计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(1):226–235 [doi: 10.15888/j.cnki.csa.008250]

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计数算法①

杨　旭, 黄　进, 秦泽宇, 郑思宇, 付国栋

(西南交通大学 电气工程学院, 成都 610097)
通信作者: 黄　进, E-mail: 2636270081@qq.com

摘　要: 针对密集场景下人群目标尺度变化大导致识别精度不高的问题, 本文提出两种多尺度特征融合结构: 注意力加权融合模块 (attention-weighted fusion module, AWF) 和自底向上融合模块 (bottom-up fusion module, BUF). 其中 AWF 模块引入注意力分支学习特征图的权重, 并将加权后的多层尺度特征进行叠加. 而 BUF 模块在处理特征图时使用空洞卷积捕获更多尺度信息, 且浅层特征图采用拼接方式融合. 经过融合模块处理的特征图具有更强的表达能力, 预测的密度图更加精准. 本文算法以 ResNet50 为骨干网络提取特征, 分别使用 AWF 和 BUF 模块进行特征融合, 在公开数据集上进行实验. 结果显示加入 AWF 模块的计数算法在 Shanghai Tech 数据集上的平均绝对误差 (MAE) 降到 45.54 (A 部分) 和 7.6 (B 部分), 均方误差 (MSE) 降到 100.28 (A 部分) 和 11.4 (B 部分), 在 UCF_CC_50 数据集上的 MAE 和 MSE 则降至 212.42 和 323.06. 而加入 BUF 模块后的算法在 Shanghai Tech 数据集上的 MAE 则为 51.6 (A 部分)、8.0 (B 部分), MSE 降到 102 (A 部分) 和 12.8 (B 部分), 在 UCF_CC_50 数据集上的 MAE 和 MSE 为 242.6 和 359.5. 实验结果表明, 本文提出的 AWF 模块和 BUF 模块都可以有效融合深层与浅层的特征信息, 优化特征图, 提高计数精度.

关键词: 人群计数; 多尺度信息; 特征融合; 注意力加权融合; 空洞卷积

引用格式: 杨旭,黄进,秦泽宇,郑思宇,付国栋.基于多尺度特征融合的人群计数算法.计算机系统应用,2022,31(1):226–235. http://www.c-s-a.org.cn/1003-3254/8250.html

Crowd Counting Algorithm Based on Multi-scale Feature Fusion

YANG Xu, HUANG Jin, QIN Ze-Yu, ZHENG Si-Yu, FU Guo-Dong

(School of Electrical Engineering, Southwest Jiaotong University, Chengdu 610097, China)

Abstract: To tackle the problem of poor recognition accuracy caused by large changes of crowd target feature in a high-density scenario, this study proposes two kinds of multi-scale feature fusion structures: attention-weighted fusion module (AWF) and bottom-up fusion module (BUF). The AWF module uses the attention branch to learn the weights of feature maps, and the weighted multi-scale features are superposed finally. The BUF module uses dilated convolution to obtain more scale information during feature processing, and the shallow feature maps are merged by stitching. The processed feature map has stronger expressive ability, and the predicted density map is more accurate. Taking ResNet50 as the backbone network for feature extraction, the algorithm presented in this study uses AWF and BUF modules for feature fusion respectively, and experiments are conducted on public datasets. The results show that the crowd counting algorithm with the AWF module can reduce the mean absolute error (MAE) to 45.54 (part A) and 7.6 (part B) and the mean square error (MSE) to 100.28 (part A) and 11.4 (part B) on the Shanghai Tech dataset. On the UCF_CC_50 dataset, the MAE and MSE are decreased to 212.42 and 323.06, respectively. Regarding the algorithm with the BUF module, the MAE is reduced to 51.6 (part A) and 8.0 (part B), and the MSE is decreased to 102 (part A) and 12.8 (part B) on the Shanghai

① 基金项目: 国家自然科学基金 (61733015); 高铁联合基金 (U1934204); 四川省重点研发计划 (2020YFQ0057)
收稿时间: 2021-03-22; 修改时间: 2021-04-19; 采用时间: 2021-04-26; csa 在线出版时间: 2021-12-17

Tech dataset. On the UCF_CC_50 dataset, the MAE and MSE are decreased to 242.6 and 359.5, respectively. Experiments indicate that the AWF module and BUF module can both effectively integrate deep and shallow feature information, thus able to optimize feature maps and improve counting accuracy.

Key words: crowd counting; multi-scale information; feature fusion; attention-weighted fusion; dilated convolution

人群计数作为计算机视觉领域下的一个重要分支, 其目的是通过算法自动统计出场景中的行人数量, 在视频监控、安防等领域有广泛应用. 尤其是在商场、车站、景点等公共场所, 进行实时的人流量统计有助于对拥挤情况的分析和异常状况的监测.

早期的人群计数算法基于目标检测进行统计, 但其精度较差, 尤其是高密度区域, 人头目标范围小导致目标检测变得困难, 因此基于目标检测计数很快被弃用. 后来出现数量回归的方法, 但该类方法不能体现人群目标的分布情况, 且计数精度并无太大提升. 直到密度图的提出给人群计数算法提供了新的思路, 该类型算法通过预测图像的密度图来统计人数, 模型学习的是图像特征到密度图的映射. 相比直接回归人群数量和基于目标检测进行统计的方法, 基于密度图的计数算法不仅能直观体现目标的分布情况, 统计精度也有很大提高, 是现在人群计数领域的主流方法.

