FF-CAM:基于通道注意机制前后端融合的人群计数

张宇倩 李国辉 雷 军 何嘉宇

(国防科技大学信息系统工程重点实验室 长沙 410073)

摘 要 单个图像中的人群计数在计算机视觉领域中备受关注,因为其在公共安全方面具有重要作用.例如,在人群聚集的场景中监控设备可以实时监测人群数量变化,对过度拥挤和异常情况进行预警以预防安全事故的发生.然而,由于受到遮挡、透视扭曲、尺度变化和背景干扰的严重影响,在单个图像中对人群计数的预测要达到较高精确度是极其困难的,其面临着巨大的挑战.在本文中,我们提出了一个名为 FF-CAM 的创新性模型来计算图像中的人群数量.它首先将主网络低层的特征图与高层的特征图合并,实现不同尺度的特征融合,且无需额外的分支或子任务,解决了由于透视导致的尺度多样性问题.随后融合的特征图被送入通道注意力模块以优化不同特征的融合过程,并进行特征通道的重新校准以充分使用全局和空间信息.此外,我们在网络的末端利用扩张卷积来获得高质量的人群密度图,扩张卷积层扩大了感受野,其输出包含更详细的空间信息和全局信息,不会降低空间分辨率.最后,我们加入基于 *SSIM* 的损失函数用于比较估计人群密度图和真值的局部相关性,以及基于回归人数的损失函数用于比较估计人群数量与真实人数之间的差异.我们的 FF-CAM 在 UCF_CC_50 数据集、ShanghaiTech 数据集和 UCF_QRNF 数据集中进行训练并测试,获得了出色的结果.在 UCF_CC_50 数据集上比现有方法的 *MAE* 提高了 4.5%,*MSE* 提高了 3.8%.

关键词 人群计数;特征融合;通道注意力;扩张卷积;高质量密度图

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2021.00304

FF-CAM: Crowd Counting Based on Frontend-Backend Fusion Through Channel-Attention Mechanism

ZHANG Yu-Qian LI Guo-Hui LEI Jun HE Jia-Yu

(*Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha* 410073)

Abstract Crowd counting has attracted much attention in computer vision owing to its contribution in public security. For example, in a crowd gathering scenario, the monitoring device can monitor changes in the number of people in real time, and provide early warning of overcrowding and abnormal conditions to prevent the occurrence of safety accidents. But on account of occlusions, perspective distortions, scale variations and background interference, it faces a great challenge to achieve high accuracy on the prediction of crowd counting in a single image. In this paper we propose a novel model to count crowds named FF-CAM. It merges the frontend feature map with the backend feature map in the baseline, achieving a fusion of various scale features without additional branches or extra subtasks. The fusion is fed into the channel-attention block to optimize the procedure, and to conduct feature recalibration to use global and spatial information. Furthermore, we utilize dilated layers to obtain a high-quality density map. The dilated convolutional layer expands the receptive field, and its output contains more detailed spatial information and global

收稿日期:2019-07-11;在线发布日期:2020-04-12.本课题得到国家自然科学基金(71673293,61806215)资助.**张宇倩**,硕士研究生,主要研究方向为计算机视觉、深度学习以及信息系统工程.E-mail:446579794@qq.com.**李国辉**(通信作者),博士,教授,博士生导师,主要研究领域为计算机视觉、信息系统工程、数据挖掘及虚拟现实技术.E-mail:guohli@nudt.edu.cn.**雷 军**,博士,讲师,主要研究方向为计算机视觉、深度学习、数据挖掘及虚拟现实技术.**何嘉宇**,硕士研究生,主要研究方向为深度学习、数据挖掘及虚拟现实技术.

information without reducing the spatial resolution. The *SSIM*-based loss function is added to compare the local correlation between the estimated density map and the ground truth, meanwhile the regression-based loss function is added to compare the difference between the estimated number and the actual number of crowd. Our FF-CAM is verified in the UCF_CC_50 dataset, the ShanghaiTech dataset and the UCF_QRNF dataset, getting brilliant estimations. Compared to state-of-the-art, *MAE* is improved by 4.5% and *MSE* is improved by 3.8% in the UCF_CC_50 dataset.

Keywords　crowd counting; features fusion; channel-attention; dilated convolutions; high-quality density map

1　引　言

近年来，随着生活水平的提高和交通的快速发展，人群计数因其在公共安全方面的贡献而备受关注. 例如，在人群聚集的场景中监控设备可以实时监测人群数量变化，预防过度拥挤和异常情况. 然而，由于受到遮挡、透视扭曲、尺度变化和背景干扰的严重影响，在单个图像中对人群计数的预测要达到较高精确性是极其困难的.

在大量的研究和努力之下，人群计数已经取得了较大的进展. 早期的工作主要是检测人群中的每个行人[1]，或使用多个人工提取的特征回归得到人数[2]. 但是在拥挤的场景中由于严重的遮挡难以准确检测到行人，故会存在较大误差. 近年来，主流的方法由直接计算人数转为生成人群密度图，进而得到总人数以解决严重遮挡问题，基于 GAN[3-4] 和基于 CNN[5-13] 的方法已经发展并且得到了明显的改善. 此外，人群密度图还包含了空间位置信息，可更好的应用于安全领域.

然而，由于距监控相机的距离不同和透视问题，同一幅图像中会存在不同大小的人群，因此人头尺度多样性是抑制计数准确度的主要难点. 一些工作[5-9]使用具有不同卷积核或是多列的卷积结构来解决尺度变化的问题，而有些方法[10-11]则是用相同大小的卷积核堆叠来替换不同的卷积核. 此外，得到的人群密度图由于背景干扰会存在较大偏差，文献[12-13]在训练过程中增加了额外的信息来强调图像中的人群以解决该问题. 但这些方法仍然存在很多不足，不能很好地解决尺度多样性的问题. Li 等人[10]证明了多列结构中，不同分支中的每列学到的是几乎相同的特征，对尺度变化的贡献很小. 当网络变得复杂时，计算量和计算复杂性急剧增加，也会导致训练速度的延迟和梯度爆炸. 基于这个问题，为了学习到不同尺度的特征，同时排除背景噪声的影响，我们考虑采用单列单卷积核的网络结构，融合低层和高层的特征图. 由于网络中不同级别的层包含不同的比例特征信息，且多个相同大小卷积核叠加后与大的卷积核具有相同的特征学习效果. 此外，不同级别的层还包含不同级别的语义信息，低层卷积可以提取细节边缘图案，有效地回归拥塞区域得到密度图，高层则可以选择性地获得有用的语义信息，将人头与背景噪声区分开来. 这样做在获得不同尺度信息的同时不增加计算量和网络结构复杂度.

另一方面，各种特征通过简单的连接难以很好地对融合的不同尺度大小的人头区域的特征进行有选择性的加强，另外，卷积层的通道容易被忽略，从而导致空间信息的不足. 而由于生成的密度值遵循逐像素预测原则，因此输出的密度图必须包含空间相干性，以呈现最近像素之间的平滑过渡. 所以我们考虑将 SE(Squeeze-and-Excitation)模块[14]引入为通道注意力模块来优化融合. Hu 等人[14]提出，SE 模块可以考虑通道的权重，进行特征重新校准以捕获空间相关性，并有选择地强调信息性强的特征. 如此一来将该模块加在特征融合之后可以优化连接过程，对学习到的不同尺度的特征图进行加权，有选择性地强调不同尺度的特征，避免直接连接造成的损失. 同时捕获的空间相关性能使最终生成的密度图呈现最近像素之间的平滑过渡，以生成高质量的人群密度图.

此外，经过池化层的特征图降低了空间分辨率，丢失了空间信息，产生的人群密度图质量不够高. 我们考虑在网络末端运用扩张卷积. Li 等人[10]证明了扩张卷积比使用卷积、池化加反卷积的方案更好地保持了特征映射的分辨率，可以包含更详细的空间

信息和全局信息，在扩大了感受野的同时不增加参数或计算量. 所以，我们运用扩张卷积可以生成高质量的人群密度图.

