网格效应, 提高信息的连续性. 4 个多尺度分支网络提取出的特征图在通道维度上进行拼接操作, 并通过一个 3×3 的卷积进行特征融合, 再送入尺度对应的 CBAM 模块中调整人群特征权重抑制复杂背景信息的干扰, 之后从 ADM 模块中输出. F_1、F_2、F_3、F_4经过语义交互融合操作后, 分别通过 ADM 模块得到特征信息S_1、S_2、S_3、S_4. 依次从上往下进行上采样操作与下一层特征进行相加融合, 最后输出人群密度特征图.

图像通过多个子网络和多列空洞卷积操作丰富了图像细节特征, 提高了人群尺度变化的鲁棒性, 但是背景干扰问题对于人群密度检测也带来了巨大的阻碍, 复杂的背景信息在很大程度上降低了人群密度检测的准确性. 为了解决复杂场景下对人群特征的提取, 本文采用了空间通道注意力机制 (CBAM)[19]增强对人群特征区域的关注度, 抑制带来干扰的背景信息.

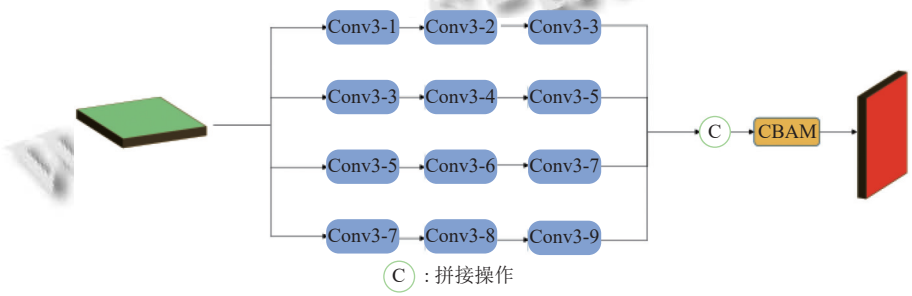


图 5 多尺度模块

CBAM 注意力模块是由空间和通道注意力组成的不仅能学习特征图的人群特征信息还能提取到空间位置信息, 相比其他的注意力模块能够获取更丰富的图像信息. 通过给定一个中间特征图, CBAM 注意力模块会通过空间和通道维度依次推断出注意力权重, 并将得出的注意力权重和输入的特征图进行相乘处理自适应融合优化, 提高人群特征区域权重. 相关公式如下:

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

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

其中, 输入特征图$F\in R^{C\times H\times W}$, 通道注意力$M_c\in R^{C\times1\times1}$, 平均池化操作为$F^c_{\text{avg}}\in R^{C\times1\times1}$, $F^c_{\max}\in R^{C\times1\times1}$为最大池化操作, σ为$Sigmoid$函数, MLP 为两层的感知机; $M_s\in R^{1\times H\times W}$为空间注意力操作, $F^s_{\text{avg}}\in R^{1\times H\times W}$为全局平均池化, $F^s_{\max}\in R^{1\times H\times W}$为全局最大池化, 然后将生成的特征图进行通道拼接, 再进行$f^{7\times7}$为7×7的卷积操作.

2.3 分割注意力模块

本文利用分割注意力模块将 MAMD 模块生成的人群密度特征图f_{den}生成注意力图来提供图像中人群空间位置信息, 提高人群密度检测准确度, 公式过程为:

$$M_{\text{att}}=Sigmoid(\text{Conv}(f_{\text{den}})) \tag{4}$$

其中, M_{att}为注意力图, $Sigmoid(\cdot)$为激活函数, 激活函数通过输出 0 到 1 的概率分数, 将背景与人群进行区分.

然后将注意力图与人群密度特征图进行逐元素相乘, 再通过 1×1 卷积融合, 从而生成高质量的密度图, 公式如下:

$$F_{\text{ref}}=\text{Conv}(f_{\text{den}}\otimes M_{\text{att}}) \tag{5}$$

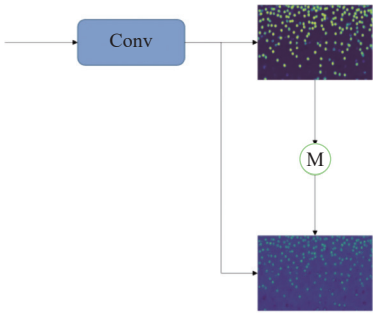
其中, F_{ref}为高质量密度图, ⊗为逐元素相乘操作, 分割注意力模块如图 6 所示.

