分类号：TP391　　单位代码：10363

密　级：公 开　　学　号:2210330167

题　目 基于深度学习的密集场景人群计数方法研究

题　目 RESEARCH ON CROWD COUNTING FOR CONGESTED SCENES BASED ON DEEP LEARNING

学　生　姓　名：林园园

校　内　导　师：杨会成（教授）

校　外　导　师：刘玉莉（高级工程师）

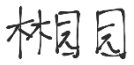
专　　业：电子信息

研　究　方　向：人工智能与系统集成

论文答辩日期：2023 年 5 月 23 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名： 

日期：　2024 年 05 月 31 日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 □，在 ____ 年解密后适用本版权书。

本学位论文属于

不保密 ☑。

学位论文作者签名：林园园　　指导教师签名：杨云成

日期：2024 年 05 月 31 日　　日期：2024 年 05 月 31 日

基于深度学习的密集场景人群计数方法研究

摘 要

在现代社会中，人群密集场景广泛存在于各种场所，如车站、商场、体育场馆等，因此准确、高效地进行人群计数具有重要意义。研究密集场景下的人群计数方法有助于更好地理解和管理人群流动情况，提高场所的安全性和管理效率。深度学习由于其强大的特征提取与表征能力，能够应用于人群空间分布的准确估计，逐渐成为了人群计数领域的主流研究方法。常规的人群计数模型由密度特征提取模块与计数模块组成，但是当前的密度特征提取方法通常存在着较大的参数量或计算量；另一方面当前人群计数模块的框架较为单一，通常基于卷积神经网络（Convolutional Neural Networks，CNN）或 Transformer 模型构建，没有将两种模型的优势耦合，难以实现场景变换和人群分布不均匀等复杂场景下人群的精确计数。针对上述问题，本文通过研究密度特征提取网络的轻量化算法，以减少计算量和参数量为优化目标提出人群计数模型的压缩方法；研究 CNN 和 Transformer 模型的融合算法，构建多任务并行的人群计数模型，实现了计数实时性和精度的有效平衡。本文主要研究内容和结论如下：

（1）**构建了人群计数的轻量化密度特征提取网络。**针对目前人群计数模型密度特征提取模块计算量、参数量过大的问题，提出了一种轻量化的密度特征提取方法。将卷积层中的非线性激活简化为线性映射，降低样本特征提取中的计算量。为了避免由此造成的特征损失，融合浅层特征图和深层特征图增强模型的表征能力，最后回归预测密度图实现人群计数。实验结果表明，本文提出的计数模

型的参数量为 0.11M，计算量为 7.41GFLOPS，对比其它人群计数模型均有所降低。与轻量级的人群计数模型相比，模型的实时性和精度上均优于前人提出的轻量化人群计数模型。但与常规的人群计数模型相比，本文提出的轻量化人群计数模型在复杂场景下的人群计数精度有待进一步提升。

（2）构建了面向密集场景人群计数的多任务并行人群计数网络。针对目前复杂背景给人群计数模型精度造成的影响，耦合 CNN 的局部特征提取优势和 Transformer 对于全局特征的捕获能力，构建了一种 CNN 和 Transformer 并行的方法进行多任务人群计数。在密度图预测网络分支采用扩张卷积网络建立从单张图像真实密度图到预测密度图的映射；在直接计数网络分支采用 Vision Transformer（ViT）建立从单张图像到人群数量的映射；利用全局平均池化来动态生成权值因子，融合密度图预测网络分支和直接计数网络分支的人数，得到最终预测人数。实验结果表明，本文提出的多任务模型在没有显著增加参数量和计算量的同时明显提升了计数精度，与常规的人群计数模型相比，该模型在取得更低的参数量和计算量的同时得到了更优异的计数精度，与轻量化人群计数模型相比，虽然没有取得最低的计算量，但是计数精度显著高于对比的轻量化人群计数模型，证明了本文提出模型计数精度的优越性。

综上所述，本文通过构建轻量化密度特征提取网络来减少模型的计算成本，同时进一步优化了模型的人群计数模块，提出了一种多任务并行的人群计数网络来提高模型的计数精度，为密集场景下实现快速人群监测和有效安全预警提供了技术支撑。

关键词： 人群计数，密集场景，深度学习，模型压缩，多任务并行

RESEARCH ON CROWD COUNTING FOR CONGESTED SCENES BASED ON DEEP LEARNING

ABSTRACT

In modern society, crowded scenes are widely found in various places, such as stations, shopping malls, stadiums, etc. So it is of great significance to accurately and efficiently count people. Studying crowd counting methods in dense scenarios can help to better understand and manage crowd flow, and improve the safety and management efficiency of venues. Due to its powerful feature extraction and representation capabilities, deep learning can be applied to the accurate estimation of the spatial distribution of populations, and has gradually become a mainstream research method in the field of crowd counting. The conventional crowd counting model is composed of a density feature extraction module and a counting module, but the current density feature extraction method usually has a large number of parameters or computational costs; On the other hand, the framework of the current crowd counting module is relatively simple, which is usually built based on Convolutional Neural Networks (CNN) or Transformer models, which does not couple the advantages of the two models, and it is difficult to achieve accurate counting of people in complex scenarios such as scene transformation and uneven population distribution. In order to solve the above problems, this paper proposes a compression method for the crowd counting model by studying the lightweight algorithm of the density feature extraction network, with the optimization goal of reducing the amount of computation and parameters, and studies the fusion algorithm

of CNN and Transformer model to construct a multi-task parallel crowd counting model, which realizes the effective balance between real-time counting and accuracy. The main research contents and conclusions of this paper are as follows:

(1) A lightweight density feature extraction network for crowd counting has been constructed. A lightweight density feature extraction method is proposed to address the issues of high computational and parameter complexity in the current crowd counting model density feature extraction module. Simplify the nonlinear activation in convolutional layers into linear mappings, reducing the computational complexity in sample feature extraction. In order to avoid feature loss caused by this, shallow and deep feature maps are fused to enhance the model's representation ability, and finally, a regression prediction density map is used to achieve crowd counting. The experimental results show that the parameter size of the proposed counting model is 0.11M, and the computational complexity is 7.41GFLOPS, which is lower compared to other crowd counting models. Compared with the lightweight crowd counting model, the real-time and accuracy of the model are better than the lightweight crowd counting model proposed by the predecessors. However, compared with conventional crowd counting models, the lightweight crowd counting model proposed in this paper needs further improvement in crowd counting accuracy in complex scenarios.

(2) **A multi-task parallel crowd counting network for crowd counting in dense scenes was constructed.** In response to the impact of complex backgrounds on the accuracy of crowd counting models, a parallel method of CNN and Transformer for multi-task crowd counting was constructed by combining the advantages of local feature extraction of CNN and the ability of Transformer to capture global features. In the

density map prediction network branch, an dilated convolutional network is used to establish a mapping from the true density map of a single image to the predicted density map; In the direct counting network branch, Vision Transformer (ViT) is used to establish a mapping from a single image to the number of people; Using global average pooling to dynamically generate weight factors, fusing density maps to predict the number of network branches and directly counting network branches, to obtain the final predicted number of people. The experimental results show that the multi-task model proposed in this paper significantly improves counting accuracy without significantly increasing the number of parameters and computation. Compared with conventional crowd counting models, this model achieves lower number of parameters and computation while achieving better counting accuracy. Compared with lightweight crowd counting models, although it does not achieve the lowest computation, its counting accuracy is significantly higher than that of the compared lightweight crowd counting models, proving the superiority of the proposed model in counting accuracy.

In summary, this paper reduces the computational cost of the model by constructing a lightweight density feature extraction network, and further optimizes the crowd counting module of the model. A parallel CNN and Transformer network is proposed to improve the counting accuracy of the model, providing technical support for fast crowd monitoring and effective security warning in dense scenes.

KEY WORDS: crowd counting, congested scenes, deep learning, model compression, multi-task parallel

目录

摘 要..... I
ABSTRACT.....III
目录.....VI
第 1 章 绪论.....1
1.1 课题研究背景及其意义.....1
1.2 国内外研究现状.....2
1.2.1 基于传统机器学习的人群计数方法.....2
1.2.2 基于深度学习的人群计数方法.....3
1.3 主要研究内容.....6
1.4 本文的技术路线与组织结构.....7
第 2 章 人群计数任务相关理论及技术概述.....9
2.1 引言.....9
2.2 深度学习网络模型.....9
2.2.1 CNN 相关基础.....9
2.2.2 经典 CNN 模型.....13
2.2.3 Transformer 模型.....15
2.3 人群计数数据集和评价指标.....19
2.3.1 人群计数数据集.....19
2.3.2 评价指标.....21
2.4 本章小结.....21
第 3 章 人群密度特征提取网络的轻量化方法研究.....22
3.1 引言.....22
3.2 轻量化密度特征提取网络构建.....23
3.2.1 轻量化线性映射块构建.....23
3.2.2 密度图生成.....25
3.2.3 特征融合.....27
3.2.4 损失函数设计.....28
3.3 实验与结果分析.....29
3.3.1 实验设置.....29
3.3.2 实验训练过程.....30
3.3.3 实验结果和对比分析.....30
3.4 本章小结.....34
第 4 章 密集场景下多任务并行的人群计数方法研究.....35
4.1 引言.....35
4.2 多任务并行人群计数网络构建.....36

4.2.1 密度图预测网络分支.....37
4.2.2 直接计数网络分支.....38
4.2.3 人数融合网络分支.....40
4.2.4 损失函数设计.....41
4.3 实验与结果分析.....42
4.3.1 实验设置.....42
4.3.2 实验训练过程.....42
4.3.3 实验结果和对比分析.....43
4.4 本章小结.....47
第 5 章 总结与展望.....48
5.1 总结.....48
5.2 展望.....49
参考文献.....50
致谢.....56
攻读硕士学位期间的研究成果.....57
1 学术论文.....57
2 发明专利.....57

第1章 绪论

本章由四个部分组成，首先对本课题的研究背景和研究意义进行介绍，其次论述了国内外的研究现状，接着简要介绍了本文的研究和创新内容，最后介绍了技术路线和组织结构。

1.1 课题研究背景及其意义

随着城镇化进程的加快，大型群众活动日益频繁，群体聚集呈常态化、多元化趋势，车站、广场、商场、景点等公共场所经常会出现人群聚集的现象。例如，在地铁高峰时段，部分站台和换乘通道乘客流量密度大，存在拥挤踩踏风险；在大型演唱会期间，极其密集的人群将会加剧安全事故的发生。由于人群拥挤引发的踩踏事故屡见不鲜，2014 年的元旦前一天，上海外滩聚集了大量市民参加跨年庆祝活动，人群相互推搡摔倒，引发了踩踏事故，造成了 85 人伤亡。其中图 1-1 展示了一些聚集人群的场景。人群计数是智能场景理解和分析中的关键任务之一，有助于公共场所人群流量的优化，可以对拥挤、踩踏等事故进行及时的预警，并辅助异常行为的监测。

图 1-1 聚集人群场景图

Fig 1-1 Diagram of the gathering crowd scene

利用人力监控的方法对人群数量进行监测通常需要较多的劳动力[1]，而且由于人力的限制，经常会出现遗漏的现象，无法实现准确的人群计数。因此使用自动化的方式取代人工对密集场景下的人群数量进行监控和预警成为了重要的课题。随着深度学习在目标分割[2][3]，场景感知[4]，行为分析[5]等相关任务中表现出了优异的性能后，人群计数任务也受到了广泛的关注，在场景单一，人群尺度分布均匀的情况下获得了较好的计数精度[6]。然而在实际应用中仍然有

很多因素影响着计数的性能，如图 1-2 所示，多变的场景和不均匀的人群尺度会造成计数精度的损失，以及当前人群计数模型密度特征提取模块参数量和计算量过大影响计数的实时性等问题，目前在人群计数领域的一些算法还不能有效的解决上述问题。因此，为了有效解决人群计数领域存在的这些问题，提高计数的精度，减少计数模型的复杂度进而有效的实现实时计数仍然是一个备受研究关注的课题。

图 1-2 复杂场景

Fig 1-2 Complex scenes

1.2 国内外研究现状

随着人工智能的发展，越来越多的学者致力于人群计数任务的研究，提出了大批计数性能良好的计数模型。针对本课题的研究内容，将从基于传统机器学习的人群计数方法和基于深度学习的人群计数方法两个方面对研究现状进行阐述。

1.2.1 基于传统机器学习的人群计数方法

基于传统机器学习的人群计数方法主要是从检测和回归两方面进行计数。检测的人群计数方法是利用一个可以进行滑动的矩形窗口在视频帧或者图像上进行滑动，检测出人群数量。例如在文献[7]中提出了一种可检测的框架，该框架对连续视频帧进行扫描检测以估计总的人群数量。利用检测的计数方法在低密度场景中行之有效，但是在高密度人群中，基于检测的方法无法识别被遮挡住的人，最终会影响计数的精度。为了解决计数场景中人群遮挡的问题，一些学者开始尝试使用回归的方法进行人群计数，回归的整体思想是从人群图像中提取到的特征的回归出人群数量[8]， Idrees 等人[9]提出构建多源特征来分析区域

中的目标，作者在文章中公开了一个 50 幅图像的人群计数数据集 UCF CC 50，该数据集对 64000 个人头都进行了标注。然而检测和回归的方法在对全局计数进行回归时忽略了图像中存在的关键空间信息，因此，为了结合图像中存在的空间信息，Lempitsky 等人[10]引入了一种学习局部补丁特征和相应对象密度图之间的线性映射的新方法。然而这些模型依赖于表征能力有限的手工特征，难以克服行人间的遮挡、形变以及场景中光照变化等问题。渐渐的，基于传统机器学习的人群计数方法逐渐被基于深度学习的人群计数方法所取代。

1.2.2 基于深度学习的人群计数方法

深度学习的方法给人群计数提供了解决问题的新思路。CNN 和 Transformer 是适用于人群特征提取的两种主流深度学习模型，CNN 使用感受野有限的卷积核来学习细节特征，但它在建模上下文依赖性方面的能力较弱。Transformer 利用多头自注意力机制，能够捕捉全局信息，但是可能会忽略一些局部特征。下面将由 CNN 和 Transformer 两个方面来介绍深度学习的人群计数方法。

（1）基于 CNN 的人群计数研究现状

CNN 在图像识别[11]、目标检测[12]和语义分割[13]等各种计算机视觉任务中表现出优异的性能。基于 CNN 的模型在人群计数任务上也取得了显著进展，在基于 CNN 的早期计数方法里，CNN 计数模型主要是通过全连接层回归。文献[14]全面介绍了基于 CNN 的计数方法。Wang 等人[15]改进了 AlexNet[11]，用于直接预测计数。HANG 等人[16]提出了一种由人群密度和人群计数交替训练的 CNN 模型。Walach[17]利用分层提升和选择性采样方法来减少计数估计误差。与现有的基于补丁的估计方法不同，Shang 等人[18]使用了一种同时估计整个输入图像的局部和全局计数的网络。但是这些模型都是通过全连接层直接回归出人数，这种计数方法无法有效捕捉空间信息，导致难以取得令人满意的结果。针对该问题，研究者开始探究密度图在人群计数中的应用，这是由于其可以在统计数量不变的前提下保持目标在空间上的分布。对于人群计数任务，CNN 的一些变体专门设计用于密度图估计。例如，CrowdNet[19]将五层深度卷积分支和三层浅卷积分支相结合来提取群组特征；MCNN[20]是一个多列 CNN 框架，它学习具有 3×3、5×5 和 7×7 不同核大小的人群密度图；切换 CNN[21]可以被认为

是 MCNN 的改进版本，它遵循多列结构，并使用额外的切换分类器选择一列来预测人群密度图；CSRNet[22]应用扩张卷积来取代后端的标准卷积，由于其不断扩大的感受野在人群计数中实现了显著的性能改进。CP-CNN[23]提出了一种上下文金字塔的卷积神经网络模型，有效的结合了人群输入图像的多尺度特征，使其在人群密度变化较大的场景下也能够生成高质量的密度图。此外，编码器-解码器 CNN 结构[24][25][26]将低级空间线索与高级语义线索相融合，也适用于人群密度估计。然而常规的人群计数模型虽然可以达到较为精确的计数精度，但这些人群计数模型的密度特征提取模块都需要大量的参数量和计算量，计数的效率不高。例如经典的 CNN 模型 ResNet-50[27]，它被广泛用于人群计数模型的密度特征提取模块进行特征的提取，然而该网络处理一个 224×224 的图像时，需要使用较大的参数量和计算量。因此，轻量化的 CNN 目前成为研究的新趋势。

（2）基于轻量化 CNN 的人群计数研究现状

近年来，人们提出了一系列研究紧凑深度神经网络的方法，如网络剪枝[28]、低比特量化[29]、知识蒸馏[30]等。文献[31]提出对神经网络中不重要的权值进行修剪，利用“l_1范数正则化”对高效 CNN 滤波器进行剪枝。除此之外，高效的神经结构设计对于用更少的参数和更低的计算量建立高效的深度网络具有非常大的潜力，基于 CNN 的轻量化模型在近期取得了优异的表现，例如，MobileNet[32]利用深度卷积和点向卷积构建了一个单元，用更大的滤波器近似原始卷积层，在大大降低了模型的参数量和计算量的同时取得了优异的性能，为模型的实用化提供了可能性。目前轻量级的卷积神经网络在人群计数领域有着广泛的应用。LCNet[33]提出了一个轻量级的端到端网络用于人群计数。该方法提出了一个尺度感知的模块，并利用该模块提取出不同尺度的特征，最后实现特征图到预测密度图的映射，使用了具有较小空间滤波器的级联卷积降低了模型的复杂度。PDDNet[34]使用所提出的轻量级金字塔卷积模块进行多尺度特征提取。LMSFFNet[35]选择 MobileViT 模块作为网络的骨干，以减少网络参数的数量和计算成本。SANet[36]采用转置卷积生成密度图。PCC-Net[37]利用对中间特征进行编码的方式，对背景进行分割。然而这些模型为了降低复杂度，大多会使用较少的卷积或忽略特征的融合，这必然会导致计数精度的损失。