最后，在人群场景中，高密度区域的局部模式和纹理特征与其他区域大不相同，但欧几里德损失建立在像素独立性假设上并忽略了它们，密度图的局部相关性未被考虑. 另外，其没有将输入图像的全局计数错误考虑进去，也与用来衡量准确度的评估指标没有直接关系. 为此，我们考虑在损失函数中加入结构相似性指数(*SSIM*)和关于回归人数的损失函数. 结构相似性指数根据局部模式计算两个图像之间的相似性，可以比较生成人群密度图与真值之间的相似性. 关于回归人数的损失函数直接衡量估计人群数量与真值之间的差异. 通过改进损失函数，网络将生成适合输入图像整体密度水平的特征，这有助于产生更准确的密度值.

基于上述讨论，我们提出了一种新型人群计数的结构：FF-CAM(Frontend-backend Fusion network through Channel-Attention Mechanism)，如图 1 所示. 我们提出的方法在 UCF_CC_50 数据集中的测试结果优于目前最先进的方法. 简而言之，我们的贡献包括以下三个方面：

(1) 我们融合了主网络低层和高层的特征图. 网络主干只有一列且只有一个大小的卷积内核，减去了额外的分支及参数量. 不同级别的卷积层不仅包含不同的语义信息，还包含不同的比例特征信息. 它们的融合可以适应由于透视效应引起的尺度变化，并且共享更多特征，同时可以排除背景干扰. 它还具有更少的参数和计算量.

(2) 我们引入了 SE 模块[14]作为 FF-CAM 的通道注意力模块. 避免直接连接造成的损失，通道注意力模块可以对融合的不同尺度大小的人头区域的特征进行有选择性的加强，由此提高网络的表达能力. 另一方面，它可以考虑通道的权重，进行特征重新校准以捕获空间相关性，使最终生成的人群密度图呈现最近像素间的平滑过渡.

(3) 我们利用一组扩张卷积[10]作为网络的末端. 其在增大了感受野的同时保证较少的参数量，包含了更详细的空间信息和全局信息，可以生成高质量的人群密度图. 此外，我们将 *SSIM*(结构相似性)和回归人数加入到损失函数中[7]. *SSIM* 可用于估计人群密度图和真值的局部一致性，基于回归人数的损失函数则衡量估计人群数量与真值之间的差异. 综合后的损失函数可以更好地衡量训练的估计值与真实值间的差异，产生更准确的密度值，提高训练准确度.

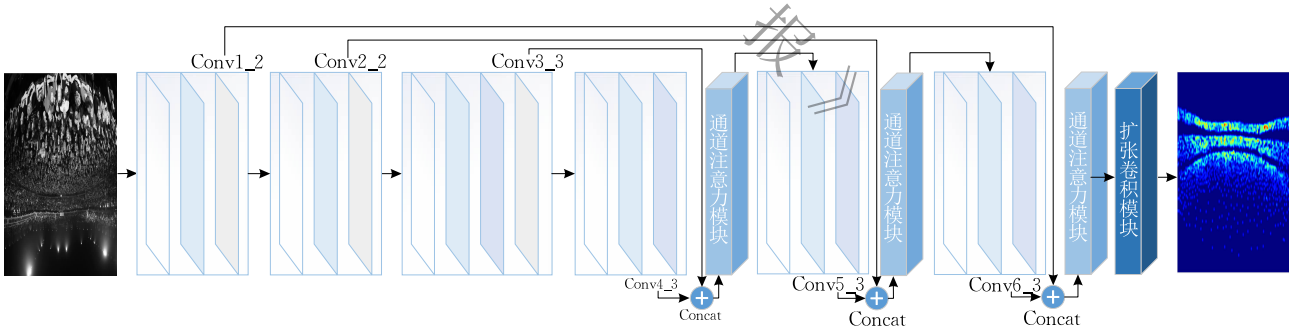


图 1 FF-CAM 网络的结构图(网络输入的是原始的拥挤人群图像，原始图像依次输入至不同的卷积层组合得到不同的特征图，Conv1_2 等即表示从不同的卷积层组合输出的特征图. 低层和高层的特征图融合(Concat)后再输入通道注意力模块. 最后经过扩张卷积模块后得到最终的人群密度图)

2 相关工作

在图像和视频中对人群进行计数已经有了很多年的发展，因为它在视频监控和公共安全中发挥着重要作用，故而受到计算机视觉领域中人们的长期关注. 但是由于遮挡，透视失真，尺度变化和背景干扰，计数精度的提高是一个相当大的挑战. 目前人群场景计数的研究大致有以下一些方法.

2.1 传统的方法

2.1.1 基于检测的方法

早期的工作主要是检测单个个体并计算数量. 2012 年，Dollar 等人[15]使用类似移动窗口的探测器来探测人体并计算图像中人的数量. Haar 小波分类器[16]用于从检测到的人体中提取低级特征，而文献[17]中则用 HOG(直方图定向梯度)分类器来提取特征. Felzenszwalb 等人[18]尝试检测身体的一些特定部分而不是整体，因为人体在拥挤的场景中总是

被遮挡.但是所有这些早期工作在非常拥挤的场景中都得到了较差的结果.

2.1.2　基于回归的方法

随着场景变得越来越拥挤，基于检测的方法存在很大限制，因此基于回归的方法被提出. Chan 等人[19]使用前景和纹理特征生成低级信息，并在学习了人群与提取的特征相对应的关系后计算出数量.随后在 2013 年，Idrees 等人[2]引入傅立叶分析和 SIFT(尺度不变的特征变换)来提取文献[19]中提出的特征.但是一些显著的特征很容易被忽视，从而导致更大的偏差.在文献[20]中，局部区域中的特征与其密度图之间的线性映射用来整合显著性信息.2015 年，由于理想线性映射增益的问题，Pham 等人[21]建议通过随机森林回归来学习非线性映射而不是线性映射.

2.2　基于深度学习的方法

随着深度学习的快速发展，卷积神经网络在人群计数领域显示出了很大的优势.

2015 年，Zhang 等人[22]训练卷积神经网络对人群密度图进行回归.他们使用密度和透视信息重新得到图像，然后使用它们微调训练好的网络并预测密度图.然而，其适用性受到透视图的要求和每个测试场景微调的限制.2016 年，Zhang 等人[9]使用多尺度卷积神经网络架构来解决人群场景中的大规模变化，并使用 1×1 卷积操作融合来自每个特定尺寸的卷积网络训练的特征图以回归得到密度图.它解决了尺度变化导致的问题.在此之后，多列[8]或多尺度[6,11,17]网络架构经常被用于人群计数问题.具体而言，Sam 等人[7]引入了一个分类器，根据密集级别选择指定的训练列.Cao 等人[8]使用尺度融合模块作为编码器来提取不同尺度的特征，并使用一组转置的卷积作为解码器来生成高质量的密度图，还提出了局部模式一致性损失函数.Zhang 等人[11]结合了多层的特征图来适应行人规模和视角的变化，引入了多任务损失，增加了相对人头数量损失函数.但是一些工作[10]则建议用相同大小的卷积核堆替换不同的卷积核.Li 等人[10]验证了使用多列卷积的有效性可能并不突出，这种分支结构中的每一列学到的都是几乎相同的特征.因此它使用 VGG16 作为基线，并在后端引入了扩张层，得到了很大的改进.此外，文献[12-13]在训练过程中增加了额外的信息以排除背景干扰.Shi 等人[12]将透视信息整合到人群密度图中，提供有关图像中人物尺度变化的附加信息，这十分有效地提高了小尺寸的人群区域的密度回归的精度.Liu 等人[13]提出了一项自监督的任务以改进人群计数网络的训练，在训练时利用未标记的人群图像以显著提高效果.它可以生成子图像的排名，其可以用于训练网络来估计一个图像是否包含比另一个图像更多的人.但额外的信息或任务可能会导致更多的资源和计算量的需求.

