Journal of Image and Signal Processing 图像与信号处理, 2024, 13(2), 130-137
Published Online April 2024 in Hans. https://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2024.132012

Hans汉斯

基于双分支自注意力的密集人群计数算法

钟德军，丁　健，易　云

赣南师范大学数学与计算机科学院，江西　赣州

收稿日期：2024年2月27日；录用日期：2024年3月27日；发布日期：2024年4月9日

摘　要

及时、准确的进行人流监控及预警是公共安全管理的迫切需求，使用基于计算机视觉的人群计数方法是满足该需求的主要方法之一。针对现有计数模型对人员前景特征和背景特征的关联不够的问题，设计基于双分支自注意力机制的密集人群计数算法。在视觉主干网络之后使用双分支自注意力模块，以促使网络关注有效的人员区域，提升主干网络的特征精炼能力。在Shanghai Tech PART B和UCF-QNRF数据集上进行大量的实验，消融实验的结果证明所提出的模块提升了人群计数的准确性。此外，实验结果表明所提出方法获得比其他经典方法更好的实验结果。

关键词

人群计数，公共安全管理，双分支自注意力，特征精炼

Dense Crowd Counting Algorithm Based on Dual-Branch Self-Attention

Dejun Zhong, Jian Ding, Yun Yi

College of Mathematics and Computer Science, Gannan Normal University, Ganzhou Jiangxi

Received: Feb. 27th, 2024; accepted: Mar. 27th, 2024; published: Apr. 9th, 2024

Abstract

The urgent need for public safety management is timely and accurate crowd monitoring and early warning. The use of crowd counting methods based on computer vision is one of the main methods to meet this need. To tackle the problem that existing counting models do not adequately correlate people's foreground features and background features, a dense crowd counting algorithm based on a dual-branch self-attention mechanism is designed. A dual-branch self-attention module is used after the visual backbone network to prompt the network to focus on effective person areas and improve the feature refining capabilities of the backbone network. A large number of experi-

文章引用: 钟德军, 丁健, 易云. 基于双分支自注意力的密集人群计数算法[J]. 图像与信号处理, 2024, 13(2): 130-137.
DOI: 10.12677/jisp.2024.132012

ments were conducted on Shanghai Tech PART B and UCF-QNRF data sets, and the results of ablation experiments proved that the proposed modules improved the accuracy of crowd counting. Furthermore, experimental results show that the proposed method obtains better experimental results than other classical methods.

Keywords

Crowd Counting, Public Safety Management, Dual-Branch Self-Attention, Feature Refining

1. 引言

人群计数在公共安全管理中发挥着重要作用，特别是在音乐会、体育赛事和庆祝活动等人群密集的场景中。如果没有恰当的管理措施，踩踏事件就很容易发生，避免踩踏事件发生的重要抓手就是及时、准确的进行人流监控及预警。由于视角遮挡和人员分布散乱，人群计数是一项具有挑战性的任务。为了解决这些问题，人们做了许多研究。其中，基于 CNN 的回归方法以人群图像为输入，生成密度图，进一步累积得到人数。同一图像中的头部大小可能会有很大差异，这对 CNN 提取尺度不变特征造成了影响。许多方法都致力于解决这个问题，包括多列网络、规模聚合模块和规模不变性体系结构等。典型的多列融合方法包括多列融合[1]和深浅网络融合[2]。典型的规模聚合模块[3] [4]按不同的核大小聚合规模不变特征。典型的规模不变性架构[5] [6]侧重于单列架构的设计。

本文基于注意力机制[7]，利用语义特征和位置特征之间的关系进行建模，在主干网络的输出后使用双分支自注意力模块。该模块通过注意力计算机制细化行人特征，提取对计数有用的特征而抑制无关特征，有效减少了网络输出中显示的误差响应。实验表明，该模块可以提高准确性和鲁棒性。在 Shanghai Tech PART B 和 UCF-QNRF 两个人群数据集上评估了我们的方法。结果表明，我们的方法获得了比其他经典方法更好的性能。