2.4 联合损失函数

本文为了提升人群密度图的质量, 采用欧几里得距离作为人群密度检测的损失函数, 用来评估网络生成密度图和人群密度图的差异, 公式如下:

$$L_1=\frac{1}{2N}\sum_{i=1}^{N}\|Z(X_i;\theta)-Z_i^{GT}\|^2 \tag{6}$$

其中, N代表训练图片数量, X_i表示输入图片, $Z(X_i;\theta)$表示X_i的预测密度图, Z_i^{GT}表示X_i的真实密度图.



Conv：1×1卷积+BN+ReLU Ⓜ：逐元素相乘

图 6 分割注意力模块

另外在注意力图模块还引入了交叉熵损失函数, 公式如下:

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N}\left(A_i^{GT}\log(P_i)+\left(1-A_i^{GT}\right)\log(1-P_i)\right) \quad (7)$$

其中, A_i^{GT}为注意力图, P_i是预测的注意力图经过*Sigmoid*激活函数得到的概率值.

本文通过联合欧几里得和交叉熵函数来对密度图进行约束, 使得网络架构收敛速度更快. 最终的损失函数由两个函数加权生成的, 公式如下:

$$LOSS = L_1 + \alpha L_2 \quad (8)$$

其中, α是调节两个误差损失的权重为 0.1, 通过实验测试得出的.

3 实验分析

本文的相关实验配置是 Windows 系统, NVIDIA 3090, 24 GB 显存, 训练框架是 PyTorch 1.13, 采用的优化器为 Adam, 初始化学习率为 0.000 1, 权重衰减为 0.000 5. 编程语言是 Python 3.7.

3.1 数据集

随着人们在人群密度检测方面不断的深入研究, 相对应的人群密度检测数据集的规模与挑战性也在不断的提高. 人群密度检测对数据集的人群的疏密程度和在进行数据集标注时的准确性要求较高. 常用的人群密度检测的公共数据集[20]有 ShanghaiTech、UCF_CC50、UCF-QNRF 等, 这些数据集在制作的过程中充分考虑到了光照、场景、人群尺度等一些因素, 而且还对图像中的人头进行标注. 本文在 ShanghaiTech、UCF_CC50 数据集上进行实验评估.

ShanghaiTech 数据集分 Part_A 和 Part_B 两部分, 总共包含了 1 198 张图片, 330 165 个头部坐标. Part_A 数据集是在互联网上的一些人群密集、场景不同的图像, Part_B 是在上海的一个繁忙的街道上进行拍摄的, 人数相对较少, 场景比较单一[21]. 总体来说, 由于 ShanghaiTech 数据集无论是在场景类型、透视角度还是人群密度都是变化多样的, 所以在该数据集上进行人群密度检测是具有一定难度的.

UCF_CC50 数据集涵盖了多个不同场景, 包含了 50 张不同大小和视角的图像. 其中数据图像中人群尺度和密度变化较大, 数据集中最少的图像中有几十人, 最高有上千人, 人数差异巨大.

3.2 评价指标

在本次实验中, 采用的是平均绝对误差 (mean absolute error, *MAE*) 和均方误差 (mean squared error, *MSE*) 两个指标进行评估实验的有效性, 定义如下:

$$MAE = \frac{1}{N}\sum_{b=1}^{N}|a_b - \hat{a}_b| \quad (9)$$

$$MSE = \sqrt{\frac{1}{N}\sum_{b=1}^{N}(a_b - \hat{a}_b)^2} \quad (10)$$

其中, N为测试集中图像的总数, a_b代表的是第 b 张图像真实的人数, $\hat{a}_b$代表的是第 b 张图像预测的人数. *MAE* 表明预测的准确性, *MSE* 表明网络的鲁棒性.

3.3 消融实验

在 CDBM 中, 需要将多个不同分辨率的图像分别输入多个子网络中进行提取图像语义信息. 为了验证输入不同图像分辨率对实验结果的影响, 本文在公共数据集上针对不同分辨率图像进行实验, 其中 2、1、0.5 分别代表原图分辨率的 2、1、0.5 倍. 实验结果如表 1 所示.