高质量真实密度图的生成和合理的网络结构设计是提高基于密度图的计数算法精度的关键. 真实密度图的制作大多采用高斯自适应核滤波, 计数网络的结构设计则有多种方式, 按特征提取分支的数量可以分为多列网络和单列网络. 多列网络利用多条并行分支提取不同尺度的特征信息, 最后汇聚多个特征, 共同预测密度图. 单列网络则只使用一条分支进行特征提取和密度图的预测, 模型结构层次更深, 提取的语义特征更高级. 多列网络存在冗余特征的缺陷, 因为各分支的结构往往相似度高, 不易控制参数让其产生差异大的多尺度特征. 单列网络虽然能提取更高级的深层语义特征, 但浅层的特征信息在传输过程中丢失, 没有得到充分利用.

针对以上两种结构存在的问题, 本文提出以单列结构 ResNet50[1] 为特征提取器, 基于注意力加权融合 (AWF) 结构与自底向上融合 (BUF) 模块为融合层的人群计数算法. ResNet50 网络采用跳跃连接方式使浅层信息能跨层传输到深层特征, 从而避免了低层特征的信息丢失. 本文算法从 ResNet50 中抽取 3 个特征尺度变化较大的层, 作为多尺度图像特征的低级表达, 然后送入融合层进一步叠加特征, 最后进行密度图的预测. 其中 AWF 模块使用 3 个分支处理不同层特征图, 在处理特征图时引入注意力支路学习特征图权重, 加强特征表达, 3 个分支处理后的特征图逐层融合作为最终特征图, 融合方式为逐元素相加. BUF 模块在 AWF 的基础上进行删减, 它从浅层依次与相邻层融合, 在最后一层使用一个注意力层加权融合. BUF 在处理浅层特征时使用了空洞卷积, 目的是在更大的感受野上学习到更多的特征信息. BUF 的两个大尺度的浅层特征采用通道拼接的方式融合, 而 AWF 模块是深层特征依次与浅层进行逐元素相加. 在公开数据集上的实验表明, 本文提出的 ResNet50+AWF 和 ResNet50+BUF 都可以显著提高人群计数精度, 并具有更强的鲁棒性.

1 相关工作

人群计数算法可以分为基于检测和基于回归两大类. 基于检测的计数原理是采用目标检测算法识别并定位人体目标, 最后统计检测到的目标数量, 如 Topkaya 等[2] 利用聚类检测识别行人进行数目统计、Li 等[3] 通过背景分割检测出图像中的人头和人体肩部来估计人群数量. 随着深度学习的兴起, 目标检测领域出现了很多高效快速的检测算法如 RCNN 系列[4]、YOLO 系列[5]. 这些检测算法在通用目标检测任务上表现良好, 但是由于人群计数任务中存在遮挡、场景复杂和人体姿态变化大等问题, 这类目标检测模型在人群计数任务上的表现很差, 因此近些年已经很少采用目标检测算法进行人群计数.

基于回归的人群计数方法可以分为数量回归和密度图回归, 前者是让模型直接学习图像特征到人群数量的映射, 即输入一张图像, 模型直接输出图像中的人数, 如 IDREES 等[6] 使用 SIFT 特征和频域分析对极度密集场景进行分块回归来估计人群数量、Chan 等[7] 使用泊松贝叶斯回归算法直接计算人群数量. 基于数量回归方法的思路通常是先提取如 HOG、SIFT、LBP

等可以表征图像信息的特征, 然后通过线性回归或高斯回归等方法学习特征到数量的映射. 基于数量回归的方法可以解决场景中人群遮挡的问题, 但是忽略了空间信息, 模型的输出只有人数, 不能体现图像中人群的分布情况, 现在也基本不再使用该类方法.

基于密度图回归的方法由 Lempitsky[8] 在 2010 年提出, 该类方法的思想是使用密度图作为图像中人群的分布表示. 密度图是一张与原始图像尺寸一致的单通道图像, 其特点是在非人头目标位置的像素值为 0, 人头目标位置的像素值之和为目标个数. 该类算法模型学习的是从图像到密度图的映射关系, 而非图像到人数的映射, 其普遍思路是采用卷积神经网络 (CNN) 提取高级图像特征, 然后解码得到图像密度图, 最后对密度图进行积分得到目标数量. 最具代表性的密度图估计方法是 Zhang 等[9] 在 2016 年提出的 MCNN 网络, 该网络采用 3 个不同卷积核大小的分支结构提取不同尺度的特征信息, 最后按通道拼接各个分支的特征图, 送入卷积层预测最终的密度图. 基于密度图的计数方法的缺陷是多尺度检测, 由于场景中的人群目标近大远小, 密集区域遮挡严重, 算法很难同时在高密度和低密度区域上取得较好结果.

为了解决多尺度检测的问题, MCNN 的思想是用不同分支提取信息, 由于各分支的差异仅在卷积核大小上, 提取的特征冗余较多, 融合后的特征并不能较好地代表多尺度信息, 因此效果不太理想, 但其多列分支提取多层特征的思想影响了后续很多工作的发展. 另一种进行多尺度检测的思路是使用密度级别分类器将图像区域划分密度等级, 对不同等级的区域使用不同模块处理. 如 Sam 等[10] 提出 Switch CNN 将人群密度定义为高低中 3 个级别, 图像被分为不同小块, 各块根据其密度级别送入相应网络进行处理. Oñoro-Rubio 等[11] 则是将输入图像转换为图像金字塔, 金字塔的各层对应着不同尺度的图像输入, 分别送入网络提取多尺度特征. 但密度级别分类算法存在级别定义不明、计算量大、分类器精度难以把控等问题.