在 2019 年，更多解决方案被提出. Wang 等人[23]构建了一个大尺度、多样化的合成人群计数数据集来预先训练他们设计的空间全卷积网络. Liu 等人[24]引入了端到端架构，该架构结合了使用多个大小的感受域得到的特征，并学习在每个图像位置的每个特征的权重. Liu 等人[25]将检测到的模糊的图像区域放大到高分辨率以进行重新训练，并添加了本地化任务.几乎所有方法都添加了额外的信息或任务来增强单一人群计数的任务.

3　主要方法论述

许多先前的方法引入了多列融合的网络结构，以减少由于透视效应导致的头部尺度变化引起的误差.它们可以融合各种不同尺寸的卷积核或不同列的各种感受野的特征图.但是不同大小的内核可能会导致更多的参数量和计算量，而多列架构可能使网络更复杂.受文献[11]的启发，我们提出基于单一大小卷积核的单列网络，通过通道注意机制融合低层和高层的特征图.该网络对于头部尺度变化和背景噪声将更具鲁棒性，同时保持结构的简洁.此外，我们网络最后的部分利用扩张卷积模块，并且将基于 *SSIM* 和基于回归人数的两个损失函数添加到综合损失函数中.

我们提出的网络结构模型如图 1 所示，该模型被称为 FF-CAM(Frontend-backend Fusion network through Channel-Attention Mechanism).我们将从四个方面详细阐述该模型.

3.1　低层-高层融合

在人群场景的采集过程中，由于同一场景下人与摄像机的距离不同，会因为透视效应导致人头大小不同，也就是存在尺度多样性的问题.为了提取不同尺度大小的特征，解决尺度多样性带来的问题，并排除背景干扰，我们提出了低层-高层特征图融合的方法.

如图 1 所示，我们网络的主干采用 VGG-16 结

构,它具有强大的特征表示能力且易于连接.我们运用 VGG-16 的前 13 层来提取多尺度的特征图.组成 FF-CAM 的所有卷积核大小均为 3×3(除一个 3×3 卷积之前的 1×1 卷积用于降低计算复杂度和最后一层 1×1 卷积层用于代替全卷积层外),多个 3×3 的卷积核堆叠与大尺度的卷积核具有相同的效果,例如 2 个 3×3 的卷积核堆叠的效果相当于 1 个 5×5 的卷积核,3 个 3×3 卷积核则相当于 1 个 7×7 的卷积核,以此类推.因此其可以学习到不同尺度的特征,但计算量要少得多,并且可以构建更深的网络.

网络中不同级别的特征层不仅包含不同级别的语义信息,还包含不同的比例特征信息.低层可以提取细节边缘图案,这对于在人群密度图中回归拥塞区域的值具有重要意义.但它无法捕捉细节,这可能会导致杂乱的背景干扰,从而导致不正确的回归.高层则可以选择性地获得有用的语义信息,因此网络可以将人群与背景噪声区分开来.

鉴于它们的特性,我们通过通道注意模块融合低层和高层的特征图,以从主干网络中获取并融合足够多的特征.

如图 1 所示,我们使用来自 VGG-16 主干网络中的 Conv1_2, Conv2_2 Conv3_3, Conv4_3 和 Conv5_3 层的特征图,其中卷积层参数设置与 VGG-16 相同.这些不同层级特征图的输入有助于提取多尺度的特征.通过最大池化层后,这些输出特征图对应的大小分别为原始输入图像的 1/2,1/4,1/8 和 1/16.首先,使用最近邻插值对 Conv4_3 输出的特征图进行上采样,并与 Conv3_3 输出的特征图融合,再将融合后的特征图输入通道注意力模块,调整两层不同特征信息融合时的权重,提高网络的表征能力.随后,Conv5_3 输出的特征图和 Conv2_2 输出的特征图的融合操作类似于 Conv4_3 和 Conv3_3,融合得到的特征图同样输入通道注意力模块.经通道注意力模块处理后的特征图输入一组卷积层:Conv1×1×512,Conv3×3×512 和 Conv3×3×512.3×3 卷积之前的 1×1 卷积用于降低计算复杂度.我们将该组卷积层输出的特征图定义为 Conv6_3 层,其同样被上采样并与 Conv1_2 的输出融合,然后以相同的方式输入到通道注意力模块.最后,输出的特征图通过扩张卷积模块后生成人群密度图.接下来我们将具体介绍通道注意力模块和扩张卷积模块,具体结构如图 2 和图 3.

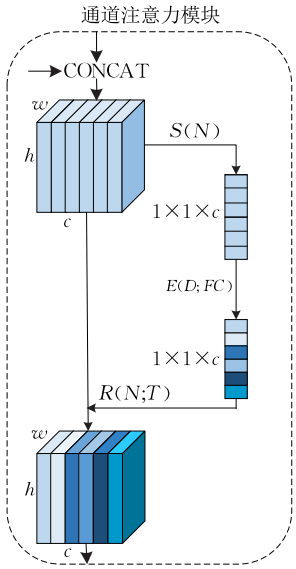


图 2 通道注意力模块的结构图(其中,CONCAT 表示两层特征图的融合,得到空间维数为 $h\times w\times c$ 的特征图)

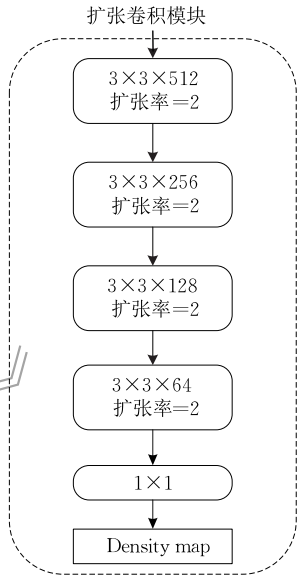


图 3 扩张卷积模块的结构图(其中扩张卷积第 1 行的参数分别表示卷积核大小和通道数)

3.2 通道注意力模块

注意力模型现在已经成为神经网络中的一个重要概念,在不同的领域中被研究和应用.文献[14]介绍了 SE 模块,它模拟了卷积特征图的通道之间的相互依赖性,从而提高了网络的表征能力.

大多数先前的工作直接组合来自不同卷积层的特征图,没有考虑融合时它们各自的权重.另一方面,由于空间信息的不足,卷积层的通道总是被忽略.SE 模块可以进行特征重新校准,选择性地强调有用信息,并且抑制不太有用的特征,网络可以学习使用全局信息.此外,它还有助于捕获空间相关性,

而无需额外的监督. 最后一点，它在计算上很轻巧. 有如此多的好处，它却只会略微增加模型复杂性和计算负担.

此外，SE 模块已被证明可以改善网络性能，并可以通过整个网络进行累积[14]. 因此，我们将 SE 块转换为我们的通道注意力模块. 具体结构如图 2 所示. 通道注意力模块包括三个过程：挤压 S、激励 E 和重新缩放 R.

首先，对两个卷积层融合后输出的特征图 N 进行挤压操作 S. 挤压操作在空间维度上聚合特征图，并通过全局平均池化层来生成通道统计量. 给定特征图的空间维数为 $h\times w\times c$，挤压操作后变为 $1\times1\times c$. 每一个通道的特征图 $N_x(x=1,2,\cdots,c)$ 对应的通道描述符 D_x 由以下公式计算：

$$D_x=S(N_x)=\frac{1}{h\times w}\sum_{i=1}^{h}\sum_{j=1}^{w}n_x(i,j) \tag{1}$$

其中，$n_x(i,j)$ 表示特征图 N_x 上第 i 行第 j 列的元素的值.

特征图 N 通过挤压操作后生成了通道描述符 $D=\{D_x,x=1,2,\cdots,c\}$. 通道描述符嵌入了通道特征响应的全局分布，因此其较低层能够利用全局感受野的信息.