表 1 不同分辨率的网络性能对比

分辨率	Part_A		Part_B		UCF_CC50	
	MAE	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*
0.5	66.1	129.3	13.0	15.8	186.8	261.7
1	63.2	113.4	9.1	11.2	195.2	270.5
2	64.7	116.8	10.2	13.0	193.3	276.6
0.5+1	61.6	103.2	9.1	11.4	179.9	258.4
0.5+2	59.8	101.4	8.0	10.7	178.9	251.1
1+2	60.1	100.3	8.7	10.9	183.5	260.6
1+0.5+2	59.2	98.7	7.2	10.2	177.4	242.8

通过对不同分辨率图像进行实验得出, 不同分辨率图像中包含不同的人群尺度信息, 利用多分辨率图像能够提取丰富的细节特征和尺度信息, 增强促进各

尺度信息交互和上下文感知能力, 提高网络检测效率, 对特征提取具有良好的增强作用. 当对分辨率为原图的 2、1、0.5 倍提取人群特征时网络的性能达到最优.

为了验证本文设计的多尺度注意力模块的有效性, 本文将去除多尺度注意力模块得到的基础网络 Basenet 和融入多尺度注意力模块的 CDBM 在公共数据集上进行对比实验, 从网络性能验证多尺度注意力模块的有效性. 实验结果如表 2 所示.

表 2 CDBM 的消融实验结果

网络	Part_A		Part_B		UCF_CC50	
	MAE	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*
Basenet	66.6	121.4	11.2	17.6	185.2	268.4
CDBM	59.2	98.7	7.2	10.2	177.4	242.8

从表 2 中可知, 融合多尺度注意力模块, 检测性能有所提升, 表明多尺度注意力模块有效地提取感兴趣的人群特征的, 较好地抑制背景信息的干扰, 利用多列空洞卷积增强尺度信息交互, 提高尺度连续变化的建模能力. 由此得出, 多尺度注意力模块可以提升多尺度特征的学习能力和网络检测的准确度.

为了验证利用本文的方法获取多分辨率图像具有良好的优越性, 本文在多个公共数据集上针对利用卷积操作获取的低分辨率图像与本文下采样的方式获取的低分辨率图像进行对比实验, 进行验证本文方法的有效性. 对比实验结果如表 3 所示.

表 3 不同方式获取多分辨率的性能对比

网络	Part_A		Part_B		UCF_CC50	
	MAE	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*
Conv-CDBM	60.4	95.5	8.9	11.2	175.8	244.6
CDBM	59.2	98.7	7.2	10.2	177.4	242.8

从表 3 可知, 利用卷积操作获取的低分辨率图像与本文采用下采样得到低分辨率图像的实验结果并没有太大区别, 但是利用卷积操作获取不同分辨率图像保留较多的高频信息会给网络造成一定的特征冗余, 增大网络的计算复杂性, 相对于本文获取的低分辨率图像保留较多的特征信息, 给后续低分辨率图像提取特征增大计算量.

3.4 基于 UCF_CC50 数据集实验结果

本文提出的方法通过与比较经典的人群密度检测算法在 UCF_CC50 数据集上的训练结果相比可以得出, 在针对小样本的 UCF_CC50 数据集中人群尺度变化极大, 人群场景变换复杂, 本文的方法具有一定优势, 训练结果如表 4 所示.

表 4 不同算法在 UCF_CC50 数据集上的 *MAE* 和 *MSE* 对比

方法	*MAE*	*MSE*
CSRNet[8]	266.1	397.5
DSSINet[9]	216.9	302.4
SFANet[12]	219.6	316.2
DUBNet[13]	243.8	329.3
SANet[14]	258.4	334.9
TEDNet[15]	249.4	354.5
DM-Count[17]	211.0	291.5
CDBM	177.4	242.8