（3）基于 Transformer 的人群计数研究现状

Transformer 主导了自然语言建模[45][46]，它利用自注意力机制来捕捉输入和输出之间的全局依赖关系。最近，许多工作[47][48][49]试图将 Transformer 应用到视觉任务中。具体而言，DETR[50]首先利用 CNN 主干提取视觉特征，然后使用 Transformer 块进行框回归和分类。ViT[51]是第一个将 Transformer 编码器[42]直接应用于图像序列补丁以实现分类任务的方法。此外，Transformer 还用于处理其他视觉任务，如多目标跟踪[52]、图像生成[53]、医学图像分割[54]。

近年来，Transformer 在人群计数的任务上也取得了优异的成果。Transformer 提供了全局感受野，更适用于基于整个图像的人数回归。CrowdFormer[55]是模仿人类自上而下的视觉感知机制，利用重叠补丁 Transformer 块预测人群密度图的代表性研究。此外，Transformer 被证明可用于弱监督人群计数，即直接回归计数，而不是密度图估计。具体而言，TransCrowd[56]是一个纯基于 Transformer 的框架，采用序列计数预测。CCST[57]继承了弱监督学习的思想，该思想使用 Swin Transformer 作为骨干，实现了性能改进。然而，Transformer 利用多头自注意力机制，能够捕获全局信息，但它可能忽略局部区域内的特性细节，从而影响计数精度。

（4）多任务人群计数研究现状

多任务学习是指在一个权值共享的学习框架中共同解决多个相关的任务。正如人群分布可以从不同的角度进行分析一样，人们提出了许多基于多任务学习的方法来解决人群计数难题。例如，在文献[38]中，利用两个监督信号来提取人群特征，其中一个是对密度水平进行分类，另一个是预测总数；Wan 等人[39]提出了一种细粒度的人群计数框架，不仅可以预测密度图，还可以识别暴力行为；Jiang 等人[40]设计了一个掩码感知网络来分割前景/背景掩码，并指导人群计数的结果；Meng 等人[41]采用图卷积网络实现了交通物体计数的自适应辅助学习。但除此之外，更多的研究[42][43][44]将定位点人数和密度图估计作为两个主要的学习目标。具体而言， DecideNet[42]自适应平衡头点定位和密度图估计相关任务之间的权值；Liu 等[43]设计了循环关注缩放网络来定位和估计不同图像分辨率下的人群分布。Wan 等人[44]引入了最优传输理论，并使用广义多任务损失进行群体代表性学习。

综上所述，在人群计数领域，已经拥有多角度的计数方法，但有一个共同

的技术方案: 特征提取和基于计数的学习。早期的研究利用人工特征来描述人群密度，而深度学习框架逐渐取代了传统的人群特征提取方法，典型的有 CNN、Transformer。就学习目标而言，现有的方法也可以分为三种类型：直接计数、密度图回归计数和多任务学习计数。通过对相关工作的综合分析，在第四章的研究中，采用 CNN 和 Transformer 的混合方法，构建了一个多任务的并行分支网络，既实现了直接计数，又实现了密度图估计计数，以解决当前人群计数任务中场景变化快以及人群分布不均导致的计数不佳的问题。

1.3 主要研究内容

当前大多数人群计数模型的密度特征提取模块存在着计数效率低，复杂度高的缺点，无法实现实时计数，且人群计数中多变的场景和变化的人群尺寸会影响计数的精度。针对目前人群计数任务存在的上述难点和挑战，本文提出了两种基于深度学习的密集场景人群计数的网络模型，其中主要研究内容如下。

（1） 构建了一种轻量化密度特征提取网络

针对当前人群计数领域实际应用中计算消耗大，实时性较差的问题，提出了一种轻量化密度特征提取方法，该方法的密度特征提取网络由本文设计的轻量化线性映射块组成，将常规卷积中的非线性激活简化为线性映射，减少了模型的计算量和参数量，为了避免线性映射造成的特征丢失，融合了浅层特征图和深层特征图增强信息的利用率。最后回归出预测密度图得到预测人数，在降低模型复杂度的同时保持了计数精度。最后在三个数据集上进行了实验，证明了该模型保持计数精度的同时，降低了参数量和计算量。

（2）构建了一种密集场景下多任务并行人群计数网络

针对当前人群计数任务背景的复杂度较高以及人群尺度变化较快从而影响了计数精度的问题，结合 CNN 和 Transformer 模型的优势，为了进一步提高人群计数的精确性，本文设计了一个多任务并行人群计数方法，该方法首先采用轻量化网络提取人群密度特征，其次优化人群计数模块，在人群计数模块采用 CNN 和 Transformer 的混合方法，进行多任务学习，在密度图预测网络分支建立真实密度图到预测密度图的映射，在直接计数网络分支建立单张图像到预测人数的映射，最后在人数融合阶段，提出了一个全新的人数融合网络分支动态

生成权值因子加权融合密度图预测网络分支和直接计数网络分支的人数。最后在三个数据集上进行了实验，该模型在不明显增长参数量和计算量的同时，显著提升了计数精度，与一众人群计数模型对比也取得了优异的计数性能。

1.4 本文的技术路线与组织结构

本文的技术路线如图 1-3 所示，图中对本文的技术路线进行了大致的解释，其中包括本文解决的问题，以及根据问题提出的解决方法，最后对提出的方法进行了实验对比。本文根据研究内容和方法进行了结构上的安排，主要分为五个部分，其内容如下：

第一章主要介绍人群计数背景以及意义，并介绍人群计数在国内外的研究现状，以及针对现有的基于深度学习的人群计数算法的一些优缺点和改进方法，最后阐述本文的主要研究内容和组织结构。

第二章介绍了人群计数中的 CNN 和 Transformer 的基础知识和概念，接着介绍了一些经典的 CNN 模型，其次介绍了人群计数数据集和实验评价指标。

第三章介绍了本文设计的轻量化密度特征提取网络模型，首先对密度特征提取网络设计的轻量化线性映射块进行介绍，其次对不同层次的线性映射块生成的特征图融合进行介绍，接着对密度图生成的方法进行介绍，其次介绍了特征提取网络的网络配置，最后在三个公开数据集上进行实验对比分析。

第四章介绍了本文设计的多任务并行人群计数网络，对人群计数的三个分支：密度图预测网络分支，直接计数网络分支，人数融合网络分支进行介绍。最后在三个公开数据集上与其他人群计数方法进行了实验对比。

最后一章对本文研究工作进行了总结，并对未来的研究方向进行了分析和探讨。

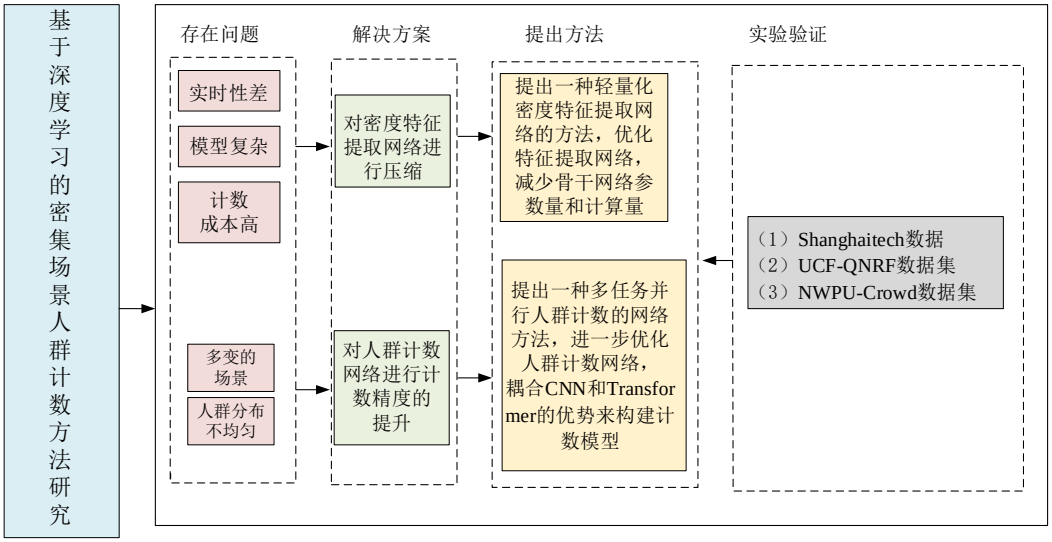


图 1-3 技术路线图

Fig 1-3 Technology roadmap

第2章 人群计数任务相关理论及技术概述

2.1 引言

本章首先对 CNN 基础知识和经典 CNN 模型进行介绍，其次介绍了 Transformer 网络，最后对相关数据集和实验评价指标进行介绍。

2.2 深度学习网络模型

2.2.1 CNN 相关基础

CNN 是深度学习中的一个重要模型，它在图像分类[27]、目标检测[58]、图像分割[3]等计算机视觉任务中广泛应用，是目前应用最普遍的模型之一。CNN 主要由卷积层、激活函数、池化层、全连接层等组成。

（1）卷积层

在 CNN 中，卷积层是必不可少的一个重要层，它的目的是利用卷积从输入的图像中提取到有效的特征。CNN 中的一层卷积层往往由多个卷积单元构成，利用反向传播的方法优化每一个卷积单元的参数。CNN 中会使用最少一种卷积操作来替代常规的矩阵乘法。卷积在图像处理的领域中是十分重要的计算法则。

在信号处理领域，往往使用一维卷积对信号进行处理，一维卷积运算操作是对两个实变的信号函数进行卷积操作，具体运算如下：对于一个输入信号列 b_s， $s=1,2,3,...,a$，经过一维卷积后的输出为：

$$d_s=\sum_{F=1}^{F}\beta_f b_{s-f+1} \tag{2-1}$$

其中 β_f 是卷积运算中的卷积核，一般情况下，卷积核的长度 F 会远小于信号长度 a。

对于二维图像 $f(i,j)$,滤波器 $h(k,l)$，二维卷积定义如公式 2-2 所示：

$$g(i,j)=f\otimes h=\sum_{k,l}f(i+k,j+l)h(k,l) \tag{2-2}$$

图像卷积的相关运算如图 2-1 所示，其中 F 为滤波器，X 为图像，O 为结果。

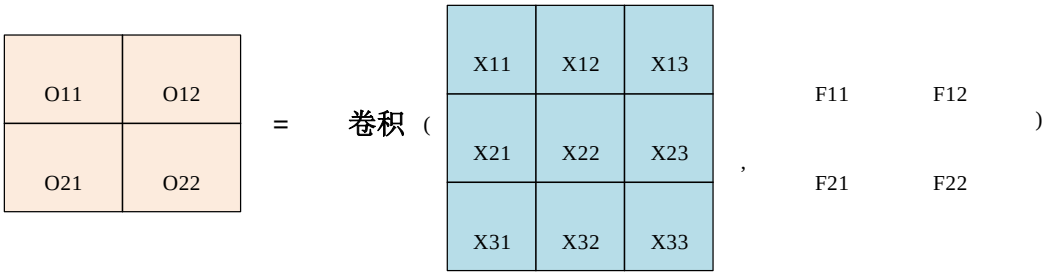


图 2-1 图像卷积相关运算

Fig 2-1 Image convolution correlation operation

$$O_{11} = F_{11}X_{11} + F_{11}X_{12} + F_{21}X_{21} + F_{22}X_{22} \quad (2\text{-}3)$$

$$O_{12} = F_{11}X_{12} + F_{12}X_{13} + F_{21}X_{22} + F_{22}X_{23} \quad (2\text{-}4)$$

$$O_{21} = F_{11}X_{21} + F_{12}X_{22} + F_{21}X_{31} + F_{22}X_{32} \quad (2\text{-}5)$$

$$O_{22} = F_{11}X_{22} + F_{12}X_{23} + F_{21}X_{32} + F_{22}X_{33} \quad (2\text{-}6)$$

假设一个输入的图像尺寸为 $H_a \times W_a \times C_a$，经过 N 个尺寸为 $F \times F$、步长为 S、零填充数为 P 的卷积操作后，输出的特征图的尺寸为 $H_b \times W_b \times C_b$，并满足公式 2-7，2-8，2-9。

$$H_b = \frac{(H_a - F + 2P)}{S} + 1 \quad (2\text{-}7)$$

$$W_b = \frac{(W_a - F + 2P)}{S} + 1 \quad (2\text{-}8)$$

$$C_2 = N \quad (2\text{-}9)$$

当步长 S =1，且零填充数 P =（F -1）/2 时，输出特征图的尺寸和输入图尺寸一致。

（2）激活函数

在 CNN 模型中，经过卷积层提取到的特征一般为线性特征，然而在实际应用中，往往需要解决一些非线性的问题，此时会在卷积层后再加上一层非线性激活函数层。常见的激活函数有逻辑回归激活函数（Sigmoid），双曲正切激活函数（Hyperbolic Tangent Function, Tanh），修正线性单元激活函数（Rectified Linear Unit, ReLu），下面依次进行介绍。

Sigmoid 函数使得输出值被限制在区间（0，1）内，可以应用在一些权重项的生成或者二分类任务，图 2-2(a)所示，在输入极小的情况下，输出为 0；在输入极大的情况下，输出为 1。数学公式如公式 2-10 所示。由图像可知在输入值

极大或极小时，输出变化十分缓慢，梯度无限接近于 0，因此在 CNN 网络层数增加后，可能会造成输出的权重参数无法更新的现象，且由于 Sigmoid 函数输出在区间（0，1）内，因此输出的结果均值非 0，会导致函数的梯度无法收敛。

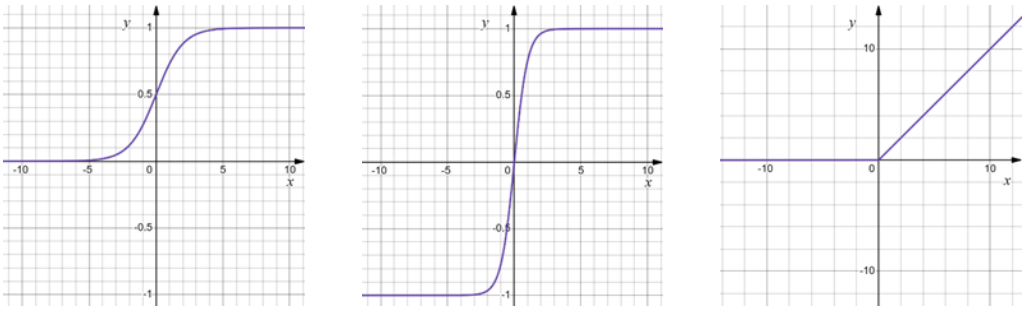
$$f(x)=\frac{1}{1+e^{-x}} \tag{2-10}$$

Tanh 函数改善了 Sigmoid 函数梯度无法收敛的问题，它将输出区间限制在（-1，1）之间，解决了输出结果均值非 0 的问题，数学公式如公式 2-11 所示，数学图像如图 2-2(b)所示。这样的改进虽然改善了在网络训练中的收敛速度，但是并没有解决梯度消失的问题。

$$f(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2-11}$$

ReLu 函数解决了上述激活函数的不足，它是一种分段函数，数学公式如公式 2-12 所示，当输入小于 0 时，输出为 0。当输入大于 0 时，输出为其本身。数学图像如图 2-2(c)所示。

$$f(x)=\max(0,x)=\begin{cases}0, & x\le 0\\ x, & x>0\end{cases} \tag{2-12}$$



(a) (b) (c)

图 2-2 激活函数

Fig 2-2 Activation function.