然后，我们将 D 送入激励操作 E，产生提取描述符 T. 它由基于非线性的两个完全连接层、一个 Relu 函数和一个 Sigmoid 函数组成. 将其表示为

$$\begin{aligned}T&=E(D;FC)=\sigma(g(D;FC))\\&=\sigma(FC_2\delta(FC_1D))\end{aligned} \tag{2}$$

其中，FC_1 是具有缩小率 k 的降维层，FC_2 是维数增加层. k 是一个超参数，它可以改变模型中块的容量和计算成本. 根据文献[14]，我们设置 $k=16$，以实现准确性和复杂性之间的良好平衡. δ 是 Relu 函数，σ 是 Sigmoid 函数. 两个完全连接层可以通过减小维度来限制模型复杂性，极大地减少了参数量和计算量. 并且其能更多地学习通道之间的非线性相互作用，可以更好地拟合通道间复杂的相关性，提高泛化性. 此外，与 one-hot 激活函数相反，Sigmoid 激活函数强调多个通道，故整个激励操作能完全捕获通道依赖性并控制每个通道的激励，获得 0～1 之间归一化的权重.

最后，通道注意力模块的输入 N 由提取描述符 T 重新加权：

$$F=R(N;T)=T\cdot N \tag{3}$$

其中，R 表示输入特征图 N 和提取描述符 T 之间的通道乘法，即通过乘法将 T 逐通道的权重加权到 N 中对应的每个通道特征图的每个特征点上，完成在通道维度上的对原始特征的重标定. 模块的最终输出 F 可以直接被送入下一层.

3.3　扩张卷积模块

在我们的网络中，输入的人群图像由最大池化层下采样再经上采样融合之后，生成的特征图为原始输入的 1/2. 特征图在经过池化层后，虽然在控制过拟合同时保持了不变性，但降低了空间分辨率，丢失了部分空间信息，产生的密度图质量不够高.

Li 等人[10]证明了扩张卷积可以比使用卷积、池化加反卷积的方案更好地保持特征映射的分辨率. 虽然反卷积层可以减轻信息的丢失，但会增加额外的复杂性，且会导致执行延迟. 基于此，我们在网络的末端利用扩张卷积层. 扩张卷积层扩大了感受野，而不增加参数或计算量. 同时，经过扩张卷积的输出可以包含更详细的空间信息和全局信息，不会降低空间分辨率. 所以，我们运用扩张卷积可以生成高质量的人群密度图，同时提高人群估计准确率.

我们在网络的末端运用扩张卷积，如图 3 表示网络末端的扩张卷积模块. 它由具有扩张率为 2 的四层扩张卷积层和一层 1×1 的卷积层组成. 每个扩张卷积层的通道数都不同，每一层后都会通过批量标准化层和 Relu 层. 1×1 卷积层用来输出最终的人群密度图，相较于全连接层其参数量更少，计算量更小. 最后，网络输出高分辨率的人群密度图.

3.4　综合损失函数

主流工作将像素上的欧几里德损失设置为训练过程中的损失函数. 在人群场景中，高密度区域的局部模式和纹理特征与其他区域（低密度区域或背景）大不相同，但欧几里德损失建立在像素独立性假设上并忽略了它们，密度图的局部相关性未被考虑. 此外，该损失函数与用来衡量准确度的 MAE 及 MSE 没有直接关系，也没有将输入图像的全局计数错误考虑进去. 为了解决上述问题，我们将基于结构相似性指数（$SSIM$）的损失函数、基于回归人数的损失函数与欧几里德损失相结合作为我们的最终损失函数，该函数可用于估计人群密度图和真值的局部一致性，并估计人群数量与真实人数之间的差异，从而使综合后的损失函数更好地表示训练产生的估计值与真实值间的差异，以生成高质量的人群密度图，提高训练准确度.

3.4.1　欧几里德损失函数

欧几里德损失用于在像素级别上衡量输出密度图与相应真值之间的差异，其定义如下：

$$L_2(\Theta)=\frac{1}{N}\sum_{i=1}^{N}\left\|F_d(I_i;\Theta)-D_i\right\|^2 \tag{4}$$

其中，Θ 表示网络训练时的一组参数，N 是训练样本的数量. $F_d(I_i;\Theta)$ 表示具有参数 Θ 的网络输入图像 I_i 后输出的估计密度图，而 D_i 是对应的真值密度图.

3.4.2 基于 *SSIM* 的损失函数

SSIM 是一种广泛用于图像质量评估领域的指标. 它根据局部模式(包括均值，方差和协方差)计算两个图像之间的相似性. *SSIM* 值的取值范围是 $[-1,1]$. 两个图像越相似，其值越大. 当两个图像相同时，它等于 1.

受 SANet[7] 启发，我们将 *SSIM* 加入损失函数. 首先，使用标准偏差为 1.5 的 11×11 归一化高斯核来估计局部统计量. 然后，权重由 $W=\{W(r)|r\in R, R=\{(-5,5),\cdots,(-5,5)\}\}$ 定义，其中 r 为中心，R 包含所有位置内核. 因此，对于每个位置 t，计算密度图 F_d 和相应的真值 D 的局部统计量.

首先计算 F_d 的局部均值 μ_{F_d} 和方差 $\sigma^2_{F_d}$：

$$\mu_{F_d}(t_{F_d})=\sum_{r_{F_d}\in R_{F_d}} W(r_{F_d})\cdot F(t_{F_d}+r_{F_d}) \tag{5}$$

$$\sigma^2_{F_d}(t_{F_d})=\sum_{r_{F_d}\in R_{F_d}} W(r_{F_d})\cdot[F(t_{F_d}+r_{F_d})-\mu_{F_d}(t_{F_d})]^2 \tag{6}$$

其次，是 D 的局部均值 μ_D 和方差 σ^2_D：

$$\mu_D(t_D)=\sum_{r_D\in R_D} W(r_D)\cdot F(t_D+r_D) \tag{7}$$

$$\sigma^2_D(t_D)=\sum_{r_D\in R_D} W(r_D)\cdot[F(t_D+r_D)-\mu_D(t_D)]^2 \tag{8}$$

由此我们可以计算 F_d 和 D 间的局部协方差 σ_{F_dD}：

$$\sigma_{F_dD}(t)=\sum_{r\in R} W(r)\cdot[F(t+r)-\mu_{F_d}(t_{F_d})]\cdot[Y(t+r)-\mu_D(t_D)] \tag{9}$$

根据这些指标，*SSIM* 逐点计算如下：

$$SSIM=\frac{(2\mu_{F_d}\mu_D+Q_1)(2\sigma_{F_dD}+Q_2)}{(\mu^2_{F_d}+\mu^2_D+Q_1)(\sigma^2_{F_d}+\sigma^2_D+Q_2)} \tag{10}$$

其中，Q_1 和 Q_2 是随机的非常小的常数，以避免被零除，我们依照文献[7]的设置来给它们赋值.

最后，基于 *SSIM* 的损失函数定义为

$$L_s=1-\frac{1}{M}\sum_t^M SSIM(t) \tag{11}$$

其中，M 是密度图中的像素总数.

3.4.3 基于回归人数的损失函数

大多数基于密度估计的计数算法通过测量预测密度图和地面实况密度图之间的每像素误差来优化其计数模型. 然而，这种方法与用来衡量准确度的评估指标 *MAE* 和 *MSE* 没有直接关系，也没有将输入图像的全局计数错误考虑进去. 为此，我们新增了另一个关于回归人数的损失函数，它直接衡量估计人群数量与真实人数之间的差异. 通过增加该损失函数，网络将生成适合输入图像的整体密度水平的特征，这有助于产生更准确的密度值. 其定义如下：

$$L_c=\|\hat{C}-C\|^2 \tag{12}$$

其中，$\hat{C}$ 和 C 分别是训练得到的人群数量和真实的人群数量.

