Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2025, 14(3), 527-535
Published Online May 2025 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2025.143052

Hans汉斯

基于多注意力特征融合的人群计数方法

李俊恩[1]，谭显洋[2]，陆许明[2]

[1]五邑大学机械与自动化工程学院，广东 江门
[2]五邑大学电子与信息工程学院，广东 江门

收稿日期：2025年2月25日；录用日期：2025年4月30日；发布日期：2025年5月12日

摘 要

针对拥挤环境下人群分布的高度不均匀、复杂背景的干扰以及遮挡问题，本文提出一种基于MobileNet V3分类模型的特征融合网络。首先从MobileNet V3网络中提取四个不同尺寸的特征图，并对每个特征图进行HAM模块操作，该模块由通道、边缘、空间注意力以及动态卷积组成，特征图通过上采样对齐分辨率，并在通道维度上拼接成综合特征图，经过1 × 1卷积压缩通道，生成最终的融合特征图用于生成密度图，完成高精度人群计数任务。该方法在ShanghaiTech、NWPU和QNRF三个具有挑战的数据集上进行了实验验证，实验结果表明，所提出的方法在计数精度和鲁棒性方面显著优于现有主流方法。

关键词

人群计数，注意力机制，多尺度特征，人群密度估计

A Crowd Counting Method Based on Multi-Attention Feature Fusion

Jun'en Li[1], Xianyang Tan[2], Xuming Lu[2]

[1]School of Mechanical and Automation Engineering, Wuyi University, Jiangmen Guangdong
[2]School of Electronics and Information Engineering, Wuyi University, Jiangmen Guangdong

Received: Feb. 25th, 2025; accepted: Apr. 30th, 2025; published: May 12th, 2025

Abstract

To address the challenges of highly non-uniform crowd distribution, complex background interference, and severe occlusions in crowded environments, this paper proposes a feature fusion network based on the MobileNet V3 classification model. The framework first extracts four multi-scale feature maps from the MobileNet V3 backbone. Each feature map undergoes processing through a Hybrid Attention Module (HAM), which integrates channel attention, edge attention, spatial attention,

文章引用: 李俊恩, 谭显洋, 陆许明. 基于多注意力特征融合的人群计数方法[J]. 人工智能与机器人研究, 2025, 14(3): 527-535. DOI: 10.12677/airr.2025.143052

and dynamic convolution operations. The processed features are then upsampled to align their spatial resolutions, concatenated along the channel dimension, and compressed via a 1 × 1 convolutional layer to generate a unified fused feature map. This fused representation is subsequently used to regress high-precision density maps for accurate crowd counting. The method is experimentally validated on three challenging datasets, namely ShanghaiTech, NWPU and QNRF, and the experimental results show that the proposed method significantly outperforms state-of-the-art approaches in both counting accuracy and robustness.

Keywords

Crowd Counting, Attention Mechanism, Multi-Scale Features, Crowd Density Estimation

1. 引言

旅游观光、庆典活动和体育赛事等大型社会活动日益增多，由于人群聚集造成的踩踏事故再次出现在大众的视野当中，如韩国梨泰园踩踏事故。人群计数和密度估计(Crowd Density Estimation)成为解决上述问题的关键，在视频监控和交通管理等安全方面应用广泛，可以帮助城市管理者、大型活动组织方实时了解人群拥挤情况，以利于早期防范群体事件、人群踩踏等。由于存在巨大的尺度变化、光照变化、复杂的背景以及人类的不规则分布，在现实中，人群计数仍然是一项非常具有挑战性的任务。因此，人群密度估计算法的研究具有重要的现实意义和应用价值。

早期的人群研究主要聚焦于基于检测的方法[1]。使用一个滑动窗口检测器来检测场景中人群，并统计相应的人数[2]。基于整体检测的传统方法，主要训练一个分类器，利用从行人全身提取的小波、HOG、边缘等特征去检测行人[3]。主要通过检测身体的部分结构，例如头、肩膀等去统计人群的数量。基于回归的方法，主要思想是通过学习一种特征到人群数量的映射。基于传统图像处理的目标计数方法所面临的最棘手的两个问题就是复杂背景的干扰和目标重叠。尽管已有一些文献提出相应的方法[4] [5]来克服这类问题，但这些方法的鲁棒性不够高，限制模型的推广使用。

