学 校 代 码 10459
学号或申请号 201812362014159
密 级

郑州大学

硕士学位论文

混合连接的跨层级聚合人群计数算法研究

作 者 姓 名：郭 强
导 师 姓 名：叶阳东 教授
学 科 门 类：工 学
专 业 名 称：计算机科学与技术
培 养 院 系：信息工程学院
完 成 时 间：2021 年 5 月

A thesis submitted to

Zhengzhou University

for the degree of Master

Research on Cross-Hierarchy Aggregation Algorithm with Hybrid Connection for Crowd Counting

By Qiang Guo

Supervisor: Prof. Yangdong Ye

Computer Science and Technology

School of Information Engineering

May 2021

摘要

随着世界人口爆炸增长和城镇化进程加快，人群大规模聚集现象屡有发生。高密度的人群聚集极易引发安全事故，进而严重影响社会正常发展秩序。因此，对真实场景下的人群计数算法研究成为计算机视觉领域的热点。人群计数广泛应用于公共区域监控、空间布局规划、道路交通管理和动植物计数等领域。

近年来，深度学习在计算机视觉领域的发展迅速，基于深度卷积神经网络的人群计数算法在计数性能方面提升明显。然而，由于真实场景中存在行人遮挡、背景杂乱、光线变化、分布不均匀和尺度变化等客观因素，人群计数算法性能受到限制。现有的人群计数方法通常采用多列或者多尺度的深度卷积神经网络来学习人群特征，只在网络的末端或者特定的网络层进行特征融合，忽略了不同网络层的特征中丰富且多样的人群尺度信息和空间语义信息。

针对人群图像中背景杂乱和尺度变化问题，为了充分利用不同网络层中的人群特征，有效提取多尺度特征信息和多层级空间语义信息，本文设计并采用混合连接的方式提出一种跨层级聚合网络（Cross-Hierarchy Aggregation Network，CHANet）的人群计数算法。首先，提出跨层级聚合（CHA）模块，采用混合连接的方式重复利用不同网络层中的人群特征并提取局部跨层级特征。CHA 模块可以有效获取人群特征中丰富的多尺度信息，保留浅层网络中的低级空间信息，融合深层网络中的高级语义信息，保证网络中信息最大量流动。使用1×1的卷积自适应地聚合局部跨层级特征，提高特征的表示能力，同时降低模型的参数量。其次，构造前端网络，采用预先训练的卷积神经网络提取人群图像的全局层级特征，具有较强的特征提取能力，加快模型训练的速度，有效缓解因数据样本不足导致的模型过拟合等问题。然后，设计后端解码器，采用转置卷积进行上采样来输出高质量的人群密度图。后端解码器进一步融合来自前端网络的全局层级特征和 CHA 模块的局部跨层级特征以获取人群特征的多尺度信息和空间语义信息，提高模型的计数性能。

本文在四个当前主流的人群计数数据集上进行实验评估。结果表明，与当前最先进的方法相比，提出的 CHANet 能够生成更高质量的密度图，具有更优越的计数性能和更好的泛化性能。模型的复杂度分析和消融实验分析验证了模型设计的合理性和有效性。

关键词：人群计数 卷积神经网络 混合连接 跨层级聚合 泛化性能

Abstract

With the explosive growth of the world's population and the acceleration of urbanization, large-scale crowd gatherings have occurred frequently. High-density crowds can easily cause safety accidents, which will seriously affect the normal development of society. Therefore, the research on crowd counting algorithms in real scenes has become a hot spot in the field of computer vision. Crowd counting is widely used in public area monitoring, spatial layout planning, road traffic management, and animal and plant counting.

In recent years, deep learning has developed rapidly in the field of computer vision, and crowd counting algorithms based on deep convolutional neural networks have significantly improved the counting performance. However, due to objective factors such as pedestrian occlusion, background clutter, light changes, uneven distribution, and scale variations in real scenes, the performance of crowd counting algorithms is limited. Existing crowd counting methods generally use multi-column or multi-scale deep convolutional neural networks to learn crowd features, and only fuse features at the end of the network or specific layers in the network, ignoring that features from different layers of the network contain rich and diverse scale information and spatial semantic information.

Aiming at the problems of background clutter and scale variations in crowd images, in order to make full use of the crowd features in different hierarchies of the network and effectively extract multi-scale information of the crowd features and multi-level spatial semantic information, this paper designs and adopts a hybrid connection to propose a novel crowd counting algorithm called cross-hierarchy aggregation network (CHANet). Firstly, a cross-hierarchy aggregation (CHA) module is proposed to reuse crowd features from different hierarchies of the network and extract local cross-hierarchy features by using the hybrid connection. The CHA module can effectively obtain the rich multi-scale information from different hierarchies, retain the low-level spatial information in the shallow hierarchies, fuse the high-level semantic information in the deep hierarchies, and ensure the maximum flow of information in the network. A 1×1 convolution is adopted to adaptively aggregate local cross-hierarchy features to improve the ability of the feature representation and reduce the amount of model parameters. Next, the front-end network is constructed to extract the global hierarchical

features of the crowd images with a pre-trained convolutional neural network. The front-end network has strong feature extraction capabilities, accelerates the speed of model training, and effectively alleviates the problem of model overfitting caused by insufficient data samples. Then, the back-end decoder is designed with transposed convolutions for up-sampling and generating high-quality density maps. The back-end decoder further fuses the global hierarchical features from the front-end network and the local cross-hierarchy features from the CHA module to obtain multi-scale information and spatial semantic information, thereby improving the counting performance of the model.

This paper conducts experimental evaluation on four current mainstream crowd counting datasets. The results show that, compared with the current state-of-the-art methods, the proposed CHANet can generate higher-quality density maps and achieve superior counting performance, as well as better generalization performance. The model's complexity analysis and ablation experiment analysis verify the rationality and effectiveness of the model design.

Key Words: crowd counting, convolutional neural network, hybrid connection, cross-hierarchy aggregation, generalization ability

目录

图目录

表目录

1　绪论

1.1　研究背景及意义

科学技术的不断进步推动人类经济社会高质量发展，城市化进程持续加快。都市圈、城市群和中心城市的发展导致了大规模人口的区域性聚集。同时，随着医疗、卫生等领域的发展和基础设施的不断完善，世界人口仍在急剧增加。越来越多的人前往广场、体育场等公共场所参加集会庆典、观看演唱会和体育赛事等大型活动，导致大规模人群聚集的现象频繁发生，如图 1.1 所示。

在景区、车站、超市和学校等公共场所中，人群密度过高将会带来很大的安全隐患。国内外因人群大规模聚集而导致的人群踩踏事故均有发生，例如，上海外滩踩踏事故，沙特米纳地区朝觐者踩踏事故和云南昆明明通小学踩踏事故等等。事故发生的客观原因在于聚集的人群数量过于庞大，超过了场所承受的最高人群密度，导致在出现意外情况时不能及时处理和疏散人群。这些问题已经引起了有关部门的高度重视。研究这些事故发生的原因发现，如果某个场所内的人群数量或密度超过了该区域的安全阈值后，就会导致人群拥挤，极易引发摔倒从而造成踩踏事故，人员伤亡的危险等级将会大大提高。这严重危害了人民生命财产安全，阻碍了城市经济发展和人们正常的生产生活。

图 1.1　公共场景中的人群图像

我国监控设备的覆盖率非常高，绝大多数路口、广场、超市、学校等公共区域均安装有公共安全视频监控设备。但是目前几乎没有实际应用的人群计数系统，

不能依靠计算机对监控捕捉的图像进行实时数据分析，监控所在区域的人群数量和密度。依靠监控人员对监控获得的视频进行判断分析，不仅消耗巨大的人力物力，同时还做不到快速、准确、有效。因此，建立有效的人群计数系统，准确地估计人群聚集频发区域内的人群数量和密度分布情况，可以指导有关部门及时有效地采取措施，控制区域内的人群数量，防止人群密度过大，对安全预警、降低安全隐患、避免踩踏事故具有十分重要的现实意义。

人群数量和密度分布预测主要可以应用于以下场景或领域：

（1）公共区域监控。针对频发的人群聚集区域，如体育场、演唱会、超市商场、车站、广场等；以及人群聚集的事件，如重大节日庆典、游行示威、旅游、集会等。这些区域和事件都会出现人群数量庞大和人群密度暴增的现象，存在巨大的安全隐患，极易导致恶性群体事故的发生。通过人群计数和密度分布预测，在人群数量或密度超过安全阈值时及时向有关部门预警并有序疏导人群，能够有效地避免人群拥挤诱发的恶性事件发生。

（2）空间布局规划。对城市重点区域如商场、学校、酒店、小区、车站、机场和交通道路等人群流量和密度分布进行分析预测，能够更加合理地规划城市发展布局，制定城市土地开发利用政策，有力支撑城市建设和经济发展。例如：对大型超市内部，通过分析购物区域的人群流量和密度分布情况，能够帮助超市管理者掌握消费者的购物偏好，制定合理的销售策略，优化产品和广告投放，配置相应的服务人员，提高超市的销售额。对车站、机场等重点交通区域内，通过对内部人群流量和密度分布的分析，能够更加合理地调整进出口、售票区和等候区的位置，更好的服务群众。同时，对不同等候区域的人数统计，还能指导优化铁路和航班运行班次。

（3）公共交通管理。人群计数方法可以应用到其他领域中，比如对道路上的车辆计数。我国公路网密集，道路交通情况复杂。截至 2019 年末，我国公路总里程 501.25 万公里[1]。同时，截止 2020 年 6 月，我国机动车保有量 3.6 亿辆[2]。道路行驶车辆密集，不仅导致交通拥堵，阻碍城市发展，更导致交通事故频发，危害人民生命安全。通过道路监控设备对行驶车辆进行计数，有利于车辆有序、平稳行驶，构建顺畅、安全的道路交通。

（4）动植物计数。人群计数方法还可以应用到动植物计数领域。在医疗生物领域，可以将人群计数方法应用到细胞计数等来统计细胞数量；在农业领域，可以对玉米、小麦等农作物进行计数来预测收成；在养殖业领域，可以对牛、羊等禽畜进行计数来掌握实时数据；在野外，可以对野生动物计数来监测野外动物的活动情况和野外环境的变化。

[1] https://www.sohu.com/a/407511407_114835
[2] https://www.chyxx.com/industry/202007/883992.html

随着深度学习（Deep Learning，DL）[1]在计算机视觉中的不断发展和应用，新颖的方法和应用不断涌现，目标检测、语义分割和图像分类等领域出现了性能优异的深度神经网络模型。人群计数领域也涌现了许多基于深度神经网络的模型和方法，相比于传统人群计数方法取得了显著的计数性能。神经网络舍弃了早先的手工特征，能够从训练数据中提取图像的抽象信息，学习到图像的特征表示。相比于手工特征，深度特征能够更好的表示图像特征。使用深度神经网络提取特征打破人工设计特征的局限性，为人群计数领域的研究发展注入新的活力。

1.2 研究现状与挑战

人群计数是人群分析领域中最重要的研究课题之一[2]，其中涉及到计算机视觉、图像处理和机器学习等多个领域的知识。人群计数的任务是对捕获的视频序列或单张图像中的人群数量进行预测。随着人工智能时代的来临，视频监控技术不断升级和发展，在社会生产和生活中扮演者越来越重要的角色。全球人口不断增加，人群聚集现象日益凸显，研究人群计数方法在智能监控领域中的应用可以有效地建立人群拥挤预警机制，快速预测公共场所内的密集人群数量，避免意外事故的发生。作为计算机视觉领域重要的科研分支，人群计数方法有着十分重要的研究价值，受到了很多研究者的关注[3][4][5][6]。它的研究领域也扩展到了多个分支，如：人群异常检测、公共安全监测、城市布局规划等，具有广泛的实际应用与商业前景。

在人群计数方法提出之前，人群数量统计借助闭路电视对某一指定的场景区域进行监视，然后依靠监控人员根据场景情况，结合经验对场景中的人群数量做出大致的估计。这种方法具有很强的主观性，既耗费人力物力，又不能快速准确地估计人群数量。传统的人群计数方法先从监控视频中对图像进行场景分割，然后人工提取特征，使用统计方法或回归方法来计算视频监控图像中的人群数量。但是，场景分割和特征提取都是由人工设定的，传统的方法不能很好的学习图像中的特征表示和语义信息，人群计数的性能受到限制。近些年，传统的人群计数方法在处理人群图像中行人轮廓不一、尺度大小变化和光线、天气因素导致的复杂环境背景等问题上难以找到有效手段，在计数性能上鲜有提升。因此，传统的人群计数方法很少出现里程碑式的研究成果。

随着智能化视频监控技术[7]的提出和发展，国内外的部分科研院校和机构对其做了很多研究[8][9][10]。2010 年，基于密度图的人群计数方法最早由 Lempitsky 和 Zisserman[11]在神经信息处理系统会议（NeurIPS）上提出，这一类的计数方法能够学习到图像特征和对应密度图之间的映射，以人群密度来计算人群数量。深度学习方法在目标检测、图像分割、模式识别等计算机视觉领域的成功应用，促

使人群计数领域研究者的视线从传统的人群计数方法转到基于深度学习的人群计数方法研究上。深度神经网络能够在前向传播的过程中学习提取数据的特征表示，并在反向传播的过程中更新模型参数，使得网络学习到需要的特征表示，具有更强的表示能力。在计算机视觉、自然语言处理、信息检索等领域，深度神经网络都取得了巨大的研究进展。除了计算机领域，在医学、经济学中对数据的处理也用到了深度神经网络。在深度神经网络模型中，卷积神经网络（Convolutional Neural Network，CNN）[12]是处理图像领域最成功的方法之一，越来越多基于 CNN 的网络模型被提出，如：VGGNet[13]、ResNet[14]和 DenseNet[15]等。这些网络模型结构灵活多变，能够很好地移植到其他模型方法中，并迁移应用到图像分类、目标检测和图像分割等众多领域。借着深度学习的热潮，大部分研究者沿用了基于密度图的方法，并提出了一批计数性能优异的深度网络模型，人群计数一度成为计算机视觉领域的热点问题[16]。国内外的工业界和学术界都对人群计数进行了深入的研究和探索，如：中国科学院大学、厦门大学、香港城市大学、上海交通大学、哥伦比亚大学、约翰斯·霍普金斯大学和腾讯、百度、海康威视、商汤科技等等。人群计数领域的文章越来越多地出现在国际顶级的会议上和权威的期刊中，如：国际计算机视觉与模式识别会议（CVPR）、国际计算机视觉大会（ICCV）、国际人工智能大会（AAAI）和模式分析与机器智能期刊（TPAMI）、计算机视觉国际期刊（IJCV）、神经网络与学习系统会刊（TNNLS）等。

本文统计了自 2015 年以来国际顶级会议录用和权威期刊收录的人群计数领域文章的数量，如图 1.2 所示。可以看到，人群计数领域发表的论文数量呈现出明显的上升趋势，尤其是近两年的论文数量爆炸式增长。同时，图 1.3 中统计了国际顶级会议和权威期刊中使用传统的人群计数方法和基于深度学习的人群计数方法的论文数量所占比重。由于深度学习的异军突起，基于深度学习的人群计数方法取得了优越的计数性能，显现出了十分具有研究价值的应用前景。

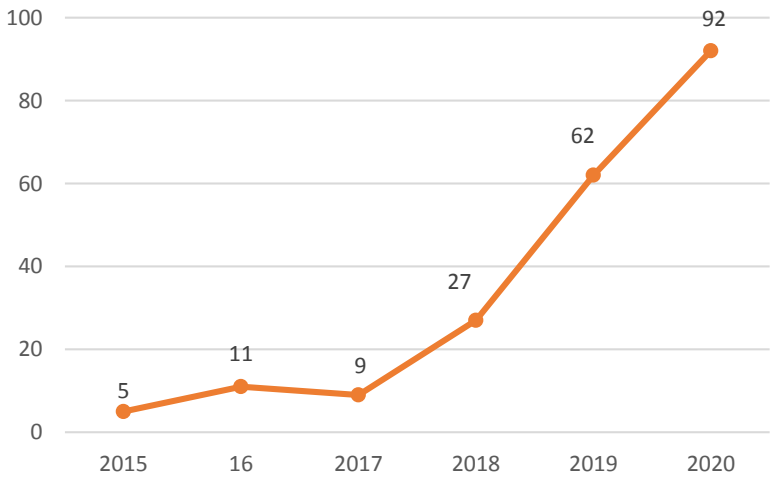


图 1.2 顶级会议和权威期刊中人群计数文章统计

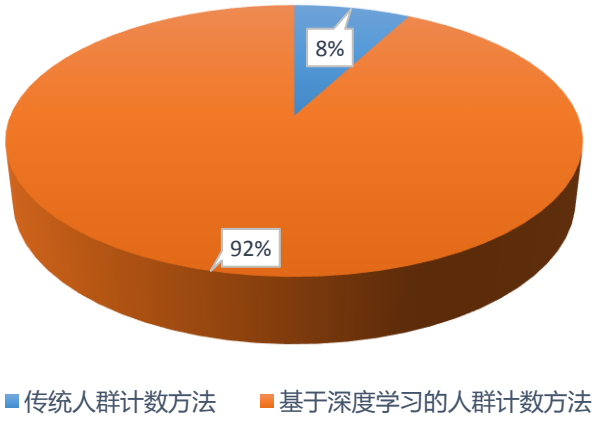


图 1.3　顶级会议和权威期刊中不同人群计数方法占比

正常的生产生活过程中，人们不可避免的出现聚集现象。因此，人群计数任务应运而生，应用于社会各个方面，并随着科学技术的不断发展而涌现出更多具有强大计数性能的人群计数方法。然而，即使对人群计数领域相关方法的研究已经取得了众多令人欣喜的成果，在实际的人群场景中，人群计数仍然面临以下几个方面的挑战需要克服：