3.4.4 综合损失函数

将基于 *SSIM* 的损失函数和基于回归人数的损失函数加入到训练过程中，最终的综合损失函数表示如下：

$$L=L_2+\alpha L_c+\beta L_s \tag{13}$$

其中，α 和 β 分别是基于回归人数的损失函数和基于 *SSIM* 的损失函数的权重，用作三个函数的平衡. 我们根据文献[7]的经验设定 $\beta=0.001$，在实验验证后设定 $\alpha=1$，具体实验见第 4.7 节.

4 实 验

我们的实验是在 4 块 TITAN Xp GPU 上进行的. 该网络基于 Pytorch 框架，我们使用 Adam 优化器来优化参数并将原始学习速率设置为 1e-5. 参数通过高斯分布随机初始化，平均值为零，标准差为 0.01. 除了输出层之外，我们还在每个卷积层之后使用批量标准化层和 Relu 层，以提高训练速度并有效地避免梯度的消失和爆炸.

4.1 真值的生成

现有的数据集一般都给定了原始图像以及其对应的人群在图像中的坐标位置及总人数. 和文献[9]一样，我们同样用高斯自适应核来生成密度图的真值. 高斯自适应核的定义如下：

$$F(x)=\sum_{i=1}^N \delta(x-x_i)\times G_{\sigma_i}(x),\ \sigma_i=\beta\overline{d_i} \tag{14}$$

其中，在真值 δ 中，对于其中任意位置 x 和每一个人头目标 x_i，$i=1,2,\cdots,N$，定义 $\delta(x-x_i)$ 是标准差为 σ_i 的高斯核，而 d_i 是 k 个最近邻的平均距离. 根据文献[9]的经验，我们设置 $\beta=0.3$，$k=3$. 对于每幅输入的人群场景图像，高斯核可将其中所有标注的人头模糊化，生成人群密度图的真值.

4.2 评估指标

大多数现有工作使用两个度量指标来衡量人群计数的准确性，平均绝对误差(*MAE*)和均方误差(*MSE*). *MAE* 表示估计的准确性，而 *MSE* 反映估计的鲁棒性. 定义如下：

$$MAE=\frac{1}{N}\sum_{i=1}^N|F_{di}-D_i| \tag{15}$$

$$MSE=\sqrt{\frac{1}{N}\sum_{i=1}^N|F_{di}-D_i|^2} \tag{16}$$

其中，N 是测试图像的数量，D_i 是第 i 个图像中的真实人群数，F_{di} 是第 i 个图像中的估计人群数.

4.3　在 UCF_CC_50 数据集上的实验

Idrees 等人[2]提出的 UCF_CC_50 数据集包括 50 个具有不同视角和分辨率的图像. 这是一个非常拥挤的数据集，平均人数达到了 1280 人，最多的一幅图片中有 4543 人. 由于包含各种人群场景且图像总数有限，这是一个非常具有挑战性的数据集. 因此，我们按照文献[2]中的标准设置执行 5 倍交叉验证，最大程度地利用样本：将数据集随机均分成五等份，以其中的四份作为训练集，剩下的一份作为测试集，共进行五次训练和测试，五次实验的结果如表 1 所示. 最后再取误差指标的平均值作为实验的最终结果.

我们将结果与最先进的方法进行比较，表 2 中列出了 *MAE* 和 *MSE* 比较的结果. 我们的 FF-CAM 的估计误差 *MAE* 和 *MSE* 在所有模型中是最小的，这表明我们得到了对 UCF_CC_50 数据集计数的最佳估计，相比于效果最好的[10]，我们的 *MAE* 提高了 4.5%，*MSE* 提高了 3.8%. 该结果验证了 FF-CAM 模型的准确性和鲁棒性.

表 1　UCF_CC_50 数据集 5 倍交叉验证结果

测试集序号	*MAE*	*MSE*
1	383.65	579.99
2	144.33	183.25
3	293.15	337.64
4	257.15	317.47
5	155.54	192.51
均值	**246.764**	**322.172**

表 2　UCF_CC_50 数据集的估计误差

方法	*MAE*	*MSE*
MCNN[9]	377.6	509.1
CMTL[8]	322.8	397.9
Switch-CNN[7]	318.1	439.2
SaCNN[11]	314.9	424.8
CSRNet[10]	266.1	397.5
FF-CAM	**246.8**	**322.2**

训练好的模型在 UCF_CC_50 数据集上得到的部分密度估计图如图 4 所示. 由图 4 可以看出，我们

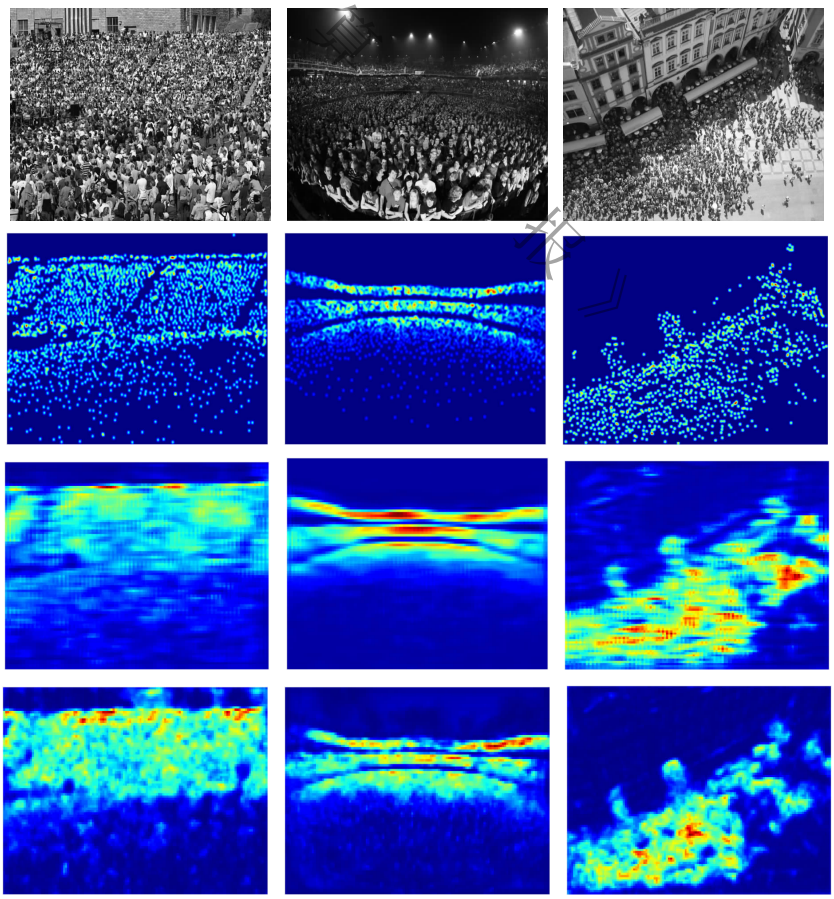


图 4　FF-CAM 模型在 UCF_CC_50 数据集上的实验对比密度图(第 1 行为原始图像，总人数分别为 1997、2960 和 1045；第 2 行为密度图的真值；第 3 行为 CSRNet[10]结构得到的密度估计图，预测总人数分别为 2100、3430 和 1185；第 4 行为 FF-CAM 得到的密度估计图，预测总人数分别为 2006、2600 和 1022)

的模型对极度拥挤的场景能进行很好的预测并生成分布较为准确的密度图，且预测人数更接近真实人数，好于 CSRNet[10] 模型. 由这些图可以看出，第二张由于透视存在人头尺度大小不一的问题，而得到的密度图很好地解决了该问题，在不同人头大小的位置生成的密疏程度不一. 第三张具有干扰的楼房背景，而得到的密度图很好地排除了干扰，未将其统计入人数.