（3）池化层

一般经过卷积层操作后，特征维度会相应变大，此时需要经过池化层对特征维度进行降低。池化层分为两种：最大池化和平均池化，如图 2-3 所示。具体操作即使用一个尺寸大小 2×2，步长为 1 的矩形窗口在特征图上滑动，取滑动区域的最大值或者平均值为输出。公式 2-13 为最大池化，在第 s 个特征图内进行最大池化操作，x_{sij} 为操作区域 L_{ab} 中的所有元素，f_{sab} 为操作区域 L_{ab} 的最

大池化输出。公式 2-14 为平均池化，在第 s 个特征图内进行平均池化操作，x_{sij} 为操作区域 L_{ab} 中的所有元素，$|L_{ab}|$ 区域 L_{ab} 中元素个数，f_{sab} 为操作区域 L_{ab} 的平均池化输出。

$$f_{sab} = \max_{(i,j)\in L_{ab}} x_{sij} \tag{2-13}$$

$$f_{sab} = \frac{1}{|L_{ab}|}\sum_{(i,j)\in L_{ab}} x_{sij} \tag{2-14}$$

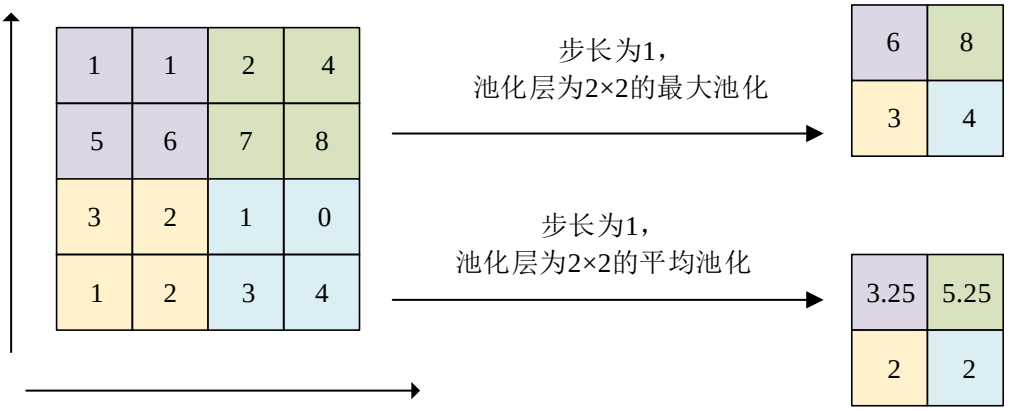


图 2-3 最大池化和平均池化

Fig 2-3 Maximum pooling and average pooling

池化层一般在两层卷积层中间使用，用于减少特征图的空间尺寸，提取出关键的特征。在人群计数任务中，输入需要进行人数估计的图片，在经过卷积层后，利用池化层对输入的特征图进行压缩降维。在 CNN 中，权重参数会随着网络的深度逐渐变大，因此在训练一个较大的 CNN 模型时，一般需要服务器或处理器进行加速，而池化层可以有效的降低网络压力，提高训练速度。

（4）全连接层

全连接层一般出现在网络的最后几层，可以起到分类器的作用，在输出时对数据进行连接并按照分类形式进行分类。全连接层将第 i 层特征图中的一些特征元素和 $i+1$ 层中的神经元进行连接，公式如式 2-15 所示。

$$Y = f(W * x + b) \tag{2-15}$$

其中 W 是权重矩阵，x 是输入，b 是偏置，f 是激活函数。权重矩阵 W 是一个矩阵，每一列代表上一层的输出特征，每一行代表下一层的输入特征，每一个元素代表两个特征之间的关系。

2.2.2 经典 CNN 模型

目前 CNN 已经广泛应用于计算机视觉等领域。其中 VGG16[59] 和 ResNet[27] 因为其强大的迁移能力和特征提取能力能够很好的适用于分类任务，也被广泛地迁移应用于人群计数任务[60]中作为特征提取器。下面对 VGG16 和 ResNet 网络进行介绍。

（1）VGG16

CNN 的一些模型在计算机视觉任务以及图像处理领域取得了巨大的成功，一些 CNN 模型也横空出世。VGG16 作为经典的模型之一，由牛津大学几何组于 2014 年首次提出。VGG16 模型网络结构简单明了，特征提取能力强大，被广泛的应用于图像处理任务[61]。VGG16 的网络结构如图 2-4 所示，该模型包含 13 个卷积层和 3 个全连接层。由图可知，所有的卷积层均采用 3×3 的卷积核，步长和填充均为 1。随着特征映射空间尺寸的变小，输出特征映射的数目逐渐增多，从 64 提升到 512。最后三个全连接层分别包含 4096、4096 和 1000 个神经元，用于将特征转化为每种类别的概率。

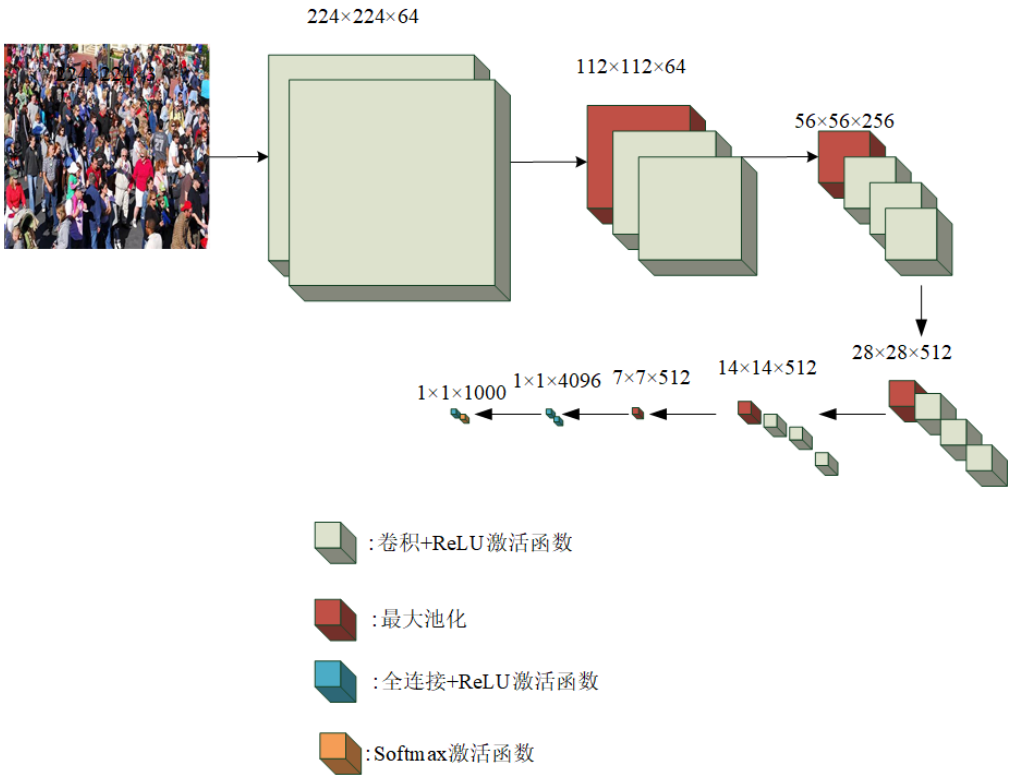


图 2-4 VGG16 的模型结构

Fig 2-4 Model structure of VGG16

VGG 模型以较深的网络结构，较小的卷积核和池化采样域，不同于此前使用大卷积核的网络结构，VGG16 所有卷积层都采用 3×3 的卷积核。使用多个小卷积核的串联达到大卷积核的感受野。使得其能够在获得更多图像特征的同时控制参数的个数，避免过多的计算量以及过于复杂的结构。VGG16 网络目前被广泛应用于各大任务的骨干网络，用来提取有效的信息特征。

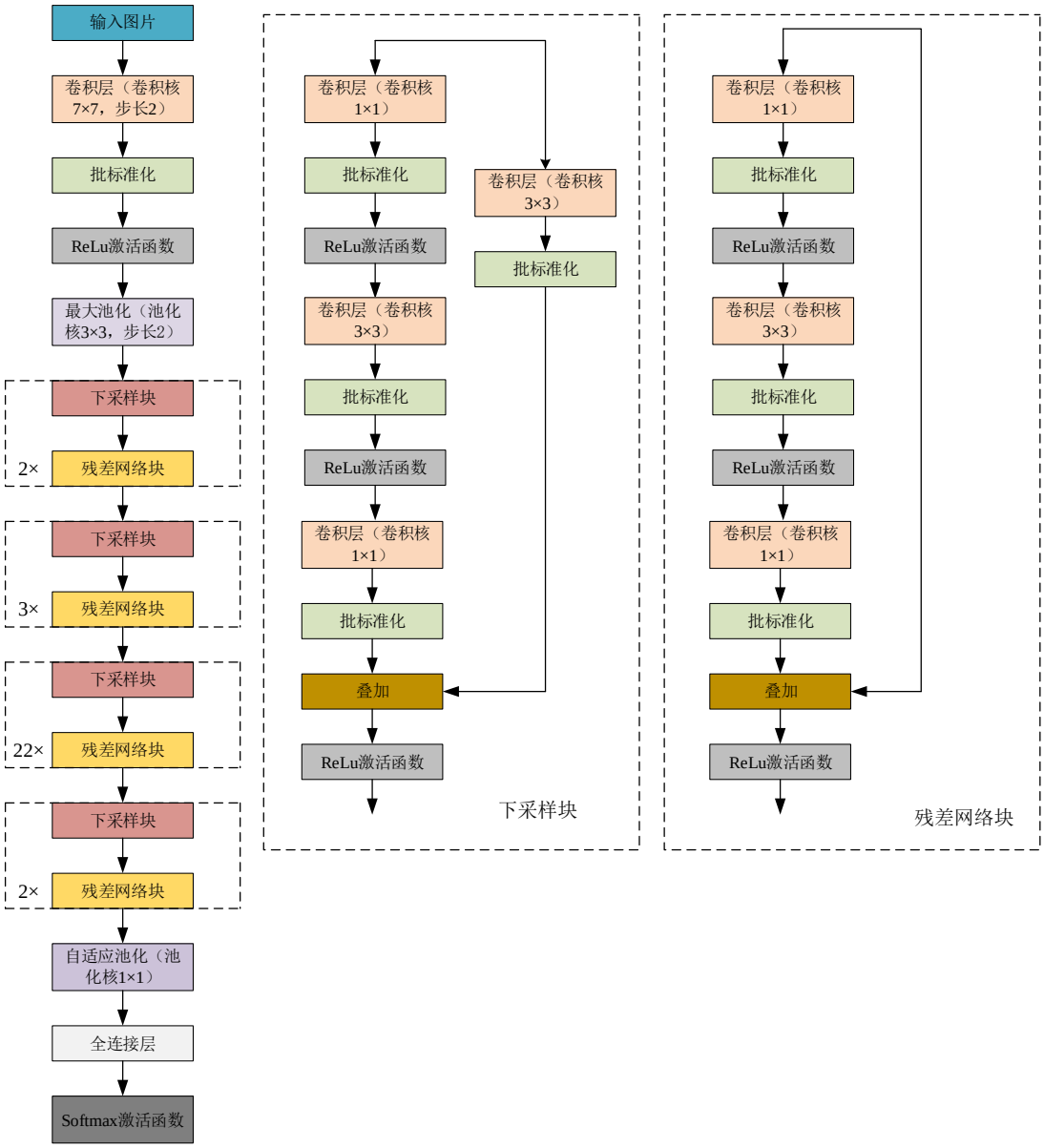


图 2-5 ResNet 的模型结构

Fig 2-5 Model structure of ResNet

（2）ResNet

ResNet 的提出是 CNN 应用于图像处理上的一个重要的里程碑，ResNet 在 ILSVRC 和 COCO2015 上都取得了最优异的成果。ResNet 网络参考了 VGG19

的网络结构，在 VGG19 网络的结构上进行修改，通过引入残差单元和短路连接机制解决了训练过程中的梯度消失和梯度爆炸的问题。如图 2-5 所示，ResNet 采用步长为 2 的卷积来进行下采样，避免了信息的丢失。ResNet 网络的设计重点旨在特征图尺寸减少时，特征图的数量会随之上升，这样可以保证提取特征的完整度。从图 2-5 可以看到 ResNet 网络每两层之间添加了短路机制，形成了残差学习，其中虚线表示特征图的数量发生变化。

2.2.3 Transformer 模型

Transformer 模型首先在文献[62]中被提出，最先被应用于自然语言处理，该模型是一个全新简单的网络架构，完全基于注意力机制，摒弃递归和卷积。该模型分为编码器和解码器两个部分。在这个模型里编码器的映射表示为（$x_1 \ldots x_n$）的输入序列到连续的序列($z_1 \ldots z_n$)。对于给定的 z，解码器单次生成输出序列符号($y_1 \ldots y_m$)。在每一步中，模型都是自动回归的，并在生成下一步时，使用前一个符号作为额外输入。网络模型图如图 2-6 所示。

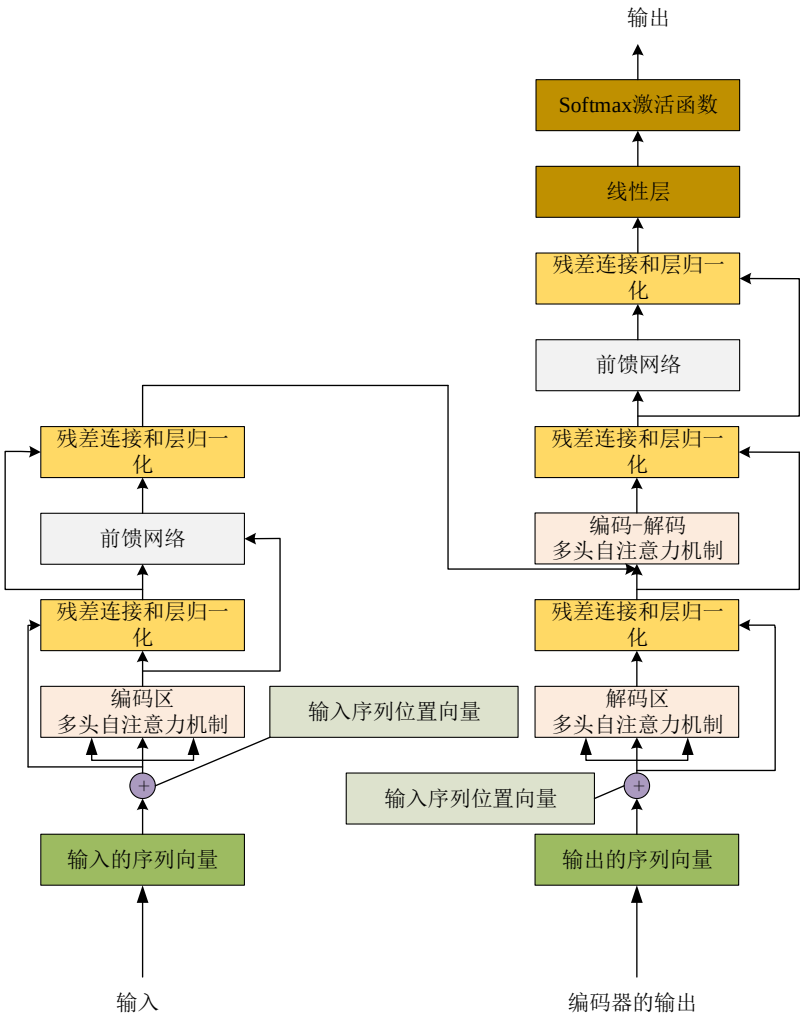


图 2-6 Transformer 的模型结构

Fig 2-6 Model structure of Transformer

图 2-6 是 Transformer 的内部结构，该模型结构分为两个块，分别为左边部分编码器块和右边部分的解码器块。上图中粉色的多头注意力机制是由多个自注意力机制组成的，多头注意力机制上方还包含一个残差连接层和层归一化，残差连接用于防止网络退化，层归一化用于对每一层的激活值进行归一化。下面对组成部分进行介绍。

（1）自注意力结构

图 2-7 为自注意力机制的结构图，在实际的模型使用中，自注意力接受的是输入或者是上一个解码器的输出，在计算的时候需要用到查询矩阵（Q），键值矩阵（K），值矩阵（V）。

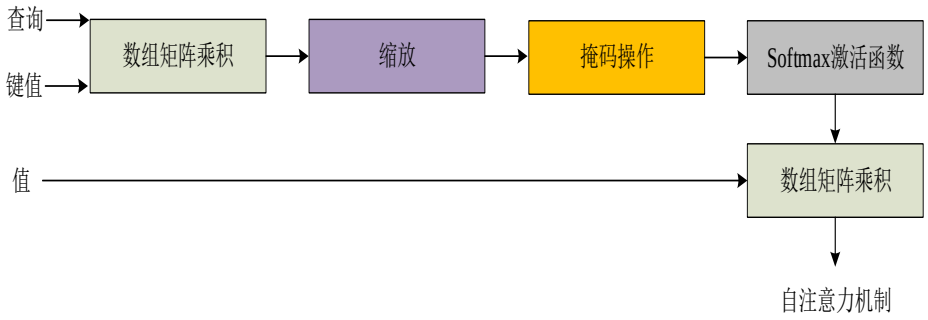


图 2-7 自注意力结构图

Fig 2-7 Self-attention structure diagram

自注意力的输入用矩阵 X 表示，则可以使用线性变阵矩阵 WQ,WK,WV 计算得到 Q,K,V 矩阵。计算如图 2-8 所示。

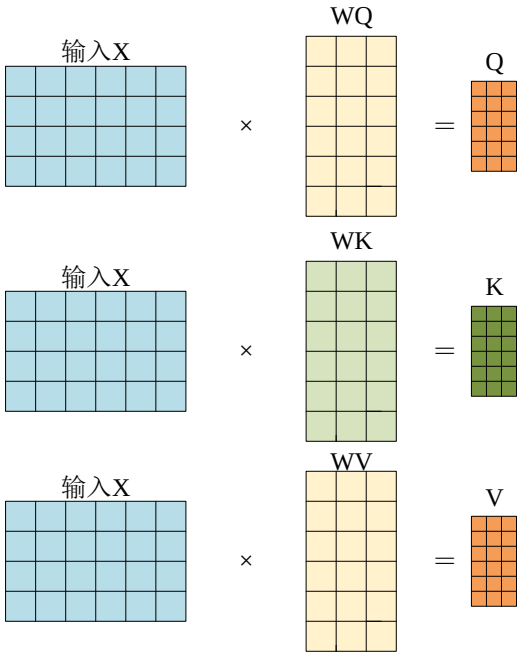


图 2-8 Q,K,V 的计算方法

Fig 2-8 Q,K,V calculation method

得到矩阵 Q,K,V 之后就可以计算自注意力的输出了，计算公式如式 2-16 所示：

$$\text{Attention}(Q,K,V)=\text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{2-16}$$

其中 Attention()为自注意力的输出，d_k 是 Q, K 矩阵的列数，即向量维度。为了得到每一行的内积，公式 2-16 中矩阵 Q 乘矩阵 K 的转置，接着除以 d_k 的平方根防止内积过大。假设输入的是一个单词数为 n 的句子，首先经过 QK^T 后，会得到一个列长度为 n 的矩阵，该矩阵可以表示为句子中单词的注意力强度，最

后使用 softmax 激活函数计算每一单词和其他单词的注意力系数。最后再乘以矩阵 V，就可以得到最终的输出，由此得到由自注意力计算输出的 Z，而多头自注意力就是由多个自注意力组合形成的，图 2-9 为多头注意力的结构图。