从表 4 来看, 本文的方法在人群尺度变化大的 UCF_CC50 数据集上的训练结果为, *MAE* 为 177.4, *MSE* 为 242.8, *MAE* 和 *MSE* 比次优的算法 DM-Count 分别减少了 34 和 48.7, 该模型性能显著提高. 在该数据集中, 人群尺度变化较大, 较好地发挥了融合多尺度注意力网络对多尺度图像的特征融合能力, 也充分地说明了基于交互的多列空洞卷积处理尺度变化大的图像的学习能力比只通过卷积神经网络特征提取的效果更好. 在 Windows 操作系统, 显卡为 3090 的设备上进行推理每张图片的时间为 42 ms 左右, 每秒可以处理 20 张左右图片, 而且在实际场景中人群密度变化并不是骤增或骤减的瞬时过程, 因此对网络模型检测效率要求相对宽松, 对于该网络模型每秒处理 20 帧图片是完全满足人群密度检测的实时性的要求.

UCF_CC50 数据集的检测结果如图 7 所示.

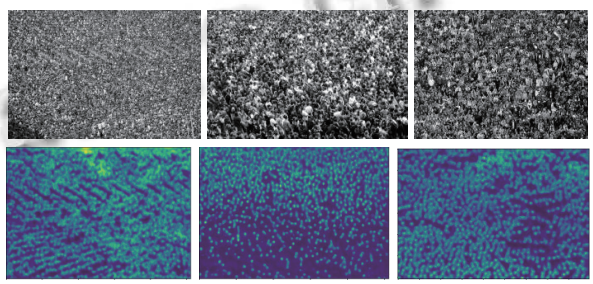


图 7 UCF_CC50 数据集的可视化结果

3.5 基于 ShanghaiTech 数据集的实验结果

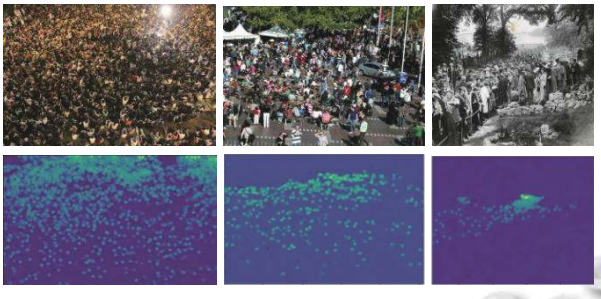
本文提出的方法在数据集 ShanghaiTech 上与多种算法进行比较, 针对 ShanghaiTech 数据集的多场景和多尺度变化情况, 该方法也能取得较好的检测结果, 结果如表 5 所示.

在表 5 中, 本文提出的融合多尺度和注意力的人群密度检测网络在 ShanghaiTech 的 Part_A 取得了比较低的 *MAE* 和 *MSE* 值, 分别是 59.2、98.7, 在 Part_B 上取得的 *MAE* 和 *MSE* 值分别为 7.2、10.2. 从表 5 可

以看出, 本文的方法对于 ShanghaiTech 的 Part_A 数据集中人群密集和人群尺度变化较大的场景下检测的结果是优于大多数其他的方法的, 它的 *MAE* 比人群密度检测精确度较高 DM-Count 减少了 0.6, 在 ShanghaiTech 的 Part_B 上的结果也取得较好的效果, 比次优的 DSSINet 网络的 *MSE* 低了 0.1. 通过实验可以看出, 基于融合多尺度特征和注意力网络可以充分地提取图像中不同尺度信息和降低背景信息带来的误差, ShanghaiTech 数据集部分检测结果如图 8 所示.

表 5 不同算法在 ShanghaiTech 数据集上的 *MAE* 和 *MSE* 对比

方法	Part_A		Part_B	
	MAE	*MSE*	*MAE*	*MSE*
CSRNet[8]	68.2	115.0	10.6	16.0
DSSINet[9]	60.6	96.0	6.8	10.3
AutoScale[10]	65.8	112.1	8.6	13.9
SFANet[12]	59.8	98.4	6.9	10.9
DUBNet[13]	64.6	106.8	7.7	12.5
SANet[14]	67.0	104.5	8.4	13.6
TEDNet[15]	64.2	109.1	8.2	12.8
DPDNet[16]	66.6	120.3	7.9	12.4
DM-Count[17]	59.7	95.7	7.4	11.8
CDBM	59.2	98.7	7.2	10.2