2. 相关工作

2.1. 注意力机制

注意力机制[7]-[14]在图像分类、目标检测领域得到了广泛的应用,它促使模型在图像中动态地分配注意力，从而更好地捕捉图像不同区域之间的全局关联性，有效增强了模型的性能。自注意力[7]关注序列中不同位置之间的关系，促使模型关注那些重要的图像区域，在密集人群计数任务中，行人与背景之间的位置关系是重要的，通过注意力机制可以细化行人信息，抑制背景信息对计数任务的干扰，以获取重要的行人特征信息，生成更准确的特征图。在深度神经网络中，不同特征映射中的不同通道通常代表不同的对象[8]。文献[9]首先提出了通道注意力的概念，并提出了 SENet [9]。通道注意力[9] [10]可以自适应地重新校准每个通道的权重，相当于一个对象选择过程，最终选择重要的通道信息。空间注意力[11] [12]是一种适应性的空间区域选择机制，可以生成跨空间域的注意力掩码，并使用它来选择重要的空间区域或直接预测最相关的空间位置。分支注意力[13] [14]是一种动态的分支选择机制，可以生成跨不同分支的注意力掩码，并使用它来选择重要的分支。

2.2. 基于 CNN 的人群计数方法

文献[1]中提出了第一个多列卷积网络(MCNN)。MCNN 是一个三层 CNN 架构，每列具有不同的感受野，MCNN 中的每一列都会生成与真实密度图形状相同的密度图，然后将各个列的输出连接起来以生成最终的密度图。文献[2]使用浅层卷积和深度卷积架构的组合来解决人群图像中的尺度变化问题。这种组合可以有效地捕获高级语义信息和低级特征，在检测大规模变化和严重遮挡的场景下的是相当可靠的。在文献[3]中，作者提出了一种新颖的编码器-解码器网络，称为规模聚合网络(SANet)。它建立在 Inception 架构的基础上，编码器通过尺度聚合模块提取多尺度特征，解码器通过使用一组转置卷积生成高分辨率密度图。文献[4]提出了一种多尺度卷积神经网络(Multi-Scale ConvolutionalNeural Network, MSCNN)，该网络基于多尺度斑点来提取与尺度相关的特征，在单列架构下获得了更好的计数性能。在文献[5]中，作者使用膨胀卷积的思路设计了一个基于膨胀卷积的人群计数网络，适用于高度拥堵的场景。该网络通过扩张的卷积层来扩展感受野，扩展了 CNN 作为后端，并且由于其纯卷积结构而易于训练。文献[6]提出了一种规模感知注意力网络来解决图像尺度变化的问题，该网络基于注意力机制，可以自动聚焦于适合图像的某些全局和局部尺度。文献[15]提出了一种端到端的级联 CNN 网络来共同学习人群计数分类和预测密度图。该网络将高级先验纳入密度估计网络，使得网络中的层能够学习全局相关的判别特征，有利于网络以更低的计数误差预测出高度精细的密度图。在文献[16]中，作者提出切换卷积神经网络，将不同的人群场景映射到它对应的密度区域。该网络将来自人群场景中的网格的 patch 传递给多个独立的 CNN 回归器，不同的 CNN 回归器具有不同的接受域，分类器将不同人群场景的 patch 传递给最佳的 CNN 回归器。文献[17]提出 TransCrowd 算法，利用 transformer 的自注意力机制，有效地提取人群的语义信息。文献[18]引入空间、通道注意力模型来估计密度图，前者对图像逐像素进行编码以提升预密度图准确性，后者在不同通道之间提取更多的判别特征以促进模型关注人群场景的核心区域。在文献[19]中，作者提出了一个端到端的人群计数框架，即检测和密度估计网络。该网络可以根据图像的实际密度情况，自适应地选择图像上不同位置的计数模式。文献[20]结合多尺度卷积神经网络(生成器)和对抗网络(鉴别器)来生成高质量的密度图，可以准确估计复杂人群场景中的人群数量。

3. 网络概述

本节将详细阐述本文的网络模型，它主要由主干网络、双分支自注意力模块和回归卷积层组成。如图 1 展示了该框架的结构图。对于每个图像 I，首先使用 VGG-16 [21]中的前 10 个卷积层作为主干网络来提取特征，其中 c、h 和 w 分别是通道、高度和宽度，Conv-128 代表输出通道数为 128 的 Conv2d 函数。