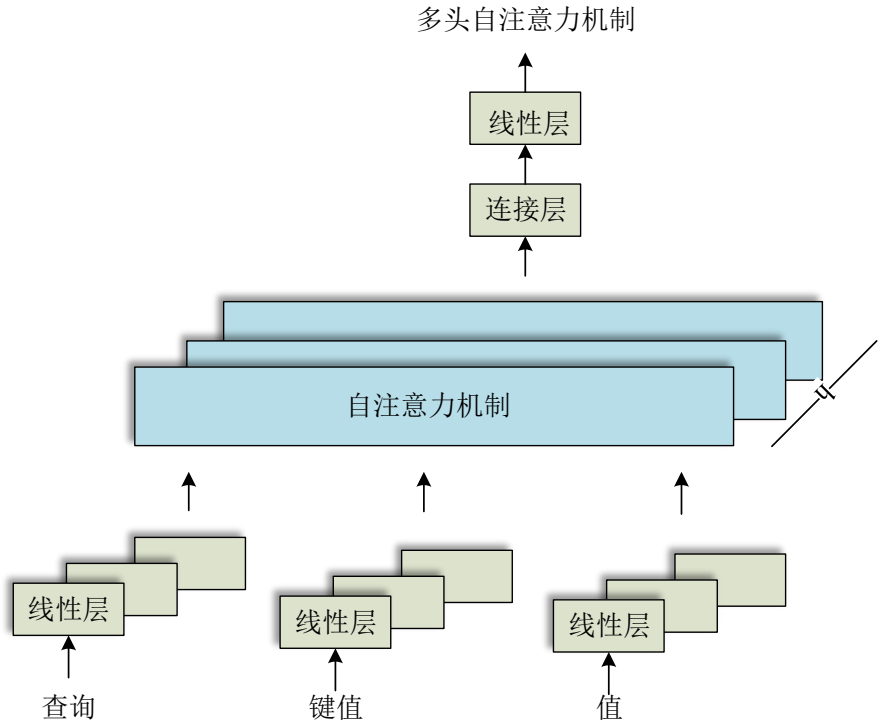


图 2-9 多头注意力的结构图

Fig 2-9 Structure diagram of Multi-Head Attention

从图 2-9 可知多头注意力包含多个自注意力，首先将输入 X 分别传送到 h 个不同的自注意力，计算得到 h 个输出矩阵 Z，将它们拼接在一起，然后传入线性层，得到多头注意力的最终输出。

（2）残差连接和层归一化

从图 2-9 可知，Transformer 的模型结构里包含残差连接和层归一化，残差连接是为了解决多层神经网络训练困难的问题，通过将一部分的前一层的信息无差的传递到下一层，可以有效的提升模型性能。层归一化对特征进行缩放，残差连接和层归一化的计算公式 2-17 所示。

$$\mathrm{Resl}() = \begin{Bmatrix} \mathrm{LayerNorm}(X + \mathrm{MultiHeadAttention}(X)) \\ \mathrm{LayerNorm}(X + \mathrm{Feed\ Forward}(X)) \end{Bmatrix} \tag{2-17}$$

其中 Resl()表示残差连接和层归一化的输出， X 表示多头注意力或者前馈层的输入， $\mathrm{MultiHeadAttention}(X)$ 表示多头注意力输出，$\mathrm{Feed\ Forward}(X)$ 表示前馈

网络输出，由于输入和输出的维度是相同的，所以可以直接相加。Feed Forward 是一个两层的全连接层，第一层的激活函数为 ReLu，第二层不使用激活函数。 $X+\text{MultiHeadAttention}(X)$， $X+\text{Feed Forward}(X)$ 是一种残差连接，通常用于解决多层网络训练的问题，可以让网络只关注当前差异的部分。

2.3 人群计数数据集和评价指标

2.3.1 人群计数数据集

本文使用的数据集有：Shanghai Tech 数据集[20]，UCF-QNRF 数据集[63]，NWPU-Crowd 数据集[64]。这些不同的数据集在拍摄角度，拍摄场景，人群密度大小，密度分布以及光照天气环境等各不相同。但是都提供了精确的人头标注。表 2-1 为这三个数据集的比较。

表 2-1 不同人群计数数据集的比较

Table 2-1 Comparison of different population counting datasets

数据集	训练集/测试集数量	分辨率	人数最大值/最小值
Shanghai TechPartA	300/182	589×868	3139/33
Shanghai TechPartB	400/316	768×1024	578/9
UCF-QNRF	1201/334	2013×2902	0/25791
NWPU-Crowd	3109/1500	2311×3383	0/20033

（1）Shanghai Tech 数据集

Shanghai Tech 数据集中有 1198 张图片，其中对 330165 人的头部中心进行了标注[65]，是一个数量较大的人群数据集。Shanghai Tech 数据集分为了 A,B 两部分。A 部分（PartA）从互联网上随机抓取了 482 张图片，示例图如图 2-10(a)所示，B 部分（PartB）从上海大都市繁忙的街道上抓取了 716 张图像，示例图如图 2-10(b)所示。两个部分之间的人群密度差异很大，A 部分数据集的人群密度明显高于 B 部分。此外 A 部分的场景变化较大，B 部分的场景较为固定。这使得人群计数的任务更加具有挑战性。A 部分和 B 部分都分为训练和测试两部分，其中 A 部分 300 张用于训练，剩下 182 张用于测试; B 部分 400 张用于训练，316 张用于测试。

（2）UCF-QNRF 数据集

UCF-QNRF 数据集包含 1535 张图像，示例图如图 2-10(c)所示，其中包含 1201 张训练集和 334 张测试集。UCF-QNRF 数据集拥有高计数人群图像和标注，该数据集图片包含多样化的视角、不同的密度和不同的光照情况以及多变的场景[66]。另一方面， UCF-QNRF 数据集存在于拍摄的真实场景中，例如建筑，天空，植物等，也使得这个数据集更加逼真和具有挑战性。

（3）NWPU-Crowd 数据集

NWPU-Crowd 数据集是目前为止人群计数任务中数量最大的数据集，示例图如图 2-10(d)所示，该数据集拥有 5109 张图片和 2133238 个标注实体。数据集被分为训练集 3109 张图片、验证集 500 张图片和测试集 1500 张图片。该数据集中包含一些人群密度极高的图片，这样可以增强模型的鲁棒性。而且该数据集的分辨率较其他数据集相比分辨率也更高。该数据集对模型的计数性能提出了更高的要求。

图 2-10 不同数据集的图片示例

Fig 2-10 Examples of images from different datasets

2.3.2 评价指标

本文评估的性能指标除了模型的参数量以及计算量外，对模型计数精度的评估利用平均绝对误差（Mean Absolute Error，*MAE*）和均方误差（Mean-square error，*MSE*）。

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \| C_e - C_g \|_1 \tag{2-18}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \| C_e - C_g \|_2^2} \tag{2-19}$$

式 2-18 和式 2-19 中，C_e 和 C_g 分别表示第 i 个人群图像的预测值和真值，N 为测试数据集中的图像总数。

2.4 本章小结

本章首先对 CNN 模型的基础知识进行介绍，其次列举了两个经典的 CNN 模型，然后介绍了 Transformer 网络。最后对本文涉及的三个人群计数数据集以及性能评价指标进行了介绍。

第3章 人群密度特征提取网络的轻量化方法研究

3.1 引言

在实际应用中，当前人群计数网络的密度特征提取模块通常都拥有复杂的网络结构和较大的计算量，无法实现实时计数，而对密集场景下的人数进行监管，预防拥挤或踩踏事故的发生，计数的实时性和计数精度几乎同等重要。为此，研究轻量化的人群计数模型是一项重要的课题。

近年来，基于CNN的人群计数受到更多学者的关注，并提出了很多高质量的计数模型。DUB-Net[67]采用复杂度较高的经典卷积神经网络ResNet-50[27]作为密度特征提取模块来提取特征，多个并行卷积来生成密度图。它通过不确定性的量化对模型的输出进行概率解释，提高了预测的质量，明确地处理了不确定的输入和特殊情况。CSRNet[22]采用复杂的骨干网络进行密度特征的提取，在后端利用卷积网络生成预测密度图。综上，为了提取更多的特征，它们一般使用了复杂的多尺度卷积作为密度特征提取网络。虽然成功捕获了不同规模的特征，但是不可避免的会产生大量的网络参数和巨大的计算负担。

针对上述问题，如何构建一个高效的轻量化人群计数模型成为了一个亟待解决的问题。目前研究中，也提出了一系列轻量级的人群计数模型。MCNN[20]每一列都使用较少的具有不同大小卷积核的卷积来预测人群密度图。TDF-CNN[68]提出了一个自顶向下的反馈卷积神经网络，利用两列不同卷积核的少量标准卷积生成密度图。然而，少量标准卷积层提取的人群特征非常有限，会导致最终计数的准确性不理想。综上，目前的轻量化人群计数网络在参数量和计算量上都有所降低，然而为了获得更快的运行速度，现有的轻量级网络常会使用较少的标准卷积或忽略特征的融合，虽然模型的复杂度降低，但是计数精度也受到了很大的影响。

为了有效的解决上述问题，本章旨在对计数模型的密度特征提取网络进行优化，设计了一种密集场景下的轻量化密度特征提取网络。人群计数模型图如图 3-1 所示。模型主要由本章设计的轻量化线性映射块堆叠组成，轻量化线性映射块利用简单的线性操作减少计算量生成特征图，再将第一个轻量化线性映

射块生成的特征图和第五个轻量化线性映射块生成的特征图进行特征融合提升网络的计数性能，最后回归出预测密度图得到最终预测人数。本章提出的模型与其余轻量化模型相比在保持了计数精度的同时参数量和计算量上均有所减少。

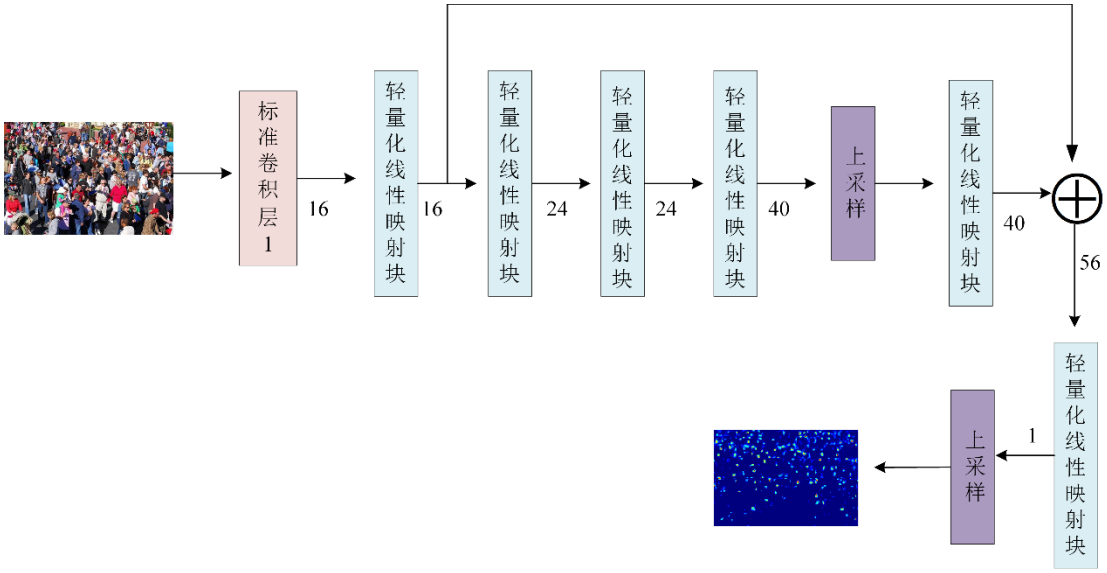


图 3-1 轻量化密度特征提取网络模型图

Fig 3-1 Lightweight density feature extraction network model diagram

3.2 轻量化密度特征提取网络构建

本章对计数模型的密度特征提取网络进行优化，设计了一个轻量化特征提取网络。特征提取网络由本章设计的轻量化线性映射块搭建而成，将第一个线性映射块和第五个线性映射块生成的特征图进行融合，最后回归出预测密度图得到最终估计人数。特征提取网络中线性映射块是由多个线性映射单元组成的，该映射单元首先使用少部分的标准卷积来生成通道数较少的本征特征图，再通过一些简单的线性映射得到新的特征图，最后再将两个特征图拼接。这种线性映射单元在没有改变输入输出特征的情况下，大大减少了模型的参数量和计算量，下面进行详细介绍。

3.2.1 轻量化线性映射块构建

本节的特征提取网络是基于轻量化线性映射块来搭建的。轻量化线性映射块是由多个轻量化线性映射单元构成的，下面将进行详细介绍。

（1）轻量化线性映射单元

轻量化线性映射块是由多个轻量化线性映射单元搭建而成，首先介绍轻量化线性映射单元，结构图如图 3-2 所示。

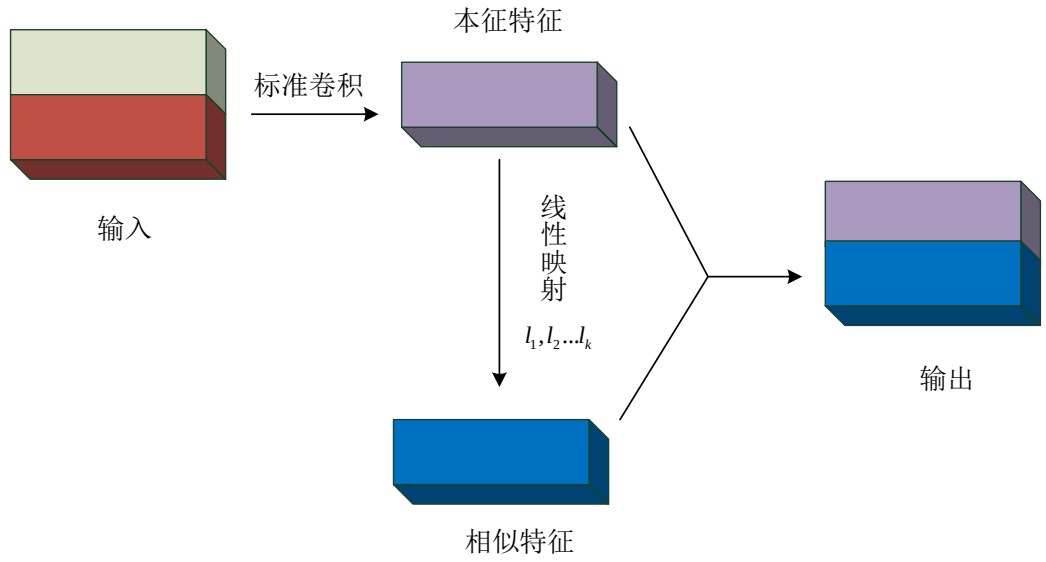


图 3-2 轻量化线性映射单元模型图

Fig 3-2 Lightweight linear mapping unit model diagram

如图 3-2 所示，映射单元一共分为 3 个步骤，首先是标准卷积生成通道数较少的本征特征图，然后利用线性映射生成相似特征图，最后将两种特征图进行拼接。下面对本征特征图的生成和相似特征图的生成进行介绍。

本征特征图的生成：在深度学习网络中特征提取时都会包括本征特征和相似特征。由于内在的本征特征较小，只需要用少量标准卷积来生成本征特征即可。具体来说假设总的特征映射为 p 个，内在本征特征为 q 个（$q < p$）。对于给定的输入图像 $I \in \mathrm{R}^{h \times w \times c}$（$h, w, c$ 分别为输入图像的长宽以及通道数），q 个内在本征特征映射 $Q \in \mathrm{R}^{h' \times w' \times q}$ 是通过一次标准卷积生成的本征特征图，如式 3-1 所示。

$$Q = I * f \tag{3-1}$$

$f \in R^{c \times k \times k \times q}$ 为使用的滤波器，$k \times k$ 为滤波核大小，$*$ 代表卷积，不包含偏置项。

相似特征图生成：对于相似特征而言，没有必要使用大量的计算来逐个生成。对上述生成的 Q，为了进一步得到想要的 p 个特征映射，对 Q 中的每个本征特征进行一系列线性操作生成 z 个相似特征：

$$z_{i,j} = l_{i,j}(q_i), \quad \forall i = 1, ..., q, j = 1, ..., z \tag{3-2}$$

其中式 3-2 的q_i为Q中的第i个内在本征特征映射，$l_{i,j}$为上述函数中生成第i个相似特征映射的第j次线性操作，$z_{i,j}$表示第i个内在本征特征映射生成的第j次相似特征。也就是说q_i可以有一个或多个相似特征。

最后就是对标准卷积生成的本征特征和线性映射的相似特征进行拼接生成最终的输出。本节的特征提取模块是基于轻量化线性映射块来搭建的。轻量化线性映射块是由多个轻量化线性映射单元构成的，下面将对其进行介绍。

（2）轻量化线性映射块

轻量化线性映射块主要由两个串联的轻量化线性映射单元组成，如图 3-3 所示第一个轻量化线性映射单元用于扩展层，扩大通道的数量，第二个轻量化线性映射单元减少了通道的数量以增加速率匹配路径。批归一化(Batch Normalization, BN)[69]和 ReLU 非线性激活函数在轻量化线性映射单元之后[70]应用。

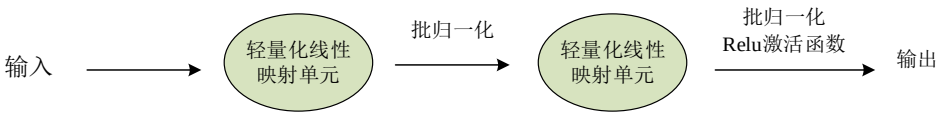


图 3-3 轻量化线性块模型图

Fig 3-3 Lightweight linear mapping block model diagram

3.2.2 密度图生成

目前在人群计数领域，基于CNN的人群计数模型在回归人数阶段，一般包括直接估计出输入图像的人数和回归出密度图再求和得到人数这两种方法。具体来说，输入的是图像，前者输出的是预测人数，后者输出人群密度图，然后通过求和得到人数。在本章中，对人群密度进行预测时，选择第二种方法，这一选择主要有两个原因:

首先，和直接生成预测的人数相比，生成的密度图会保留更多有效直观的信息。例如，在密度图中可以直观的观察到输入图像中的人群分布情况。这些分布情况在实际场景中有重要的意义。如果在输出的密度图中观察到某一小区域的人群密度远远高于其他区域，则可能表明该区域发生了人数过多的情况。其次，在使用CNN学习密度图的过程中，会加强卷积网络对不同大小人头的适应性。因此，当输入的图片透视效果变化较大时，也可以保证计数的准确性。