4.4 在 ShanghaiTech 数据集上的实验

ShanghaiTech 数据集是一个多样且拥挤的数据集，由 Zhang 等人[9] 提出. 该数据集包括 Part A 和 Part B 两部分，Part A 是从网上收集而来，共有 482 张图片；Part B 则是从上海的拥挤繁忙的街道上收集而来，共有 716 张图片. 两个部分都是十分拥挤的数据集，Part A 平均人数达到了 501 人，最多的一幅图片中有 3139 人. 而 Part B 相对不那么拥挤，平均人数为 124 人，最多的一幅图片中有 578 人. 在 Part A 数据集中，300 张图片用来训练，剩下的 182 张则用来测试. Part B 数据集里的 400 张图片用来训练，316 张用于测试.

表 3 中列出了我们将估计结果的误差 *MAE* 和 *MSE* 与最先进的方法进行比较的结果. 从表中可以看出，我们的方法在 Part B 数据集中测试的结果优于其他的方法，*MAE* 和 *MSE* 分别提高了 2.8% 和 1.3%. 这说明我们的方法在 Part B 数据集上表现得很好，证明了 FF-CAM 的优越性. 同时其在 Part A 数据集上的 *MSE* 提高了 4.5%，说明模型的鲁棒性较强. 但 *MAE* 则略差于 CSRNet[10]，这反映出我们的方法可能需要更多的训练和实验来提高其预测的准确性.

表 3 ShanghaiTech 数据集的估计误差

方法	Part A		Part B	
	MAE	*MSE*	*MAE*	*MSE*
MCNN[9]	110.2	173.2	26.4	41.3
Switch-CNN[7]	90.4	135.0	21.6	33.4
SaCNN[11]	86.8	139.2	16.2	25.8
CSRNet[10]	68.2	115.0	10.6	16.0
FF-CAM	**71.0**	**109.8**	**10.3**	**15.8**

图 5 和图 6 展示了训练好的模型在 ShanghaiTech 数据集上估计得到的部分密度估计图. 可以看出，我们的模型在这两个部分的数据集上都有较好的表现，

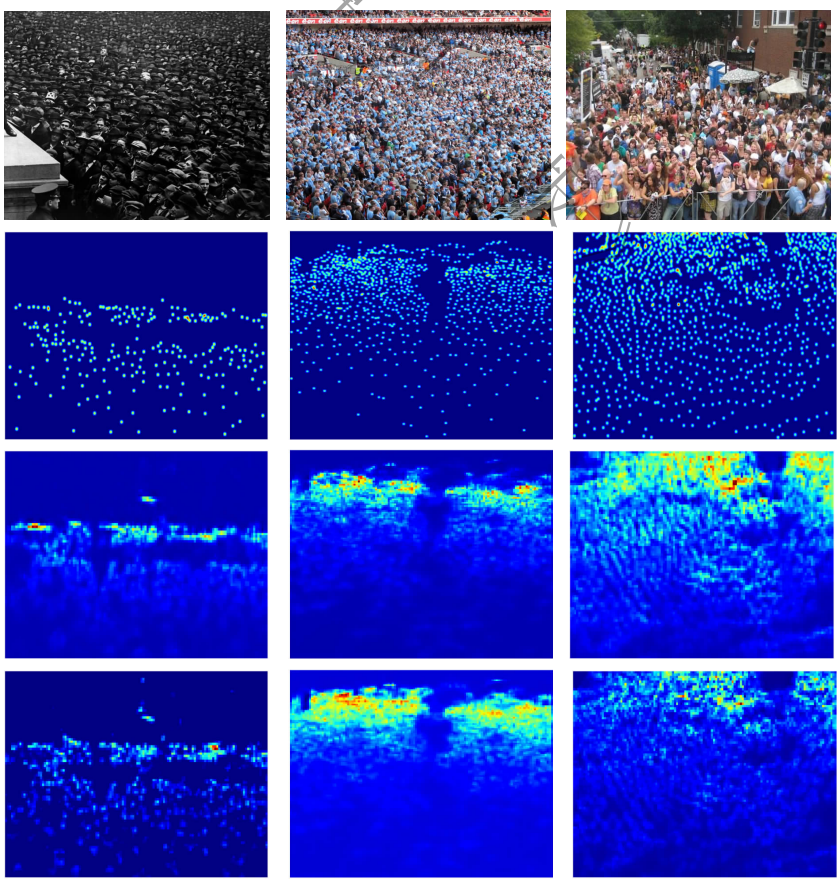


图 5 FF-CAM 模型在 ShanghaiTech A 数据集上的实验对比密度图（第 1 行为原始图像，总人数分别为 239、1005 和 1174；第 2 行为密度图的真值；第 3 行为 CSRNet[10] 得到的密度估计图，预测总人数分别为 379、741 和 1448；第 4 行为 FF-CAM 得到的密度估计图，预测总人数分别为 346、870 和 1402）

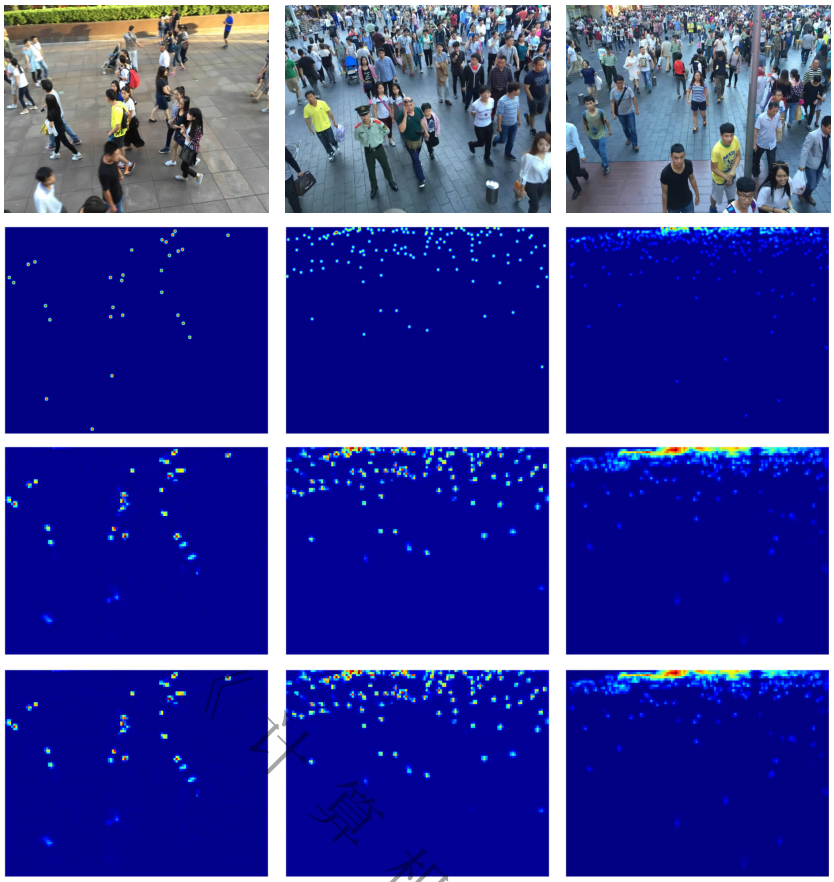


图 6　FF-CAM 模型在 ShanghaiTech B 数据集上的实验对比密度图(第 1 行为原始图像，总人数分别为 28、130 和 467；第 2 行为密度图的真值；第 3 行为 CSRNet[6] 得到的密度估计图，预测总人数分别为 24、117 和 418；第 4 行为 FF-CAM 得到的密度估计图，预测总人数分别为 27、121 和 429)

生成了分布较为准确的密度图，预测的结果更接近于真值，且分辨率也较高. 比较图 5 和图 6，ShanghaiTech Part A 数据集极度拥挤，而 ShanghaiTech Part B 数据集则相对稀疏，这说明在极度拥挤的数据集上我们的网络还需要更多的图片进行训练以提高模型的准确度.