主干网络进行特征提取后，使用双分支自注意力模块，通过注意力机制细化行人特征，促使模型捕捉密集人群中的重要人体信息，提取对计数有用的特征而抑制无关特征，以生成更准确的特征图，提升密度图的准确性，增强模型的鲁棒性。最后，模型使用回归卷积层进行解码，接着使用大小为 8 的缩放因子进行上采样，得到与输入图片尺寸大小相同的密度特征图。

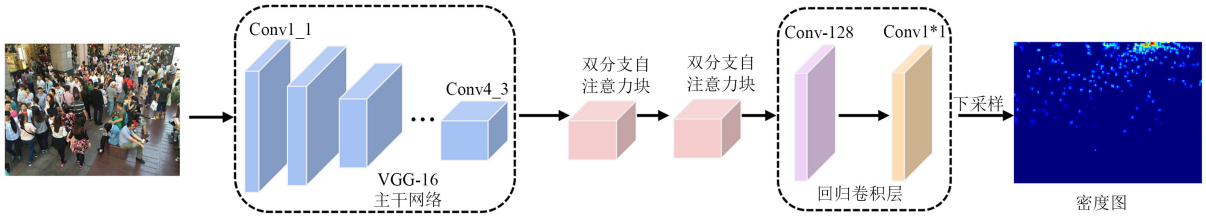

Figure 1. Model structure diagram
图 1. 模型结构图

3.1. 主干网络

主干网络用于从图像中提取基本特征，以供后续网络模块使用。在人群计数任务中，通常会使用预训练过的卷积神经网络作为骨干网络，并通过迁移学习来微调其参数，以更好地适应特定任务。以图 2 中展示的主干网络为例，其输入是原始的人群图像，经过一系列卷积核大小为 3 × 3，通道数分别为 64、128、256 和 512 的卷积层处理，逐步提取图像的特征信息。最终，这些特征汇聚到一个输出层，输出层的图像尺寸为原始图像尺寸的 1/64。这里将 VGG16 [21]的前 10 个卷积层作为骨干网络，并获取在 ILSVRC2012 数据集上预训练得到的权重作为骨干网络卷积层的权重。