最近一些研究着重设计特殊的网络结构来提升网络性能, 如 Babu 等[12] 提出的 TDF-CNN 采用一种自下而上和自上而下交叉的特殊结构来获取更精确的特征图. Li 等[13] 提出的 CSRNet 使用了多层空洞卷积以加强特征图的表达能力. Zhang 等[14] 采用注意力机制, 其设计的 RA-Net 叠加全局注意力特征与局部注意力特征, 并提出拼接、相加和卷积 3 种融合方式. Guo 等[15] 提出的 DADNet 在网络中引入空洞卷积、可形变卷积与注意力机制, 网络结构更加复杂, 计数精度有一定提升.

针对多尺度问题, 本文从特征融合角度出发, 设计了两种特殊融合结构, 基于注意力加权融合 (AWF) 模块与自底向上融合 (BUF) 模块. AWF 模块采用多分支并行融合的方式汇聚不同层次的特征, 相比单分支网络能捕捉到更多的信息, 为了避免多层分支信息的冗余, 采用注意力机制为每层特征分配权重, 从而提高特征图的精度. BUF 模块中将不同尺度的特征信息逐层融合. 为了避免浅层信息丢失, BUF 模块引入空洞卷积增强特征表达能力, 同时在最后一层使用注意力分支学习特征权重, 进一步改善特征图. 因此, 相比其他多分支结构或单分支网络, AWF 和 BUF 模块的引入可以获取更高效的特征图, 降低密度图的预测误差. 在各个公开人群数据集上进行实验, 与未加入融合模块的网络和其他的人群计数算法进行对比, 结果表明, 本文提出的注意力加权融合方式确实可以进一步提升网络性能, 降低计数误差.

2 网络模型设计

如图 1 所示, 本文设计人群算法的整体架构由骨干网络、融合模块和解码模块 3 部分构成. 算法使用骨干网络提取多尺度图像特征, 然后将不同层次的输出特征送入融合模块进行处理, 解码模块则负责最终密度图的预测, 密度图的积分值即为预测人数.

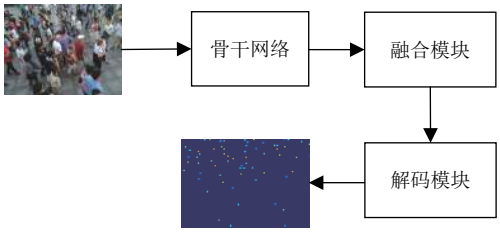


图 1 算法结构

骨干网络是网络的特征提取部分, 通常选取一些基础神经网络结构. 该部分提取的是较为浅层的图像特征, 因此本文在骨干网络后面增加了融合模块, 其作用是将浅层特征进一步融合, 融合后的特征图具有更强的特征表达能力. 但融合特征图是一个多通道的特征图, 同时它的分辨率远远小于输入图像, 所以最后将

融合特征图送入解码模块进行降维和分辨率的放大, 最终得到一个与输入图像分辨率相同的单通道密度图.

2.1 骨干网络

多列卷积网络存在训练困难和特征图冗余大的问题, 而单列卷积网络在人群计数任务上的有效性已经在 CSRNet 模型上得到证实, CSRNet 以经典的单列卷积网络 VGG[16] 为骨干提取特征, 使用空洞卷积作为解码器就能超越大多数多列网络的性能. 但 VGG 完全由卷积-池化-激活模块顺序叠加, 在较深的特征层上会忽视掉很多浅层信息, 因此本文选取引入跨层连接方式的 ResNet50 为骨干网络.

ResNet 网络的基本结构是残差块, 其特点是浅层信息通过跨层连接方式直接叠加到主干分支的输出上. 具体连接方式如图 2, 输入为 x, 主干支路输出为 $F(x)$, 则整个残差块的输出为 $y=F(x)+x$. 残差块的特殊连接方式可以让网络在较深层中保留浅层特征, 解决网络过深时梯度消失导致网络性能退化的问题.

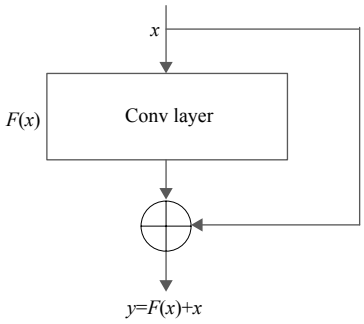


图 2 残差块

ResNet 可以看作多个残差块串联的网络, 根据卷积层和全连接层的个数进行不同的配置. 本文使用的是 ResNet50, 它的各层参数原始配置如表 1 所示, 其中 n×Residual 层表示 n 个残差块的串联, 每个残差块的主干支路是 3 个参数不一的卷积层. ResNet50 分为 6 个部分, 除去头尾的卷积池化和全连接层, 中间 4 部分由重复次数不一的残差块串联组成, 各层残差块的重复次数分别是 3、4、6、3, 表 1 中的网络输入图像尺寸设置为 256×256, output 列表示每一层的特征图输出大小.

原始 ResNet50 是分类模型, 因此网络最后一层采用了全连接层和 Softmax 函数预测类别输出, 而在人群计数网络中需要的输出是特征图而非类别, 故删除最后的全连接层和激活函数, 同时为了提取不同层次级别的特征信息, 本文抽取出网络中 layer2、layer3、layer4 的输出作为 3 层初级图像特征送入融合模块, 各层的输出尺寸可以通过各层的卷积核步长控制.

最终采用作为特征提取模块的 ResNet50 结构如图 3 所示, 原始 ResNet50 配置中的 layer4 输出特征图分辨率太小, 因此将 layer2 中 3×3 卷积核的步长大小修改为 1, 最终输出特征图的形状格式为 512×64×64、1024×32×32 和 1024×16×16.