（1）行人遮挡。由于硬件设备和拍摄角度等原因，从监控设备获取的视频序列和图像都是二维的人群场景，当人群发生聚集时，人与人之间的距离非常近，相邻的行人之间会存在躯体遮挡、覆盖等现象，导致不能正确识别被遮挡的行人，影响最终的计数结果。尤其是在特别密集的人群场景中，靠肉眼进行观察计数时，图像远景处的行人会被近景处的行人遮挡大小不一的躯体，导致行人轮廓不完整、形状不同，从外观很难辨别出图像中某一位置是否存在行人。因此，行人遮挡大大增加了人群计数的难度，始终影响人群计数方法的计数性能。虽然已经有研究者针对行人遮挡问题提出了解决方案，但是结果仍然不够理想。

（2）背景杂乱。不同区域、不同位置的摄像装置拍摄出来的人群聚集场景各不相同。例如：在公园、旅游景区、野外等地点，拍摄出来的人群背景可能存在树木、花草、湖泊等；在城市街道、广场等区域，拍摄出来的人群背景可能有高楼建筑、车辆、广告牌等；在演唱会、节日庆典、游行示威等场合，拍摄出来的人群背景会出现标语、横幅、旗帜、座椅等。这些不同的背景特征存在很大的差异，在训练数据受限的情况下，人群计数方法很难区分和学习不同的人群背景特征，进而出现识别混乱的现象，在没有行人的背景位置识别出行人。除了拍摄场景不同导致人群背景杂乱外，不同的天气环境同样严重影响人群计数方法的计数性能。在雨、雪、雾等特殊气象条件下，人群图像将会出现严重的模糊现象，对人群计数方法是一个巨大挑战。

（3）光线变化。在实际的社会生产生活中，人群计数还面临着复杂光线变化的影响。人群图像在拍摄过程中，受拍摄时间、天气、地点等因素的影响，太阳光照出现变化，对于人群图像特征产生巨大改变。例如：白天和晚上、晴天和阴天、阳光下和阴影中等。此外，在演唱会、商场、文艺晚会等有灯光照射的场所，对于人群在图像中的外观和色彩等方面产生不可估计的影响。背光拍摄或对光拍摄也会对收集的人群图像数据造成影响。复杂的光线变化，需要人群计数方法具有良好的适应性和鲁棒性。

（4）分布不均匀。人群聚集现象具有随机性、不规律性。在人群聚集的场景中，某些区域人群十分密集，其他区域可能几乎没人。例如：在体育场内，四周坐满了密密麻麻的观众，而中心区域只有参加比赛的运动员；在游行、集会等场合，广场、道路等空旷地带挤满了人群，而周围的建筑、树木位置不可能有人。在这些人群图像中，人群密度分布极为不均匀，对于人群计数方法而言很难学习到准确的特征表示，生成的人群密度图出现异常情况，导致预测的人群数量与真实数量存在偏差，影响人群计数性能。

（5）尺度变化。由于摄像装置安装的角度、倾斜的幅度和高低远近位置不同，从三维立体空间拍摄获取的二维人群视频序列或图像势必受到摄像装置透视效果的干扰。具体表现为在获取的一张人群图像中，由于摄像装置透视形变问题的存在，致使处在不同景深位置的行人在图像中成像时所占的像素面积大小不一样，导致获得的表示行人特征的向量也不一样。透视效果给人群图像带来的问题表现为人群尺度大小不同，远离摄像装置的图像位置，人群轮廓变得很小，显得十分密集，所占的图像区域反而不大，靠近摄像装置的图像位置，人群轮廓非常清晰，视觉上会相对稀疏，所占的图像区域相对较大。人群尺度变化是由于摄像硬件装置不可避免地带来的问题，是人群计数领域一直研究的重难点问题。如何适应不同场景、同一场景的不同区域所带来的人群尺度变化，对于提升人群计数性能具有十分重要的影响。

为了克服人群计数中面临的这些挑战，研究者从数据集预处理、图像特征表示和模型框架设计等多个方向，使用不同的方法策略，提出了很多有效的人群计数方法。

1.3 本文研究内容

本文主要针对人群计数中背景杂乱和尺度变化问题进行研究，结合现有的使用多列和多尺度的特征融合方法来解决人群尺度变化问题的思路，设计混合连接的方式，提出一种跨层级聚合网络（Cross-Hierarchy Aggregation Network，CHANet）的人群计数算法。本文的主要研究内容如下：

（1）针对当前人群计数领域面临的复杂问题，特别是对于背景杂乱和尺度变化问题，为了提取网络中不同尺度的人群特征信息和丰富的空间语义信息，本文采用稠密连接和残差连接设计一种混合连接的方式，并使用混合连接的方式提出一种跨层级聚合（CHA）模块。采用稠密连接将当前网络层的特征和前面网络层的特征进行聚合，提取不同网络层中的局部跨层级特征，获取特征中的多尺度信息；采用残差连接跳跃地将前端网络层中的信息和当前网络层的信息融合，丰富网络中信息的流动。混合连接的方式保证了模型可以重复利用不同网络层的特征，保留了浅层网络中的低级空间信息，融合了深层网络中的高级语义信息，保证网络中信息最大量流动，有效提取网络中的多尺度特征信息和多层级空间语义信息，使得模型具有强大的人群特征表示能力，提高人群计数性能。

（2）模型的前端网络采用具有强大特征提取能力和迁移性能的 VGG-16[13] 网络来提取人群图像的全局层级特征。使用在 ImageNet[17]上预先训练的网络参数来初始化前端网络，从而保证前端网络在模型训练阶段具有较强的特征提取能力，同时加快了模型训练的速度，有效缓解了训练过程中因数据样本不足导致的模型过拟合等问题，提高了模型的性能。

（3）模型的后端解码器采用转置卷积进行上采样来生成高质量的人群密度图。转置卷积具有可训练的网络参数，和普通卷积有着相似的本质，网络能够根据学习到的特征进行参数优化，更合理地恢复图像或特征的像素值，还原成输入图像的尺寸大小。后端解码器进一步融合来自前端网络的全局层级特征和 CHA 模块的局部跨层级特征以获取人群特征的多尺度信息和空间语义信息，提高密度图的质量。

（4）采用高斯滤波器的方法将数据集中的原始标签生成真实密度图来训练网络模型。目前主流的人群计数数据集大多采用点标注的方法标注人群图像，将图像中人头中心位置坐标的像素点值标记为 1，代表一个人，未标记的像素点值为 0。这种点标注的方法将图像中人群数量离散为单个的点，网络模型很难学习到人群的特征表示。使用高斯核函数将标注的像素点进行模糊化处理，把离散的标注点转换成热力图形式的连续密度图，从而使网络模型在训练过程中学习到更多的特征表示，提高模型的计数效果。

（5）在四个主流的人群计数数据集上进行了大量充分的实验。实验结果表明本文提出的模型具有更优越的计数性能，能够生成高质量的人群密度图。模型泛化实验的结果表明模型具有较强的泛化性能，能够适应不同场景下的计数任务。对模型复杂度进行对比分析，结果表明模型在合理计算量和参数量的基础上，计数性能得到了显著提升，验证了模型设计的合理性。对模型进行消融实验分析，结果验证了模型设计的有效性。

1.4 本文组织结构

本文共分为五个章节，具体内容安排如下：

第一章：绪论。首先，简要阐述了人群计数领域的研究背景、应用领域和意义。然后，对当前人群计数领域的研究现状做了简单的介绍，并分析了目前人群计数领域面临的挑战。接着，介绍了本文的主要研究内容和贡献。最后，给出全文的组织结构。

第二章：相关工作。首先，介绍了深度学习的基础知识，包括卷积神经网络、网络参数优化和常用的深度学习框架。然后，简要概述了现有的人群计数方法的分类，并回顾了其中主要的人群计数模型。最后，介绍了卷积神经网络在人群计数中的应用。

第三章：跨层级聚合网络的人群计数模型。首先，给出 CHANet 的研究动机。然后，详细介绍了 CHANet 的网络框架、核心模块和后端解码器。接着，阐述了本文生成真实密度图的方法和原理。最后，给出 CHANet 在训练过程中使用的损失函数。

第四章：实验结果及分析。首先，介绍了人群计数算法的评价标准。然后，给出本文的实验设置，包括模型训练细节、数据集和数据扩充方法。接着，通过实验对 CHANet 的计数性能和生成密度图质量进行了对比分析，并给出了可视化的计数结果。最后，对 CHANet 的泛化性能和复杂度进行了对比分析，并通过消融实验验证了模型的有效性。

第五章：总结和展望。总结本文所做的核心工作，并展望了人群计数未来的研究方向和内容。

2 相关工作

2.1 深度学习

由于深度神经网络在计算机视觉等领域的成功应用，其在特征提取和表示上具有远超传统手工特征的优势，受到了研究者的青睐。近年来，使用深度学习特别是基于卷积神经网络的人群计数方法层出不穷，展现了优异的计数性能。本文提出的 CHANet 同样采样卷积神经网络设计实现。本节介绍了深度学习中卷积神经网络、网络参数优化和深度学习框架等基础知识。

2.1.1 卷积神经网络

为了处理网格类结构的数据，卷积神经网络这种特殊的神经网络被提出，在图像处理方面具有无与伦比的优势[18]。卷积神经网络模仿生物视觉的神经系统机制设计构建，它具有三个重要的特性，分别是稀疏交互（Sparse Interactions）、参数共享（Parameter Sharing）和等变表示（Equivariant Representations）。常见的卷积神经网络中主要含有以下几个部分：输入输出、卷积层（Convolutional Layer）、激活函数（Activation Function）和池化层（Pooling Layer）等。

（1）输入输出。卷积神经网络的输入可以是多维数组形式的数据，如一维向量，二维矩阵等。在单幅图像的人群计数中，不考虑图像通道数的情况下，输入的是二维矩阵数据，输出的是预测的人群密度图。

（2）卷积层。作为卷积神经网络中的最重要的组成部分，卷积层的设计思想来源于数字信号处理中的卷积运算。本质上，卷积层的卷积操作是一种加权的线性运算。卷积核（Kernel），又叫滤波器（Filter），是卷积层中执行卷积操作的关键组件。卷积核的大小固定了对特征进行采样操作的范围，一般来说远小于输入特征的尺寸大小。卷积层中卷积核的个数越多，输出的特征通道数越大。在卷积过程中，卷积核在输入数据上平滑移动并进行运算，每次运算后滑动的长度称为步长（Stride）。有时为了保证输出特征的尺寸大小，需要在输入特征四周填充尺寸大小不同的随机数值来对边缘特征进行采样，称为填充（Padding），一般使用0来填充。图 2.1 展示的是卷积层的运算过程：输入是一个5×5的二维矩阵，卷积核大小 k 是3×3，步长 s 为2，填充大小 p 为1，得到的输出矩阵尺寸为3×3。输出尺寸的计算公式如下所示：

$$S_{out} = \frac{S_{in} + 2p - k}{s} + 1 \tag{2.1}$$

其中S_{in}和S_{out}分别表示输入和输出特征的尺寸大小。

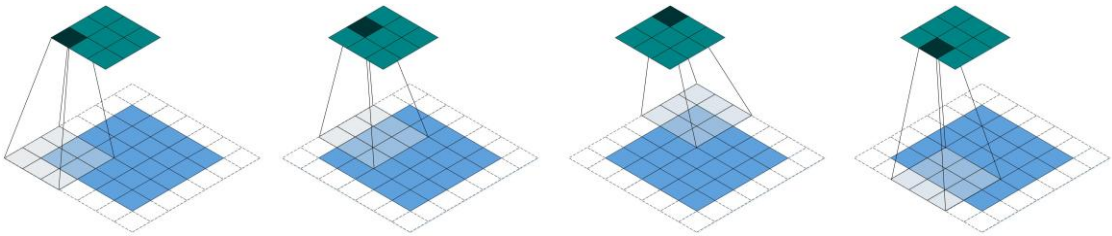


图 2.1　卷积层运算过程

（3）激活函数。在卷积神经网络中，卷积运算只具备线性运算的能力，每一层输出的特征和输入的特征都是线性关系。这样无论网络有多少层，都是多个线性的组合，理论上和一个卷积层的线性运算是等效的，网络的表征能力大大降低，很难拟合非线性问题。因此，在卷积神经网络中，非线性的激活函数同样不可缺少。Sigmoid、Tanh、ReLU 等非线性函数是常见的激活函数。

本文中使用 ReLU 作为 CHANet 中的激活函数，其数学公式如下所示。从图 2.2 的函数图像可以看出，ReLU 既符合非线性的要求，又在一定取值范围内具有线性的特点，使得 ReLU 相比于 Sigmoid 和 Tanh 函数计算简单，加快了收敛速度，解决了梯度消失的问题。

$$ReLU(x) = max(0, x) \tag{2.2}$$

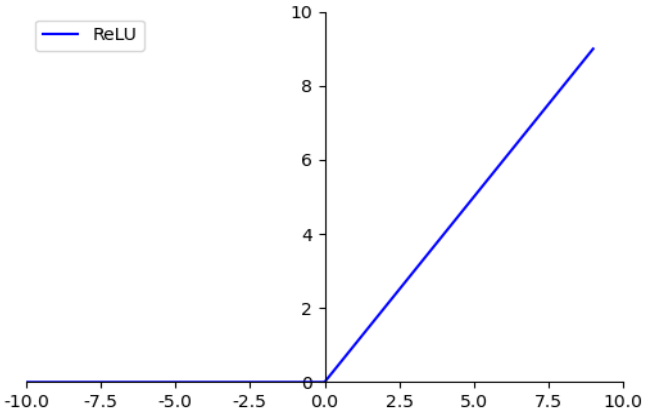


图 2.2　ReLU 函数图像

（4）池化层。也称作下采样层。池化层能够减小特征的空间尺寸并保留特征重要的空间信息，从而降低网络中的参数量并减缓过拟合现象，是卷积神经网络中极为重要的一部分。因此，池化层经常出现在卷积神经网络中。常用的池化层运算有两种，分别是最大池化和平均池化。如图 2.3 所示，最大池化是对池化窗口区域内的特征值取最大值，平均池化则是求平均值。图中池化窗口的大小为 2×2，步长为2。

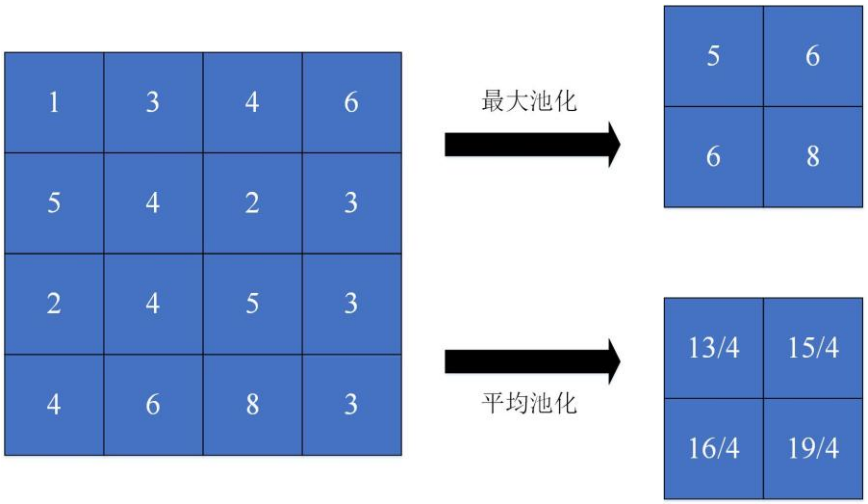


图 2.3　池化层运算过程

2.1.2　网络参数优化

深度学习算法在大多数情况下都涉及到优化，深度神经网络的训练过程就是一个模型参数优化的过程。模型参数优化是指在网络的训练过程中对网络中的可训练参数进行调整，使得网络能够更好的拟合训练数据，输出的预测值更加接近真实值。通过反向传播，深度神经网络计算网络中可训练参数的梯度，使用基于梯度的优化算法学习网络中的参数。优化算法从某一起点出发，沿着梯度方向向下寻找最优的参数来最小化网络的损失函数。优化的起点即网络初始的参数值，在一定程度上影响着网络的收敛性能和训练速度。深度学习中常用的优化算法有随机梯度下降（Stochastic Gradient Descent，SGD）、动量（Momentum）、AdaGrad[19]、AdaDelta[20]、Adam[21]等。