目前人群计数领域大部分仍是基于密度图回归的方式来获得更高的计数效果，大多选用多分支。Hydra CNN [6]和 MCNN [7]是两种十分经典的目标计数模型。两者都是通过多列网络分支来学习多尺度目标，不同的是 Hydra CNN 是对输入图像的不同尺度进行处理，而 MCNN 采用不同大小的卷积核来获取不同大小的感受野。Li 等[8]提出的 CSRNet 基于 VGG-16 主干网络，并引入空洞卷积提取更大感受野的特征，从而生成高质量的密度图以实现精准的人群计数。Cao 等[9]提出的 SANet (Scale-Adaptive Network)采用多列卷积模块提取不同尺度的特征，并通过自适应融合层动态整合感受野信息，从而生成精细的密度图，以提升在人群目标尺度变化下的人群计数精度。Liu 等[10]提出的 CAN (Context-Aware Network)通过引入上下文感知机制，自适应地分配不同区域的权重，以捕获全局和局部特征。Gao 等[11]提出的 PCC-Net 基于空间卷积网络，引入透视感知机制以校正不同区域的尺度变化，网络通过选择性地继承和整合不同尺度的信息，从而更好地处理大规模和小规模目标的分布。LMSFFNet [12]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。Shu 等[13]通过将人群计数任务从空间域转换到频域，利用频域特征来有效捕捉人群分布和密度变化。Liang 等[14]提出利用 Transformer 模型，通过全局上下文信息的学习来解决人群定位任务。该方法采用端到端的变换器架构，能够有效捕捉长距离依赖关系并增

强模型在复杂场景中的定位能力。Han 等[15]提出的 Steerer 模型通过选择性继承学习(Selective Inheritance Learning)来解决尺度变化问题，采用了一种新的特征学习策略，以同时提高人群计数和定位任务的准确性。Tian 等[16]提出的 CCTrans 通过引入 Transformer 架构简化了人群计数任务，该方法利用自注意力机制捕捉全局依赖关系。当前的人群计数模型通过减少参数量和计算量来提升运行速度，但往往过度简化了特征融合过程，忽视了不同特征的有效结合。尽管[15] [16]在准确度上表现出色，但其高复杂度和大量参数量增加了部署和训练的难度。

针对上述问题，本文提出一种基于 MobileNet V3 [17]分类模型的特征融合网络，主要研究贡献如下：

1) 提出一个结构简单的人群计数模型，该模型仅有一个主干和一个多尺度特征融合结构组成，大幅度减少了模型的参数量和计算量；

2) 设计了多层注意力模块，通过空间、通道、边缘注意力和动态卷积，有效捕捉人群的关键特征。

2. 网络结构

本文提出的基于 MobileNet V3 分类模型的多注意力特征融合网络(Dynamic Attention Convolutional Network, DACNet)的结构如图 1 所示。该网络分为三部分：特征提取层、注意力模块、特征融合模块。