4.5　在 UCF_QNRF 数据集上的实验

UCF_QNRF 数据集由 Idrees[26] 等人提出，同样是一个多样且拥挤的数据集，但图片总数量有 1535 张，人的总数多达 1 251 642，远多于其他两个数据集. 其是从三个不同的数据集来源收集而来，包含了全球各个场景，且同时拥有拥挤和稀疏的人群场景. 我们取 1201 张图片用来训练，剩下的 334 张则用于测试.

表 4 中列出了我们将估计结果的误差 *MAE* 和 *MSE* 与最先进的方法进行比较的结果. 从表中可以看出，我们方法的 *MAE* 提高了 13.3%，这说明预测效果有了明显提升，估计误差较小. 但 *MSE* 则略逊于现有方法，可能是预测结果还不够稳定，存在少量误差较大的图片.

表 4　UCF_QNRF 数据集的估计误差

方法	*MAE*	*MSE*
MCNN[9]	277	426
CMTL[8]	252	514
Switch-CNN[7]	228	445
Idrees 等人[26]	132	191
FF-CAM	**114.5**	**200.5**

图 7 展示了训练好的模型估计得到的部分密度估计图. 可以看出，我们的模型对图 7 后两张的估计值较 Switch-CNN[7] 更为准确，且生成的密度图的分布也更加精准，分辨率更高，这反映出我们模型对拥挤和相对稀疏的场景都能进行很好的预测并生成分布较为准确的密度图，较接近于真值. 同时我们可以看到三幅图都具有房屋和树木的背景干扰，预测生成的密度图则避免了此干扰，进一步验证了模型的抗干扰性. 但是，第一张图的估计值相对真值有一定的偏差，是我们模型的测试中少量的误差较大的图片，这也可以解释模型的 *MSE* 略逊于现有方法. 下一步需要更多的训练来提高模型的鲁棒性，排除大的误差.

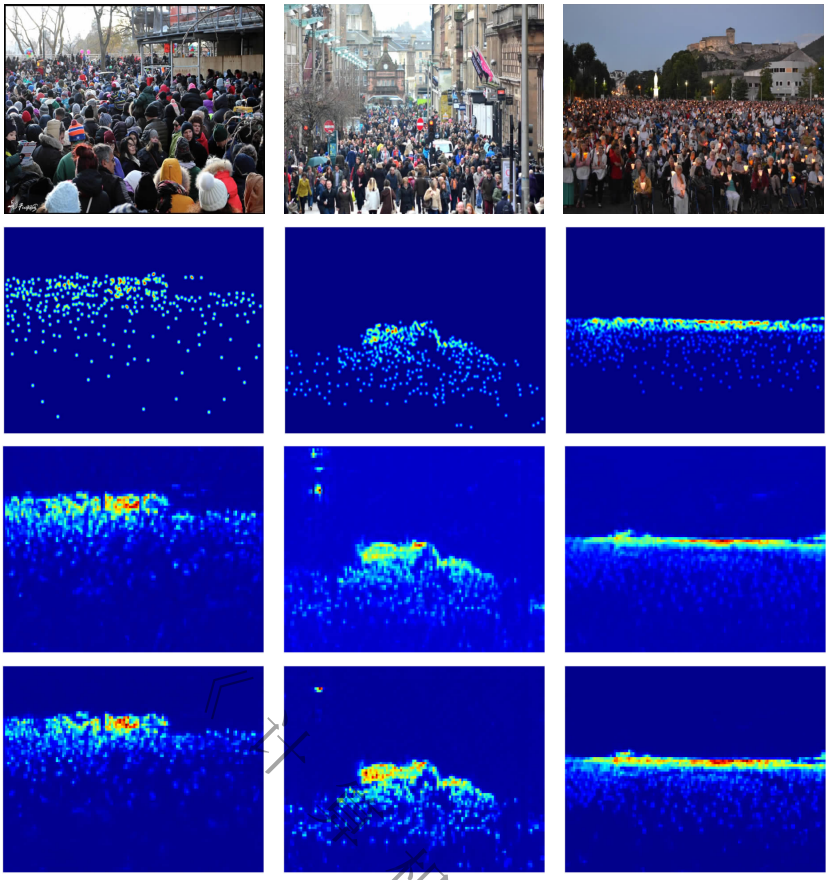


图 7 FF-CAM 模型在 UCF-QNRF 数据集上的实验对比密度图(第 1 行为原始图像,总人数分别为 349、435 和 1017;第 2 行为密度图的真值;第 3 行为 Switch-CNN[7] 得到的密度估计图,预测总人数分别为 365、477 和 1069;第 4 行为 FF-CAM 得到的密度估计图,预测总人数分别为 393、440 和 1017)

4.6 消融实验

我们在 ShanghaiTech Part A 数据集上进行了消融实验来验证 FF-CAM 结构的有效性,图 8 给出了消融实验的结果对比.

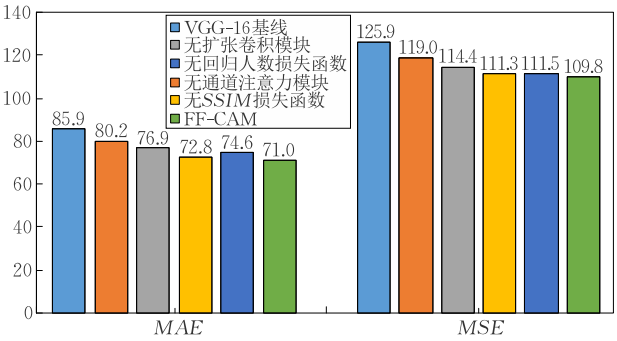


图 8 消融实验结果对比图

我们首先在 VGG-16 基线上进行了训练和测试的实验.从图 8 可以看出,FF-CAM 的估计误差明显优于 VGG-16 基线的结果.与 VGG-16 网络相比,FF-CAM 模型的 *MAE* 提高了 17.3%,*MSE* 提高了 12.7%,证明 FF-CAM 的网络结构很好地提高了预测精度.

随后我们在保持 FF-CAM 的其他结构不变时,分别去掉其中的通道注意力模块,扩张卷积模块,基于 SSIM 的损失函数和基于回归人数的损失函数,进行训练并测试.每一个消融实验得到的 *MAE* 和 *MSE* 的对比如图 8.

在去掉所有的通道注意力模块后,模型的 *MAE* 下降了 11.4%,*MSE* 下降了 7.7%,验证了通道注意力模块对整个模型的增益.

在去掉扩张卷积模块后,模型的 *MAE* 下降了 7.6%,*MSE* 提高了 4.0%,证明了扩张卷积的有效性.

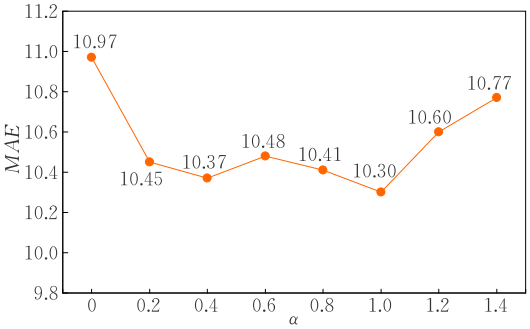
相对于其他模块,基于 *SSIM* 的损失函数和基于回归人数的损失函数对整个模型的影响较小,但去掉后模型的 *MAE* 和 *MSE* 也有所下降,说明其在一定程度上提高了预测精度.具体来说,基于回归人数的损失函数提高效果略高于基于 *SSIM* 的损失函数.