假设所有的人群都提供逐点标注，即图像中的个体都进行了点标注（每个头部一个点），标记了 M 个头像的输入图像可以表示为一个函数：

$$P(x)=\sum_{i=1}^{M}\delta(x-x_i) \tag{3-3}$$

x 表示人群图像中的图像坐标，x_i 表示 x 中第 i 个注释点的坐标，δ 表示冲激函数。

为了将式 3-3 中的函数转换为连续密度函数，可以将该函数与高斯核 G_σ [71] 进行卷积，使密度为 $F(x)=P(x)*G_\sigma(x)$。然而，在真实的三维场景中，这些图像中的人头并不都是独立存在的，在密集场景下，一个小区域内都可能存在很多的标注人头。并且由于透视失真，在近处的人头所占区域会大一些，远处人头所占的区域会小一些。如果使用高斯核 G_σ 来生成密度图，会造成计数精度的损失，因此本章中使用几何自适应核来生成密度图。

对于给定图像中的每个头像 x_i，将其与 c 个最近的人头的距离表示为 $\left\{l_1^i,l_2^i,...,l_c^i\right\}$，平均距离为 $\bar{l}^i$，如式 3-4 所示。因此与 x_i 所对应的区域大小可以大致表示为图片地面上的一个半径 $\bar{l}^i$ 的圆。

$$\bar{l}^i=\frac{1}{c}\sum_{j=1}^{c}l_j^i \tag{3-4}$$

为了估计邻近 x_i 的人群数量，需要将 $P(x)$ 与方差为 σ_i 的高斯核进行卷积。密度公式如式 3-5 所示，β=0.3，$*$表示卷积。图 3-4 是本章使用的数据集中的部分图片利用几何自适应高斯核得到的密度图。

$$F(x)=P(x)*G_{\sigma_i}(x)\quad \sigma_i=\beta\bar{l}^i \tag{3-5}$$

(a) Shanghaitech数据集部分图片　(b) 由(a)图生成的真实密度图　(c) UCF-QNRF数据集部分图片　(d) 由(c)图生成的真实密度图

图 3-4 部分数据集图片几何自适应高斯核得到的真实密度图

Fig 3-4 Some of the datasets images use geometric adaptive Gaussian kernel to obtain ground real density maps

3.2.3 特征融合

在 CNN 计数网络中，特征融合可以充分利用特征图中的数据，提高模型的计数性能和鲁棒性。在特征提取阶段，深层次的特征图和浅层次的特征图都拥有不同尺度的特征信息，使用特征融合可以帮助网络更全面的利用输入的特征信息，减少特征的丢失。

在本章的特征提取阶段使用特征图连接的方式对特征进行融合，特征图是一个有着长，宽，通道数的三维矩阵，使用连接的方式融合特征可以将不同通道数的特征图进行融合，具体操作如下，假设需要进行连接的两个特征图分别为 $I \in [H,W,\acute{C}]$, $I_1 \in [H,W,C]$, 则经过连接后的特征图为 $I_2 \in [H,W,\acute{C}+C]$。示意图如图 3-5 所示。

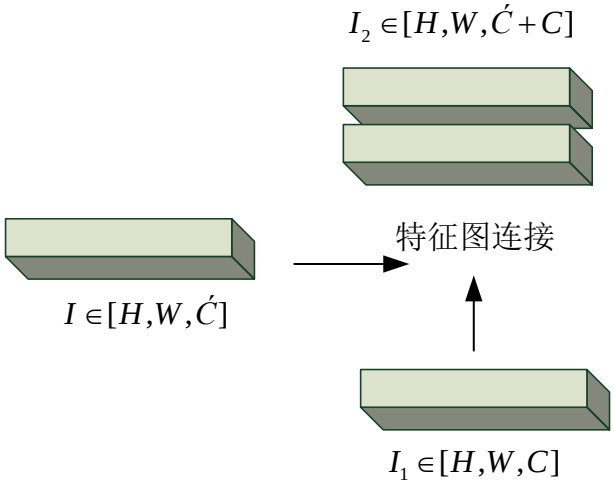


图 3-5 特征融合示意图

Fig 3-3 Schematic diagram of feature fusion

表 3-1 为轻量化密度特征提取网络模型的结构配置。第一个上采样层为了匹配特征图融合的尺寸，第二个上采样层为了提高计数精度，使输出的密度图恢复至原始输入分辨率大小。步长指卷积核在图片中每一步移动的距离。标准卷积层 1 表示卷积核为 3×3 的标准卷积。

表 3-1 轻量化密度特征提取模型的结构配置

Table 3-1 Structural configuration of a lightweight density feature extraction model

输入尺寸	操作方式	卷积核尺寸	输出通道	步长
224×224×3	标准卷积层 1	3	16	2
112×112×16	轻量化线性映射块	3	16	1
112×112×16	轻量化线性映射块	3	24	2
56×56×24	轻量化线性映射块	3	24	1
56×56×24	轻量化线性映射块	5	40	2
56×56×40	上采样层	无	无	无
112×112×40	轻量化线性映射块	5	56	1
112×112×56	轻量化线性映射块	3	1	1
112×112×1	上采样层	无	无	无

3.2.4 损失函数设计

上述介绍了轻量级人群计数的密度特征提取网络。为了提高计数精度，本章的模型在训练时采用欧几里得距离[72][73]来度量估计的密度图与真实密度图之间的差值，损失函数定义如式 3-6：

$$L_1(\theta)=\frac{1}{2N}\sum_{i=1}^{N}\| D(X_i;\theta)-D_g \|_2^2 \tag{3-6}$$

其中θ是模型中的一组可学习参数。N 为训练图像的个数。L_1 为估计密度图与真值密度图之间的损失。X_i 是输入图像，D_g 是图像 X_i 的真实密度图。$D(X_i;\theta)$ 表示由模型生成的估计密度图。

由于该模型训练的最终结果是希望预测的人数能够接近于真实的人数，因此还提出了一个真实人数和估计人数之间的损失，首先将估计密度图 $D(X_i;\theta)$ 求和得到估计人数 C_e，定义如式 3-7 所示：

$$C_e=\sum_{i=0}^{W}\sum_{j=0}^{H}D(X_i;\theta) \tag{3-7}$$

其中 W,H 分别表示估计密度图的宽和长。真实人数和估计人数之间的损失 L_2 如式 3-8 所示。

$$L_2(\theta)=\frac{1}{2N}\sum_{i=1}^{N}\| C_e-C_g \|_2^2 \tag{3-8}$$

其中 L_2 为估计人数与真实人数之间的损失。C_g 是真实人数，C_e 表示估计人数。

最终损失函数由上述两个损失函数加权求和得到，权重因子超参数为γ，如式 3-9 所示：

$$L_{loss} = L_1(\theta) + \gamma L_2(\theta) \tag{3-9}$$

3.3 实验与结果分析

3.3.1 实验设置

本文中所研究的人群计数任务所用的工作站是两个 NVIDIA RTX3090 显卡，处理器为 Intel Core-I7。此外在软件上，所使用的编程语言为 python，版本为 3.7。模型的训练和测试阶段，全部都是在 pytorch 深度学习的框架上进行的，版本为 1.6。

在训练阶段，为了减少内存和增强模型的鲁棒性和多样性，首先使用随机裁剪从源人群图像中生成训练补丁。随机裁剪是使用一个固定大小为 224×224 的矩形方框在输入图片中随机位置处生成，由于裁剪时位置是随机的，所以每次裁剪的结果都不一样，在训练过程采用随机裁剪是为了增加训练过程中模型的鲁棒性和多样性。如图 3-6(a)所示，白色的方框是每次迭代后在图片中随机位置处生成的，每迭代一次，就会产生一个随机的矩形方框，从而就会产生一个随机的子图。顺序裁剪是指将输入图像按照固定尺寸，有顺序的进行裁剪，如图 3-6(b)所示，每次迭代生成一个规律的子图。使用随机裁剪的方式因为增加了训练的多样性，因此要比顺序裁剪进行训练的方法效果更好。

接下来，使用Adam[74]算法对本章提出的模型框架进行优化，默认学习率为 1e-5，权值衰减为 1e-4，批大小为 8，训练轮数为 500。损失函数为 3.2.4 节定义的L_{loss}。本节在第二章介绍的Shanghai Tech、UCF-QNRF和NWPU-Crowd三个主流人群统计数据集上，对所提出的模型的框架的性能进行了评估。

（a）随机裁剪示意图　　（b）顺序裁剪示意图

图 3-6 不同裁剪方法示意图

Fig 3-6 Diagram of different clipping methods

3.3.2 实验训练过程

本文提出的模型没有任何预训练模型可以使用，所以需要重新训练，训练时输入为选定的人群数据集，输出为模型的权重 λ，训练的大致过程如下。

（1）初始化训练轮数 $i=0$，设置训练轮数为 500，当 i 达到 500 或达到设定的 MAE 或 MSE 时，训练结束。

（2） 每执行一次第一步，将会得到预测的密度图 $D(X_i;\theta)$ 和真实的密度图 D_g，预测人数 C_e，真实人数 C_g。

（3） 根据第 3.2.4 定义的损失函数计算出损失：$L_{loss}(D(X_i;\theta),D_g,C_e,C_g)$，$\triangle\lambda = L_{loss}$。

（4）计算当前训练的 MAE 和 MSE。

（5） 当训练的次数达到了一个指定更新学习率的轮次时，更新学习率 $l_r = 0.5l_r$。

（6） 更新模型的权重 $\lambda = \lambda - l_r * \triangle\lambda$

（7） 重复上述除（1）以外的所有步骤，当训练达到步骤（1）的条件时，停止训练。

3.3.3 实验结果和对比分析

第一部分对比本章提出的轻量化计数模型与其他人群计数模型的参数量和计算量。第二部分验证本章所提出的轻量化的人群计数模型的计数能力，在三

个计数数据集上进行了实验，并将实验结果与多个经典的人群计数算法模型进行了对比。比较的经典人群计数模型分为两大类，一类是常规的人群计数模型，另一类是轻量级人群计数模型。实验结果如表 3-2 所示。

表 3-2 不同人群计数模型方法在 3 个数据集的实验结果

Table 3-2 Experimental results of different population estimation model methods in 3 datasets

人群计数模型方法	PartA		PartB		UCF-QNRF		NWPU-Crowd		Params (M)	计算量（GFLOPS）
	MAE	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*		
CP-CNN[23]	73.6	**106.4**	20.1	30.1	-	-	-	-	68.4	-
CSRNet[22]	**68.2**	115.0	10.6	16	120.3	208.5	121.1	**387.8**	16.26	325.3
CAN[74]	**62.3**	**100**	**7.8**	**12.2**	**107.0**	**183.0**	**106.3**	**386.5**	18.1	193.58
DUB-Net[67]	**64.4**	**106.8**	**7.7**	**12.5**	**105.6**	**180.5**	-	-	18.05	-
SUA-Fully[75]	**66.9**	125.6	12.3	17.9	119.2	213.3	-	-	15.85	-
STNet[76]	**52.9**	**83.6**	**6.3**	**10.3**	**87.9**	**166.4**	-	-	15.56	
MCNN[20]	110.2	173.2	26.4	41.3	277.0	426.0	232.5	714.6	0.13	11.87
TDF-CNN[68]	97.5	145.1	20.7	32.8	-	-	-	-	0.13	-
SANet[36]	75.3	122.5	10.5	17.9	152.6	247.0	190.6	491.4	0.91	71.45
LCNet[33]	93.3	149.0	15.3	25.2	-	-	-	-	0.86	-
PCC-Net[37]	73.5	124.0	11.0	19.0	148.7	247.3	167.4	566.2	0.55	72.80
PDDNet[34]	72.6	112.2	10.3	17.0	130.2	246.6	-	-	1.1	-
LMSFFNet[35]	85.85	139.9	**9.2**	**15.1**	112.8	201.6	-	-	4.58	14.9
Our model	70.1	107.4	10.1	16.0	110.3	187.7	119.8	461.5	**0.11**	**7.41**

由于UCF-QNRF数据集和NWPU-Crowd数据集出现的较晚，且图片人群密度跨度较大，对于人群计数任务有着很大的挑战性，因此一些人群计数模型还未在这两个数据集上进行实验。而常规的人群计数模型一般不会在实验中进行参数量和计算量的计算，且一些轻量化的计数模型也仅给出了参数量的计算。因此在表 3-2 中仅列出了对比模型中进行了实验的数据。

由表 3-2 可知本章提出的轻量级人群计数模型的参数量仅有 0.11MB，计算量仅有 7.41GFLOPS。对比其余轻量级的计数模型是最低的。在计数性能上，与轻量化计数模型相比，在PartB数据集上，本文模型的 *MAE* 仅比最新的LMSFFNet模型高 0.9， *MSE* 仅高 0.9。而本文模型比LMSFFNet模型的参数量低 4.47，计算量低 7.49。在其余数据集上，本文模型与其他轻量化人群计数模型

相比都取得了最佳的 *MAE* 和 *MSE* 。

与常规的人群计数模型相比，在PartA数据集上，本文模型的 *MAE* 比模型CSRNet ，CAN，DUB-Net，SUA-Fully 和STNet 高 1.9，7.8，5.7，3.2 和 17.2，*MSE* 比模型CP-CNN，CAN，DUB-Net和STNet高 1.0，7.4，0.6 和 23.8。在PartB数据集上，本文模型的 *MAE* 分别比模型CAN，DUB-Net和STNet高 2.3，2.4 和 3.8，*MSE* 分别比模型CAN， DUB-Net和STNet高 3.8， 3.5 和 5.7。在UCF-QNRF数据集上本文模型的 *MAE* 分别比模型CAN，DUB-Net和STNet高 3.3，4.7 和 22.4，*MSE* 分别比模型CAN， DUB-Net和STNet高 4.7， 7.2 和 21.3。在NWPU-Crowd数据集上本文模型的 *MAE* 比模型CAN高 13.5，*MSE* 分别比模型CAN和CSRNet高 73.7 和 75。

虽然和常规的人群计数模型相比，在三个数据集上没有取得最佳的 *MAE* 和 *MSE* ，但是本章提出的模型的参数量和计算量与这些常规的人群计数模型相比大大减少。且本章提出的模型对比其他轻量化的人群计数模型，计数性能更为优异，模型的参数量和计算量也更低。因此本章的实验结果证明本章提出的模型在保持了一定的计数精度的同时提高了计数效率。

三个数据集上的可视化结果如图 3-7 所示。从左到右分别依次为输入的测试集图片，真实密度图和估计密度图，第一行和第二行为PartA数据集上部分实验结果，第三行和第四行为PartB数据集上部分实验结果，第五行和第六行为UCF-QNRF数据集上部分实验结果，第七行和第八行为NWPU-Crowd数据集上部分实验结果。

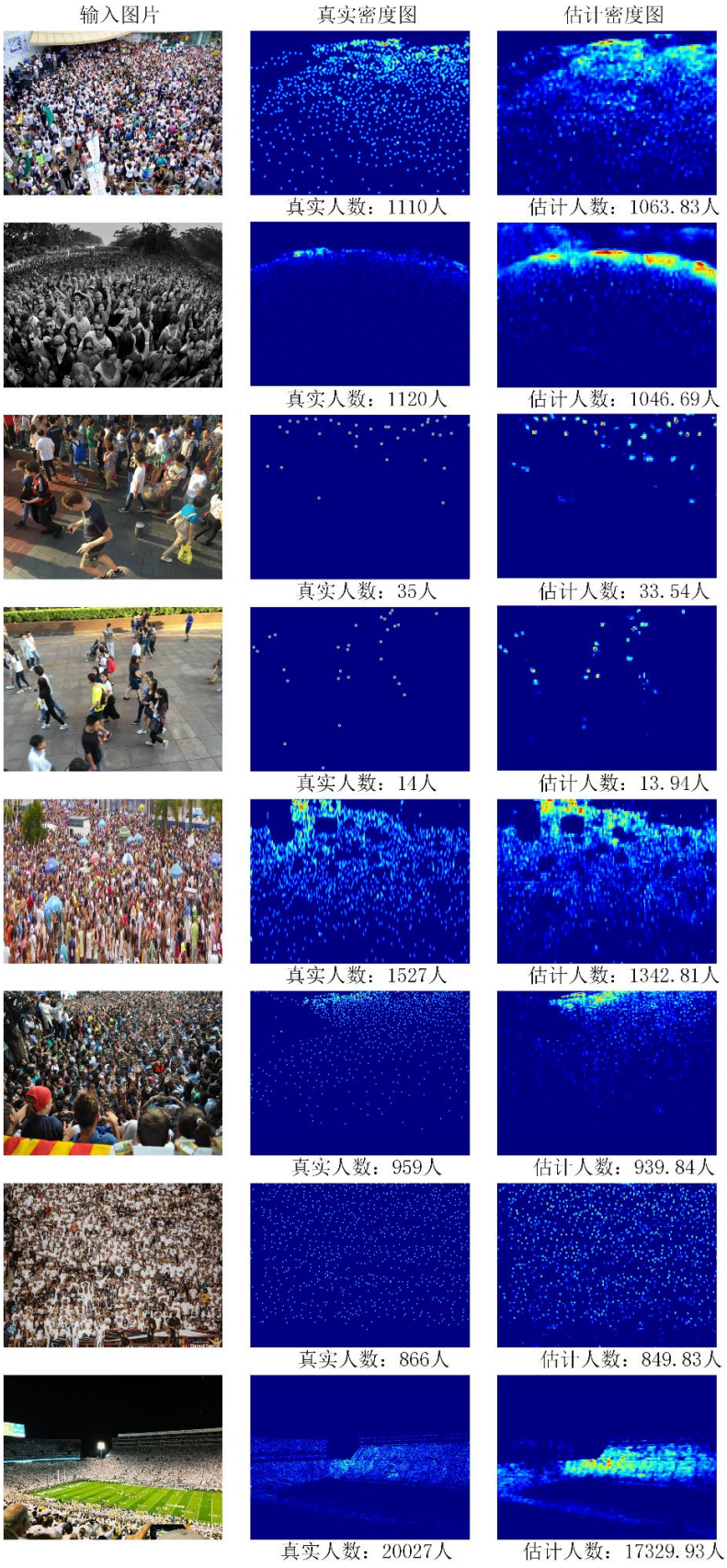


图 3-7 不同数据集上部分实验结果

Fig 3-7 Some experimental results on different datasets

3.4 本章小结

本章提出了一种轻量化密度特征提取网络，在保证模型复杂度降低的同时保留了模型的计数性能。针对目前人群计数模型密度特征提取模块计算量和参数量过大的问题，设计了一个由轻量化线性映射块构建的密度特征提取网络。特征提取网络利用简单的线性映射提取密度特征，通过特征融合减少了特征的丢失，在保持了计算精度的同时降低了参数量和计算量。最后使用密度图回归的方式回归预测密度图。最后在实验阶段，将本章提出的轻量化计数模型在三个主流的数据集上进行了实验，并与多个经典人群计数模型进行对比。由实验的指标可知，本章提出的轻量化计数模型与其他轻量级计数模型相比，参数量和计算量是最低的，取得的计数性能也较为优异。与常规的计数模型相比，虽然计数能力不是最优异的，但是提高了计数效率，更具备实用性。