(a) ShanghaiTech 的 Part_A 数据集的可视化结果

(b) ShanghaiTech 的 Part_B 数据集可视化结果

图 8 ShanghaiTech 数据集可视化结果

4 结论

本文主要提出了一种融合多尺度和注意力机制的人群密度检测模型, 该模型利用多列子网络提取不同分辨率图像信息, 通过融合模块进行交互融合, 获取图像特征信息, 并利用多尺度注意力模块获取多尺度信息, 抑制背景信息的干扰, 可以减少由人群尺度不同和背景干扰引起的人群检测误差, 能够更加准确地对密集人群进行密度估计. 在 ShanghaiTech、UCF_CC50 数据集上进行实验, 结果表明该网络具有良好的预测能力, 在一定程度上减少了检测误差.

参考文献

1 余鹰, 李剑飞, 钱进, 等. 基于多尺度特征融合的抗背景干扰人群计数网络. 模式识别与人工智能, 2022, 35(10): 915–927. [doi: 10.16451/j.cnki.issn1003-6059.202210005]

2 Bai S, He ZQ, Qiao Y, *et al*. Adaptive dilated network with self-correction supervision for counting. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 4593–4602.

3 Wang X, Lv RR, Zhao Y, *et al*. Multi-scale context aggregation network with attention-guided for crowd counting. Proceedings of the 15th IEEE International Conference on Signal Processing. Beijing: IEEE, 2020. 240–245.

4 Yang YF, Li GR, Wu Z, *et al*. Reverse perspective network for perspective-aware object counting. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 4373–4382.

5 Han K, Wang YH, Tian Q, *et al*. GhostNet: More features from cheap operations. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 1577–1586.

6 张世辉, 赵维勃, 王磊, 等. MSIANet: 多尺度交互注意力人群计数网络. 电子与信息学报, 2023, 45(6): 2236–2245.

7 Wang WX, Liu QL, Wang W. Pyramid-dilated deep convolutional neural network for crowd counting. Applied Intelligence, 2022, 52(2): 1825–1837. [doi: 10.1007/s10489-021-02537-6]

8 Li YH, Zhang XF, Chen DM. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1091–1100.

9 Liu LB, Qiu ZL, Li GB, *et al*. Crowd counting with deep structured scale integration network. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 1774–1783.

10 Xu CF, Liang DK, Xu YC, *et al*. AutoScale: Learning to scale for crowd counting. International Journal of Computer Vision, 2022, 130(2): 405–434. [doi: 10.1007/s11263-021-01542-z]

11 Zand M, Damirchi H, Farley A, *et al*. Multiscale crowd counting and localization by multitask point supervision. Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022. 1820–1824.

12 Zhu L, Zhao ZJ, Lu C, *et al*. Dual path multi-scale fusion networks with attention for crowd counting. arXiv: 1902.01115, 2019.

13 Oh M, Olsen PA, Ramamurthy KN. Crowd counting with decomposed uncertainty. Proceedings of the 34th AAAI Conference on Artificial Intelligence, the 32nd Innovative Applications of Artificial Intelligence Conference, the 10th AAAI Symposium on Educational Advances in Artificial Intelligence. New York: AAAI Press, 2020. 11799–11806.

14 Cao XK, Wang ZP, Zhao YY, *et al*. Scale aggregation network for accurate and efficient crowd counting. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 757–773.

15 Jiang XL, Xiao ZH, Zhang BC, *et al*. Crowd counting and density estimation by trellis encoder-decoder networks. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 6133–6142.

16 Lian DZ, Chen XN, Li J, *et al*. Locating and counting heads in crowds with a depth prior. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 9056–9072. [doi: 10.1109/TPAMI.2021.3124956]

17 Wang BY, Liu HD, Samaras D, *et al*. Distribution matching for crowd counting. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 135.

18 Song QY, Wang CA, Jiang ZK, *et al*. Rethinking counting and localization in crowds: A purely point-based framework. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 3345–3354.

19 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. LNCS 11211: Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 3–19.

20 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proceedings of the 3rd International Conference on Learning Representations. San Diego, 2015.

21 Wan J, Wang QZ, Chan AB. Kernel-based density map generation for dense object counting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(3): 1357–1370. [doi: 10.1109/TPAMI.2020.3022878]

(校对责编: 牛欣悦)