表 1 ResNet50 的各层参数配置

layer	type	filters	kernel size	output
layer0	Convolutional	64	7×7	128×128
	Max pool	—	3×3	64×64
layer1	3×Residual	64	1×1	64×64
		64	3×3	
		256	1×1	
layer2	4×Residual	128	1×1	32×32
		128	3×3	
		512	1×1	
layer3	6×Residual	256	1×1	16×16
		256	3×3	
		1024	1×1	
layer4	3×Residual	512	1×1	8×8
		512	3×3	
		2048	1×1	
layer5	Average pool, full-connect, Softmax			1×1

2.2 注意力加权融合模块

从骨干网络中提取出 3 个层级的特征图后, 对多层特征的处理通常有两种方法: 先各自预测一个输出, 然后整合不同级别的输出结果; 先融合各层特征然后进行预测输出. 本文设计的融合模块都采用第二种方法, 将提取到的不同层级特征送入一个融合模块后再进行预测.

本文设计的第一种融合模块为注意力加权融合, 结构如图 4 所示, 其中 Conv2D 表示特征图的处理部分, 由卷积层和 SE 模块[17] 串联而成, Up 表示双线性上采样.

AWF 模块的输入是 3 个尺寸和通道数都不同的特征图, 每个特征图送入对应的处理分支, 3 个分支的输出特征图逐次叠加. 每个处理分支由两列 (1×1 卷积)→(BN 层)→(3×3 卷积) →(BN 层)→(1×1 卷积) →(BN 层)→(ReLU 激活层) →(SE 模块) 的串联结构组成. 其中一列输出为多通道, 另一列为注意力分支, 输出为单通道, 将两列特征图相乘后作为该分支的输出.

SE 模块如图 5 所示, 它的作用是在特征图通道数

较多的层中调整通道权重, SE 模块使用额外一个分支学习每个通道的权重对原始特征图按通道进行重新组合. 融合结构中使用了两次上采样层, 目的是便于相邻层特征图叠加. 各层特征经过各自分支处理后进行叠加的方式是按元素相加, 融合后的特征图尺寸和浅层较大的特征图尺寸相同, 即 64×64, 通道数为 512.

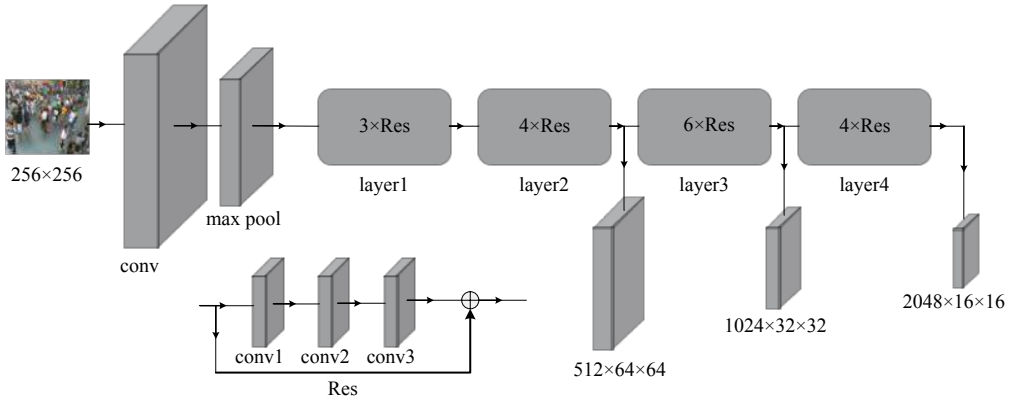


图 3 特征提取模块

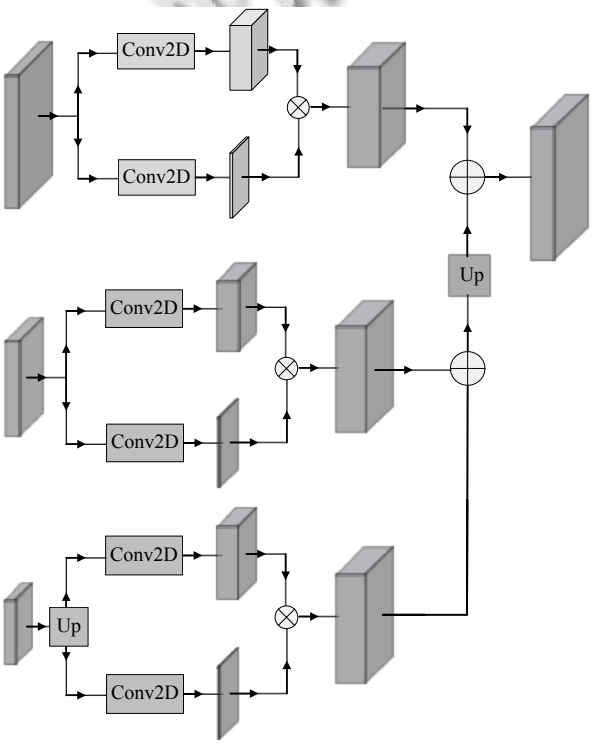


图 4 注意力加权融合

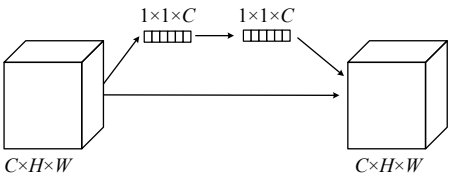


图 5 SE 模块

注意力加权融合结构在处理输入特征图时引入一个与主列结构类似的注意力分支, 目的是学习特征图的权重映射. 通过为特征图逐像素分配权重的方式, 实现对信息的聚焦, 权重系数大小体现该位置特征的重要程度, 让模型更关注人群目标区域, 而不是图像背景等无关区域, 因此加权融合后的特征具有更强的表达能力.