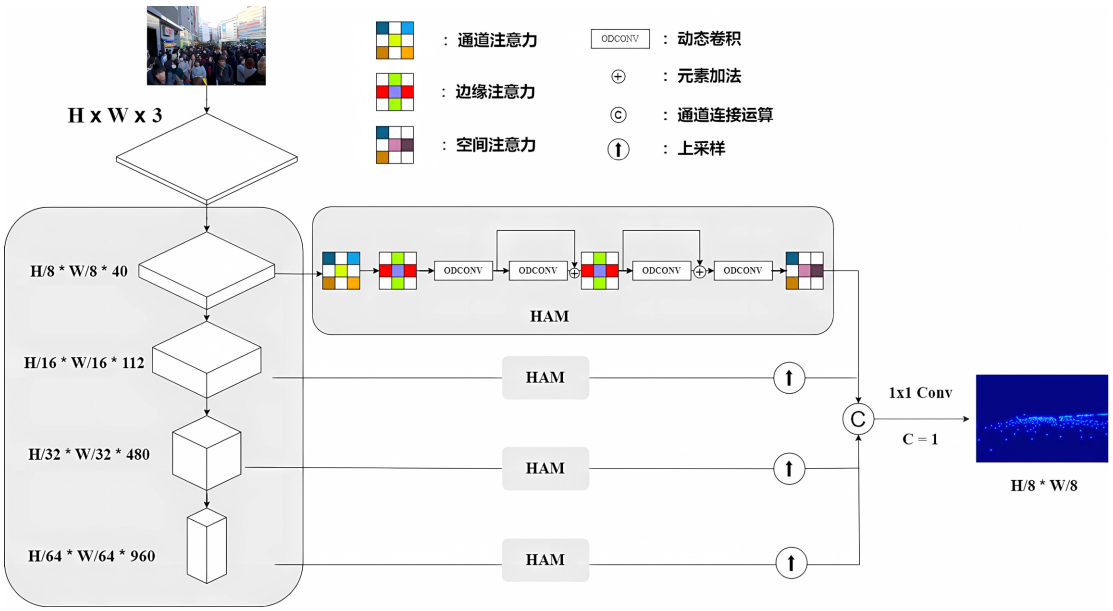

Figure 1. DACNet architecture diagram
图 1. DACNet 结构图

2.1. Backbone

主干网络是模型中最开始的若干层网络，使用的是 MobileNet V3 模型，相较于同类模型，其 large 版本仅需 5.4 M 参数。在需要复杂推理的人群计数任务中，使用 MobileNet 能够大大减少模型的参数量和计算复杂度。MobileNet 通过引入自适应卷积、SE 模块等，采用多种正则化技术，并通过优化的网络结构减少了过拟合的风险，能够更好地捕捉人群计数任务中的复杂场景特征，提升模型对不同人群密度、背景噪声等变化的适应能力。

2.2. HAM 模块

HAM (Hybrid Attention Module)混合注意力模块，多种注意力机制并行处理特征图。输入的图像先经

过前端网络，即 MobileNet V3 网络进行初级特征提取，输出通道分别为 40、112、480、960，输入到 HAM 模块。该模块的基础卷积为动态卷积[18]，在模块前后使用通道[19]和空间注意力[20]模块，卷积过程穿插使用边缘注意力[21]，专注相关特征。

注意力模块

图 2(a)通道注意力捕获特征图不同通道之间的相互依赖关系，通过全局平均池化获取每个通道的全局信息，根据重要性使用一个共享的全连接层为每个通道分配权重，抑制信息较少的特征，其计算如公式(1)所示：

$$M_C(F)=\sigma\left(\mathrm{MLP}\left(\mathrm{AvgPool}(F)\right)\right)+\mathrm{MLP}\left(\mathrm{MaxPool}(F)\right) \tag{1}$$

其中，M_C 为通道注意力权重矩阵，F 为输入特征图，σ 为 Sigmoid 激活函数，MLP 为全连接层，AvgPool 和 MaxPool 为全局平均池化和全局最大池化。

图 2(b)边缘注意力通过一个卷积层对特征图进行边缘检测，使用 sobel_x 和 sobel_y 分别计算水平方向和垂直方向的边缘生成注意力图，网络能够根据图像中的边缘信息自适应地调整每个像素点的特征贡献，从而更好地捕捉到重要区域的信息。其计算如公式(2)所示：

$$M_E(F)=\sigma\left(f_{attn(F)}\times\left(1+\alpha\times\frac{\sqrt{\left(F*G_x\right)^2+\left(F*G_y\right)^2}}{C}\right)\right) \tag{2}$$

其中，M_E 为边缘注意力权重矩阵，f_{attn} 为基础注意力图(两层卷积 + BN + ReLU + Sigmoid)，α 为可学习的边缘增强系数，G_x、G_y 为 Sobel 卷积核，C 为边缘幅度图的归一化常数。