本文使用 Adam 优化算法来优化网络模型的参数。Adam 是当前深度学习领域中广泛使用的参数优化算法之一，结合了动量和自适应学习率的算法。Adam 算法记录了梯度的一阶矩和二阶矩，并用来动态地调整每个参数的学习率。它的主要优点是加入了偏置矫正机制，每一次迭代学习率都处于一个确定的区间内，从而使参数更新较为平稳。

2.1.3　深度学习框架

越来越多的研究者致力于探索和研究深度学习领域，促使深度学习框架的开发、应用和发展。同时，深度学习框架的不断更新也极大促进了深度学习领域的研究。目前，深度学习领域的研究者使用的深度学习框架不尽相同，如 Caffe[22]、TensorFlow[23]、Keras[24]、PyTorch[25]等等，如图 2.4 所示[26]。这些深度学习框架具有完整的使用文档、简洁的代码结构、丰富的算法模块，在计算机视觉、自然

语言处理、生物医学等领域取得了成功的应用。

图 2.4　常见深度学习框架

本文提出的网络模型在 PyTorch 框架上设计实现。Facebook 的人工智能研究院（FAIR）在 2017 年开源了 PyTorch 深度学习框架。PyTorch 基于动态图设计，具有强大的 GPU 加速的张量计算，同时包含自动求导的深度神经网络。在灵活性、易用性和速度三个方面，PyTorch 在目前的深度学习框中达到了较高水平。PyTorch 在设计上力求最少的封装，符合人类思维模式，代码易于理解。这样的设计能够让使用者更好的实现自己的想法，不用被框架本身所限制。同时，PyTorch 有着十分活跃的社区，使用者可以进行提问和交流。

2.2　人群计数方法

由于人群计数方法的研究有着十分重要的现实意义，越来越多的研究者致力于探索和研究更加实用、有效、快速、准确的人群计数方法。随着更多新的人群计数方法的提出和发展，人群计数方法种类逐渐增多，对人群计数方法的研究也更加深入。研究者从不同的角度、不同的背景知识，采用不同的理论方法和解决思路为人群计数方法的研究提供了更多研究方向。本节根据当前人群计数方法的研究成果，将这些方法划分为三类：基于检测的人群计数、基于回归的人群计数和基于密度图的人群计数，并逐一进行介绍。

2.2.1　基于检测的人群计数

在研究人群计数的早期阶段，大部分人群计数的研究工作聚焦于基于检测的方法。使用一个滑动窗口检测器来检测场景中行人的躯体轮廓，并统计检测框的

个数来计算相应的人群数量，如图 2.5 所示[27]。

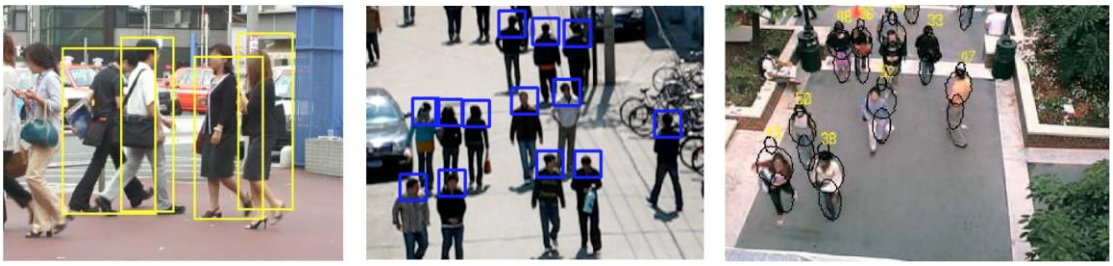


图 2.5　基于检测的人群计数

基于检测的方法根据检测部位不同又可以分为基于整体的检测和基于局部的检测。基于整体的检测方法首先提取整个图像的特征，然后使用从行人整体提取的边缘特征[28]、Shapelet 特征[29]、Haar 特征[30]、HOG 特征[31]等去检测行人。常用的算法分类器有 SVM[32]，boosting[33]，随机森林[34]等。基于整体的检测方法在计数对象稀疏时取得了较好的性能，但在面对密集人群时，由于人群遮挡覆盖，计数效果会明显下降。研究者开始探索密集人群场景下的有效计数方法。在大多数密集人群场景中，行人被遮挡的区域往往是躯干和四肢部位，而头部、肩膀等位置不容易被遮挡。因此，基于局部的检测方法[35][36][37]被提出用来处理密集人群计数问题。这种方法相比基于整体的检测方法，在效果上有略微的提升。

基于检测的方法强烈依赖于目标的特征。此外，采用滑动窗口进行目标检测的算法会消耗大量计算成本。滑动窗口在图像中移动检测，如果为了得到细粒度特征信息，就必须使用较小的步幅来移动窗口，导致计算量大大增加；如果窗口一次移动的步幅过大，虽然会降低计算量，但会造成检测的图像细节信息丢失，影响计数性能。

值得一提的是，近两年仍然有研究者致力于探索基于检测的方法并取得了很好的结果。Laradji 等人[38]提出了一种不需要估计检测对象大小和尺寸的计数方法，设计的损失函数能够激励网络仅使用点标注就能输出每个对象实体的检测块。Liu 等人[39]提出了一个深度检测网络模型，使用点标注数据来检测行人头部的大小和位置，然后计算人群数量，避免了边界框昂贵的标记成本。

当计算高密度人群数量时，由于行人间严重的遮挡，检测器很难训练。在这种情况下，基于检测的方法性能下降明显。基于回归的方法避免了对检测器的依赖，通常具有更高的性能，在人群计数中受到越来越多的关注。

2.2.2　基于回归的人群计数

基于检测的方法性能经常受到人群严重遮挡等问题的限制。为了解决这个问题，研究者提出了基于回归的人群计数方法[40][41][42][43]。与基于检测的方法相比，

基于回归的方法进一步提高了计数性能。这类方法不需要找到图像中行人的具体位置，而是首先定义感兴趣区域（Region of Interest，ROI），确定场景的透视图，然后提取图像中的前景、边缘、纹理等特征，从原始特征学习到对应人群数量的映射关系，如图 2.6 所示[27]。

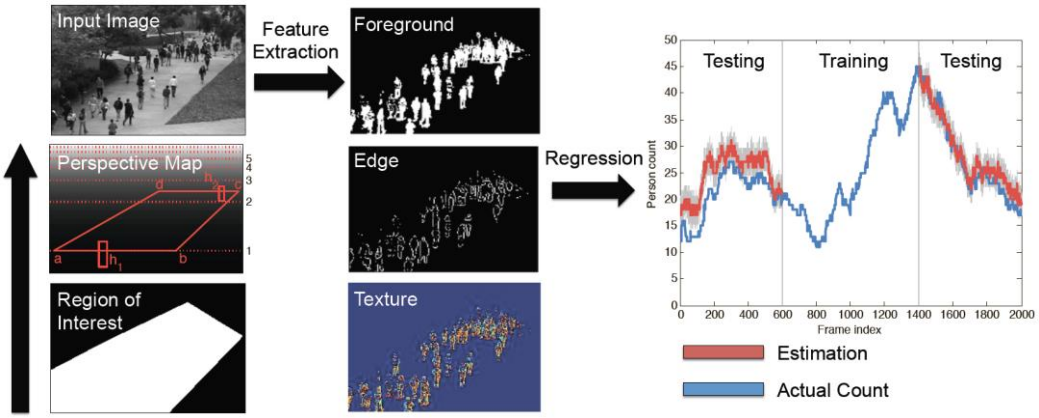


图 2.6　基于回归的人群计数

基于回归的方法可以总结为两个步骤：特征提取和回归建模。在建立回归模型前必须要解决图像特征表示的问题，提取合适的特征对于基于回归的方法来说至关重要。基于回归的方法使用手工提取的特征，如：纹理特征（Gray-level Co-occurrence Matrix，GLCM）、加速鲁棒特征（Speed Up Robust Features，SURF）、梯度直方图特征（Histogram of Oriented Gradient，HOG）等。常用的回归模型包括：线性回归（Linear Regression，LR），岭回归（Ridge Regression，RR），高斯过程回归（Gaussian Process Regression，GPR）等。

Chan 等人[40]提出了一个隐私保护系统来估计不均匀人群的大小，通过回归提取的手工特征来统计人群数量。Chan 等人[41]提取前景和纹理特征中的低级信息并使用贝叶斯泊松回归来统计人群数量。Chen 等人[42] 提出的模型能够自动学习相互依赖的低级特征和多维结构化输出之间的函数映射，发现不同特征对于不同空间位置的人群数量的内在重要性。Idrees 等人[43]依赖多源信息，如低置信度头部检测、纹理元素和频域分析等来估计人群数量。

虽然基于回归的方法减轻了对检测框的依赖，行人遮挡的问题在一定程度上得到了缓解。但在密集人群图像中，这类方法的计数性能仍然受到制约，不能很好解决高密度的行人遮挡问题。此外，这类方法仍然严重依赖于手工特征，选择合适的特征提取方法是基于回归方法的关键瓶颈。因此，研究者在基于回归的方法上继续深入研究，提出了基于密度图的人群计数方法。

2.2.3 基于密度图的人群计数

随着研究的深入，研究者发现视频图像中人群的空间信息和分布情况对于人群数量统计十分重要，而基于检测的方法和基于回归的方法都忽略这一问题。这两类方法只对图像中的人群数量进行了预测，并没有学习到人群的空间地理位置，无法准确预测人群的空间位置信息和密度分布情况。在实际的人群监控预警中，这些方法并不能有效的为有关部门疏散人群提供指导性的决策和部署。

在基于回归的人群计数方法的基础上，Lempitsky 和 Zisserman[11]开创性地提出了密度图的概念，引起了研究者的广泛关注。它通过学习图像中的局部特征和对应的密度图之间的一种映射关系来预测人群数量。Rodriguez 等人[44]证实，使用密度图计数可以极大地提高计数性能。由于密度图不仅反映了人群的空间分布信息，而且提高了计数的准确性，基于密度图的方法逐渐成为一类主流的人群计数方法。

如图 2.7 所示，左侧是一张人群图像，中间是图像对应的真实密度图，右侧是通过学习图像特征和真实密度图之间的映射关系，模型最终输出的预测密度图。通过对密度图中各个像素点的值求和可以计算出图像中人群的数量。基于密度图的人群计数方法不仅能够更准确计算出图像中的人群数量，同时还能学习到人群重要的空间分布信息。

Pham 等人[45]提出了一种基于图像块的公共场景人群密度估计方法，将结构化学习框架应用于随机决策森林进行人群计数。Xu 等人[46]采用随机森林作为回归模型，提取大量丰富的特征集进行人群密度估计，提高了人群数量预测的性能。

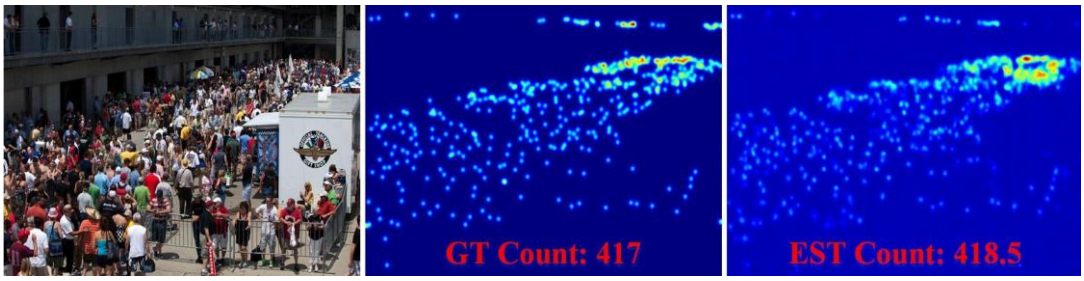


图 2.7 基于密度图的人群计数

2.3 卷积神经网络在人群计数中的应用

随着深度学习的快速发展，卷积神经网络强大的特征提取能力对研究者产生了巨大的吸引力。深度特征舍弃了设计好的人工特征，能够学习到更好的图像特征表示。使用卷积神经网络提取特征打破了人工设计特征具有的局限性。在基于密度图的人群计数方法上，Fu 等人[47]于 2015 年首次提出将卷积神经网络应用到人群计数领域，提高了模型估计的速度和准确度。同年，Wang 等人[48]采用卷积

神经网络设计了一种端到端的回归模型，自动学习有用的图像特征进行人群密度估计，计数性能取得了巨大的提升。基于密度图的深度神经网络模型成为人群计数领域的主流。

Zhang 等人[3]提出一种跨场景的人群计数方法来解决未知场景下人群计数性能下降的问题。采用深度卷积神经网络提取人群特征并使用人群密度和人群数量两个相关的学习目标来交替训练模型，使得模型能够获得较好的局部最优解。使用一种数据驱动的方法来对训练好的模型进行微调，提高了人群计数的性能。Zhang 等人[49]提出一种多列卷积神经网络结构的人群计数模型（Multi-Column Convolutional Neural Network，MCNN），可以处理不同尺寸和像素大小的图像。核心思想是使用三列不同卷积核大小的网络来学习不同尺寸大小的人群特征，最后融合图像中的多尺度特征并生成人群密度图。

多列结构在 MCNN 中的成功应用一定程度上影响了后来的人群计数模型的结构，很多人群计数模型中都能看到多列的结构。Sindagi 等人[50]提出一种上下文金字塔卷积神经网络（CP-CNN）的人群计数模型，通过全局上下文估计（Global Context Estimator，GCE）和局部上下文估计（Local Context Estimator，LCE）融合全局和局部上下文信息来生成高质量的人群密度并进行人群数量估计。密度图估计（Density Map Estimator，DME）用来生成高维特征图，后端的融合卷积网络（Fusion-CNN）将高维特征图与全局、局部上下文信息进行融合。由于全局和局部上下文估计模块是相互独立训练的，训练模型所需要的消耗很大。Sam 等人[4]提出了 Switch-CNN，在多列结构的基础上加入一个开关分类器，通过学习不同图像块中的人群密度把人群密度按等级分类并用来训练这个开关分类器。根据开关分类器对输入的图像进行密度等级分类，然后把图像送入具有不同卷积核大小的独立的 CNN 回归器来生成对应的密度图。这样就可以使用不同尺度的网络结构来学习对应密度等级的图像特征，提高人群数量预测的准确性。Deb 和 Ventura[5]采用扩张卷积设计了一个多列聚合的卷积神经网络模型（Aggregated Multicolumn Dilated Convolution Network，AMDCN）以提高多列网络结构对多尺度人群特征信息的提取。AMDCN 使用 6 列卷积神经网络，每一列具有不同的卷积核大小和扩张率，聚合模块使用相同卷积核大小、不同扩张率的一组扩张卷积层来融合不同列中的特征信息。Cao 等人[51]提出一个尺度聚合的编码器-解码器网络模型（Scale Aggregation Network，SANet），用于精确和高效地估计人群数量。解码器中使用了一种尺度聚合模块来提取多尺度特征，尺度聚合模块中通过不同卷积核大小的卷积提取特征并进行融合，打破多列结构的独立性，增加了更多可能的特征组合形式，增强了输出特征的表示能力和尺度多样性。Cheng 等人[52]提出了一种新的多列互学习策略（Multi-column Mutual Learning，McML）来

指导现有的多列网络模型提高特征表示的尺度不变性。现有的多列网络中不同列之间往往表现出几乎相同的尺度特征，严重影响计数精度，导致过拟合。McML可以估计不同列之间的互信息，近似地表示不同列之间特征的尺度相关性，通过最小化互信息，引导每一列学习不同的尺度特征。McML 交替优化每一列，同时保持其他列固定在每个小批量训练数据上，异步更新优化参数，有效减少冗余，提高泛化性能。