2.3 自底向上融合模块

AWF 模块的处理方式是各层级的输入特征分别进行优化再进行相加融合, 本文设计的另一种融合模块, 自底向上融合 (BUF) 模块则按照由浅到深的顺序依次融合 3 个特征, 其具体结构如图 6 所示. $f1$、$f2$、$f3$ 分别是 ResNet50 中的 layer2、layer3 和 layer4 输出, $f1$ 通过一个卷积块处理后与 $f2$ 按通道方向进行拼接, 再进行卷积运算得到 $f4$. $f3$ 与 $f4$ 通过加权叠加, 权重系数 $a1$ 为 $f3$ 卷积输出的单通道特征值.

和 AWF 模块中的卷积类似, 各个 conv 块由 (1×1 卷积) →(BN 层)→(3×3 卷积) →(BN 层)→(1×1 卷积) →(BN 层)→(ReLU 激活层) 的串联结构构成, 1×1 卷积作用是通道变换, 3×3 卷积负责降采样和特征学习. $f1$、$f2$、$f3$ 本身是骨干网络的不同层级输出, 如果采用普通卷积运算后拼接, 可能产生较多的冗余, 因此 3×3 部分采用的是空洞率为 2 的空洞卷积, 目的是在更大的感受野上获取更多的特征信息. 另外 $f3$ 和 $f4$ 叠加时只使用一个卷积块学习特征权重, 而不是分别学习权重系

数, 目的是减少参数量. $f3$ 和 $f4$ 共用权重系数, 最终输出 $output = a_1{}^* f_3 + (1-a_1)^* f_4$, 其尺寸和输入的深层特征图尺寸相同, 是 16×16, 通道数为 2048.

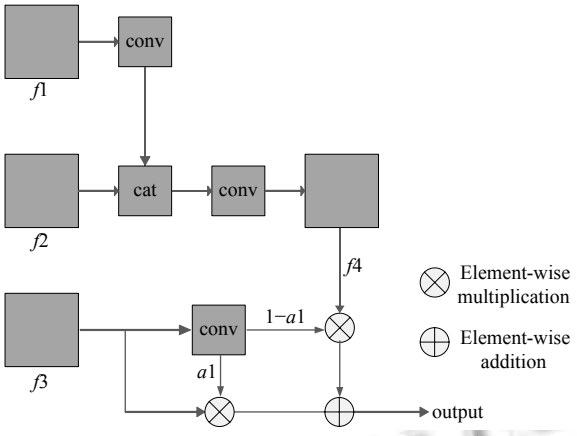


图 6 自底向上融合

BUF 模块使用空洞卷积, 在不增加参数量的情况下能捕获更多上下文信息, 减少传输过程中的浅层信息丢失, 同时只在最后一层使用卷积层学习特征图权重, 相比 AWF 使用多个分支学习权重的方法, BUF 具有更小的参数量, 结构更加简洁.

2.4 解码模块

多尺度特征经过 AWF 模块或 BUF 模块融合, 最后进入解码模块预测图像密度图.

解码模块的结构如图 7 所示, 输入 f 是融合后的特征图, 经过两个 conv 层和上采样层 Up 后输出最终的密度图. conv 层的作用是降低特征通道数, 由 1×1 卷积和 ReLU 激活函数构成. Up 层是对特征图进行双线性插值, 使输出密度图的分辨率与输入图像一致.

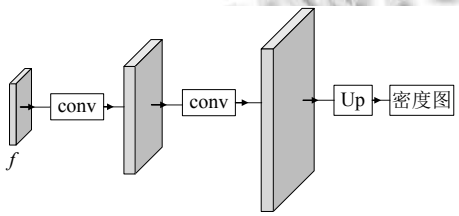


图 7 解码模块

3 实验分析

3.1 密度图标签

人群计数数据集的标注是对图像中人头中心坐标的点标注, 实验需要预先将数据集的点标签转换为密度图标签. 本文采用自适应高斯核滤波器对坐标点位置进行处理, 假设一个人头大小是 3×3 像素, 如图 8(a) 所示. 使用滤波模糊后的密度图如图 8(b), 各个像素点的概率值之和为 1, 对整张图的所有人头区域进行滤波, 即可得到完整图像的密度图标签.

0	0	0
0	1	0
0	0	0

(a) 单个人头中心位置

0.011 3	0.083 8	0.011 3
0.083 8	0.619 3	0.083 8
0.011 3	0.083 8	0.011 3

(b) 单个人头中心的密度图

图 8 高斯滤波器生成密度图标签

像素位置 x^i 处存在人头, 则点标签的对应处为 1, 可以用脉冲函数 $\delta(x-x^i)$ 表示, 一张图的人头数量是 N, 那么图像可以表示为 $H(x)=\sum_{i=1}^{N}\delta(x-x^i)$.

选择自适应高斯核滤波器 $G_\sigma=\exp\left(\frac{x-x^i}{2\sigma}\right)$, 其中方差 $\sigma=\beta\bar{d}$, 超参数 $\beta=0.3$, $\bar{d}$ 表示位置的 k 个邻近人头点的平均距离, 即根据目标之间的密集程度自动调整方差大小, 从而生成不同的高斯模板. 高斯滤波后的密度图为 $F(x)=H(x)^*G_\sigma$, 对其积分就可得到图像中的人头数量.