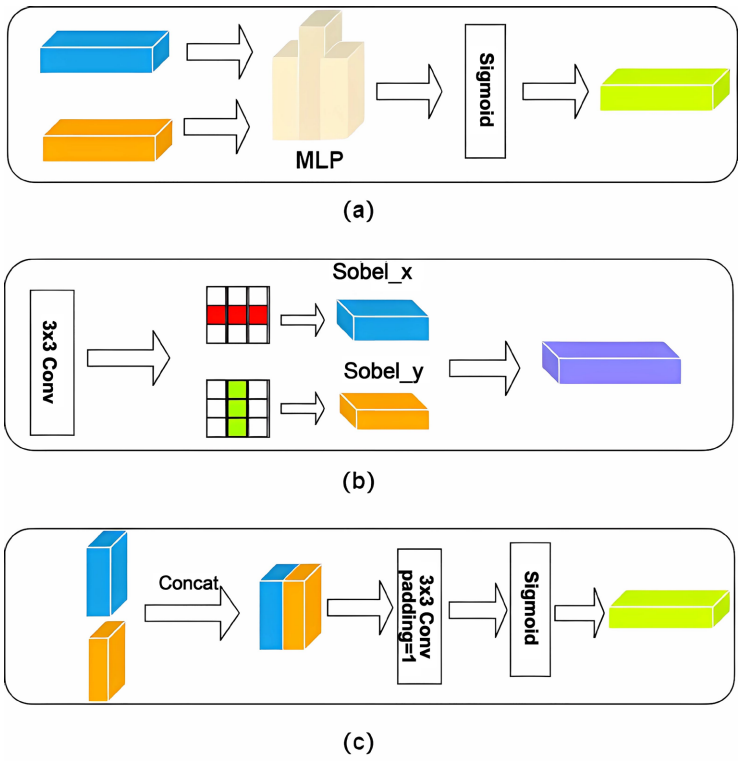

Figure 2. (a) Channel attention; (b) Edge attention; (c) Spatial attention
图 2. (a) 通道注意力；(b) 边缘注意力；(c) 空间注意力

图 2(c)空间注意力聚焦于空间维度的重要区域，对特征图进行平均池化，使用一个卷积层来生成空间注意力图分配权重。其计算如公式(3)所示：

$$M_S(F)=\sigma\left(f^{7\times7}\left(\left[\mathrm{AvgPool}(F);\mathrm{MaxPool}(F)\right]\right)\right) \tag{3}$$

其中，M_S为空间注意力权重矩阵，F 为输入特征图，$f^{7\times7}$为 7 × 7 的卷积核，[;]为通道维度上的拼接。

2.3. 动态卷积

动态卷积是由多个并行卷积核的注意力动态组合捕获多个维度特征，其中有输入通道、输出通道、内核大小、内核数量。动态卷积利用平均池化和自注意力组合来确定注意力权重，卷积核进行自适应参数调整，更好地与输入数据特征保持一致。

2.4. 损失函数

本文采用 DM-Count [22]中提出的 Loss 函数，由计数损失、最优传输损失和总变化损失组成。

计数损失是人群计数中的一个关键损失，其量化了模型预测的人数与实际标记的人数之间的差距，从而有助于模型更好地收敛。定义如公式(4)所示：

$$\ell_c(z,\hat{z})=\left|\|z\|_1-\|\hat{z}\|_1\right| \tag{4}$$

其中，$z\in R_+^n$ 表示点注释的矢量化二进制映射，$\hat{z}\in R_+^n$ 由神经网络返回的矢量化预测密度图，$\|\cdot\|1$ 表示向量的 L1 范数。

最优传输损失用于测量两个概率分布之间的差异，其利用 Sinkhorn 算法进行迭代，使模型学习更接近真实分布的概率密度，定义如公式(5)所示：

$$\ell_{ot}(\hat{z})=\left\langle\frac{\beta^*}{\|\hat{z}\|_1}-\frac{\langle\beta^*,\hat{z}\rangle}{\|\hat{z}\|_1^2},\hat{z}\right\rangle \tag{5}$$

其中，β^* 是 OT 损失中 Sinkhorn 算法迭代的解，$\langle\cdot,\cdot\rangle$代表两个向量的卷积。

变化损失用于采用L1 范数量化预测密度图和真实密度图中相邻像素之间的变化差异，如公式(6)所示：

$$\ell_v(z,\hat{z})=\|z\|_1\cdot\|z-\hat{z}\|_1 \tag{6}$$

总体损失函数是计数损失、OT 损失和 TV 损失的和，如公式(7)所示：

$$\ell_{all}=\ell_c+\lambda_1\ell_{ot}+\lambda_2\ell_v \tag{7}$$

其中，λ_1,λ_2 为可调参数，对损失进行加权。

评估指标：在测试时，本文采用平均绝对误差(Mean Absolute Error, MAE)和均方误差(Mean Square Error, MSE)。其定义如公式(5)、(6)所示：

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}\left|C_i-C_i^{GT}\right| \tag{8}$$

$$\mathrm{MSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|C_i-C_i^{GT}\right|^2} \tag{9}$$