Li 等人[53]认为MCNN中的多列结构效果并不明显，并用实验验证多列CNN学习到了近乎相同的尺度特征。针对高度拥挤场景下的人群计数，提出了一种使用数据驱动的深度神经网络模型（CSRNet）。CSRNet 主要有两个部分组成，前端使用 VGGNet 提取人群特征，后端采用扩张卷积取代池化操作来扩大卷积核的感受野。扩张卷积能够更好的捕捉到人群头部特征的细粒度，提高了模型生成密度图的质量。Zhang 等人[54]认为人群图像中像素之间存在相互依赖，独立的像素预测可能存在噪声和不一致问题。为了解决这一问题，提出了一个含有自注意力机制的关系注意力网络（Relational Attention Network，RANet）来捕捉像素之间的相互依赖型。RANet 通过局部自注意力模块（Local Self-Attention，LSA）和全局自注意力模块（Global Self-Attention，GSA）计算像素之间的短期和长期依赖性来增强自注意力机制，并进一步引入一个关系模块来融合 LSA 和 GSA，实现含有丰富信息的聚合特征表示。

Jiang 等人[55]提出了一个格子状编码器-解码器网络（Trellis Encoder-Decoder network，TEDnet）来生成高质量的密度图。TEDnet 在不同的编码阶段合并多个解码路径来层级地聚合特征，提高了特征的表示能力。采用稠密跳跃连接交错地联通网络，充分地融合多尺度特征。此外， 提出了一种新的组合损失函数来提高特征图之间局部一致性和空间相关性的相似度。Ding 等人[56]认为基于密度图的人群计数方法需要考虑密度分布的正确性问题，即生成的密度图中可能存在假阴性（False Negatives）和假阳性（False Positives）的问题。针对该问题，提出了一种对称的网络模型，由编码器和解码器构成，融合子编码器和子解码器中的特征图来生成更合理的密度图。同时，提出使用块绝对误差（Patch Absolute Error，PAE）作为评价标准来度量密度图准确性。Zhou 等人[10]提出了一种多尺度生成对抗网络（Multiscale Generative Adversarial Network，MS-GAN）用于生成任意人群密度场景下的高质量密度图。MS-GAN 使用多尺度卷积神经网络融合多层级特征来检测人群的尺度变化，由多尺度生成器生成的密度图通过训练好的对抗网络来解决低质量密度图和真实密度图之间的二分类任务。

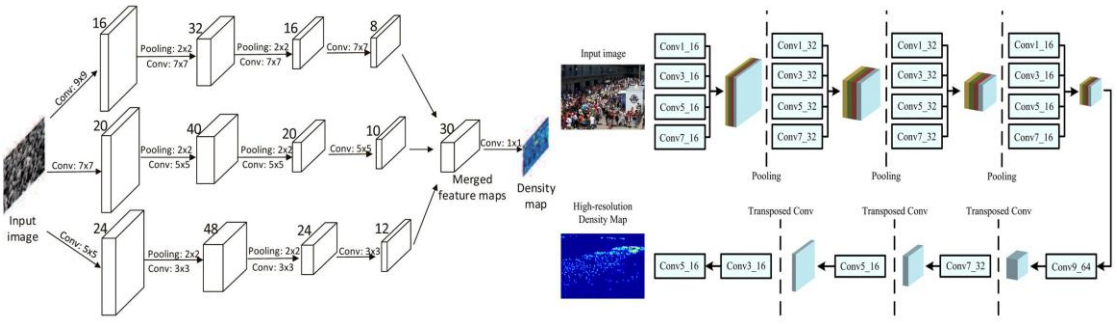
3　跨层级聚合网络的人群计数模型

本章主要介绍采用混合连接的卷积神经网络设计提出的人群计数模型——跨层级聚合网络（Cross-Hierarchy Aggregation Network，CHANet）。首先，给出CHANet 的研究动机。然后，详细阐述 CHANet 的网络框架、跨层级聚合（CHA）模块和后端解码器。接着，介绍本文采用的生成真实密度图的方法，最后给出模型的损失函数。

3.1　研究动机

现有的多列结构的人群计数方法一般被用来解决人群尺度变化问题，具有代表性的多列结构的人群计数模型是 MCNN，如图 3.1 中（a）所示。通过采用多列不同卷积核大小的卷积神经网络来提取不同尺度的人群特征，在网络的末端进行特征融合以获取多尺度特征，学习更优的人群特征表示，进而提高计数性能。然而，不同尺度特征的组合受到限制，网络能够表示的尺度范围相对较小。同时，多列的网络结构增加了网络的参数量和训练成本。此外，不同列之间可能学习到相近或相同的尺度特征，导致模型出现过拟合问题。

多尺度聚合的人群计数方法通过不同卷积核大小的卷积层提取尺度特征并在在特定的网络层融合前端的多尺度特征，具有代表性的多尺度聚合的人群计数模型是 SANet，如图 3.1 中（b）所示。多尺度聚合的方法打破多列结构只在网络末端进行尺度融合的特性，增加了更多可能的尺度组合种类，一定程度上增强了网络对多尺度特征的表示能力。然而，这一类方法没有充分利用不同网络层中的特征，降低了网络中的信息流动，一些对人群特征表示有益的信息可能在网络中丢失，进而导致生成的密度图质量下降，影响模型的计数性能。



（a）MCNN的多列结构　　（b）SANet的多列结构

图 3.1　具有代表性的多列和多尺度网络模型

Hou 等人[57]指出，深层网络提取高级特征中丰富的语义信息，浅层网络提取低级特征中大量的空间信息。提出在网络旁侧输出端（Side Output）采用短连接（Short Connection）重复利用不同网络层的特征以提高网络的表示能力。然而，特征重用的网络层和连接方式受到了限制。单一的连接方式影响了网络中的信息提取和流动，也没有充分利用每一层网络提取的特征。

在人群计数中，不同网络层提取的特征中不仅包含了多尺度信息，同时也含有大量的空间信息和语义信息。多尺度信息可以有效缓解人群尺度变化的问题；空间信息有利于学习人群的空间分布和位置关系，语义信息可以辨别图像中的人群前景特征和背景特征，有效缓解人群背景杂乱的问题。

针对人群计数中背景杂乱和尺度变化问题进行研究，结合现有的多列结构和多尺度聚合的人群计数方法中存在的不足，本文提出 CHANet。设计一种混合连接的方式将稠密连接和残差连接融入模型中，保证了模型可以重复利用不同网络层的特征，既保留了浅层网络中的空间信息，又融合了深层网络中的语义信息，同时保证网络中信息最大量地流动，有效提取网络中的多尺度特征信息和多层级语义信息，使得模型具有强大的人群特征表示能力，提高人群计数性能。

3.2　整体网络框架

本文提出的 CHANet 将稠密连接和残差连接以混合连接的方式融入网络模型，采用基于密度图的方法输出人群图像对应的预测密度图，并通过计算密度图来更精确估计给定图像中的人群数量。CHANet 是一个可训练的端到端的模型，该网络模型可以以任意分辨率的图像作为输入并输出对应的人群密度图。图像中的人群数量可以通过对密度图中所有像素点上的值进行求和得到。

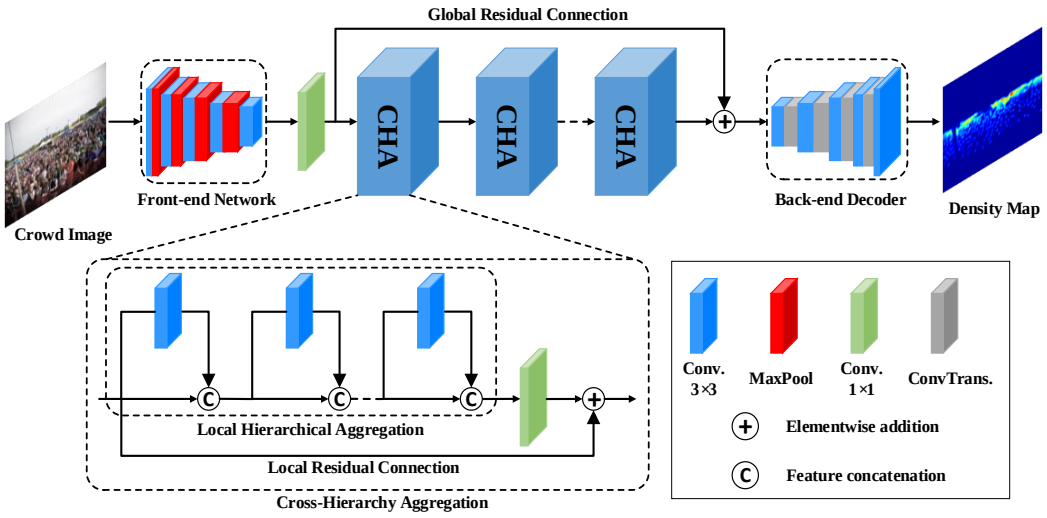


图 3.2　跨层级聚合网络的人群计数模型

CHANet 模型整体框架如图 3.2 所示。该网络模型主要由三部分组成：前端网络、若干 CHA 模块和后端解码器。由于 VGG-16 具有非常强大的特征表示能力和迁移性能，本文采用 VGG-16 的前十三层作为前端网络，从输入的人群图像中提取前端特征。前端网络的详细设计如表 3.1 中的前两列所示，Conv 表示卷积层，MaxPool 表示最大池化层。前端网络有 13 个卷积层，卷积核的大小为3×3，步幅为 1，特征通道数从 64 到 512，每经过一次最大池化层，通道数翻倍，直至通道数为 512 为止。有 4 个最大池化层，池化窗口的大小为2×2，步幅为 1。给定图像$I \in R$，尺寸大小为$H \times W$。经过前端网络的处理得到的特征图尺寸为$512 \times H/16 \times W/16$。为了减少模型中需要优化的参数，加快网络训练速度，在前端网络后边加入一个卷积核为1×1的卷积层，将特征通道数降为 128。最终得到前端网络的输出特征：

$$f_c = F_c(F_v(I)) \tag{3.1}$$

其中F_v表示前端网络。F_c表示前端网络后添加的单个卷积层，具有较少数量的卷积核来聚合来自F_v的跨通道特征信息，并减少通道的数量。f_c作为全局层级特征进行下一步操作。

表 3.1　CHANet 模型前端网络和后端解码器结构

网络层	前端网络	网络层	后端解码器
1-2	3×3-1-64 Conv	1	3×3-1-128 Conv
	2×2-2 MaxPool	2	2×2-2-128 ConvTrans
3-4	3×3-1-128 Conv	3	3×3-1-128 Conv
	2×2-2 MaxPool	4	2×2-2-128 ConvTrans
5-7	3×3-1-256 Conv	5	3×3-1-128 Conv
	2×2-2 MaxPool	6	2×2-2-128 ConvTrans
8-10	3×3-1-512 Conv	7	3×3-1-128 Conv
	2×2-2 MaxPool	8	2×2-2-128 ConvTrans
11-13	3×3-1-512 Conv	9	1×1-1-1 Conv

假设本文提出的 CHANet 模型设计了 H 个 CHA 模块，每个 CHA 模块中使用了 C 个卷积层（C 仅计算卷积核大小为3×3的卷积层个数），在实验章节 4.6.1 将对模型中超参数 C 和 H 的值进行验证分析。则第 h 个 CHA 模块的输出f_h可以表示为：

$$\begin{aligned} f_h &= F_{CHA,h}(f_{h-1}) \\ &= F_{CHA,h}\left(F_{CHA,h-1}(\dots(F_{CHA,1}(f_c))\dots)\right) \end{aligned} \tag{3.2}$$

其中$F_{CHA,h}$表示第 h 个 CHA 模块的操作。$F_{CHA,h}$是 CHA 模块中一系列网络层的函数操作，如卷积层、激活函数（ReLU）和批标准化（Batch Normalization，BN）[58]。由于f_h是通过从第 h 个 CHA 模块中聚合每个网络层的特征而生成的，因此将f_h称为局部跨层级聚合特征。通过 H 个 CHA 模块提取跨层级聚合的特征，得到f_H之后，采用一个全局残差连接来直接聚合全局层级特征f_c和f_H。CHA 模块的具体细节将在本章 3.3 中进行介绍。

最后，CHANet 使用后端解码器进一步提取跨层级聚合的多层级特征，融合人群特征中的多尺度信息和语义信息，并重构图像信息，生成人群图像的高质量密度图。后端解码器将来自全局残差连接的特征作为输入，将最终生成的密度图尺寸调整为原始图像的尺寸大小，其详细设计将在本章 3.4 中进行介绍。最后，采用单个 1×1 的卷积输出高质量密度图。CHANet 的最终输出为：

$$f_{map} = F_{CHAN}(I) \tag{3.3}$$

其中F_{CHAN}表示本文提出的 CHANet 模型的所有操作，f_{map}表示输入人群图像的相应密度图。

3.3　跨层级聚合模块

本节将详细介绍 CHANet 模型中的核心模块 CHA，该模块由局部层级聚合和跨层级残差连接构成。两种不同的特征融合方式从不同的网络层获得图像的多尺度特征信息和多层级语义信息，丰富了网络模型中的信息流通，使得 CHANet 捕捉更多有益的人群特征信息，保留浅层网络中人群特征的空间信息，同时融合深层网络中人群特征的语义信息，增强模型对人群特征的表示能力，提高模型的计数性能。

3.3.1　局部层级聚合

深度卷积神经网络中不同网络层提取的层级特征包含不同信息。浅层网络能够学习到图像的低层空间特征信息，深层网络能够学习到图像的高层语义特征信息。当卷积神经网络层数加深之后，网络浅层中的信息或者梯度会随着网络层增加出现消失和“washout”的现象。为了缓解这一问题，提出了稠密连接的方法。稠密连接最早出现于 Huang 等人[15]提出的 DenseNet，其主要思想是特征重复利用，核心模块如图 3.3 所示。稠密连接将当前网络层的特征以前馈的方式和所有前面网络层的特征堆叠起来作为当前网络层的输出特征传递给下一层网络。稠密连接的方法加强了网络中的特征传播和重复利用，保证了网络层之间丰富的信息

流动，有效避免信息和梯度在深层网络传递中出现消失的现象，缓解了梯度消失的问题，使得网络更容易训练。

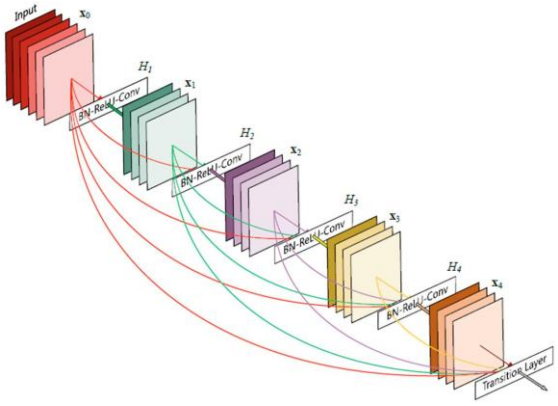


图 3.3　稠密连接核心模块

因此，本文采用稠密连接的方式将不同网络层的特征聚合在一起，以提取不同网络层中多样的人群特征，获得丰富的人群尺度信息和空间语义信息，并将包含这些信息的特征传递给下一个网络层。

图 3.4 详细展示了 CHA 模块中的网络细节，包含输入输出的特征尺寸、特征通道数的变化以及卷积核的数量等。规定第 h 个 CHA 模块的输入和输出分别为f_{h-1}和f_h。当 h 的值为 1 时，f_{h-1}等于f_0，表示第一个 CHA 模块的输入。从图 3.2 中可以看到，f_0就是前端网络输出的全局层级特征f_c。f_{h-1}和f_h具有相同的特征通道数和尺寸大小，都是 128×H/16×W/16。在局部层级特征聚合的过程中，每经过一个卷积层的特征堆叠，特征通道数增加 128，特征尺寸大小不变。