然而由最新的研究可知，卷积网络对于全局特征的提取能力较弱，下一章将优化人群计数模块，来克服卷积网络在计数任务上的这种缺陷，尽可能的提高计数的精确性。

第4章 密集场景下多任务并行的人群计数方法研究

4.1 引言

针对当前的人群计数模块网络较为单一，难以应对复杂场景下人数精确的估计，本章在第三章的基础上耦合 CNN 和 Transformer 的优势构建了多任务并行的人群计数网络。

CNN 和 Transformer 是适用于人群特征提取的两种主流深度学习模型，在最新的研究中， CNN[14][77]的弱监督计数性能相对较差，这是由于卷积核的感受野有限，并且无法对全局上下文信息进行建模，而全局上下文信息在人群计数中起着至关重要的任务[51][74]。此外，CNN 无法建立图像块之间的相互作用，因此无法学习人群和背景之间的对比特征。与 CNN 相比，Transformer 由于其注意力机制可以充分学习不同图像块之间的对比度特征，这种特征可以用于区分人群和背景区域。一些研究人员提出了用于人群计数的 Transformer 架构[56]，有效的捕捉人群全局上下文信息。然而，Transformer 会将输入图像直接划分为一系列的补丁标记，这样会导致人数的高估或者低估。此外，由于纯 Transformer 模型缺乏 CNN 的归纳偏置能力，因此需要更多的参数才能获得卓越的性能。综上所述，CNN 使用感受野有限的卷积核来学习面向局部的特征，但它在建模长上下文依赖性方面的能力相对较弱。Transformer 利用多头注意力机制，能够捕获全局信息，但它可能会忽略局部区域内的细节特征。简而言之，CNN 主要关注人群场景中的密度分布，而 Transformer 则直接捕捉不均匀的人群分布。因此，可以考虑一种将 CNN 和 Transformer 混合的计数模型。

目前在人群计数模块有两种方法进行近似计数，点注释生成密度图计数或对于整个图像直接估计个体的总数[56]，示意图如图 4-1 所示，图 4-1(a)利用点注释生成真实密度图，然后设计回归器生成预测密度图，最后利用损失函数测量预测密度图与真实密度图之间的误差。图 4-1(b)使用从图像到计数的角度直接回归人群图像的总数。但是点注释的方法难以在高拥挤区域中检测头部点，而直接估计人数的方法会依赖全局视角，在人数稀疏的区域反而会产生噪声，且不适合感受野有限的 CNN 模型。因此，密度图估计和直接估计应当是相辅相成

的。

基于此，在本章提出了一个多任务的并行人群计数网络，模型图如图 4-2 所示。该方法的密度特征提取网络采取第三章优化后的特征提取网络，在保持精度的同时减少一些模型的参数量和计算量。本章主要优化人群计数模块，将密度特征提取网络提取到的特征图馈送到人群计数模块，具体来说，人群计数模块分为三个分支，在密度图预测网络分支使用 CNN 网络生成预测密度图进行人群计数。在直接计数网络分支使用 Transformer 网络直接回归图像中人群的总数。最后在人数融合网络分支，本章设计了动态因子生成器用来自适应地融合密度图预测网络分支和直接计数网络分支估计的结果。最后在三个人群数据集上进行了实验对比分析，证明了本章提出模型在计数精度上的优越性。

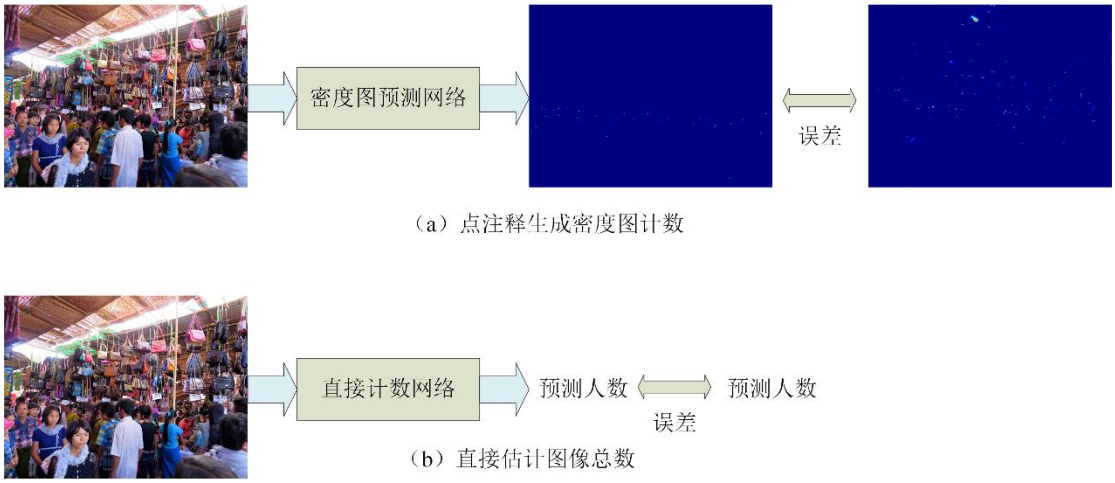


图 4-1 两种人群计数方法

Fig 4-1 Two methods of estimating the number of people

4.2 多任务并行人群计数网络构建

在本章中提出了一种密集场景下多任务并行的人群计数网络，结合 CNN 和 Transformer 的方法来预测密集场景下的个体数量。首先在密度特征提取模块采用第三章的特征提取网络来提取特征生成特征图，再将提取到的特征图馈送到人群计数模块，人群计数模块是并行的三个分支。三个分支分别密度图预测网络分支，直接计数网络分支和人数融合网络分支。具体操作为：特征图馈送到三个分支后，在密度图预测网络分支回归估计密度图求和得到预测的人数。在直接计数网络分支依赖 Transformer 的全局特征提取能力直接生成人数，在人数

融合网络分支动态地生成权值项。最后利用生成的权值项动态融合密度图预测网络分支和直接计数网络分支生成的人数，得到预测的最终人数。图 4-2 为模型基本结构图。下面将详细介绍各个模块的组成。

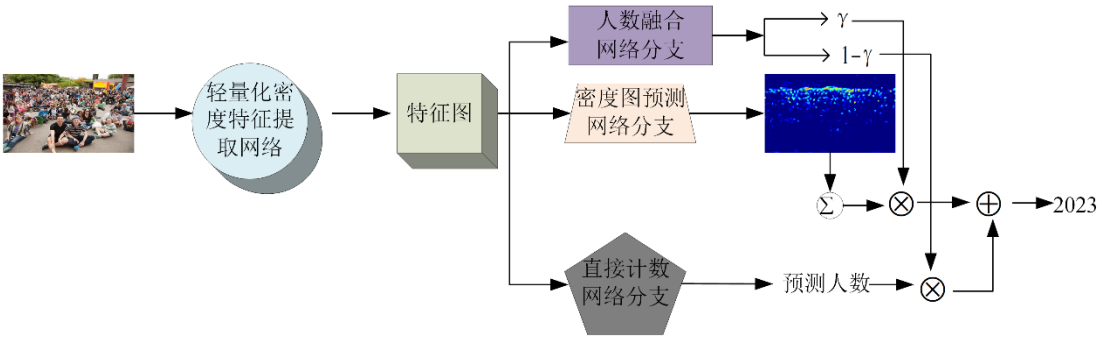


图 4-2 多任务并行人群计数网络模型结构图

Fig 4-2 Structure diagram of the network model for multi-task parallel crowd count estimation

4.2.1 密度图预测网络分支

密度图预测网络分支采用多个扩张卷积进行堆叠。从密度特征提取模块提取特征，再送入构建的密度图预测网络分支来生成估计密度图，最后对估计的密度图求和得到最终的估计人数。

（1）扩张卷积

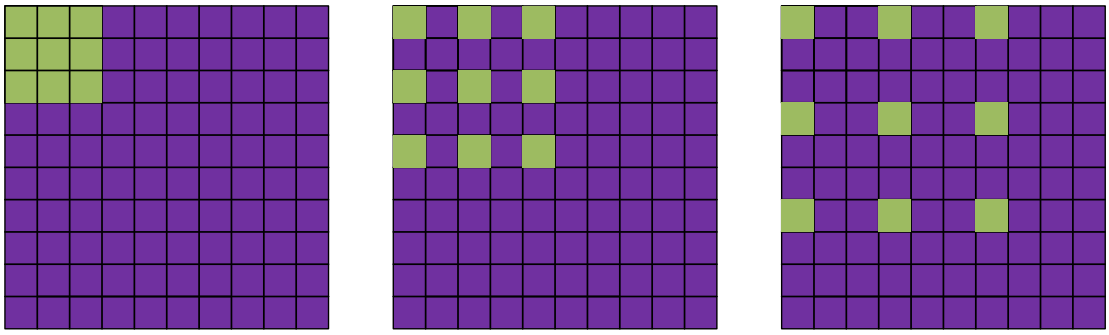
与标准的卷积相比，扩张卷积扩大了感受野。能够捕获更深层次的位置导向信息。二维扩张卷积的定义如式 4-1 所示。

$$F(a,b)=\sum_{i=1}^{M}\sum_{j=1}^{N} f(a+r\times i,b+r\times j)w(i,j) \tag{4-1}$$

其中 $F(a,b)$ 是输入 $f(a,b)$ 与滤波器 $w(i,j)$ 扩张卷积的输出，a,b 分别为长度和宽度，r 表示扩张卷积的扩张速率。如果 $r=1$，则扩张卷积变为标准卷积。

目前，最大池化层和平均池化层大多被用于进行特征压缩和防止过拟合，但它们也极大损耗了特征图的分辨率，这会导致提取到的特征映射图的一部分空间信息丢失。然而，反卷积层虽然可以减轻信息丢失，但它较高的复杂度会导致时延的情况发生。近年来，扩张卷积层在计算机视觉任务[14]上的准确率显著提高，它利用较小的卷积核就可以替代大卷积核的标准卷积或池化层，如图 4-3 所示。这种特性不仅减少了模型的参数量和计算量，还扩大了感受野。虽然可以使用更多的卷积层来产生更大的感受野，但是引入的操作也会更多。在扩

张卷积中，一个具有 $s\times s$ 滤波器的小尺寸核被扩大到 $s+(s-1)(r-1)$，扩张率为 r，从而可以在保持相同分辨率的情况下灵活地聚合多尺度上下文信息。如图 4-3 所示，使用不同扩张率下的 3×3 卷积核得到的感受野也不同，正常卷积得到 3×3 感受野，r=2 和 r=3 的扩张卷积分别得到 5×5 和 7×7 感受野。



（a）r=1　　（b）r=2　　（c）r=3

图 4-3 不同的扩张率下的 3×3 卷积

Fig 4-3 3×3 convolution at different expansion rates

（2）密度图预测网络分支搭建

在密度图预测网络分支，使用两层扩张率为 2 的 3×3 卷积核的扩张卷积和两层 1×1 卷积核的标准卷积。在每层卷积后都添加ReLu激活函数。图 4-4 是密度图预测网络分支的结构示意图。

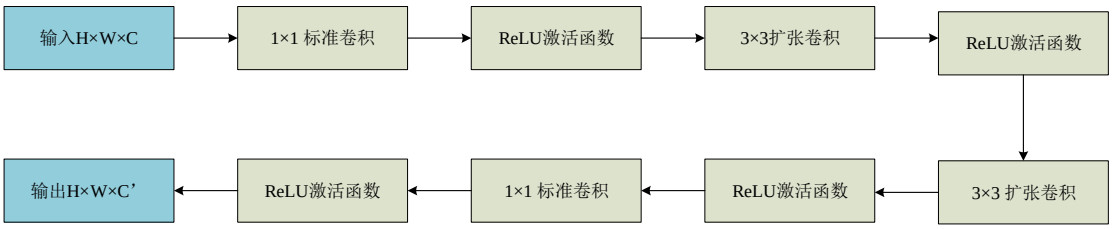


图 4-4 密度图回归模块的结构图

Fig 4-4 Structure diagram of the density map regression module

4.2.2 直接计数网络分支

Transformer 模型已经成为自然语言处理的首选模型。然而在计算机视觉领域，卷积架构仍然占据指导地位。ViT[51]是 2020 年 Google 团队提出的将 Transformer 应用在图像分类的模型，虽然不是第一篇将 Transformer 应用在视觉任务的论文，但是因为其模型“简单”且效果好，可扩展性强，成为了 Transformer 在计算机视觉领域应用的里程碑著作。

ViT 的模型如图 4-5 所示：

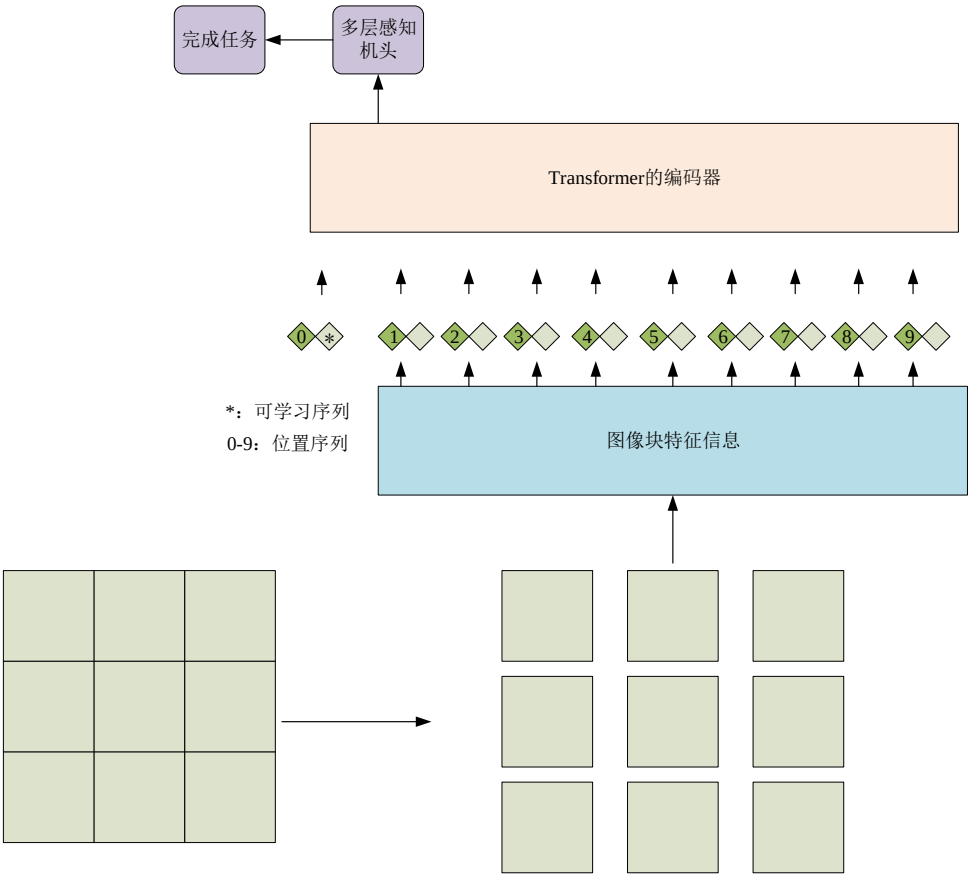


图 4-5 ViT 模型概述

Fig 4-5 Overview of the ViT model

由图 4-5 可知，将输入的特征图分割成固定大小的补丁，线性嵌入每个补丁，并添加位置嵌入，再将结果向量序列馈送到标准的 Transformer 编码器。为了执行分类任务，使用标准方法向序列中添加一个额外的可学习的“分类令牌”（classification token）。

Transformer 编码器接受作为输入的令牌嵌入的 1D 序列。为了处理 2D 图像，将图像 $x \in R^{H \times W \times C}$ 重塑为一个平坦的 1D patch 序列 $x \in \mathrm{R}^{N \times (P^2 \cdot C)}$，其中 (H,W) 为原始图像的分辨率，C 为通道数，(P,P) 为每个图像块的分辨率。$N = HW / P^2$ 得到的块数，这也是 Transformer 的有效输入序列长度。使用标准的可学习的 1D 位置嵌入添加到补丁嵌入中以保留位置信息。最后将 1D 序列送入 Transformer 模型的编码器，ViT 中没有解码器。