其中，N 为测试图像的数量，C_i 为第 i 张图像中的预测人数，C_i^{GT} 为第 i 张图像中的真实人数，由预测的密度图和标注的真实密度图求和得到。

3. 网络结构

3.1. 实验环境

本文所使用的实验框架为 PyTorch 1.12.1 + python 3.8 + cuda 12.4 + anaconda 3，硬件配置为 NVIDIA GeForce RTX 4070 Ti GPU，Intel(R) Core(TM) i7-12700K CPU。模型训练时采用批量大小为 16 的 Adam 优化器对参数进行优化，初始学习率设置为 1e−5，学习率会根据 CosineAnnealingLR 衰减，最小值设为 1e−5。每个训练图像都会被裁剪成原图大小的一半，然后以 0.5 的概率进行水平翻转，以增强训练数据。

3.2. 实验结果

3.2.1. 实验数据集

本文选择公开数据集 Shanghai Tech、NWPU 和 UCF-QNRF 进行实验与评估。ShanghaiTech 共有 1486 张图片，分为 Part A 和 Part B，Part A 为高密度图像，多为城市街道、大型公共场所等复杂环境，人头标注较密集；Part B 对模型场景适应能力要求较低，多为公园、街道等开放环境，人头标注相对稀疏。NWPU 共 5159 张图像，超过 218 万个标注点，场景多样，包括城市街道、商场、体育场等不同场景，并且包含不同的天气和光照条件。UCF-QNRF 数据集包含 1535 张图像，总计约 125 万人头标注，该数据集的特点是场景多样性、高密度与低密度并存、复杂背景和一些极端情况。

3.2.2. 不同人数估计模型方法在 3 个数据集的实验结果与分析

不同人数估计模型方法在 3 个数据集的实验结果如表 1 所示，数据集可视化样例如图 3 所示。表 1 展示了不同人数估计模型在 Part A、Part B、NWPU-Crowd 和 QNRF 数据集上的算法性能对比，评估指标包括 MAE (平均绝对误差)、MSE (均方误差)、参数量(M)和计算量(GFLOPS)。由实验结果可知，本文提出的模型 DACNet 在保持高准确度的同时，展现了卓越的计算效率，具体分析如下。

Table 1. Comparison of algorithm performance for different crowd estimation models on datasets
表 1. 不同人数估计模型方法在数据集中的算法性能对比

Model	Part A		Part B		NWPU-Crowd		QNRF		Params (M)	计算量 (GFLOPS)
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE		
MCNN [7]	110.2	173.2	26.4	41.3	232.5	714.6	277.0	426.0	**0.13**	11.87
CSRNet [8]	68.2	115.0	10.6	16.0	121.1	387.8	120.3	208.5	16.26	**325.3**
SANet [9]	67.0	104.5	8.4	13.6	190.6	491.4	152.6	247.0	0.91	71.45
CAN [10]	62.3	100.0	7.8	12.2	106.3	386.5	107.0	183.0	18.1	193.58
PCC-Net [11]	73.5	124.0	11.0	19.0	167.4	566.2	148.7	247.3	0.55	72.8
LMSFFNet [12]	85.85	139.9	9.2	15.1	-	-	112.8	201.6	4.58	14.9
CHFL [13]	57.5	94.3	6.9	11.0	76.8	343.0	-	-	21.51	108.27
CLTR [14]	56.9	95.2	6.5	10.6	61.9	246.3	-	-	40.95	28.73
STEERER [15]	54.5	86.9	**5.8**	**8.5**	54.3	238.3	-	-	64.42	94.4
CCTrans [16]	**52.3**	**84.9**	6.2	9.9	**38.6**	**87.8**	**82.8**	**142.3**	**103.58**	99.44
DAC Net	58.9	95.3	**7.7**	12.5	51.6	124.2	94.5	170.9	9.72	**2.56**

在 Part A 数据集上，DACNet 的 MAE 为 58.9，MSE 为 95.3，优于大部分基线模型，如 MCNN (110.2/173.2) 和 LMSFFNet (85.85/139.9)，甚至比更复杂的模型 CSRNet (68.2/115.0)表现更佳。在 Part B 数据集上，

DACNet 达到了 7.7 MAE 和 12.5 MSE，同样超越了许多方法，如 PCC-Net (11.0/19.0)和 CAN (7.8/12.2)。在 NWPU-Crowd 数据集上，DACNet 实现了 51.6 MAE 和 124.2 MSE，相比传统模型 MCNN (232.5/714.6)和 PCC-Net (167.4/566.2)，有显著提升。在 UCF-QNRF 数据集上，DACNet 实现了 94.5 MAE 和 170.9 MSE，优于绝大部分表中模型。轻量化与高效性方面，DACNet 参数量仅为 9.72 M，远低于 CSRNet (16.26 M)和 CAN (18.1 M)。计算量仅为 2.56 GFLOPS，相比 MCNN (11.87 GFLOPS)和 LMSFFNet (14.9 GFLOPS)，大幅降低了计算复杂度。综上所述，DACNet 在确保高精度的同时，实现了轻量化设计与极低的计算开销，充分证明了其在精准度与效率之间的平衡优势，非常适合应用于实时性要求高或资源受限的实际场景。