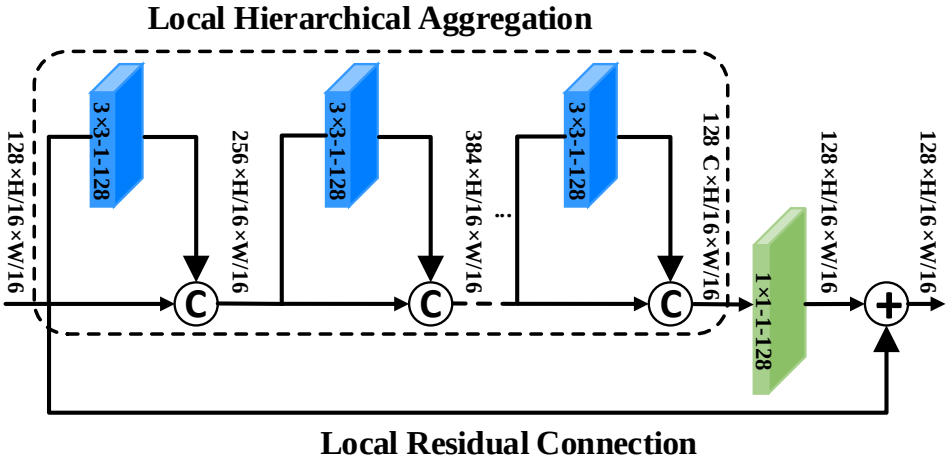


图 3.4　CHA 模块的具体结构设计

C 表示每一个 CHA 模块中卷积核为3×3的卷积层的个数。因此，第 h 个 CHA 模块的第 c 个卷积层的输出计算如下所示：

$$f_{h,c} = [f_{h,c-1}, \sigma(F_{h,c}(f_{h,c-1}))] \tag{3.4}$$

其中$F_{h,c}$表示第 h 个 CHA 模块中的第 c 个卷积操作，σ是跟在卷积之后的运算，如 ReLU 和 BN。[，]表示以“concat”的方式聚合来自不同网络层的特征。前一层网络的特征$f_{h,c-1}$通过当前层网络$F_{h,c}$和σ的操作提取当前层级的特征信息，并与前一层网络直接传递的特征进行聚合，获得不同网络层级中的特征信息，丰富了网络中的信息流动，有效地保留了不同网络层级中包含的多尺度特征信息。经过 C 个卷积层的层级特征聚合，得到了局部层级聚合特征$f_{h,C}$。$f_{h,C}$在进行特征聚合之后，特征通道数为 $128 \cdot C$，这导致网络参数量剧增，不利于网络的优化和模型的学习。更重要的是，$f_{h,C}$并没有将聚合而来的特征进行深层的融合以提取从不同层级获得的信息。因此，采用了一个1×1的单个卷积融合来自局部层级聚合的跨层级特征以自适应地提取跨层级特征中的有用信息。单个卷积的操作如下所示：

$$f'_{h,C} = F_{1\times1}(f_{h,C}) \tag{3.5}$$

其中$F_{1\times1}$表示 CHA 模块中1×1的卷积操作。$F_{1\times1}$可以使得 CHA 模块的输出f_h与输入f_{h-1}保持相同的特征通道数，从而降低了模型的计算复杂度，增强了网络的非线性建模能力，并确保能够在模型中构建跨层级残差连接以进一步加强网络中的特征传播和信息流动。

3.3.2　跨层级残差连接

理论上来讲，随着卷积网络的层数增加，模型能够更好地拟合训练数据并降低训练误差。然而在实际应用中，在网络层数加深之后，模型开始收敛，训练误差反而变大。为了解决这一问题，He 等人[14]提出了残差网络 ResNet，在残差网络中使用了残差连接，核心模块是残差块（Residual Block），如图 3.5 所示。

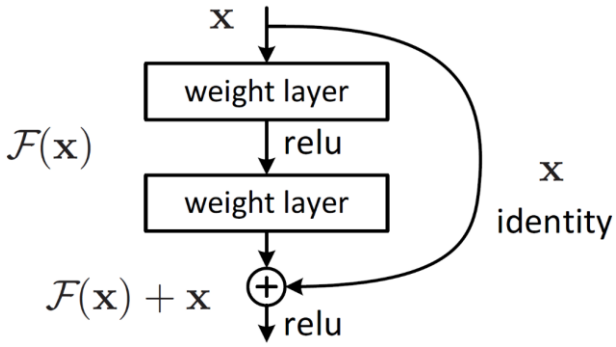


图 3.5　残差块的结构设计

残差连接利用恒等映射的思想，将跨网络层的输入特征x和输出特征$\mathcal{F}(x)$进行加法运算，使得网络中的特征信息传播更加丰富，简单有效地解决了深层网络退化问题。同时，残差连接不增加额外的网络参数，也不增加计算复杂度，使得

网络更容易优化，在一定程度上加快了网络的收敛。

受此启发，本文采用了残差连接的思想设计了跨层级残差连接。在 CHA 中的残差连接，将其称为局部跨层级残差连接；在 CHA 模块外的残差连接，将其称为全局跨层级残差连接。

局部跨层级残差连接：每个 CHA 模块都有 C 个卷积层进行局部层级特征聚合，并通过单个卷积将聚合的特征信息进一步融合，同时将特征通道数和尺寸调整到与输入特征相同。为了进一步加强特征的重复利用和信息的最大流动以获取更加充分的多尺度特征信息和空间语义信息，在每个 CHA 模块中引入了一个局部跨层级残差连接，该连接跳过整个局部层级特征聚合将聚合特征$f'_{h,C}$与 CHA 模块的输入特征以按像素相加的方式进行连接。因此，第 h 个 CHA 模块的最终输出表示为：

$$f_h = f_{h-1} + f'_{h,C} \tag{3.6}$$

其中f_{h-1}是第 h 个 CHA 模块的输入，$f'_{h,C}$是局部层级特征聚合的输出。CHA 模块中混合使用了稠密连接和残差连接，这两种连接方法以不同的方式重复利用不同网络层中的特征。交叉、跨越的混合连接方式，保证了网络中的特征信息的最大流动，有效地将浅层网络中的信息传递给深层的网络，极大地保留了网络中有用的人群特征信息，提高了网络的特征表示能力和计数性能。

全局跨层级残差连接：为了更好利用从前端网络提取的全局层级特征，保留有用的全局人群尺度信息和空间信息，采用全局跨层级残差连接来融合全局层级特征f_c和从 CHA 模块中学习到的跨层级聚合特征。将经过 H 个 CHA 模块提取出来的特征表示为f_H，特征f_c和特征f_H以按像素相加的方式进行特征的融合，表示为如下形式：

$$f_g = f_c + f_H \tag{3.7}$$

可以看到，在 CHA 模块中有针对性地使用了稠密连接和残差连接两种特征融合方法进行混合连接，重复利用不同网络层中的特征，有效加强了网络中的信息流动，充分保留了浅层网络中有利的人群特征信息。网络模型利用学习到的浅层和深层网络中的信息能够更好的表示人群特征，从而提高计数性能。

产生的融合特征f_g将被送到 CHANet 的后端解码器，后端解码器能够进一步融合提取f_g中的特征信息并重构特征图的尺寸，以生成和对应图像相同分辨率的高质量密度图。

3.4　后端解码器

在深度神经网络中，通常会在两个相邻的卷积层之间加入一个池化层，如本文提出的 CHANet，其前端网络中就用了四个最大池化层。池化层能够对特征进

行压缩，去除特征中的冗余信息，减少网络参数量，加快网络优化并防止过拟合。同时，池化层造成 CHANet 输出的密度图分辨率降低，仅有原始图像的 1/16，严重影响生成的密度图质量。因此需要在后端对网络的输出进行上采样，提高生成密度的分辨率。插值法的上采样方法使用的是已经设计好的参数去恢复图像中的像素值，网络自身并没有学习如何最优的恢复所需要的值。转置卷积（Transposed Convolution）[59]，又称为反卷积（Deconvolution），可以代替插值法进行上采样。转置卷积和普通的卷积有着相同的本质，将特征图进行补 0，使用具有可训练参数的卷积核对特征图进行“反向”卷积操作，扩大特征图的尺寸。图 3.6 为转置卷积的操作过程，以尺寸为2×2的特征图为例，先将边缘补 0，转置卷积的卷积核为3×3，步长为 1，然后逐步进行操作，最终获得尺寸为4×4的特征图。

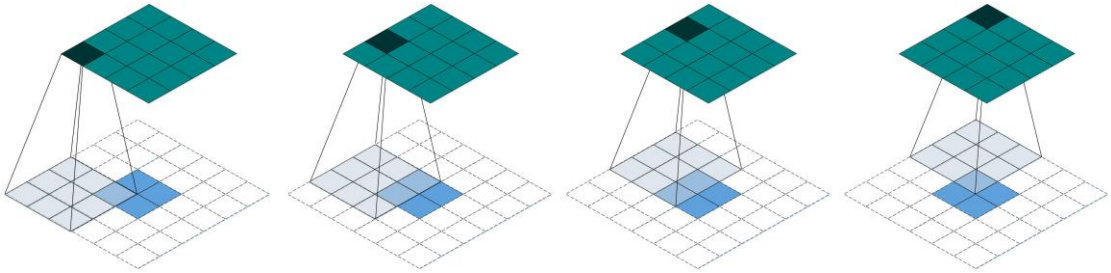


图 3.6　转置卷积的运算过程

后端解码器主要由卷积和转置卷积两种操作组成，将一个3×3、步长为 1 的卷积层和一个2×2、步长为 2 的转置卷积层作为一个组。这样使用转置卷积层可以重构特征图的尺寸，提高特征图的分辨率；使用卷积层可以进一步提取尺寸修复后的特征图中的信息。在后端解码器中一共采用四组卷积层和转置卷积层交替进行，并在最后使用不包含 ReLU 和 BN 的单个1×1的卷积生成人群密度图。

表 3.1 的后两列展示了后端解码器的细节设计，ConvTrans 表示转置卷积，每经过一个转置卷积操作，特征图的尺寸大小将会扩大两倍。经过四个转置卷积之后，生成的密度图分辨率将还原成输入图像的大小。前八层网络的特征通道数都为 128，最后单个卷积的通道数为 1，最终输出尺寸为$1 \times H \times W$的密度图。

3.5　真实密度图

本文提出的 CHANet 采用基于密度图的方法进行人群数量预测，不仅能够预测图像中人群的数量，同时还能从生成的密度图中获得人群的密度大小和空间分布情况。现有的主流人群计数数据集使用的是点标注的标签，只给出了图像中人头中心像素点的二维坐标作为原始标签。如果只使用原始的数据标签进行训练，将导致模型难以学习到准确的人群特征表示，生成的预测密度图也会非常稀疏，进而影响计数结果的准确性。

本文采用高斯滤波函数对原始数据标签进行处理，生成人群图像对应的真实密度图。使用真实密度图来训练网络学习人群特征，生成对应的预测密度图。具体地，每一个标注的人头中心像素点的值为 1，代表一个人，未标注的像素点的值为 0，图像中的总人数为标注点的总数。高斯滤波函数将人头中心像素点进行模糊处理，生成一个以该像素点为中心，高斯核大小为边的正方形，正方形区域所有像素点的值的和为 1，表示为一个人。使用高斯滤波函数将离散的标注点转换成连续的密度图。生成的真实密度图以热力图的方式呈现，并通过求和所有像素点的值得到图像中的总人数。

对于在像素点x_i标注的头部，它可以由函数$\delta(x-x_i)$来表示，则一张图像中所有的人头标注点可以表示为：

$$H(x)=\sum_{i=1}^{N}\delta(x-x_i) \tag{3.8}$$

其中N为图像中的总人数。使用高斯滤波函数对其进行模糊处理生成真实密度图标签的过程如下：

$$Z^{gt}=\sum_{i=1}^{N}\delta(x-x_i)\times G_\sigma(x) \tag{3.9}$$

其中$G_\sigma(x)$表示带参数σ的高斯核函数，遵循已有工作[60][61][62]，本文将σ设置为 4。图 3.7 展示了使用高斯核函数生成真实密度图的图像样例。图中第一列是从数据集中随机选择的人群图像，第二列是人工对相应图像进行点标注的可视化结果，第三列是根据数据原始标签生成的真实密度图。

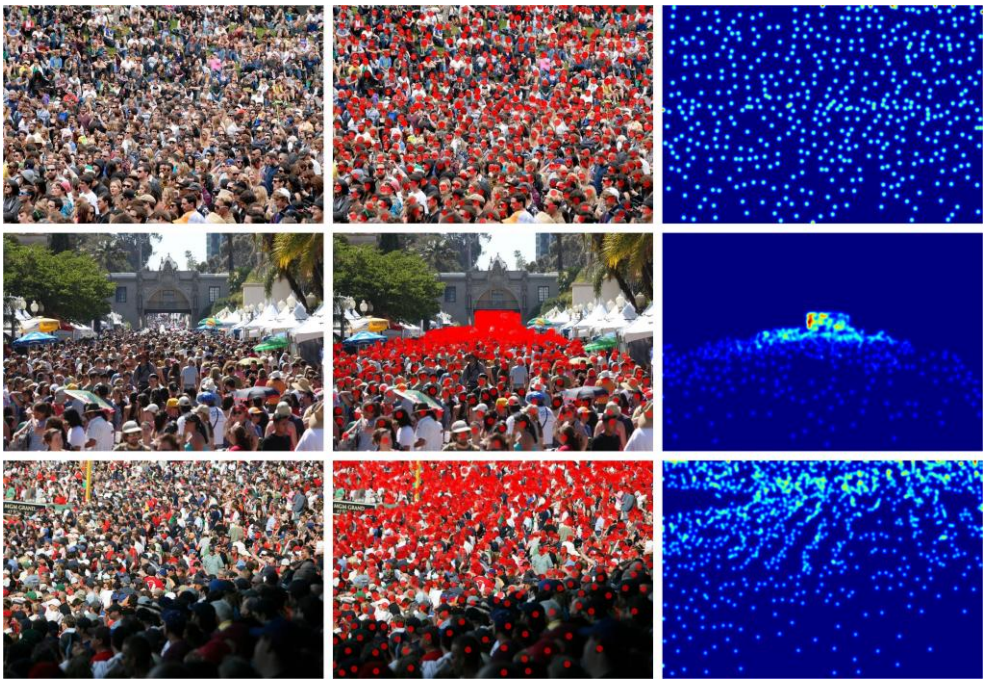


图 3.7　真实密度图的可视化

3.6　损失函数

损失函数一般和优化问题有关，用来评估模型的预测值和真实值之间的误差距离。在神经网络模型的训练过程中，通过最小化损失函数求解和优化模型。因此损失函数在模型的训练过程中起到了至关重要的作用。

回归问题的主要任务是让模型预测出一个更接近真实值的数值，人群计数的主要任务就是预测一张图像中的人群数量，因此人群计数属于回归问题。本文提出的 CHANet 模型输出的是人群密度图，通过人群密度图来预测人群数量。损失函数评估的是 CHANet 预测出来的人群密度图和真实的人群密度图之间的误差距离。回归问题中通常使用绝对值误差（即 L1 损失）和均方误差（即 L2 损失）作为损失函数。L2 损失使用平方操作放大预测值和真实值的误差距离，对偏离真实值的预测敏感，有利于提高模型的稳定性和鲁棒性。同时，L2 损失函数是平滑的，在模型优化过程中有利于计算梯度。因此，遵循已有工作[49][53][61]，本文选择使用 L2 损失作为 CHANet 的损失函数来对模型进行优化。L2 损失实质上计算的是欧几里得距离，其定义如下：

$$L(\theta)=\frac{1}{2S}\cdot\sum_{i=1}^{S}\left\|Z(X_i;\theta)-Z_i^{gt}\right\|_2^2 \tag{3.10}$$

其中θ是 CHANet 模型中所有需要优化的参数，S是训练过程中的批量大小。

$Z(X_i;\theta)$是模型输出的预测密度图，其中X_i表示的是模型的输入。Z_i^{gt}表示的是输入X_i对应的真实密度图。

4 实验结果及分析

本章为了验证提出的 CHANet 模型具有优越的计数性能，使用 ShanghaiTech 数据集[49]、UCF-QNRF 数据集[63]、WorldExpo'10 数据集[3]以及 Beijing BRT 数据集[64]等四个在人群计数领域中惯用的数据集进行了充足的实验验证和分析。本章给出了人群计数常用的评价标准，说明了模型训练的详细设置，介绍了实验使用的人群计数数据集和数据处理方法，展示了所提模型与当前一流的人群计数方法在四个数据集上的计数结果比较，分析了本模型生成密度图的质量评估，评估了模型的泛化性能，给出了模型的复杂度分析，并对模型的超参数 C 和 H 以及模型设计的有效性进行了验证分析。

4.1 评价标准

在人群计数领域，最重要的任务是预测图像中的人群数量。因此，预测的人群数量越接近真实人群数量，人群计数方法的计数性能就越好。本文采用平均绝对误差（Mean Absolute Error，MAE）和均方根误差（Root Mean Squared Error，RMSE）作为人群计数的评价标准，对提出的 CHANet 模型的计数性能进行评价。MAE 和 RMSE 的公式定义如下：

$$MAE = \frac{1}{S} \cdot \sum_{i=1}^{S} |N_i - \widehat{N}_i| \tag{4.1}$$

$$RMSE = \sqrt{\frac{1}{S} \cdot \sum_{i=1}^{S} |N_i - \widehat{N}_i|^2} \tag{4.2}$$

其中S是测试集中图像的数量，N_i和$\widehat{N}_i$分别表示第i张图像中的真实人群数量和模型预测的人群数量。MAE 计算的是预测人群数量和真实人群数量之间的平均数量误差，评价人群计数模型的计数准确性。RMSE 测量的是预测人群数量和真实人群数量之间的平均偏离误差。从公式可以看出，RMSE 对偏离真实人群数量的误差比较敏感，用来评价人群计数模型的稳定性和鲁棒性。总的来说，MAE 和 RMSE 的值越低越好。在本文的实验中，统一使用 MAE 和 RMSE 来对人群计数方法的计数性能进行评价。