消融实验结果表明,分别去掉各个模块后预测精度都有一定的下降,这说明每个模块都对网络性能有一定的提升作用,验证了我们提出的方法的有

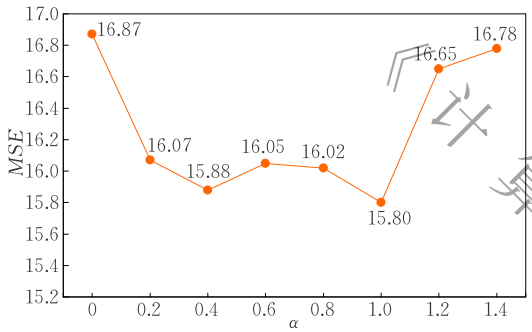
效性和合理性.

4.7　参数实验

我们在 ShanghaiTech Part B 数据集上对综合损失函数中参数 α 的取值进行了消融实验，来验证最优取值的参数有效性，图 9 给出了参数实验的结果对比.



(a) *MAE*结果对比图



(b) *MSE*结果对比图

图 9　参数 α 消融实验的结果对比图(其中，横轴表示 α 的取值变化，纵轴表示评估指标值的变化)

由图 9 可看出，误差评估指标 *MAE* 和 *MSE* 关于不同参数 α 取值的曲线先递减后递增，当 $\alpha=1$ 时误差最小，故取 $\alpha=1$.

5　结　论

在本文中，我们提出了一个用于人群计数的 FF-CAM 框架. 它基于单列网络，只利用单一大小的卷积内核，但性能卓越. 我们提出了主干网络的低层和高层的特征图融合，然后输入到通道注意力模块，最后将得到的特征图馈送到扩张卷积模块中以产生高分辨率的密度图. 我们的 FF-CAM 准确、稳定、简洁，并具有良好的泛化能力. 其在 UCF_CC_50 数据集和 ShanghaiTech Part B 数据集上的测试结果优于现有的方法. 在接下来的工作中，我们将在人群计数的其他公开数据集上进行训练和测试，并与最先进的方法进行比较，以检验我们提出的网络在不同的环境和疏密场景下的性能.

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. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(10): 1986-1998

[2] Idrees H, Saleemi I, Seibert C, Shah M. Multi-source multi-scale counting in extremely dense crowd images//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Oregan, Portland, 2013: 2547-2554

[3] Yang Jianxing, Zhou Yuan, Kung Sun-Yuan. Multi-scale generative adversarial networks for crowd counting//Proceedings of the IEEE International Conference on Pattern Recognition. Beijing, China, 2018: 1051-4651

[4] Olmschenk G, Tang H, Zhu Z. Crowd counting with minimal data using generative adversarial networks for multiple target regression//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe, USA, 2018: 1151-1159

[5] Sindagi V A, Patel V M. Generating high-quality crowd density maps using contextual pyramid CNNs//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 1879-1888

[6] Cao X, Wang Z, Zhao Y, Su F. Scale aggregation network for accurate and efficient crowd counting//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 734-750

[7] Sam D B, Surya S, Babu R V. Switching convolutional neural network for crowd counting//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 6

[8] Sindagi V A, Patel V M. CNN-based cascaded multitask learning of high-level prior and density estimation for crowd counting//Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance. Lecce, Italy, 2017: 1-6

[9] Zhang Y, Zhou D, Chen S, et al. Single image crowd counting via multi-column convolutional neural network//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 589-597

[10] Li Y, Zhang X, Chen D. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1091-1100

[11] Zhang L, Shi M, Chen Q. Crowd counting via scale-adaptive convolutional neural network//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe, USA, 2018: 1113-1121

[12] Shi Miaojing, Yang Zhaohui, Xu Chao, Chen Qijun. Revisiting

perspective information for efficient crowd counting// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 7271-7280

[13] Liu X, van de Weijer J, Bagdanov A D. Leveraging unlabeled data for crowd counting by learning to rank//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7661-7669

[14] Hu Jie, Shen Li, Sun Gang. Squeeze-and-excitation networks //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7132-7141

[15] Dollar P, Wojek C, Schiele B, Perona P. Pedestrian detection: An evaluation of the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(4): 743-761

[16] Paul V, Jones M J. Robust real-time face detection. International Journal of Computer Vision, 2004, 57(2): 137-154

[17] Dalal N, Triggs B. Histograms of oriented gradients for human detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005: 886-893

[18] Felzenszwalb P F, Girshick R B, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645

[19] Chan A B, Vasconcelos N. Bayesian Poisson regression for crowd counting//Proceedings of the IEEE 12th International Conference on Computer Vision. Kyoto, Japan, 2009: 545-551

[20] Lempitsky V, Zisserman A. Learning to count objects in images//Proceedings of the Advances in Neural Information Processing Systems. Cambridge, USA, 2010: 1324-1332

[21] Pham V-Q, Kozakaya T, Yamaguchi O, Okada R. COUNT forest: Co-voting uncertain number of targets using random forest for crowd density estimation//Proceedings of the Computer Vision IEEE International Conference on IEEE Computer Society. Washington, USA, 2015: 3253-3261

[22] Zhang C, Li H, Wang X, Yang X. Cross-scene crowd counting via deep convolutional neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 833-841

[23] Wang Qi, Gao Junyu, Lin Wei, Yuan Yuan. Learning from synthetic data for crowd counting in the wild//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 8190- 8199

[24] Liu W, Salzmann M, Fua P. Context-aware crowd counting// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5099-5108

[25] Liu Chenchen, Weng Xinyu, Mu Yadong. Recurrent attentive zooming for joint crowd counting and precise localization// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1217-1226

[26] Idrees H, Tayyab M, Athrey K, et al. Composition loss for counting, density map estimation and localization in dense crowds//Proceedings of the European Conference on Computer Vision. Berlin, Germany, 2018: 544-559

ZHANG Yu-Qian, M. S. candidate. Her research interests include computer vision, deep learning and information systems engineering.

LI Guo-Hui, Ph. D., professor, Ph. D. supervisor. His research interests include computer vision, information system engineering, data mining and virtual reality technology.

LEI Jun, Ph. D., lecturer. His research interests include computer vision, deep learning, data mining and virtual reality technology.

HE Jia-Yu, M. S. candidate. His research interests include deep learning, data mining and virtual reality technology.

Background

The paper focuses on the crowd counting in single images in computer vision. Nowadays in the top computer vision conferences, novel frameworks are proposed to solve challenges and improve the estimation accuracy on the common datasets. State-of-the-art works are introduced in the paper. Our paper improves the estimation errors on one dataset, superior to state-of-the-art, and the estimation errors on other two datasets also get great results.

The study is supported by the National Natural Science Foundation of China (Nos. 71673293, 61806215).

The two National Natural Science Foundation of China focus on the analysis of crowd group behavior in public complex place. It is the most concerned issue in public security management. However, the crowd behavior in public open area is complex. Crowd behavior is various in different scenarios, thus it is difficult to model them directly. Our previous research found that the crowd behavior could be described by crowd collectiveness when the public safety was

considered. The crowd behavior differences in different scenarios and the difficult problems in unified modeling can be resolved by extracting the general crowd collectiveness. We regard the crowd system in public place as a complex system. From the perspective of visual data observation, we investigate the methods to mine and discover the universal crowd collectiveness in public complex place. In addition, we analyze the evolution of the crowd collectiveness, which present the generation mechanism of crowd event and abnormal behaviors. According to this idea, we define four collectiveness that can be quantitatively descripted and measured: dynamic crowd collectiveness, static crowd collectiveness, conflictiveness in crowd collectiveness, and stability in crowd collectiveness. The collectiveness is measured by the new graph-based learning method. Robust collectiveness map is generated by multi-view learning method. By analyzing the changes of crowd collectiveness in the timeline, the occurrence and evolution of crowd collectiveness event can be represented in spatio-temporal dimension. By analyzing the evolution of crowd collectiveness, we also research the problem of exploring abnormal crowd collectiveness motion and the recognition problem of different crowd collectiveness motions.

This paper collects different crowd scenarios, and can provide the number of people and the density of the scenario in single images. It can help us to analyze the evolution of the crowd collectiveness and the crowd behavior in public open area.