3.2 损失函数

损失函数用于衡量网络输出与真实标签之间的误差, 本文采用的是欧几里得距离损失 L_e 与密度一致性损失 L_c[18] 加权求和作为网络整体损失函数, 其计算公式如下.

$$L=L_e+\lambda L_c \tag{1}$$

其中, 系数 λ 是常数, 通常取 100, 两种损失的表达式分别如式 (2) 和式 (3) 所示.

$$L_e=\frac{1}{N}\sum_{i=1}^{N}\left\|F\left(X^i,\Theta\right)-Gt^i\right\|^2 \tag{2}$$

$$L_c=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{s}\frac{1}{k_j^2}\left\|P_{\text{avg}}\left(F\left(X^i,\Theta\right)\right)-P_{\text{avg}}\left(Gt^i\right)\right\|_1 \tag{3}$$

其中, N 表示数据集中的图像数量, $F\left(X^i,\Theta\right)$ 表示图像 X^i 的预测密度图, Gt^i 表示该图像的真实密度图, s 表示尺度级别数, P_{avg} 表示平均池化操作, k 表示池化后的

输出大小. L_e 损失逐像素计算预测密度图与真实密度图的误差, 是像素级的误差. 密度一致性损失衡量的则是局部区域的误差, 其比较的是将池化后的特征图差异. 文中设置了 3 个尺度级别, 即 s=3, 平均池化的输出尺寸依次是 1×1、2×2 和 4×4.

3.3 评价指标

对模型的性能评价采用平均绝对误差 *MAE* 和均方误差 *MSE* 两个指标, 其计算公式如下:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|z_i - z_i'| \tag{4}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|z_i - z_i'|^2} \tag{5}$$

其中, N 表示测试的图片数量, z_i 表示预测的人数, z_i'表示该图像的实际人数. *MAE* 即在测试集上预测的人数与实际人数的平均误差, 反映了模型的准确性. *MSE* 则是度量预测人数与实际人数的偏离程度, 可以反映模型的鲁棒性. 两个指标值都是越低越好.

3.4 实验结果

为了验证算法的有效性, 本文选择了公开数据集 Shanghai Tech[9] 和 UCF_CC_50[6] 作为实验数据, 数据集信息如表 2 所示, 其中 Shanghai Tech 数据集分为 A、B 两部分. A 部分 (SHHA) 图像分辨率不统一, 场景人群密度较大, B 部分 (SHHB) 图像分辨率均为 768×1024, 人群密度较低. UCF_CC_50 的图像数量较少, 但图像分辨率不统一, 人群密度极大.

表 2 数据集信息

数据集	图片数量	总人数	平均
SHHA	482	241 677	501
SHHB	716	88 488	123
UCF_CC_50	50	63 974	1 279

3 个数据集涵盖了低分辨率、高分辨率、低人群密度和高人群密度场景图像, 能够较好评估算法的性能. Shanghai Tech 数据集已经划分好训练集和测试集, 其中 SHHA 中的训练集共 300 张图像, 测试集 182 张图像, SHHB 的训练集和测试集图像数分别是 400 和 316. 而 UCFCC_50 仅为 50 张零散图像, 为了便于对比, 采用业界通用的交叉验证划分方法, 即每次选择不同的 10 张图像作测试集, 剩余图像作训练集, 将 5 次的测试结果均值作为最终的测试结果.

实验环境基于 Ubuntu 18.04 操作系统, 采用 PyTorch 深度学习框架, CPU 和 GPU 配置为 Intel(R) Xeon(R) Silver 4116 和 NVIDIA Tesla K80.

在数据处理阶段, 本文使用了分块训练. 具体操作是使用滑窗法裁剪数据集, 并在裁剪过程中进行随机翻转扩充训练集. 裁剪和翻转操作可以扩大数据集并让训练集的图像分辨率相同, 方便训练. 训练完成后, 在模型性能测试阶段, 对测试集上的图像也进行裁剪, 最后叠加各块的密度图结果.

另外, 对于训练集和测试集都进行了归一化处理, 即先计算整个数据集图像的均值和方差, 按照该均值和方差对每张图像进行归一化处理, 再送入网络训练和测试. 由于高斯滤波器生成的密度图的像素值很小, 范围 0–1. 如果直接进行训练, 很容易出现梯度弥散现象, 因此实验中对数据集的密度图标签进行等比例放大, 放大系数为 100.

融合算法的骨干网络部分是 ResNet50, 在初次训练时的初始化为采用预训练权重初始化. 预训练权重可以加快网络收敛, 减少训练时间. 而网络的后半部分是自定义的融合结构和解码模块, 采用的初始化方式是 kaiming 初始化和 0 初始化方式, 其中卷积核初始化是 kaiming 初始化, 其余层如 BN 层等采用 0 初始化. 训练初始学习率为 0.000 1, 衰减方式为按轮次衰减, 每隔 10 轮学习率衰减为上一次的 0.95 倍, 总训练轮数为 500 轮, 优化损失采用 Adam 算法.