本文还采用峰值信噪比（Peak Signal to Noise Ratio，PSNR）和结构相似性（Structural Similarity，SSIM）来评价人群计数方法生成的预测密度图的质量。PSNR 和 SSIM 的公式定义如下：

$$PSNR = 10 \cdot \log_{10}(\frac{MAX_G^2}{MSE}) \tag{4.3}$$

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}|G(i,j) - E(i,j)|^2 \tag{4.4}$$

$$SSIM = \frac{(2\mu_G\mu_E + c_1)(2\sigma_{GE} + c_2)}{(\mu_G^2 + \mu_E^2 + c_1)(\sigma_G^2 + \sigma_E^2 + c_2)} \tag{4.5}$$

$$c_1 = (k_1 L)^2，\ c_2 = (k_2 L)^2 \tag{4.6}$$

其中MAX_G^2表示密度图可能的最大像素值，如果以 8 位二进制来表示密度图中的像素，则MAX_G^2等于 255。$G(i,j)$表示真实密度图，$E(i,j)$表示预测密度图。m、n表示密度图的尺寸。MSE 表示$G(i,j)$和$E(i,j)$的均方误差。μ_G、σ_G^2和μ_E、σ_E^2分别表示$G(i,j)$和$E(i,j)$的均值和方差，σ_{GE}表示 $G(i,j)$和$E(i,j)$的协方差。c_1和c_2为两个常数，避免除 0。k_1和k_2为默认值，分别设置为 0.01 和 0.03[65]。L 是像素点的取值范围。

PSNR 用来评价预测密度图的画质，SSIM 从亮度、对比度和结构三个方面评价预测密度图和真实密度图的相似程度。总的来说，PSNR 和 SSIM 的值越大越好，SSIM 的最大值为 1。

4.2 实验设置

4.2.1 模型训练细节

本文提出的 CHANet 在开源的深度学习框架 PyTorch 框架上实现，在 Ubuntu16.04 系统上进行实验，所有训练和测试过程均在单个 GPU 上运行，配置为 NVIDIA TITAN V 12GB。CHANet 的前端网络使用 VGG-16 在 ImageNet 数据集上预先训练的权重进行参数初始化，因为 ImageNet 数据集中的数据量庞大，在其上面训练的 VGG 网络具有强大的特征提取能力并且在其他分类和回归任务中表现出良好的迁移性能。模型中的其他可训练参数采用均值为 0，标准差为 0.01 的 Gaussian 分布来随机初始化权重。实验采用 Adam 优化算法，以 0.00001 的学习率来优化模型中所有的参数。训练轮数设置为 1000，批量大小设置为 8。

4.2.2 实验数据集

自 UCSD 数据集[40]公开以来，人群计数数据集的数量开始逐渐增加，数据集的规模和图像质量也在不断提高，为人群计数领域中的方法研究和实际产业应用产生了巨大的推进作用。这些公开的数据集为人群计数领域的研究者提供了统

一的对比标准，对于提出的人群计数方法的有效性能够进行规范化的验证。在本文实验中，主要使用以下四个数据集，分别是 ShanghaiTech 数据集、UCF-QNRF 数据集、WorldExpo'10 数据集以及 Beijing BRT 数据集。表 4.1 中给出了四个数据集的具体的属性信息，其中 Part_A 和 Part_B 是 ShanghaiTech 数据集中的两个子数据集。图 4.1 展示了来自上述四个数据集中的样例图像。

表 4.1 人群计数数据集属性

数据集	分辨率	图像数量	标注数量	最少人数	最多人数	平均人数
Part_A	多样	482	241,677	33	3,139	501
Part_B	768×1024	716	88,488	9	578	124
UCF-QNRF	多样	1,535	1,251,642	49	12,865	815
WorldExpo'10	576×720	3980	199,923	1	253	50
Beijing BRT	360×640	1,280	16,795	1	64	13

（1）ShanghaiTech 数据集。2016 年，来自于上海科技大学（ShanghaiTech University）的 Zhang 等人[49]在国际顶级计算机视觉与模式识别大会 CVPR 上公开了目前最流行的 ShanghaiTech 数据集。这个数据集共包含 1198 幅图像，标注了 330165 人。整个数据集由 Part_A 和 Part_B 两个独立的子数据集组成。Part_A 是从互联网上随机下载的任意分辨率的人群图像，共有 482 张图像，场景中的人群相对稠密。其中，有 300 张人群图像用于训练网络模型，剩余 182 张用于测试网络模型。不同于 Part_A，Part_B 由上海街道视频监控系统以768×1024的固定分辨率拍摄得到，共有 716 张图像，场景中的人群相对稀疏。其中，有 400 张人群图像用于训练网络模型，剩余 316 张用于测试网络模型。

（2）UCF-QNRF 数据集。UCF-QNRF 数据集[63]出现于 2018 年的欧洲计算机视觉大会 ECCV，由佛罗里达大学（University of Central Florida）的计算机视觉研究中心开源。这个数据集从网络上收集了一共 1535 张来自不同人群场景的图像，其中有演唱会、大型集会、体育馆、旅游景区、交通枢纽等等。由于不是从监控摄像机或模拟人群场景中收集的图像，这些图像具有不同的分辨率，最低的分辨率为377×300，最高的分辨率为9999×6666。同时，人群密度、人群分布情况和人群背景也有很大不同，这使得该数据集更接近于真实情况下的人群计数场景。数据集中总共标注了 1251642 人，单个图像中的人数最少的有 49 人，最多的有 12865 人，因此对于目前最好的人群计数方法来说，这个数据集也是一个巨大的挑战。数据集包含一个训练集和一个测试集，训练集中有 1201 幅图像，

测试集中有 334 幅图像。

图 4.1　数据集中的样例图像

（3）WorldExpo’10 数据集。来自上海交通大学和香港中文大学的 Zhang 等人[3]于 2015 年在国际顶级计算机视觉与模式识别大会 CVPR 上公布了 WorldExpo’10 数据集。这个数据集是一个大规模跨场景人群计数数据集，所有图像均来自于 2010 年上海世界博览会的监控视频，其中包括 108 个监控摄像机拍摄的 1132 段带注释的视频序列。由于监控摄像机具有不同角度和方位的鸟瞰场景，所以数据集中包含了各种各样的人群场景。数据集分为训练集和测试集两部分。训练集中的图像来自于 103 个监控场景的 1127 个一分钟长的视频序列中的标记帧，每个视频有 3 个标记帧，两个标记帧之间的时间间隔为 15 秒，共有 3380 个标记帧。测试集中的图像来自于 5 个不同监控场景的 5 个一小时长的视频序列

中的标记帧。每个测试场景中有 120 个标记帧，两个标记帧之间的时间间隔为 30 秒。训练集中的场景和测试集中的场景都不相同。每个场景都提供了透视图和感兴趣区域（ROI）。该数据集总共标注了 199923 个人，平均每帧有 50 人。

（4）Beijing BRT 数据集。2018 年，厦门大学的 Ding 等人[64]在国际顶级声学、语音与信号处理会议 ICASSP 上开源了 Beijing BRT 数据集。这个数据集中的图像来自于北京市公交车站摄像机拍摄的监控视频。由于是固定摄像机实时拍摄，因此每张图像的分辨率大小为 640×360，并且时间跨度从早上到晚上。图像中包含了太阳光线变化、阴影和尺度变化，涵盖了真实场景下的人群聚集现象，十分具有挑战性。该数据集共有 1280 张图像，其中 720 张图像是训练集，另外 560 张图像是测试集。与前三个数据集相比，该数据集中所有图像均来自交通运输领域，因此 Beijing BRT 数据集对于研究智能交通领域中公共交通调度、交通路线规划、人流量分析等具有十分重要的现实意义。

4.2.3 数据扩充方法

人群计数数据集通常图像数量少，规模小，因此会出现样本数据不平衡问题，并在训练过程中出现过拟合现象。如果没有大量多样的训练数据，深度卷积神经网络很难获得更好的结果，因此本文设计使用一种数据扩充策略来为训练阶段生成更多样化的数据。首先，尺度问题一直困扰着人群计数的准确性，所以实验采用随机缩小和放大图像尺度的方法来提高模型对于尺度的学习。其次，一般来讲，数据扩充的方法还有水平翻转和垂直翻转，本文采用随机水平翻转图像来增加数据量，因为垂直翻转颠倒了人头和脚的空间位置，导致网络可能学习到错误的空间信息，不利于计数的准确性。然后，由于本文提出的 CHANet 中使用了 BN，并将批量大小设置为 8，实验中需要将图像裁剪成相同尺寸，综合考虑硬件资源和训练效果，实验中将 ShanghaiTech 数据集、UCF-QNRF 数据集和 WorldExpo’10 数据集中的图像随机裁剪的尺寸为400×400，Beijing BRT 数据集中所有图像均为相同尺寸且其中宽小于 400，故使用原尺寸进行训练。最后，所有数据集中的图像均采用 ImageNet 数据集中归一化的均值和标准差进行归一化操作。

4.3 实验结果对比及分析

本节使用前面介绍的 MAE 和 RMSE 评价标准来评估本文提出的模型 CHANet 的计数准确性和模型的稳定性，并使用 PSNR 和 SSIM 评估生成密度图质量，最后展示模型对四个数据集中测试集的样例进行预测所生成的人群密度图。在本章实验结果的表格中，统一将 CHANet 模型的实验结果全部加粗显示，CHANet 模型中超参数 C 和 H 取值为 5 和 3。每一个评价标准中最好的实验结果

用下划线标出。

4.3.1 计数结果对比与分析

为了验证本文提出的模型 CHANet 在人群计数中的有效性，本文将在前述四个数据集上进行实验，并与近几年提出的一流的人群计数方法进行对比。这些方法包括 Zhang et al.[3]、MCNN[49]、FCN[66]、CP-CNN[50]、D-ConvNet[67]、SANet[51]、CSRNe t[53]、CL-CNN[63]、DR-ResNet[64]、L2R[68]、RANet[54]、TEDnet[55]、ANF[69]、CAN[70]、DADNet[61]、McML[52]、FMLF[56]、SFCN[72]、DUBNet[71]等。

表 4.2 列出了在 ShanghaiTech 数据集上的实验结果比较。在 Part_A 数据集上，本文提出的 CHANet 与表中其他方法相比较，都取得了最好的 MAE 和 RMSE，实现了最佳的计数效果。在 Part_B 数据集上，本文所提的 CHANet 与表中其他方法相比较，取得了最好的 MAE 和较好的 RMSE。与表中 DUBNet 相比，CHANet 在 Part_A 上的 MAE 和 RMSE 分别降低了 13.6%和 10.5%；在 Part_B 上的 MAE 和 RMSE 分别降低了 14.3%和 10.4%。实验结果表明，本文提出的 CHANet 具有很好的计数性能，在不同的尺度和场景下能够更准确地估计人数。

表 4.2 在 ShanghaiTech 数据集上的实验结果

Method	Venue	Part_A		Part_B	
		MAE	RMSE	MAE	RMSE
MCNN	CVPR’2016	110.2	173.2	26.4	41.3
D-ConvNet	CVPR’2018	73.5	112.3	18.7	26.0
CSRNet	CVPR’2018	68.2	115.0	10.6	16.0
SANet	ECCV’2018	67.0	104.5	8.4	13.6
L2R	TPAMI’2019	73.6	112.0	13.7	21.4
TEDnet	CVPR’2019	64.2	109.1	8.2	12.8
CAN	CVPR’2019	62.3	100.0	7.8	12.2
ANF	ICCV’2019	63.9	99.4	8.3	13.2
RANet	ICCV’2019	59.4	102.0	7.9	12.9
DADNet	MM’2019	64.2	99.9	8.8	13.5
McML	MM’2019	59.1	104.3	8.1	10.6
SFCN	IJCV’2020	64.8	107.5	7.6	13.0
DUBNet	AAAI’2020	64.6	106.8	7.7	12.5

CHANet	**KBS'2021**	**55.8**	**95.6**	**6.6**	**11.2**

表 4.3 列出了 CHANet 与表中其他 9 种方法在 UCF-QNRF 数据集上的实验结果比较。与第二好的 SFCN 相比，CHANet 在 MAE 上降低了 3.8%，取得了最好的计数结果。在 UCF-QNRF 数据集上的结果表明，CHANet 在尺度变化差异巨大并且非常稠密的人群场景中表现出优异的计数性能，能够很好地处理不同背景下的人群图像。

表 4.4 列出了在 WorldExpo'10 数据集上 CHANet 与其他一流方法的实验结果比较。与前两个数据集相比，WorldExpo'10 数据集中的人群相对稀疏，同时测试数据集和训练数据集中的场景和拍摄角度不同。CHANet 在在第一个测试场景中取得了最好的 MAE 结果，并在其他四个测试场景中取得了较低的计数误差。总的来看，这五个场景的平均 MAE 是 6.7，比次优的 CAN 降低了 9.5%。实验结果表明，本文提出的 CHANet 能够很好地适应相对稀疏的场景，对于跨场景的人群计数具有稳定的性能。

表 4.5 列出了在 Beijing BRT 数据集上的实验结果比较。本文提出的 CHANet 获得了最好的 MAE 和 RMSE，这表明 CHANet 能够很好地适应真实生活场景中光线的巨大变化。

上述实验结果表明，与当前一流的人群计数方法相比，本文提出的 CHANet 在四个主流人群计数数据集上均取得了优异的计数结果和稳定的性能。在不同人群场景下，CHANet 对于密集和稀疏人群的计数性能都处于一流水平。

表 4.3 在 UCF-QNRF 数据集上的实验结果

Method	Venue	MAE	RMSE
CL-CNN	ECCV'2018	132	191
L2R	TPAMI'2019	124	196
TEDnet	CVPR'2019	113	188
CAN	CVPR'2019	107	183
ANF	ICCV'2019	110	174
RANet	ICCV'2019	111	190
DADNet	MM'2019	113	189
DUBNet	AAAI'2020	106	181
SFCN	IJCV'2020	102	171

CHANet	**KBS'2021**	**98**	**177**

表 4.4 在 WorldExpo'10 数据集上的实验结果（MAE）

Method	Venue	s1	s2	s3	s4	s5	Average
Zhang et al.	CVPR'2015	9.8	14.1	14.3	22.2	3.7	12.9
D-ConvNet	CVPR'2018	1.9	12.1	20.7	8.3	2.6	9.1
CSRNet	CVPR'2018	2.9	11.5	8.6	16.6	3.4	8.6
SANet	ECCV'2018	2.6	13.2	9.0	13.3	3.0	8.2
L2R	TPAMI'2019	3.8	17.5	13.8	12.7	5.2	10.5
TEDnet	CVPR'2019	2.3	10.1	11.3	13.8	2.6	8.0
CAN	CVPR'2019	2.9	12.0	10.0	7.9	4.3	7.4
ANF	ICCV'2019	2.1	10.6	15.1	9.6	3.1	8.1
McML	MM'2019	2.8	11.2	9.0	13.5	3.5	8.0
FMLF	TITS'2020	2.8	12.1	9.4	15.6	3.5	8.7
CHANet	**KBS'2021**	**1.4**	**11.6**	**8.7**	**8.9**	**2.8**	**6.7**

表 4.5 在 Beijing BRT 数据集上的实验结果

Method	Venue	MAE	RMSE
MCNN	CVPR'2016	2.24	3.35
ResNet-14	CVPR'2016	1.48	2.22
FCN	VISAPP'2017	1.74	2.43
CSRNet	CVPR'2018	1.68	2.35
DR-ResNet	ICASSP'2018	1.39	2.00
FMLF	TITS'2020	1.34	2.02
CHANet	**KBS'2021**	**1.09**	**1.71**

4.3.2 密度图质量评估与分析

为了验证本文提出的 CHANet 生成密度图的质量，使用 PSNR 和 SSIM 对 CHANet 生成的预测密度图进行了图像质量评估。表 4.6 列出了 CHANet 在不同数据集上的质量评估结果。由于人群计数领域的研究者习惯使用 ShanghaiTech

Part_A 数据集进行对比实验，表 4.7 列出了在 ShanghaiTech Part_A 数据集上 CHANet 和其他具有代表性方法在 PSNR 和 SSIM 上的结果比较。结果表明，本文提出的 CHANet 在保持优异的计数性能的同时，能够生成高质量的密度图。图像质量评估结果验证了 CHANet 中采用转置卷积设计的后端解码器能够有效地恢复密度图的分辨率，提高密度图的质量。

表 4.6 在四个数据集上 CHANet 生成密度图的质量评估

Dataset	PSNR	SSIM
Part_A	28.26	0.83
Part_B	32.61	0.93
UCF-QNRF	30.34	0.87
WorldExpo'10	34.26	0.93
Beijing BRT	37.24	0.97