在本节的直接计数网络分支采用几个堆叠的 Transformer 编码器来直接生成人数，模型结构如图 4-6 所示。

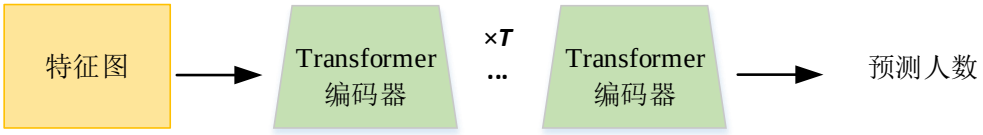


图 4-6 直接计数网络分支

Fig 4-6 Count network branches directly

直接计数网络分支的模型参数配置如表 4-1 所示。

表 4-1 直接计数网络分支的模型参数

Table 4-1 Model parameters that count network branches directly

	头数量	输入通道	输出通道
编码器 1	6	112	112
编码器 2	6	112	112
编码器 3	6	112	112
编码器 4	6	112	1

其中表 4-1 中头数量表示编码器中多头注意力机制的自注意力模块数量。

4.2.3 人数融合网络分支

在上述的密度图预测网络分支和直接计数网络分支都通过不同的方式生成了预测的人群数量，为了融合上述两个分支预测的人数，设计了人数融合网络分支，它可以自适应地融合密度图预测网络分支和直接计数网络分支估计的结果。利用全局平均池化对主干特征图C_f进行下采样，然后通过全连接网络对动态因子γ进行回归。人数融合网络分支对应的运算可表示为:

$$\gamma_1 = F_{gap}(C_f) \tag{4-2}$$

$$\gamma_2 = F_{fc}(\gamma_1) \tag{4-3}$$

$$\gamma = Sigmoid(\gamma_2) \tag{4-4}$$

其中式 4-2 F_{gap}表示全局平均池化，式 4-3 中F_{fc}表示全连接网络，又因为输出的动态因子应为 0-1 的数。所以式 4-4 中$Sigmoid()$是将动态因子γ正则化到数值区间[0,1]的 Sigmoid 激活函数。模型的结构图如图 4-7 所示。

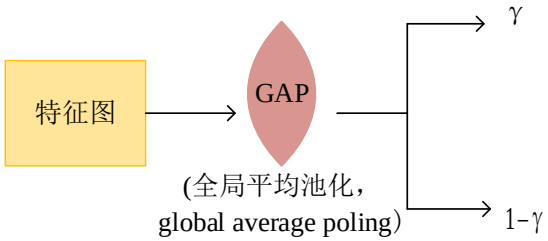


图 4-7 人数融合网络分支

Fig 4-7 The number of people is integrated into the network branch

在动态地生成权值因子后，就需要融合密度图预测网络分支和直接计数网络分支的人数。首先将密度图预测网络分支得到的估计密度图C求和得到估计的人数C_{den}，再利用人数融合网络分支融合两个分支估计的人数，具体计算公式如式 4-5，式 4-6 所示。

$$C_{den}=\sum_{i=0}^{W}\sum_{j=0}^{H}C(i,j) \tag{4-5}$$

$$D=\gamma C_{den}+(1-\gamma)T_{den} \tag{4-6}$$

其中式 4-5 中W为密度图宽度，H为密度图高度，式 4-6 中T_{den}是直接计数网络分支的估计的人数，γ为生成的权值项，D为最终预测人数。

4.2.4 损失函数设计

由于本章提出的模型最终预测人数的结果是结合密度图预测网络分支与直接计数网络分支得到的，所以在本章中该模型的损失函数设定如式 4-7 所示：

$$L_{loss}=\lambda_1 L_{den}+L_{cnt}+\lambda_2 L_{vit} \tag{4-7}$$

其中L_{den}表示预测密度图与真实密度图之间的损失函数，L_{cnt}表示最终预测人数与真实人数之间的损失函数，L_{vit}表示直接计数网络分支的预测结果与真实人数之间的损失函数。其中λ_1，λ_2为每个目标分量的权重超参数。

L_{den}利用像素级损失测量真实密度图与预测密度图之间的偏差，如式 4-8 所示：

$$L_{den}=\| C-\check{C} \|_2^2 \tag{4-8}$$

其中$\check{C}$为真实密度图，C为预测密度图。

L_{cnt}测量最终预测人数与真实人数之间的偏差，如式 4-9 所示：

$$L_{cnt} = \| D - \check{D} \|_1 \tag{4-9}$$

其中$\check{D}$为真实人数，D为预测人数。

L_{vit}测量直接计数网络分支的预测结果与真实人数之间的偏差，如式 4-10 所示：

$$L_{vit} = \| T_{den} - \check{D} \|_1 \tag{4-10}$$

其中T_{den}为直接计数网络分支的预测结果，$\check{D}$为真实人数。

4.3 实验与结果分析

4.3.1 实验设置

本章中模型的软件硬件配置均和第三章实验设置一样。在训练阶段，首先使用随机裁剪和水平翻转从源人群图像中生成训练补丁。随机裁剪大小设置为224×224，以减少内存使用。接下来，使用Adam算法对本章提出的模型框架进行优化，默认学习率l_r为 1e-5，权值衰减为 1e-4，批大小为 8，训练轮数为 500。损失函数为 4.2.4 节定义的L_{loss}。

本节在第二章介绍的ShanghaiTech、UCF-QNRF和NWPU-Crowd三个主流人群统计数据集上，对所提出的模型的框架的性能进行了评估。

4.3.2 实验训练过程

本文提出的模型没有任何预训练模型可以使用，所以需要重新训练，训练时输入为选定的人群数据集，输出为模型的权重λ，训练的大致过程如下。

（1）初始化训练轮数$i=0$，设置训练轮数为 500，当i达到 500 或达到设定的*MAE*或*MSE*时，训练结束。

（2）每执行一次第一步，将会得到预测密度图C和真实密度图$\check{C}$，最终预测人数D，真实人数$\check{D}$，直接计数网络分支预测人数T_{den}。

（3）根据第 4.2.4 节定义的损失函数计算出损失：$L_{den}(C,\check{C})$，$L_{loss}(C,\check{C},D,\check{D},T_{den})$，$\triangle\lambda_1 = L_{den}$，$\triangle\lambda_2 = L_{loss}$。

（4）计算当前训练的*MAE*和*MSE*。

（5）当训练的次数达到了一个指定更新学习率的轮次时，更新学习率

$l_r = 0.5l_r$。

（6）更新模型的权重 $\lambda_1 = \lambda_1 - l_r * \triangle\lambda_1$， $\lambda_2 = \lambda_2 - l_r * \triangle\lambda_2$。

（7）重复上述除（1）以外的所有步骤，当训练达到步骤（1）的条件时，停止训练。

4.3.3 实验结果和对比分析

在本小节中，评估了本章提出的计数模型的性能，并与其它人群计数的主流方法进行了比较。比较的计数模型分为两类，一类是常规的人群计数模型，另一类是轻量化人群计数模型。接着进行消融实验，证明本章提出模型各网络分支的优越性。

由表 4-2 可知，本章提出的模型对比第三章的模型参数量和计算量在没有显著增加的情况下，计数的精度实现了明显的提升。和常规的人群计数模型相比，本章模型取得了最低的参数量和计算量，且该模型仅在PartA数据集上的 *MAE* 相较于人群计数模型CAN低 1.6，与其余常规的人群计数相比，均取得了最低的 *MAE* 和 *MSE* 。与轻量化的人群计数模型相比，本章提出的模型虽然没有取得最低的参数量，但计数精度相较于轻量化人群计数模型取得了显著提升。

其中可视化结果如图 4-8。图 4-8 从左到右依次为输入图片，真实密度图和密度图预测网络分支生成的预测密度图，第一行和第二行为PartA数据集上部分实验结果，第三行和第四行为PartB数据集上部分实验结果，第五行和第六行为UCF-QNRF数据集上部分实验结果，第七行和第八行为NWPU-Crowd数据集上部分实验结果。

图 4-8(a)的真实人数为 500 人，由密度图预测网络分支估计的人数为 484.89 人，由直接计数网络分支估计的人数为 510.07 人，人数融合网络分支生成的动态权值因子为 0.48，最终预测人数为 498 人。图 4-8(b)的真实人数为 164 人，由密度图预测网络分支估计的人数为 162.91 人，由直接计数网络分支估计的人数为 169.03 人，人数融合网络分支生成的动态权值因子为 0.66，最终预测人数为 165 人。图 4-8(c)的真实人数为 234 人，由密度图预测网络分支估计的人数为 231.02 人，由直接计数网络分支估计的人数为 243.22 人，人数融合网络分支生成的动态权值因子为 0.51，最终预测人数为 237 人。图 4-8(d)的真实人数

为 194 人，由密度图预测网络分支估计的人数为 190.32 人，由直接计数网络分支估计的人数为 194.22 人，人数融合网络分支生成的动态权值因子为 0.57，最终预测人数为 192 人。图 4-8(e)的真实人数为 3652 人，由密度图预测网络分支估计的人数为 3010.07 人，由直接计数网络分支估计的人数为 3928.80 人，人数融合网络分支生成的动态权值因子为 0.31，最终预测人数为 3644 人。图 4-8(f)的真实人数为 3302 人，由密度图预测网络分支估计的人数为 2998.84 人，由直接计数网络分支估计的人数为 3453.63 人，人数融合网络分支项生成分支生成的动态权值因子为 0.34，最终预测人数为 3299 人。图 4-8(g)的真实人数为 544 人，由密度图预测网络分支估计的人数为 531.76 人，由直接计数网络分支估计的人数为 550.72 人，人数融合网络分支生成的动态权值因子为 0.46，最终预测人数为 542 人。图 4-8(h)的真实人数为 94 人，由密度图预测网络分支估计的人数为 93.01 人，由直接计数网络分支估计的人数为 98.22 人，人数融合网络分支生成的动态权值因子为 0.81，最终预测人数为 94 人。

表 4-2 不同人群计数模型方法在 3 个数据集的实验结果

Table 4-2 Experimental results of different number of people estimation model methods in three datasets

人群计数 模型方法	PartA		PartB		UCF-QNRF		NWPU-Crowd		Params (M)	计算量 （GFLOPS）
	MAE	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*		
CP-CNN[23]	73.6	106.4	20.1	30.1	-	-	-	-	68.4	-
CSRNet[22]	68.2	115.0	10.6	16	120.3	208.5	121.1	387.8	16.26	325.3
CAN[74]	**62.3**	100	7.8	12.2	107.0	183.0	106.3	386.5	18.1	193.58
DUB-Net[67]	64.4	106.8	7.7	12.5	105.6	180.5	-	-	18.05	-
SUA-Fully[75]	66.9	125.6	12.3	17.9	119.2	213.3	-	-	15.85	-
MCNN[20]	110.2	173.2	26.4	41.3	277.0	426.0	232.5	714.6	**0.13**	11.87
TDF-CNN[68]	97.5	145.1	20.7	32.8	-	-	-	-	**0.13**	-
SANet[36]	75.3	122.5	10.5	17.9	152.6	247.0	190.6	491.4	**0.91**	71.45
LCNet[33]	93.3	149.0	15.3	25.2	-	-	-	-	**0.86**	-
PCC-Net[37]	73.5	124.0	11.0	19.0	148.7	247.3	167.4	566.2	**0.55**	72.80
PDDNet[34]	72.6	112.2	10.3	17.0	130.2	246.6	-	-	**1.1**	-
LMSFFNet[35]	85.85	139.9	9.2	15.1	112.8	201.6	-	-	**4.58**	14.9
Our model	63.9	**94.0**	**6.9**	**11.3**	**84.2**	**143.5**	**78.9**	**340.3**	7.21	**7.90**

本节对 UCF-QNRF 数据集进行消融分析，因为它涵盖了不同人群密度水平

的场景，计数范围从 0 到 12865。本章提出的模型主要特点可以概括为两个方面:（1） CNN 和 Transformer 的混合用于特定任务的学习。（2）人数融合网络分支动态生成权值因子来融合密度图预测网络分支和直接计数网络分支的人数，生成最终预测人数。为了进一步证明这两个方面在解决人群计数任务上的有效性。在实验中设计了比较模型 ModelA，ModelB, ModelC，其中比较模型的密度特征提取模块相同，人群计数模块如表 4-3 所示，计数性能如表 4-3 所示。

表 4-3 消融实验的实验结果

Table 4-3 Experimental results of ablation experiments

	网络分支		最终结果融合			
模型	密度图预测网络分支	直接计数网络分支	平均融合	人数融合网络分支	*MAE*	*MSE*
ModelA	✓	×	×	×	105.4	179.5
ModelB	×	✓	×	×	94.3	161.5
ModelC	✓	✓	✓	×	89.2	154.8
Our Model	✓	✓	×	✓	**84.2**	**143.5**

实际上，本文设计的多任务并行人群计数模型中的密度图预测网络分支和直接计数网络分支都可以用来独立的进行人群计数，本章首先在 UCF-QNRF 数据集上评估了单个网络分支的性能。ModelA 包含密度图预测网络分支，它的 *MAE* 和 *MSE* 分别为 105.4 和 179.5，ModelB 包含直接计数网络分支，它的 *MAE* 和 *MSE* 分别为 94.3 和 161.5。接着 ModelC 合并密度图预测网络分支和直接计数网络分支，最终预测人数采用平均融合的方法，此时的 *MAE* 和 *MSE* 分别为 89.2 和 154.8。可以发现，两个分支相互促进， *MAE* 和 *MSE* 都有所下降。证明结合密度图预测网络分支和直接计数网络分支要优于单个网络分支的性能。最后，本章的模型合并密度图预测网络分支和直接计数网络分支，最终预测人数采用人数融合网络分支动态生成权值因子。此时， *MAE* 和 *MSE* 相比于 ModelC 也有所下降。因此实验证明动态生成权值因子要优于平均融合。

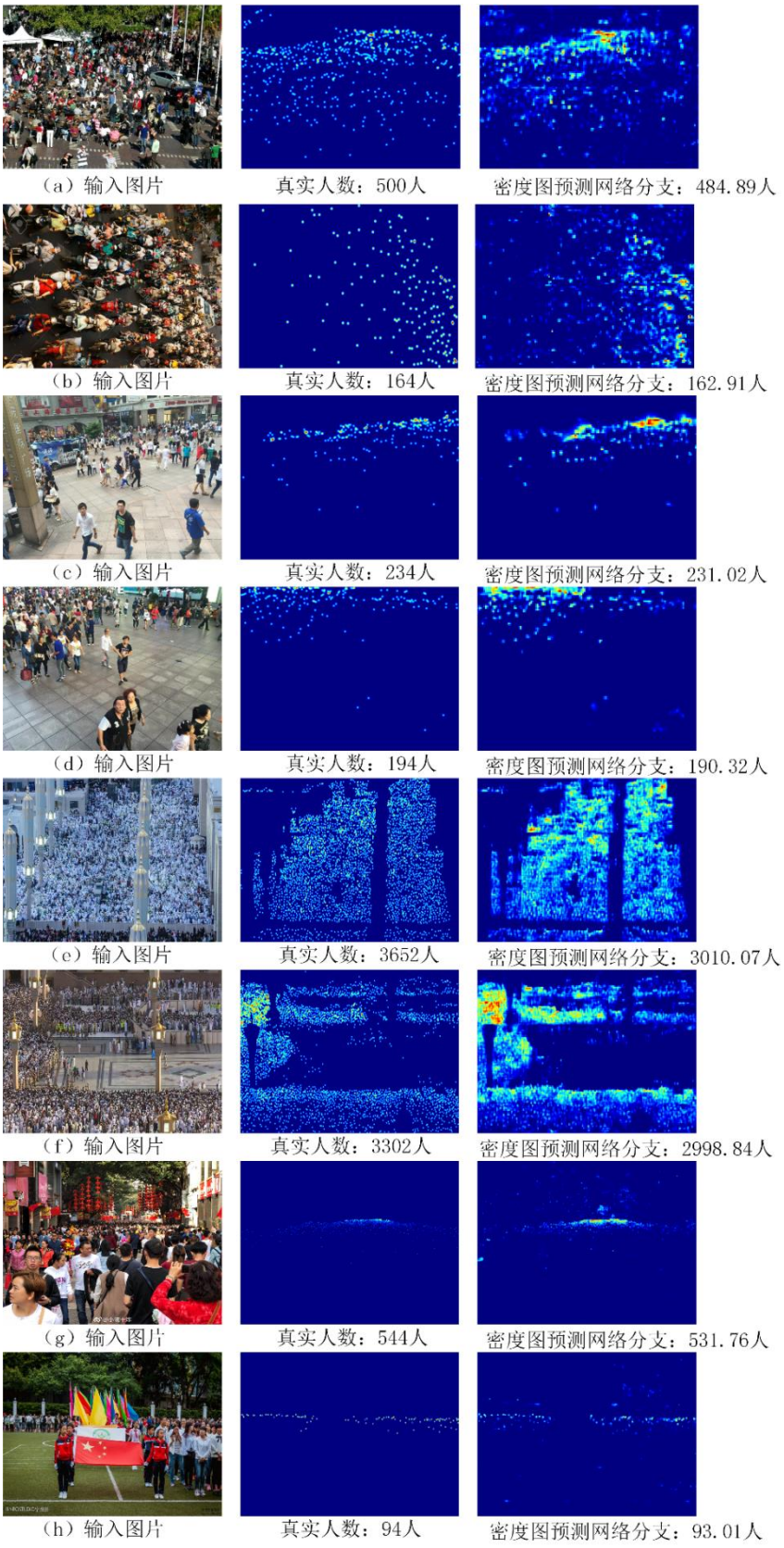


图 4-8 数据集上部分实验结果

Fig 4-8 Some of the experimental results on the dataset

4.4 本章小结

本章优化了计数网络中的人群计数模块，提出了一种多任务并行的人群计数网络模型，提出的模型框架在相应的数据集上都取得了优异的计数性能。利用 CNN 网络和 Transformer 网络各自的优点，设计了一个并行 CNN 和 Transformer 的多任务学习模型。将输入图片经过密度特征提取模块提取到的特征图馈送到三个分支，在密度图预测网络分支采用堆叠的扩张卷积来回归密度图求和得到估计人数，在直接计数网络分支使用 ViT 来直接估计人数。最后提出了人数融合网络分支，使用全局平均池化来动态生成权值项，自适应融合密度图预测网络分支和直接计数网络分支的估计人数，得到最终的预测人数。 最后在实验阶段，将本章提出的计数模型在三个主流的数据集上进行了实验，并与多个人群计数模型进行对比。由实验的指标可知，本章提出的模型在没有显著增加参数量和计算量的同时，取得了优异的计数性能。