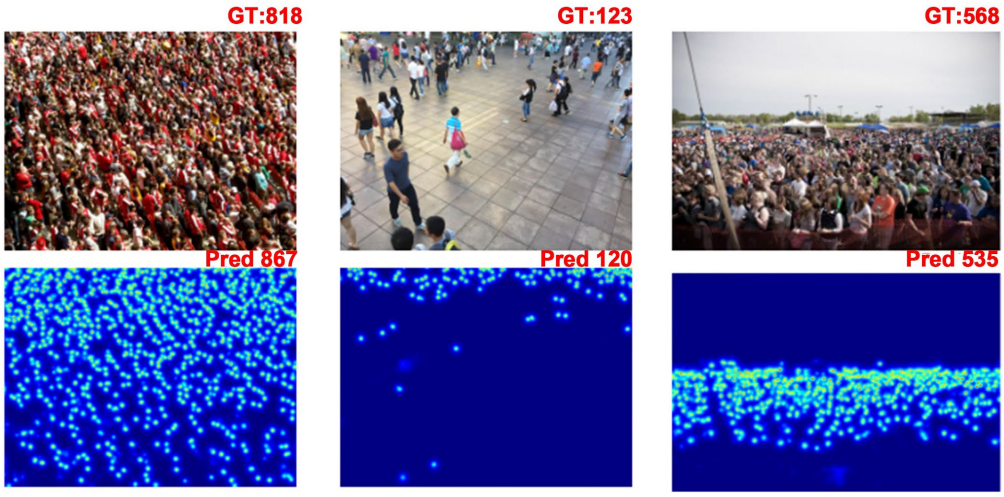

Figure 3. Dataset visualization sample
图 3. 数据集可视化样例

3.3. 消融实验

为了评估各种优化在网络设计中的有效性，本文基于 ShanghaiTech 下的 Part A 数据集进行了 5 组消融实验。实验结果如表 2 所示，引入动态卷积、残差模块和边缘注意力后，动态卷积通过自适应卷积核提高了模型的局部特征表达能力，能够更好地处理具有不同特征分布的图像；残差模块的引入使得信息能够在深层网络中更容易传播，增强了网络的表达能力，并且提升了模型在复杂任务中的表现；边缘注意力模块通过引导网络关注图像的边缘区域，增强了模型对物体轮廓的识别能力。在所有优化模块联合优化的情况下，模型表现出了最大的精度提升，MAE 进一步降低(从 64.5 降至 58.9)。

Table 2. Ablation experiment
表 2. 消融实验

动态卷积	残差模块	边缘注意力	通道注意力	空间注意力	E_{MA}	E_{RMS}
√			√	√	64.5	109.4
√	√		√	√	62.7	102.6
√		√	√	√	60.3	97.1
√	√	√	√	√	58.9	95.3

4. 总结

本文提出了一种基于 MobileNet V3 的特征融合网络(DACNet)。该方法以轻量化的 MobileNet V3 作

为主干网络，并对每个特征图分别应用通道、空间和边缘注意力机制进行加权输出。通过动态卷积在四个维度上捕获特征，有效增强了模型对复杂场景的适应能力。本文在两个不同的人群数据集上进行了相关实验，并通过对比分析验证了模型在人群计数任务中的优越性。通过消融实验，进一步验证了各个设计模块对模型性能的贡献，充分证明了所提出的模块设计在提升模型整体效果方面的合理性与有效性。

基金项目

本研究得到广东省电子信息(半导体)领域重点专项(项目编号：2022ZDZX1033)的支持。

参考文献

[1] Dollar, P., Wojek, C., Schiele, B. and Perona, P. (2012) Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**, 743-761. https://doi.org/10.1109/tpami.2011.155

[2] Dalal, N. and Triggs, B. (2005) Histograms of Oriented Gradients for Human Detection. 2005 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR*'05), San Diego, 20-25 June 2005, 886-893. https://doi.org/10.1109/cvpr.2005.177

[3] Felzenszwalb, P.F., Girshick, R.B., McAllester, D. and Ramanan, D. (2010) Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**, 1627-1645. https://doi.org/10.1109/tpami.2009.167

[4] Xu, C.Y. (2014) Research on Automated Cervical Cytological Smears Interpretation Method. Ph.D. Thesis, Chongqing University.