表 4.7 在 ShanghaiTech Part_A 数据集上密度图的质量评估对比

Method	Venue	PSNR	SSIM	MAE	RMSE
MCNN	CVPR'2016	21.42	0.52	110.2	173.2
CP-CNN	ICCV'2017	21.72	0.72	73.6	106.4
CSRNet	CVPR'2018	23.79	0.76	68.2	115.0
SANet	ECCV'2018	23.36	0.78	67.0	104.5
ADCrowdNet	CVPR'2019	24.48	0.88	63.2	98.9
TEDnet	CVPR'2019	25.88	0.83	64.2	109.1
ANF	ICCV'2019	24.10	0.78	63.9	99.4
DADNet	MM'2019	24.16	0.81	64.2	99.9
CHANet	**KBS'2021**	**28.26**	**0.83**	**55.8**	**95.6**

4.3.3 计数结果可视化

为了更直观地展示本文提出的 CHANet 对于图像中人群数量的预测结果和生成的密度图，给出了来自上述四个数据集中的测试样例的可视化结果。如图 4.2 所示，本文从四个数据集的测试集中分别选取了两张图像，真实人数从几个人到几千人。第一行到第五行的图像依次来自于 ShanghaiTech Part_A、ShanghaiTech Part_B、UCF-QNRF、WorldExpo'10 和 Beijing BRT 数据集。第一列和第四列是

输入的人群图像，其中 WorldExpo’10 中的图像用绿色实线标出了 ROI，图像下方给出了真实人数和预测人数的误差；第二列和第五列是人群图像的真实密度图，图像下方给出了对应的真实人数；第三列和第六列是 CHANet 生成的预测密度图，图像下方给出了对应的预测人数。为了更精确地表达计数的准确性，预测人数中使用了小数，真实情况下人群数量是一个整数。从图中可以看出，对于不同环境下的人群场景，无论人群是稀疏还是稠密，CHANet 都具有十分优异的计数性能，预测密度图中人群分布和数量相当接近真实的密度图，能够明显地区分图像中的人群分布区域和背景区域，表明 CHANet 很好地学习到了人群特征的表示，较为准确地预测了人群数量。

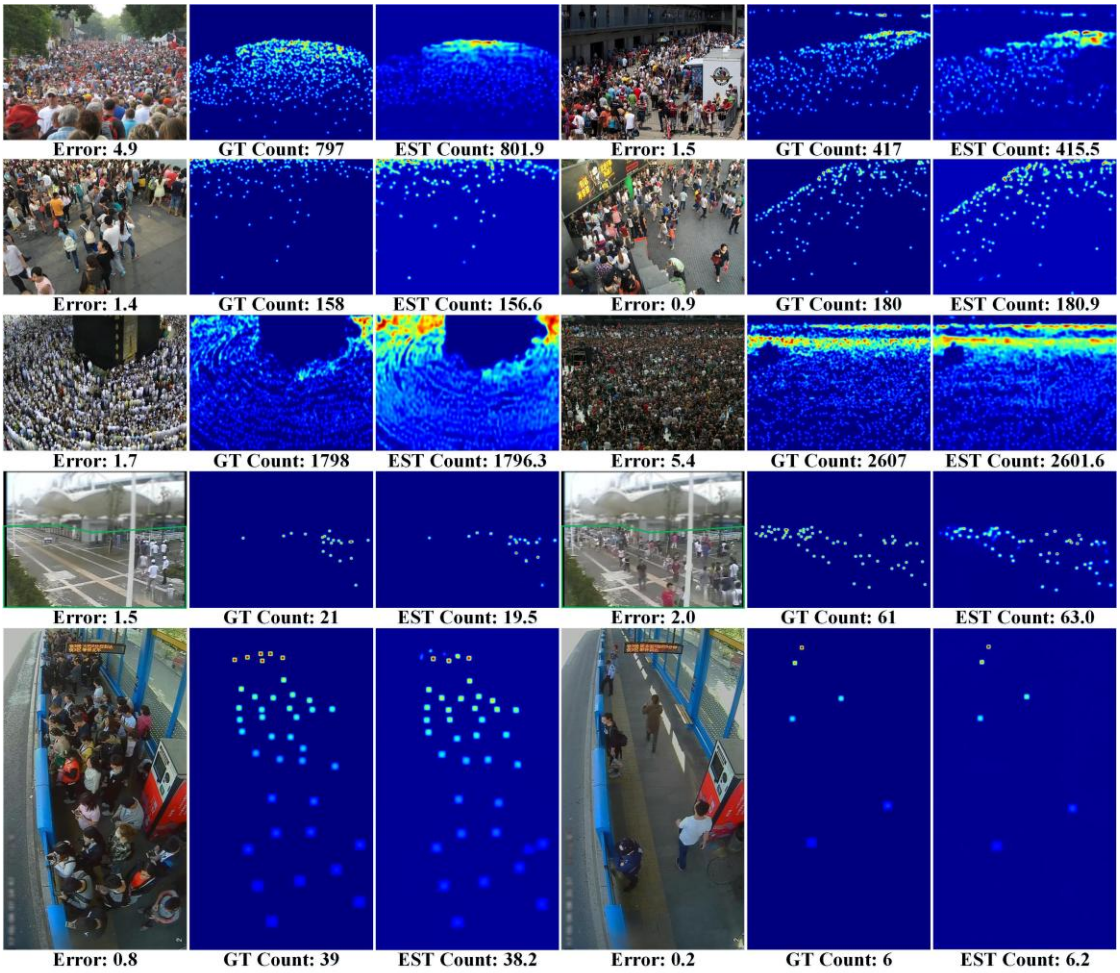


图 4.2　计数结果的可视化

4.4　模型泛化性能对比分析

不同数据集的场景下，人群尺度变化不同将导致在一个数据集上学习到的人群尺度信息不能适应另一个数据集；背景环境不一致也将导致模型的计数性能下降。为了验证模型具有很好的泛化性能，在 ShanghaiTech 数据集上对本文提出的

CHANet 进行测试并与 MCNN、Cascaded-MTL[60]、CSRNet、D-ConvNet 、DSPNet[62]等方法进行对比。具体地，将 Part_A 和 Part_B 中的训练集和测试集分别合并，使用 Part_A 的全部图像作为训练集来训练模型，使用 Part_B 的全部图像作为测试集来测试模型；反之，使用 Part_B 作为训练集，使用 Part_A 作为测试集。两组实验分别标记为“Part_A→Part B”和“Part_B→Part A”。同时，其他实验设置保持不变。

表 4.8 给出了模型泛化性能的结果。从实验数据可以看到，CHANet 在四项指标中取得三项最优的结果。这表明本文提出的 CHANet 能够很好地从不同的网络层提取有用的人群尺度信息和背景信息，更好地拟合不同场景中的人群特征，具有优异的泛化性能。

表 4.8 在 ShanghaiTech 数据集上的模型泛化性能对比

Method	Venue	Part_A→Part B		Part_B→Part A	
		MAE	RMSE	MAE	RMSE
MCNN	CVPR’2016	85.2	142.3	221.4	357.8
Cascaded-MTL	AVSS’2017	40.5	77.0	224.0	417.0
D-ConvNet	CVPR’2018	49.1	99.2	140.4	226.1
CSRNet	CVPR’2018	16.8	28.5	131.6	210.3
DSPNet	ESWA’2020	15.1	26.2	120.5	194.6
CHANet	**KBS’2021**	**12.4**	**20.9**	**112.4**	**219.9**

4.5 模型复杂度对比分析

本节对提出的CHANet模型的复杂度进行实验分析，并与MCNN、CP-CNN、CSRNet、ADCrowdNet[73]、RAZNet[74]、CAN 等方法进行对比。

4.5.1 计算量对比分析

对提出的 CHANet 模型的计算复杂度进行了分析。在人群计数的实际应用中,往往对于计数的时效性有很高的要求。同时，实际应用中的设备受到场地和预算等的限制，对模型的计算复杂度有一定要求。因此很有必要对模型的计算复杂度进行分析。遵循已有文献[75][76]，使用浮点运算次数（floating point of operations，FLOPs）来测量计算复杂度。如表 4.9 所示，GFLOPs 是 Giga FLOPs 的缩写，表示十亿浮点运算次数。由于 CP-CNN 和 ADCrowdNet 等方法没有公布官方的代

码，表中使用“>”来估计这些方法的最小 FLOPs 值。表中 RAZNet 的浮点运算次数最多，是 MCNN 的 40 多倍。过高的计算量严重限制了 RAZNet 的实际应用。相反，CHANet 的浮点运算次数是除了 MCNN 之外最少的，同时具有最佳的计数性能。这表明 CHANet 的计算复杂度适中，具有较高的实际应用价值。

4.5.2 参数量对比分析

对模型的参数量进行了分析。表 4.10 给出了本文提出的 CHANet 和其他方法的实验结果比较，表中 Params (M)表示以百万为数量级的参数量。从实验结果看出，MCNN 拥有最少的参数数量，模型占用空间最小。除了 MCNN 外，CSRNet 和 CAN 的参数数量和占用空间相对较少。CHANet 比 CSRNet 多 28%的参数，但 MAE 和 RMSE 分别降低了 18.2%和 16.9%。与 CAN 相比，CHANet 多了 19.8%的参数，但 MAE 和 RMSE 分别降低了 10.4%和 4.4%。实验结果表明 CHANet 的参数量处于可接受的量级，同时大幅度提升了模型的计数性能。

表 4.9 在 ShanghaiTech Part_A 数据集上的计算量对比

Method	Venue	GFLOPs	MAE	RMSE
MCNN	CVPR’2016	4.34	110.2	173.2
CP-CNN	ICCV’2017	>72.85	73.6	106.4
CSRNet	CVPR’2018	66.19	68.2	115.0
RAZNet	CVPR’2019	182.15	65.1	106.7
ADCrowdNet	CVPR’2019	>97.74	63.2	98.9
CAN	CVPR’2019	70.16	62.3	100.0
CHANet	**KBS’2021**	**64.90**	**55.8**	**95.6**

表 4.10 在 ShanghaiTech Part_A 数据集上的参数量对比

Method	Venue	Params (M)	MAE	RMSE
MCNN	CVPR’2016	0.13	110.2	173.2
CP-CNN	ICCV’2017	68.4	73.6	106.4
CSRNet	CVPR’2018	16.26	68.2	115.0
RAZNet	CVPR’2019	101.7	65.1	106.7
ADCrowdNet	CVPR’2019	37.7	63.2	98.9
CAN	CVPR’2019	18.1	62.3	100.0

CHANet	**KBS'2021**	**22.58**	**55.8**	**95.6**

实际上，模型计数性能和复杂度之间存在平衡。综上所述，本文提出的CHANet在适当增加了模型复杂度的基础上，显著提高了人群计数的性能，模型的设计具有合理性。

4.6　消融实验分析

4.6.1　超参数 C 和 H 的验证

对 CHANet 中的超参数 C（代表 CHA 模块中3×3的卷积层的个数）和 H（代表模型中 CHA 模块的个数）进行研究和实验，确定 C 和 H 的取值，从而获得 CHANet 最优的模型结构。

在 ShanghaiTech Part_A 数据集上对于 C 和 H 进行实验验证。根据经验，将 C 和 H 初始设置为 3。首先，将 H 固定来验证 C 的值，如图 4.3 所示，CHANet 在 C 为 5 的时候取得了最好的 MAE 和 RMSE。然后，将 C 的值固定为 5 来验证 H 的值，如图 4.4 所示，CHANet 在 H 为 3 的时候取得了最佳的计数性能。实验中，使用 CHANet 中的前端网络（VGG-16 网络的前 13 层）作为基准（Baseline）进行比较。综上所述，本文提出的 CHANet 采用了 3 个 CHA 模块，每个 CHA 模块中有 5 个卷积层。

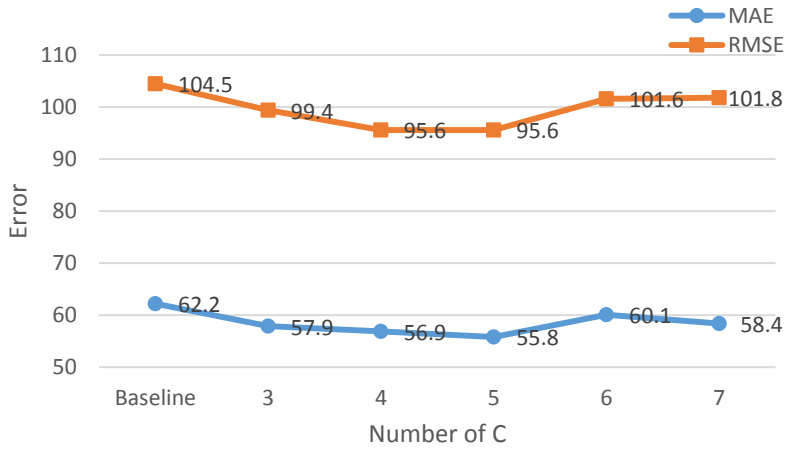


图 4.3　验证超参数 C 的实验结果

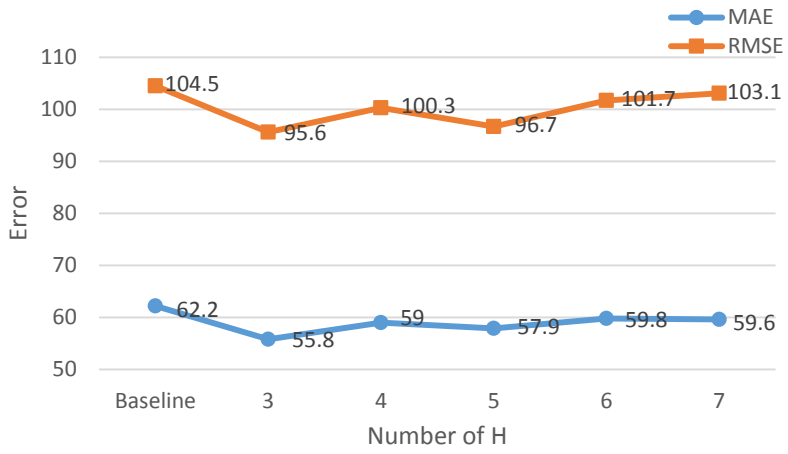


图 4.4　验证超参数 *H* 的实验结果

4.6.2　模型有效性验证

为了验证 CHANet 模型设计的有效性，进行了消融实验分析。实验使用了 4.6.1 中提到的 Baseline 作为基准，对比了 CHANet 在分别去掉稠密连接（CHANet w/o dense）和残差连接（CHANet w/o residual）的情况下的计数结果，如表 4.11 所示。

表 4.11　在 ShanghaiTech Part_A 数据集上的消融实验

Method	MAE	RMSE
Baseline (VGG-16)	62.2	104.5
CHANet w/o dense	58.0	97.3
CHANet w/o residual	57.4	102.6
CHANet	**55.8**	**95.6**

相比于 CHANet w/o dense，CHANet 在加入稠密连接后的 MAE 和 RMSE 分别降低了 3.8%和 1.7%，计数结果优于去掉稠密连接的结果。这表明 CHANet 的稠密连接可以重复利用不同网络层中的特征，有效地提取和融合网络中人群特征的多尺度信息和多层级空间语义信息。相比于 CHANet w/o residual，CHANet 在加入残差连接后的 MAE 和 RMSE 分别降低了 2.8%和 6.8%，计数结果优于去掉残差连接的结果。这表明 CHANet 中的残差连接能够有效地融合当前网络层和前面网络层的特征信息，更好地拟合人群特征。

同时可以看到，稠密连接对于 MAE 的下降幅度则高于残差连接，而残差连接对于 RMSE 的下降幅度明显高于稠密连接。这也证明了稠密连接对于人群尺度信息的提取和融合是有效的，能够提高计数性能，因而对 MAE 的降低更明显；残差连接能够将前面网络层的信息直接传递到后面，在一定程度上使得网络提取

的信息更加平滑，提高了模型的稳定性，因而对 RMSE 的降低更明显。

相比于 Baseline (VGG-16)，CHANet 在 MAE 和 RMSE 上分别降低了 10.3% 和 8.5%，计数结果提升幅度好于单独加入稠密连接或残差连接，表明了本文融合稠密连接和残差连接的混合连接方式的有效性。实验结果验证了 CHANet 模型设计的合理性和有效性。

5 总结与展望

5.1 全文工作总结

本文结合现有的使用多列和多尺度融合的人群计数方法解决人群尺度变化问题的思路，采用混合连接的方式，设计提出了一种跨层级聚合网络（CHANet）的人群计数算法。本文的主要工作总结如下：