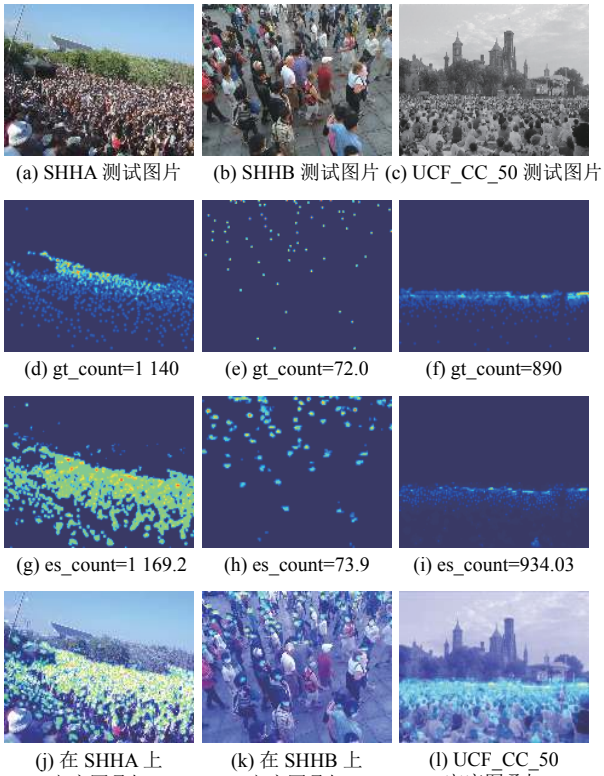
为了验证本文算法的有效性, 实验中与近年的不同类型网络进行对比, 在各个数据集上的测试结果如表 3 所示. 其中 MCNN 是多列网络, 各列由简化的 VGG 结构组成; MSCNN[19] 则是由若干个多列结构 Inception 模块串联而成; RANet 以 Hourglass 结构[20] 为特征提取器; CSRNet 和 PACNN[21] 是以 VGG 为骨干的单列计数网络; DUBNet[22] 是基于 ResNet50 的多输出网络. 实验中对比的模型基本涵盖了当前人群计数领域的各种类型方法, 结果较为客观全面.

从表 3 可以看出, 相对其他方法, 引入融合模块的算法在高低密度场景下都有较好表现. 在 SHHA 和 SHHB 数据集上, 引入 AWF 模块后的 *MAE* 和 *MSE* 都达到最优. 在高密度的 UCF_CC_50 上, *MAE* 降低到 212, 超过其他算法. 同时 AWF 模块在 Shanghai Tech 数据集上的误差略优于 BUF 模块, 但在 UCF_CC_50 数据集上 AWF 和 BUF 的误差相差较大, 整体而言 AWF 模块性能更优.

表 3 实验结果对比

方法	SHHA		SHHB		UCF_CC_50	
	MAE	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*
MCNN[2016]	110.2	173.2	26.4	41.3	377.6	509.1
MSCNN[2017]	83.8	127.4	17.7	30.2	363.7	468.4
CSRNet[2018]	68.2	115.0	10.6	16.0	266.1	397.5
RANet[2019]	59.4	102.0	7.9	12.9	239.8	319.4
PACNN[2019]	66.3	106.4	8.9	13.5	267.9	357.8
DUBNet[2020]	64.6	106.8	7.7	12.5	243.8	329.3
ResNet50+AWF	45.539	100.283	7.6	11.4	212.42	323.06
ResNet50+BUF	51.6	102.0	8.0	12.8	242.6	359.5

图 9 和图 10 分别展示了两种算法在各个数据集上的预测结果, 其中图 9 表示引入 AWF 模块的计数算法, 第 1 行是 3 个数据集中的原始图像, 第 2 行表示对应的真实密度图和标记人数, 第 3 行表示预测结果, 第 4 行是预测密度图与原始图像叠加的结果比较. 图 10 表示使用 BUF 模块的结果, 3 张测试的图像和图 9 相同, 因此图 10 只显示了引入 BUF 的模型预测密度图与其在原始图像上的叠加效果. 从图 9 和图 10 的结果可以看出, 在 3 个数据集上, 两种模型都能取得较好的预测效果.



(a) SHHA 测试图片 (b) SHHB 测试图片 (c) UCF_CC_50 测试图片

(d) gt_count=1 140 (e) gt_count=72.0 (f) gt_count=890

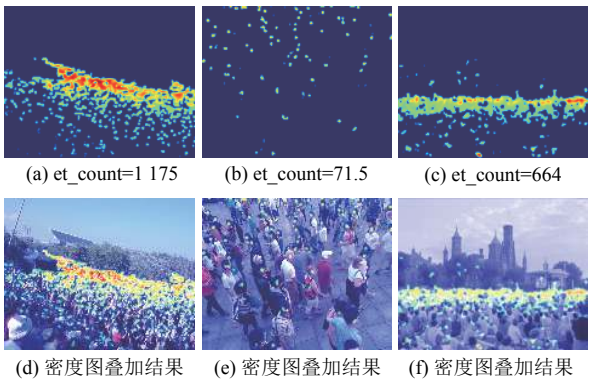
(g) es_count=1 169.2 (h) es_count=73.9 (i) es_count=934.03

(j) 在 SHHA 上密度图叠加 (k) 在 SHHB 上密度图叠加 (l) UCF_CC_50 密度图叠加

图 9 AWF 模块测试图片结果

3.5 消融实验

为进一步测试 AWF 和 BUF 结构的性能, 在 SHHA 数据集上进行消融实验. 选取人群计数领域常用的骨干结构 VGG16, 在其基础上加入 AWF 和 BUF 模块, 对比 ResNet50 进行实验. 各种模型的解码模块与第 3.4 节实验中的设置相同, 各自测试结果如表 4 所示.