[5] Maitra, M., Kumar Gupta, R. and Mukherjee, M. (2012) Detection and Counting of Red Blood Cells in Blood Cell Images Using Hough Transform. *International Journal of Computer Applications*, **53**, 13-17. https://doi.org/10.5120/8505-2274

[6] Oñoro-Rubio, D. and López-Sastre, R.J. (2016) Towards Perspective-Free Object Counting with Deep Learning. *Computer Vision—ECCV* 2016 14*th European Conference*, Amsterdam, 11-14 October 2016, 615-629. https://doi.org/10.1007/978-3-319-46478-7_38

[7] Enzweiler, M. and Gavrila, D.M. (2009) Monocular Pedestrian Detection: Survey and Experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 2179-2195. https://doi.org/10.1109/tpami.2008.260

[8] Li, Y., Zhang, X. and Chen, D. (2018) CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 1091-1100. https://doi.org/10.1109/cvpr.2018.00120

[9] Cao, X., Wang, Z., Zhao, Y. and Su, F. (2018) Scale Aggregation Network for Accurate and Efficient Crowd Counting. *Computer Vision—ECCV* 2018 15*th European Conference*, Munich, 8-14 September 2018, 757-773. https://doi.org/10.1007/978-3-030-01228-1_45

[10] Liu, W., Salzmann, M. and Fua, P. (2019) Context-Aware Crowd Counting. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 5099-5108. https://doi.org/10.1109/cvpr.2019.00524

[11] Gao, J., Wang, Q. and Li, X. (2020) PCC Net: Perspective Crowd Counting via Spatial Convolutional Network. *IEEE Transactions on Circuits and Systems for Video Technology*, **30**, 3486-3498. https://doi.org/10.1109/tcsvt.2019.2919139

[12] Yi, J., Shen, Z., Chen, F., Zhao, Y., Xiao, S. and Zhou, W. (2023) A Lightweight Multiscale Feature Fusion Network for Remote Sensing Object Counting. *IEEE Transactions on Geoscience and Remote Sensing*, **61**, Article ID: 5902113. https://doi.org/10.1109/tgrs.2023.3238185

[13] Shu, W., Wan, J., Tan, K.C., Kwong, S. and Chan, A.B. (2022) Crowd Counting in the Frequency Domain. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 19618-19627.

[14] Liang, D., Xu, W. and Bai, X. (2022) An End-to-End Transformer Model for Crowd Localization. *Computer Vision—ECCV* 2022 17*th European Conference*, Tel Aviv, 23-27 October 2022, 38-54. https://doi.org/10.1007/978-3-031-19769-7_3

[15] Han, T., Bai, L., Liu, L. and Ouyang, W. (2023) Steerer: Resolving Scale Variations for Counting and Localization via Selective Inheritance Learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, 2-3 October 2023, 21848-21859.

[16] Tian, Y., Chu, X. and Wang, H. (2021) Cctrans: Simplifying and Improving Crowd Counting with Transformer. arXiv: 2109.14483.

[17] Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L., Tan, M., *et al*. (2019) Searching for MobileNetV3. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 1314-1324. https://doi.org/10.1109/iccv.2019.00140

[18] Li, C., Zhou, A. and Yao, A. (2022) Omni-Dimensional Dynamic Convolution. arXiv: 2209.07947.

[19] Hu, J., Shen, L., Albanie, S., Sun, G. and Wu, E. (2020) Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 2011-2023. https://doi.org/10.1109/tpami.2019.2913372

[20] Wang, X., Girshick, R., Gupta, A. and He, K. (2018) Non-Local Neural Networks. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7794-7803. https://doi.org/10.1109/cvpr.2018.00813

[21] Chen, L., Zhang, X. and Yang, Y. (2019) Edge Attention for Visual Question Answering. arXiv: 1911.12294. https://doi.org/10.48550/arXiv.1911.12294

[22] Wang, B., Liu, H., Samaras, D. and Nguyen, M.H. (2020) Distribution Matching for Crowd Counting. *Advances in Neural Information Processing Systems*, Vol. 33, 1595-1607.