（1）为了提取不同网络层中丰富的人群尺度信息和空间语义信息，本文采用稠密连接和残差连接两种方法设计了混合连接的方式，并使用混合连接的方式提出了一种跨层级聚合（CHA）模块。跨层级聚合模块使用稠密连接将不同网络层的特征进行聚合，有效提取网络中的局部跨层级特征，获取特征中的多尺度信息；使用残差连接跳跃地将当前网络层的特征信息和浅层网络层中的特征信息进行融合。混合连接的方式重复利用网络中的多层级特征，从而保证网络中信息最大量地流动，有效提取网络中的多尺度特征信息和多层级空间语义信息，使得模型具有强大的人群特征表示能力，提高了人群计数性能。

（2）采用在 ImageNet 上预先训练 VGG-16 网络作为前端网络来提取人群图像的全局层级特征。VGG-16 网络具有强大特征提取能力和迁移性能，使得前端网络在模型训练阶段具有较强的特征提取能力，加快了模型训练的速度，有效缓解了因数据样本不足导致的模型过拟合等问题，提高了模型的性能。

（3）后端解码器采用转置卷积进行上采样来生成高质量的人群密度图。转置卷积具有可训练的网络参数，在网络训练过程中根据学习到的特征进行参数优化，更合理地恢复图像或特征的像素值，还原成输入图像的尺寸大小，后端解码器进一步融合来自前端网络的全局层级特征和 CHA 模块的局部跨层级特征，提高了生成密度图的质量。

（4）采用高斯滤波器的方法将数据集中的标签生成真实密度图来训练网络模型。本文提出的模型采用的是基于密度图的人群计数方法，需要将数据标签转换成密度图的形式来训练优化网络。使用高斯核函数将标注数据进行模糊化处理，把离散的标注点转换成连续的密度图，从而使网络模型在训练过程中学习到更多的特征表示，提高了模型的计数效果。

（5）在四个主流的人群计数数据集上进行了大量充分的实验。实验结果表明本文提出的模型具有更优越的计数性能，能够生成高质量的人群密度图。模型泛化实验的结果表明模型具有较强的泛化性能，能够适应不同场景下的计数任务。对模型复杂度进行对比分析，证明了模型设计的合理性。模型的消融实验分析，验证了模型设计的有效性。

5.2 未来展望

作为计算机视觉领域重要的研究分支，人群计数取得了更多突破性的成就，并由此演化更多的研究方向。人群计数的研究发展很好地契合了城市智能化的发展需求，有着巨大的实际意义和应用价值。人群计数不仅可以应用到公共区域安全监控，同时还可以迁移到城市空间布局规划，公共交通管控和动植物计数等跨学科领域。

近年来，基于密度图的人群计数方法的提出和深度神经网络的发展使得人群计数方法在性能上出现了巨大的提升。然而在人群计数领域中仍然有需要进一步研究和探索的空间。

（1）无监督的人群计数方法。本文提出的方法是有监督的计数方法，需要庞大的标签数据才能提升计数性能。人群计数数据集中可能要对百万、千万级以上的行人进行标注，有些行人仅靠肉眼很难看清。因此大量、准确的标注数据十分困难，人工标注耗费大量财力物力和时间。目前在半监督领域已经有一些研究成果，然而效果并不理想。因此，探索无监督的人群计数方法能够有效缓解对数据集标签的依赖，对于真实场景和跨场景的人群计数具有重要的研究意义。

（2）视频中的人群计数方法。目前的人群计数方法大多是对单幅图像或者视频中某一帧的静止的人群进行计数，很少有研究视频中运动的人群计数方法。这是因为在视频中行人的位置和数量一直在移动，需要加入时间因素，涉及到行人重识别领域，对计算量有很大需求。此外，还缺少相关的数据集。视频中的人群计数方法研究对于视频监控领域的实时人群计数具有十分重要的应用价值。

（3）轻量型的人群计数方法。人群计数方法虽然取得了优异的计数性能，但是模型的复杂度也随之增加。复杂的模型设计不仅增加了训练的成本和难度，也制约了模型在实际应用中的部署。因此，需要借助知识蒸馏、模型压缩、剪枝等方法，在不影响模型性能的基础上，降低模型的复杂度，实现人群计数模型轻量化，有利用人群计数方法的应用落地。

参考文献

[1] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[2] Teixeira T, Dublon G, Savvides A. A survey of human-sensing: Methods for detecting presence, count, location, track, and identity[J]. ACM Computing Surveys, 2010, 5(1): 59-69.

[3] Zhang C, Li H, Wang X, et al. Cross-scene crowd counting via deep convolutional neural networks[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 833-841.

[4] Sam D B, Surya S, Babu R V. Switching convolutional neural network for crowd counting[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 5744-5752.

[5] Deb D, Ventura J. An aggregated multicolumn dilated convolution network for perspective-free counting[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018: 195-204.

[6] Jiang X, Zhang L, Xu M, et al. Attention Scaling for Crowd Counting[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 4705-4714.

[7] 黄凯奇, 陈晓棠, 康运锋, 等. 智能视频监控技术综述[J]. 计算机学报, 2015, 38(06): 1093-1118.

[8] 时增林, 叶阳东, 吴云鹏, 等. 基于序的空间金字塔池化网络的人群计数方法[J]. 自动化学报, 2016, 42(006): 866-874.

[9] Sheng B, Shen C, Lin G, et al. Crowd counting via weighted VLAD on a dense attribute feature map[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 28(8): 1788-1797.

[10] Zhou Y, Yang J, Li H, et al. Adversarial learning for multiscale crowd counting under complex scenes[J]. IEEE Transactions on Cybernetics, 2020: 1-10.

[11] Lempitsky V, Zisserman A. Learning to count objects in images[C]. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems, 2010: 1324-1332.

[12] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[13] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]. In: Proceedings of the 3rd International Conference on Learning Representation, 2015.

[14] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[15] Huang G, Liu Z, Maaten L van der, et al. Densely connected convolutional networks[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2261–2269.

[16] 曾鑫. 面向多尺度人群计数的深度神经网络算法研究[D].郑州大学,2020.

[17] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009: 248–255.

[18] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(06): 1229-1251.

[19] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization[J]. Journal of Machine Learning Research, 2011, 12(7).

[20] Zeiler M D. Adadelta: an adaptive learning rate method[J]. arXiv preprint, 2012: 1212.5701.

[21] Kingma D P, Ba J. Adam: A method for stochastic optimization[C]. In: Proceedings of the 2nd International Conference on Learning Representation, 2015.

[22] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]. In: Proceedings of the 22nd ACM International Conference on Multimedia, 2014: 675-678.

[23] Abadi M, Barham P, Chen J, et al. TensorFlow: A system for large-scale machine learning[C]. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, 2016: 265-283.

[24] Chollet F. Keras: The python deep learning library[J]. Astrophysics Source Code Library, 2018: ascl: 1806.022.

[25] Paszke A, Gross S, Chintala S, et al. Automatic differentiation in pytorch[C]. In: Proceedings of the Conference on Neural Information Processing Systems, 2017.

[26] 陈云. 深度学习框架 PyTorch：入门与实践[M]. 北京：电子工业出版社，2018: 2.

[27] Loy C C, Chen K, Gong S, et al. Crowd counting and profiling: Methodology and evaluation[M]. Modeling, Simulation and Visual Analysis of Crowds, 2013: 347-382.

[28] Wu B, Nevatia R. Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors[C]. In: Proceedings of the IEEE International Conference on Computer Vision, 2005: 90-97.

[29] Sabzmeydani P, Mori G. Detecting pedestrians by learning shapelet features[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007: 1-8.

[30] Viola P, Jones MJ. Robust real-time face detection[J]. International Journal of Computer Vision, 2004, 57: 137–154.

[31] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005: 886-893.

[32] Gao C, Liu J, Feng Q, et al. People-flow counting in complex environments by combining depth and color information[J]. Multimed Tools Appl, 2016, 75: 9315–9331.

[33] Viola P, Jones M J, Snow D. Detecting pedestrians using patterns of motion and appearance[C]. In: Proceedings of the IEEE International Conference on Computer Vision, 2003: 734-741.

[34] Gall J, Yao A, Razavi N, et al. Hough forests for object detection, tracking, and action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33: 2188-2202.

[35] Lin S F, Chen J Y, Chao H X. Estimation of number of people in crowded scenes using perspective transformation[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2001, 31(6): 645-654.

[36] Li M, Zhang Z, Huang K, et al. Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection[C]. In: Proceedings of the 19th International Conference on Pattern Recognition, 2008:1-4.

[37] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010: 1627-1645.

[38] Laradji I H, Rostamzadeh N, Pinheiro P O, et al. Where are the blobs: Counting by localization with point supervision[C]. In: Proceedings of the European Conference on Computer Vision, 2018: 547-562.

[39] Liu Y, Shi M, Zhao Q, et al. Point in, box out: Beyond counting persons in crowds[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 6469-6478.

[40] Chan A B, Liang Z S J, Vasconcelos N. Privacy preserving crowd monitoring: Counting people without people models or tracking[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008: 1-7.

[41] Chan A B, Vasconcelos N. Bayesian poisson regression for crowd counting[C]. In: Proceedings of the IEEE International Conference on Computer Vision, 2009: 545-551.

[42] Chen K, Loy C C, Gong S, et al. Feature mining for localised crowd counting[C]. In: Proceedings of the 23rd British Machine Vision Conference, 2012: 21.1-21.11.

[43] Idrees H, Saleemi I, Seibert C, et al. Multi-source multi-scale counting in extremely dense crowd images[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013: 2547-2554.

[44] Rodriguez M, Laptev I, Sivic J, et al. Density-aware person detection and tracking in crowds[C]. In: Proceedings of the IEEE international conference on computer vision, 2011: 2423-2430.

[45] Pham V Q, Kozakaya T, Yamaguchi O, et al. Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation[C]. In: Proceedings of the IEEE International Conference on Computer Vision, 2015: 3253-3261.

[46] Xu B, Qiu G. Crowd density estimation based on rich features and random projection forest[C]. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2016: 1-8.

[47] Fu M, Xu P, Li X, et al. Fast crowd density estimation with convolutional neural networks[J]. Engineering Applications of Artificial Intelligence, 2015, 43: 81-88.

[48] Wang C, Zhang H, Yang L, et al. Deep people counting in extremely dense crowds[C]. In: Proceedings of the 23rd ACM International Conference on Multimedia, 2015: 1299-1302.

[49] Zhang Y, Zhou D, Chen S, et al. Single-image crowd counting via multi-column convolutional neural network[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern

Recognition, 2016: 589-597.

[50] Sindagi V A, Patel V M. Generating high-quality crowd density maps using contextual pyramid cnns[C]. In: Proceedings of the IEEE International Conference on Computer Vision, 2017: 1861-1870.

[51] Cao X, Wang Z, Zhao Y, et al. Scale aggregation network for accurate and efficient crowd counting[C]. In: Proceedings of the European Conference on Computer Vision, 2018: 734-750.

[52] Cheng Z, Li J, Dai Q, et al. Improving the learning of multi-column convolutional neural network for crowd counting[C]. In: Proceedings of the 27th ACM International Conference on Multimedia, 2019: 1897–1906.

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

[54] Zhang A, Shen J, Xiao Z, et al. Relational attention network for crowd counting[C]. In: Proceedings of the IEEE International Conference on Computer Vision, 2019: 6787–6796.

[55] Jiang X, Xiao Z, Zhang B, et al. Crowd counting and density estimation by trellis encoder-decoder networks[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 6133–6142.

[56] Ding X, He F, Lin Z, et al. Crowd density estimation using fusion of multi-layer features[J]. IEEE Transactions on Intelligent Transportation Systems, 2020: 1-12.

[57] Hou Q, Cheng M M, Hu X, et al. Deeply supervised salient object detection with short connections[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(4): 815-828.

[58] Loffe S, Szegedy C. Batch Normalization: Accelerating deep network training by reducing internal covariate shift[C]. In: Proceedings of the 32nd International Conference on Machine Learning, 2015: 448-456.

[59] Zeiler M D, Krishnan D, Taylor G W, et al. Deconvolutional networks[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010: 2528-2535.

[60] Sindagi V A, Patel V M. CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting[C]. In: Proceedings of the 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, 2017: 1-6.

[61] Guo D, Li K, Zha Z J, et al. DADNet: Dilated-attention-deformable convnet for crowd counting[C]. In: Proceedings of the 27th ACM International Conference on Multimedia, 2019: 1823-1832.

[62] Zeng X, Wu Y, Hu S, Wang R, Ye Y. DSPNet: Deep scale purifier network for dense crowd Counting[J]. Expert Systems with Applications, 2020, 141: 112977.

[63] Idrees H, Tayyab M, Athrey K, et al. Composition loss for counting, density map estimation and localization in dense crowds[C]. In: Proceedings of the European Conference on Computer Vision, 2018: 532-546.

[64] Ding X, Lin Z, He F, et al. A deeply-recursive convolutional network for crowd counting[C]. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal

Processing, 2018: 1942-1946.

[65] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[66] Marsden M, McGuinness K, Little S, et al. Fully convolutional crowd counting on highly congested scenes[C]. In: Proceedings of the International Conference on Computer Vision Theory and Applications, 2017: 27–33.

[67] Shi Z, Zhang L, Liu Y, et al. Crowd counting with deep negative correlation learning[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 5382-5390.

[68] Liu X, Weijer J van de, Bagdanov A D. Exploiting unlabeled data in CNNs by self-supervised learning to rank[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019: 1862–1878.

[69] Zhang A, Yue L, Shen J, et al. Attentional neural fields for crowd counting[C]. In: Proceedings of the IEEE International Conference on Computer Vision, 2019: 5713–5722.

[70] Liu W, Salzmann M, Fua P. Context-aware crowd counting[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 5099–5108.

[71] Oh M, Olsen P A, Ramamurthy K N. Crowd counting with decomposed uncertainty[C]. In: Proceedings of the 34rd AAAI Conference on Artificial Intelligence, 2020: 11799–11806.

[72] Wang Q, Gao J, Lin W, et al. Pixel-Wise Crowd Understanding via Synthetic Data[J]. International Journal of Computer Vision, 2020, 129: 225–245.

[73] Liu N, Long Y, Zou C, et al. ADCrowdNet: An attentioninjective deformable convolutional network for crowd understanding[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 3225–3234.

[74] Liu C, Weng X, Mu Y. Recurrent attentive zooming for joint crowd counting and precise localization[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 1217-1226.

[75] He K, Sun J. Convolutional neural networks at constrained time cost[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 5353–5360.

[76] Wang Q, Gao J, Lin W, et al. NWPU-Crowd: A large-scale benchmark for crowd counting and localization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

致谢

时光轮转，冬去春来。再回首，已是三个春秋。眉湖水波潋滟，杨柳新绿初现，梅花正值烂漫，郑大的春天一如初见。三年的研究生生涯，往事如昨，历历在目。短暂的三年，收获了很多，成长了很多。

值此论文完成之际，我要特别感谢导师叶阳东教授。感谢叶老师选择我作为他的学生，也庆幸自己能够成为叶老师的学生。老师崇高的师风师德、渊博的专业知识、严谨的治学态度、精辟的人生感悟令我钦佩不已。在学习上，老师高标准、严要求，引导我如何去做研究，如何向内求。行下易，行上难。做学问就是一个向上求索的过程。老师在潜移默化中提高了我对自己的要求，同时也坚定了我对科研的追求。在生活上，老师也教会了我很多做人做事的道理，对于我以后的人生很有益处。老师孜孜不倦的教诲使我在短暂的三年里学到了知识，锤炼了品质，磨砺了性格。得遇良师，幸甚！感谢师恩！

感谢数据挖掘与机器学习实验室的姬波老师、卢红星老师、朱真峰老师、娄铮铮老师、吴云鹏老师、闫小强老师、田侦老师和吴宾老师，他们严谨的科研作风、精益求精的工作态度，使我深受感触，以他们为榜样，不断激励自己。感谢博士师兄胡世哲、张明明、李辉、孙中川和师姐王有为在学术和生活上的指点和帮助。感谢曾鑫师兄在科研和学术上给予的指引和提供的帮助。感谢侯振泉、鲁博仁、夏春管、何旭蔚、王若彬师兄和张雪、张曦师姐的帮助和鼓励。感谢毛奕桥、刘豪森、徐富国、王博、崔佳彬等人的三年同窗，感谢张麒、方海川、钟李红、史凯远等师弟师妹的一路同行。

特别感谢我的父母对我人生的支持，感谢他们的爱与付出。感谢妻子在背后的默默付出和陪伴，感谢两个小女儿给予的快乐和动力，也感谢妹妹的帮助和照顾。感谢他们的支持和鼓励，使我顺利完成学业。

最后，感谢参加论文审阅和答辩的各位专家和学者。

郭强

二零二一年五月