(a) et_count=1 175 (b) et_count=71.5 (c) et_count=664

(d) 密度图叠加结果 (e) 密度图叠加结果 (f) 密度图叠加结果

图 10 BUF 模块测试图片结果

表 4 消融实验

方法	*MAE*	*MSE*
VGG16	72.6	116.9
VGG16+AWF	64.2	109.6
VGG16 + BUF	66.4	115.2
ResNet50	54.9	105.4
ResNet50+AWF	45.539	100.283
ResNet50 + BUF	51.6	102.0

ResNet50 在 SHHA 数据集上的 *MAE* 可以达到 54.9, 误差低于 VGG16 的 72.6, 表明 ResNet50 的效果优于 VGG16, 可见使用了残差块的 ResNet50 能够更好的提取网络特征. 而两种网络在加入 AWF 模块和 BUF 模块后的 *MAE* 均有 8~10 的提升, 表明两种模块对多层特征的复用是有效的, AWF 和 BUF 可以灵活嵌入各种骨干网络, 融合多尺度特征, 提升网络性能, 其中 AWF 的结构比 BUF 更复杂, 参数量更大, 但效果更好.

4 结语

本文针对人群计数网络在密集场景下检测困难, 多尺度信息利用不足等问题, 提出两种多尺度特征图融合方式: 注意力加权融合 AWF 与自底向上融合 BUF. AWF 模块在特征图处理过程中引入注意力分支学习特征图权重, 加权后的特征图再进行逐层叠加融合. BUF 模块由浅到深依次融合多层特征, 较浅的两层特征处理后按照通道顺序拼接, 然后与最深层的特征进

行加权融合. 在几个公开数据集上进行实验, 结果表明两种融合结构都能有效提高网络性能. 以 ResNet50 为骨干网络, AWF 为融合层的人群计数算法在 Shanghai Tech 数据集上的绝对误差降低到 45.539 和 7.6, 均方误差降低至 100.283 和 11.4, 远超其他计数网络. 在高密度的 UCF_CC_50 数据集上, *MAE* 降低至 212.42, 优于其他算法, 引入 BUF 融合层的算法效果略差于 AWF, 但相对现有方法仍有较大提升. 在 SHHA 数据集上的消融实验表明两种模块的有效性, 它们不仅可以嵌入 ResNet50, 在其他骨干网络中嵌入也可以实现降低计数误差的效果.

AWF 和 BUF 模块的有效性证明了从特征融合的角度来设计网络结构对人群计数算法性能的提升是有效的, 但新模块的加入带来了参数量的增长. 在未来的工作中, 本文将进一步研究更高效的改进方式, 设计更轻量高效的即插即用模块.

参考文献

1 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

2 Topkaya IS, Erdogan H, Porikli F. Counting people by clustering person detector outputs. 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Seoul: IEEE, 2014. 313–318. [doi: 10.1109/AVSS.2014.6918687]

3 Li M, Zhang ZX, Huang KQ, *et al*. Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection. 2008 19th International Conference on Pattern Recognition. Tampa: IEEE, 2008. 1–4. [doi: 10.1109/ICPR.2008.4761705]

4 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 580–587. [doi: 10.1109/CVPR.2014.81]

5 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788. [doi: 10.1109/CVPR.2016.91]

6 Idrees H, Saleemi I, Seibert C, *et al*. Multi-source Multi-scale counting in extremely dense crowd images. Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 2547–2554. [doi: 10.1109/CVPR.2013.329]

7 Chan AB, Vasconcelos N. Bayesian Poisson regression for crowd counting. 2009 IEEE 12th International Conference on Computer Vision. Kyoto: IEEE, 2009. 545–551. [doi: 10.1109/ICCV.2009.5459191]

8 Lempitsky V, Zisserman A. Learning to count objects in images. Proceedings of the 23rd International Conference on Neural Information Processing Systems. Vancouver: ACM, 2010. 1324–1332.

9 Zhang YY, Zhou DS, Chen SQ, *et al*. Single-image crowd counting via multi-column convolutional neural network. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 589–597. [doi: 10.1109/CVPR.2016.70]

10 Sam DB, Surya S, Babu RV. Switching convolutional neural network for crowd counting. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 4031–4039. [doi: 10.1109/CVPR.2017.429]

11 Oñoro-Rubio D, López-Sastre RJ. Towards perspective-free object counting with deep learning. 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 615–629. [doi: 10.1007/978-3-319-46478-7_38]

12 Sam D B, Babu RV. Top-down feedback for crowd counting convolutional neural network. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018. 7323–7330.

13 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. [doi: 10.1109/CVPR.2018.00120]

14 Zhang AR, Shen JY, Xiao ZH, *et al*. Relational attention network for crowd counting. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 6787–6796. [doi: 10.1109/ICCV.2019.00689]

15 Guo D, Li K, Zha ZJ, *et al*. DADNet: Dilated-attention-deformable ConvNet for crowd counting. Proceedings of the 27th ACM International Conference on Multimedia. Nice: ACM, 2019. 1823–1832. [doi: 10.1145/3343031.3350881]

16 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:

1409.1556, 2014.

17 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7132–7141. [doi: 10.1109/CVPR.2018.00745]

18 Dai F, Liu H, Ma YK, *et al*. Dense scale network for crowd counting. arXiv: 1906.09707, 2019.

19 Zeng LK, Xu XM, Cai BL, *et al*. Multi-scale convolutional neural networks for crowd counting. 2017 IEEE International Conference on Image Processing (ICIP). Beijing: IEEE, 2017. 465–469. [doi: 10.1109/ICIP.2017.8296324]

20 Newell A, Yang KY, Deng J. Stacked hourglass networks for human pose estimation. 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 483–499. [doi: 10.1007/978-3-319-46484-8_29]

21 Shi MJ, Yang ZH, Xu C, *et al*. Revisiting perspective information for efficient crowd counting. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7271–7280. [doi: 10.1109/CVPR.2019.00745]

22 Oh MH, Olsen P, Ramamurthy K N. Crowd counting with decomposed uncertainty. Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2020. 11799–11806. [doi: 10.1609/aaai.v34i07.6852]