第5章 总结与展望

5.1 总结

近年来，随着城市居住人口和人群密度的逐渐增加，踩踏事故时有发生。利用计算机视觉技术在图像或视频中准确估计人群数量的应用愈加广泛。在一些公共集会或体育赛事上，参与的人数是活动规划或者空间设计的重要信息。越来越多的学者致力于深度学习的人群计数研究，并提出了大量高质量的计数方法。然而在实际应用中，仍然有很多问题影响着人群计数的性能。人群计数模型主要由密度特征提取模块和人群计数模块两部分构成，而大多数性能较好的计数方法都是使用参数量或计算量较大的网络作为主干来进行密度特征的提取，虽然可以达到较好的计数性能，但是在模型的训练过程中，这些方法消耗了很多资源，实用性不高。其次，当前人群计数网络大多采用单一的 CNN 网络和 Transformer 网络来进行计数，而最新研究可知，CNN 网络在提取局部特征时的能力较强，但是对全局特征提取能力较弱；Transformer 模型利用多头自注意力机制能够捕获全局特征，但它可能会忽略局部区域内的细节特征。这些单一的人群计数网络难以应对当前计数场景的多变性和人群尺度分布不均匀导致计数精度降低的问题。因此现有的人群计数模型在计数任务上还有着严峻的挑战。针对当前人群计数模型中存在的上述问题，本文通过提出一种轻量化密度特征提取网络来降低模型的复杂度，在此基础上，进一步构建 CNN 和 Transformer 的多任务并行网络来提高计数精度。本文所做的具体工作如下：

（1）构建了一种轻量化密度特征提取网络。针对人群计数模型拥有复杂的主干网络导致资源消耗过大，无法实时计数的问题，本文提出了一种轻量化密度特征提取网络模型。该网络利用简单的轻量化线性映射生成特征图，通过浅层特征图和深层特征图的融合减少特征损失，提高预测密度图的精度，最后回归估计密度图求和得到预测人数。最后在三个主流的人群数据集上进行了实验，并与常规人群计数模型和轻量化计数模型进行对比，证明了本文提出的轻量化计数模型在保持计数精度的同时，提高了计数效率。

（2）构建了一种密集场景下的多任务并行人群计数网络。针对当前计数背

景变化较快以及人群尺度分布不均匀导致人群计数精度不佳的问题，优化了计数模型的人群计数模块，设计了一种多任务并行人群计数方法。在密度特征提取模块采用第三章设计的轻量化特征提取网络进行密度特征提取，将提取到的特征图馈送到三个分支，三个分支各自学习不同的任务。密度图预测网络分支利用堆叠的扩张卷积来回归密度图，直接计数网络分支采用 ViT 直接生成人数，人数融合网络分支提出了使用全局平均池化动态生成权值项来融合密度图预测网络分支和直接计数网络分支的人数。最后在三个人群数据集上进行了实验，并与经典的网络计数模型进行对比，实验结果表明，在不显著增长参数量和计算量的同时，计数精度取得了明显的提升，证明了计数模型实现了优异的计数性能。

5.2 展望

本文研究了基于深度学习的人群计数方法，针对现有的人群计数算法存在的一些问题，提出了一些针对性的解决方法。然而人群计数任务还存在着很多严峻的挑战，但是研究工作过程中由于工作量和时间的关系，还有不少亟待解决的问题，在以后需要作进一步的研究。

（1）目前人群计数大多着重利用头部标注再结合高斯核进行建模，但是这样会忽略身体的其他部分，因此后续的研究工作可能可以对行人的身体部分进行建模，实现更加有效的密度生成。

（2）本文提出的多任务并行人群计数模型虽然取得了优异的计数性能，最终预测人数也采用了自适应融合的方法，但是在密度图预测网络分支和直接计数网络分支这两个分支中没有实现两个相关任务之间的信息传递，因此可能会有一些特征的丢失。因此在后续的研究中考虑加入两个分支间的信息传递。

参考文献

[1] 胡耀聪. 基于深度卷积网络的人群计数算法研究[D].合肥：安徽大学, 2017.

[2] DONG L, PARAMESWARAN V, RAMESH V, et al. Fast Crowd Segmentation Using Shape Indexing[C]// International Conference on Computer Vision (ICCV). Rio de Janeiro, Brazil: IEEE, 2007: 1-8.

[3] KANG K, WANG X. Fully Convolutional Neural Networks for Crowd Segmentation[J]. Computer Science, 2014, 49(1):25–30.

[4] TANG X, WANG X, ZHOU B. Understanding Collective Crowd Behaviors: Learning a Mixture model of Dynamic Pedestrian-Agents[C]// Conference on Computer Vision and Pattern Recognition (CVPR). Providence, RI, USA: IEEE, 2012: 2871-2878.

[5] SHAO J, CHEN C L, WANG X. Scene-Independent Group Profiling in Crowd[C]// Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, USA: IEEE, 2014:2227-2234.

[6] 王陆洋. 基于卷积神经网络的图像人群计数研究[D].合肥：中国科学技术大学, 2020.

[7] KONG D J, GRAY D, TAO H. A viewpoint invariant approach for crowd counting[C]// International Conference on Pattern Recognition (ICPR). Hong Kong, China: IEEE, 2006: 1187-1190.

[8] CHEN K, LOY C C, GONG S, et al. Feature mining for localised crowd counting[EB/OL]. https://api.semanticscholar.org/CorpusID:1910869, 2023-11-03.

[9] IDREES H, SALEEMI I, SEIBERT C, et al. Multi-source multi-scale counting in extremely dense crowd images[C]// Conference on Computer Vision and Pattern Recognition (CVPR). Portland, OR, USA: IEEE, 2013: 2547–2554.

[10] LEMPITSKY V, ZISSERMAN A. Learning to count objects in images[C]// In Advances in Neural Information Processing Systems (NIPS). Vancouver, Canada: IEEE, 2010:1324–1332.

[11] KRIIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[12] REN S Q, HE K M, GIRSHICK R. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[13] CHEN L C, PAPANDREOU G, KOKKINOS I. Semantic image segmentation with deep convolutional nets and fully connected CRFS[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[14] SINDAGI V A, PATEL V M. Hierarchical attention-based crowd counting network[J]. IEEE Transactions on Image Processing, 2020, 29: 323-335.

[15] WANG C, HANG H, YANG L, et al. Deep people counting in extremely dense crowds[C]// Proceedings of the 23rd ACM international conference on Multimedia. New York NY United

States: ACM, 2015: 1299–1302.

[16] HANG C, LI H S, WANG X G. Cross-scene crowd counting via deep convolutional neural networks[C]// Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015: 833–841.

[17] WALACH E, WOLF L. Learning to count with cnn boosting[C]// European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands: Springer, 2016: 660–676.

[18] SHANG C, AI H, BAI B. End-to-end crowd counting via joint learning local and global count[C]// IEEE International Conference on Image Processing (ICIP). Phoenix, AZ, USA: IEEE, 2016: 1215–1219.

[19] BOOMINATHAN L, KRUTHIVENTI S S, BABU R V. CrowdNet: A deep convolutional network for dense crowd counting[C]// International Multimedia Conference (MM). New York, United States: ACM, 2016: 640–644.

[20] ZHANG Y, ZHOU D, CHEN S, et al. Single-image crowd counting via multi-column convolutional neural network[C]// Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016:589-597.

[21] SAM D B, SURYA S Babu, BADU R. Switching convolutional neural network for crowd counting[C]// Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 5744–5752.

[22] LI Y, ZHANG X, CHEN D. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes[C]// Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE, 2018: 1091-1110.

[23] SINDAD V A, PATEL V M. Generating high-quality crowd density maps using contextual pyramid CNNs[C]// International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 1861–1870.

[24] JIANG X, XIAO Z, ZHANG B, et al. Crowd counting and density estimation by trellis encoder-decoder networks[C]// Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 6133–6142.

[25] WANG P, GAO C, WANG Y. Mobilecount: An efficient encoder-decoder framework for real-time crowd counting[J]. Neurocomputing, 2020, 407: 292–299.

[26] SONG Q, WANG C, WANG Y, et al. To choose or to fuse? scale selection for crowd counting[J]. AAAI Technical Track on Computer Vision II, 2021, 35(3): 2576–2583.

[27] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 770–778.

[28] HAN S, MAO H Z, DALLY W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding[J]. Fiber, 2015,56(4): 3-7.

[29] RASTEGARI M, ORDONEZ V, REDMON J, et al. XNOR-Net: Imagenet classification using binary convolutional neural networks[C]// European Conference on Computer Vision

(ECCV). Amsterdam, the Netherlands: Springer, 2016: 525–542.

[30] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network [EB/OL]. https://doi.org/10.48550/arXiv.1503.02531, 2015-03-09.

[31] LI H, KADAV A, DURDANOVIC I, SAMET H, et al. Pruning filters for efficient convnets [EB/OL]. https://doi.org/10.48550/arXiv.1608.08710, 2017-03-10.

[32] HOWARD A G, ZHU M L, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications [EB/OL]. https://doi.org/10.48550/arXiv.1704.04861, 2017-04-17.

[33] MA X, DU S, LIU Y. A lightweight neural network for crowd analysis of images with congested scenes[C]// IEEE International Conference on Image Processing (ICIP). Taipei, Taiwan: IEEE, 2019: 979–983.

[34] LIANG L H, ZHAO H L, ZHOU F B, et al. PDDNet: lightweight congested crowd counting via pyramid depth-wise dilated convolution[J]. Applied Intelligence, 2022, 53(9): 10472-10484.

[35] YI J, SHEN Z L, CHEN F, et al. A Lightweight Multiscale Feature Fusion Network for Remote Sensing Object Counting[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5902113.

[36] CAO X, WANG Z, ZHAO Y, et al. Scale aggregation network for accurate and efficient crowd counting[C]// Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 734-750.

[37] GAO J, WANG Q, LI X. PCC Net: Perspective crowd counting via spatial convolutional network[J]. IEEE Transactions on Circuits and Systems for Video Technology,2020, 30(10)**:** 3486–3498.

[38] HU Y, CHANG H, NIAN F, et al. Dense crowd counting from still images with convolutional neural networks[J]. Journal of Visual Communication and Image Representation, 2016, 38: 530–539.

[39] WAN J, KUMAR N S, CHAN A B. Fine-grained crowd counting. IEEE transactions on image processing, 2021, 30: 2114–2126.

[40] JIANG S, LU X, LEI Y, et al. Mask-aware networks for crowd counting. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 20(9): 3119–3129.

[41] MENG Y, BRIDGE J, ZHAO Y, et al. Transportation object counting wit-h graph-based adaptive auxiliary learning[J]. IEEE Transactions on Intelligent Transportation Systems. 2023, 24(3): 3422–3437.

[42] LIU J, GAO C, MENG D, et al. DecideNet: Counting varying density crowds through attention guided detection and density estimation[C]// Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE, 2018: 5197–5206.

[43] LIU C, WENG X, MU Y. Recurrent attentive zooming for joint crowd counting and precise localization[C]// Conference on Computer Vision and Pattern Recognition (CVPR). Long

Beach, CA, USA: IEEE, 2019: 1217-1226.

[44] WAN J, LIU Z, CHAN A B. A generalized loss function for crowd counting and localization. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Virtual: IEEE, 2021: 1974–1983.

[45] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]// Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186.

[46] LIU Y H, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[C]// Chinese National Conference on Computational Linguistics (CCL). Huhhot, China: Chinese Information Processing Society of China, 2021: 1218-1227.

[47] CARION N, MASSA F, SYNNAEYE G, et al. End-to-end object detection with transformers[C]// European Conference on Computer Vision (ECCV). Tel Aviv, Israel: Springer, 2020: 213–229.

[48] CHEN H T, WANG Y H, GUO T Y, et al. Pretrained image processing transformer [EB/OL]. https://doi.org/10.48550/arXiv. 2012.00364, 2020-12-03.

[49] DAI Z G, CAI B L, LIN Y G, et al. UP-DETR: Unsupervised pre-training for object detection with transformers[C]// Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2021: 1601-1610.

[50] ZHU X Z, SU W J, LU L W, et al. Deformable DETR: Deformable transformers for end-to-end object detection[C]// International Conference on Learning Representations (ICLR). Virtual: Ithaca, 2021: 1041-1056.

[51] DOSOVITSKIY A, BETER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]// International Conference on Learning Representations (ICLR). Virtual: Ithaca, 2021: 1909-1929.

[52] MAO W A, GE Y T, S C H, et al. TFPose: Direct human pose estimation with transformers [EB/OL]. https://doi.org/10.48550/arXiv.2103.15320, 2021-03-29.

[53] WANG X P, YESHWANTH C, NIEßNER M. Sceneformer: Indoor scene generation with transformers [EB/OL]. https://doi.org/10.48550/arXiv.2012.09793, 2021-04-02.

[54] VALANARASU J M J, OZA P, HACIHALILOGLU I, et al. Medical Transformer: Gated axial-attention for medical image segmentation[C]// International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Strasbourg, France: Springer, 2021: 36-46.

[55] YANG S, GUO W, REN Y. Crowdformer: An overlap patching vision transformer for top-down crowd counting[C]// International Joint Conference on Artificial Intelligence (IJCAI). ShenZhen China: Morgan Kaufmann, 2022: 1545–1551.

[56] LIANG D, CHEN X, XU W, et al. Transcrowd: weakly-supervised crowd counting with transformers[J]. Science China Information Sciences, 2022, 65(6): 160104.

[57] LI B, ZHANG Y, XU H, et al. CCST: crowd counting with swin transformer[J]. The Visual Computer: International Journal of Computer Graphics, 2022, 39(7):2671-2682.

[58] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]// Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 779–788.

[59] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]// International Conference on Learning Representations (ICLR). San Diego, CA, USA: OpenReview.net, 2015: 1-14.

[60] KONG D J, GRAY D, TAO H. A viewpoint invariant approach for crowd counting[C]// International Conference on Pattern Recognition (ICPR). Hong Kong, China: IEEE, 2006: 1187-1190.

[61] 夏殷锋. 基于卷积神经网络的人群密度估计方法研究[D]. 合肥：中国科学技术大学, 2021.

[62] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// International Conference on Neural Information Processing Systems (NIPS). Long Beach California USA: ACM, 2017: 6000-6010.

[63] IDREES H, TAYYAB M, ATHREY K, et al. Composition loss for counting, density map estimation and localization indense crowds[C]// European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018:532-546.

[64] WANG Q, GAO J, LIN W, et al. Nwpu crowd: A largescale benchmark for crowd counting and localization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(6): 2141-2149.

[65] 陆金刚，张莉. 基于多尺度多列卷积神经网络的密集人群计数模型[J]. 计算机应用, 2019, 39(12): 3445-3449.

[66] 贾云舒. 基于深度学习的人群计数方法研究[D]. 成都：电子科技大学, 2022.

[67] OH M, OLEN P A, RAMAMURTHY K N. Crowd counting with decomposed uncertainty[C]// AAAI Conference on Artificial Intelligence (AAAI). New York, NY, USA: AAAI, 2020: 11799–11806.

[68] SAM D B, BABU R V. Topdown feedback for crowd counting convolutional neural network[C]// AAAI Conference on Artificial Intelligence (AAAI). New Orleans, LA, USA: AAAI, 2018: 7323–7330.

[69] IOFFE S, SZEGED C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]// International Conference on Machine Learning (ICML). Lille, France: ACM, 2015: 448-457.

[70] SANDLER M, HOWARD A, ZHU M L, et al. Mobilen-etv2: Inverted residuals and linear bottlenecks[C]// Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE, 2018: 4510–4520.

[71] 付宇豪. 基于卷积神经网络的人群计数算法研究与应用[D]. 北京：北京工业大学, 2019.

[72] 陆金刚. 面向人群数量估计的多尺度卷积神经网络模型研究[D]. 苏州：苏州大学, 2020.

[73] 刘彦博，贾瑞生，徐志峰. 基于尺度空间金字塔网络的人群计数算法[J]. 中国科技论文, 2021, 16(3): 276-280.

[74] LIU W, SALZMZNN M, FUA P. Context aware crowd counting[C]// Conference on Computer Vision and Patter-n Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 5099–5108.

[75] MENG Y D, ZHANG H R, ZHAO Y T, et al. Spatial uncertainty-aware semi-supervised crowd counting[C]// International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 15549–15559.

[76] WANG M, CAI H, HAN X, et al. STNet: Scale tree network with multi-level Auxiliator for crowd counting[J]. IEEE Transactions on Multimedia, 2022, 25: 2074-2084.

[77] LEI Y, LIU Y, ZHANG P. Towards using count level weak supervision for crowd counting[J]. Pattern Recognition, 2021, 109: 107616.

致谢

在安徽工程大学的七年时间里，我要真诚的感谢我的导师杨会成老师，老师在学习和生活上对我的帮助，让我受益匪浅，还要感谢陪我们度过硕士生涯的胡耀聪，徐可老师，感谢您对我的帮助。其次我要感谢我的同门李雯婷，帅真，感谢你们的陪伴，让我的学习生涯充满欢声笑语，感谢师弟师妹让 402 实验室时刻充满朝气，感谢我的爸爸妈妈和姐姐，感谢你们每时每刻都站在我的身后鼓励我，最后祝愿母校安徽工程大学越来越好。

攻读硕士学位期间的研究成果

1 学术论文

[1] LIN Y Y, YANG H C, HU Y C, et al. Crowd counting in congested scene by CNN and Transformer[C]// 2023 7th International Conference on Electronic Information Technology and Computer Engineering (EITCE). Xiamen, China: ACM, 2023: 1092-1095.

[2] 林园园，杨会成，胡耀聪.基于轻量化卷积神经网络的人数估计算法研究[J]. 重庆工商大学自然科学版学报.（已录用）.

[3] HU Y C, LIN Y Y, YANG H C, et al. CLDE-Net: Crowd Localization and Density Estimation based on CNN and Transformer Network[J]. IEEE Intelligent Transportation. Systems Transactions Scholar One Manuscripts Administrative Center.（已录用）

2 发明专利

[1] 胡耀聪，林园园，杨会成，等. 基于 CNN 与 Transformer 的人数估计方法[P].（已受理，申请号：202311026153.9）