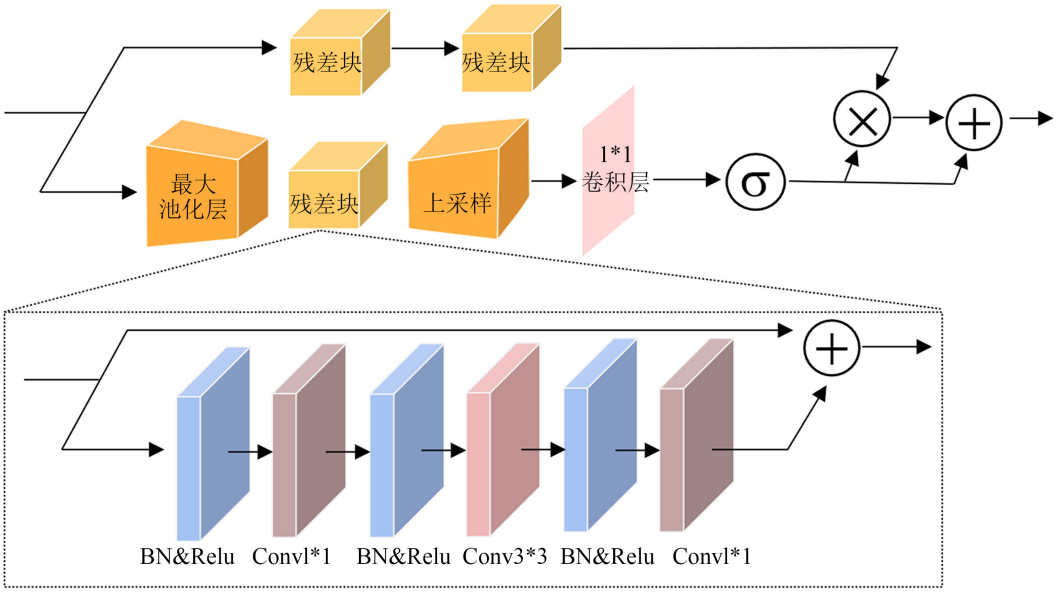

Figure 2. Structure diagram of dual-branch self-attention block
图 2. 双分支自注意力块结构图

3.2. 双分支自注意力模块

为了建模人员前景特征和背景特征的关联，引入自注意力特征关联，设计双分支注意力模块，如图 2。注意力机制可以对语义特征和位置特征之间的关系进行建模，本文在主干网络的输出后添加两层自注意力模块。自注意力机制的作用机理是在输入特征之间建议一对一的关联，减少对其它外部信息的依赖。通过重新分配权重来实现对自身重要输出特征的关注，从而实现提炼内部特征之间模式相关性的功能。自注意力模块被用来提炼主干网络输出的特征，通过建模前景和背景特征之间的关联，使得算法更加关注有助于计数的特征，忽略与计数任务无关的背景特征，从而增加模型的通用性。

每个双分支自注意力模块由两个子路径组成，其中一个子路径由两个残差块串接构成，输出多维特征图 F_1；另一个子路径由下采样层、残差块、上采样层、1 × 1 卷积层和 Sigmoid 激活层构成。上采样层的输出为 F_2，则输出的自注意力权重 A 可表示为 $A=\text{Sigmoid}\left(\text{Conv1}\left(F_2\right)\right)$。两个子路径中用到的残差块均采用在 ResNet [22]中提出的残差结构。采用的激活层起到注意力筛选的作用，为另一个子路径输出的特征分配权重。在注意力模块的最后，两个子路径的输出被逐元素相乘，然后累加原特征 F_1 作为输出。每个注意力块输出的特征图 F_{out} 可表示为 $F_{out}=A\times F_1+F_1$。

双分支自注意力块有效促进了模型提取有用的特征，其跳跃连接内的残差单元可以加快训练阶段的收敛速度。通过实验发现叠加使用两个双分支自注意力模块可以获得更好的特征关联效果，且性能最佳。

4. 实验及结果分析

4.1. 实验数据集

Shanghai Tech Part B [1]数据集由 716 张图像组成(400 张图像用于训练，316 张图像用于测试)，它的注释包含人头位置信息和人数，是在上海市区繁忙的街道上拍摄的。

UCF-QNRF [23]数据集包含 1535 张具有挑战性的图像(1201 张图像用于训练，334 张图像用于测试)，共有 1,251,642 个注释，是目前此类数据集中注释最全面的数据集。图像中对象的最小和最大数量分别为 49 和 12,865。这个大规模数据集涵盖了不同的位置、视角、透视效果和一天中的不同时间。

4.2. 评估指标

本文使用平均绝对误差(MAE)和均方根误差(RMSE)作为评估计数方法的指标，其定义如下：

$$\mathrm{MAE}=\frac{1}{M}\sum_{i=1}^{M}\left|N_i^{gt}-N_i\right| \tag{1}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(N_i^{gt}-N_i\right)^2} \tag{2}$$

其中 M 是样本图像的数量。N_i^{gt} 和 N_i 分别是第 i^{th} 图像的真实值和估计数。MAE 更多地衡量方法的准确性，RMSE 更多地衡量方法的稳健性。两者越低代表性能越好。

4.3. 实验设置

本文实验训练环境为 Ubuntu18.04，使用 Pytorch 和 C^3 框架[24] [25]作为训练学习的基础框架。模型采用 ImageNet 预训练的 VGG16 的前 10 层卷积网络作为主干网络，输入图像的像素为 224 × 224。训练时采用权重衰减为 1×10^{-4} 的 Adam 优化器来优化模型参数，初始学习率为 1×10^{-5}，学习率的衰减率为 0.995。

4.4. 与其他方法的比较

实验对比了本文算法与其他主流人群计数模型的性能，包括 MCNN [1]、CMTL [15]、Switch-CNN [16]、TransCrowd [17]、SCAR [18]、DecideNet [19]、MS-GAN [20]。表 1 和表 2 展示了各类模型在两个数据集上的实验结果。从表 1、表 2 中可以看出，使用双分支自注意力模块在两个数据集上均取得了较好的效果。

Table 1. Comparison with other methods on the Shanghai Tech PART B dataset
表 1. 在 Shanghai Tech PART B 数据集上与其他方法的对比

Methods	MAE	RMSE
CMTL [15]	20.0	31.1
Switch-CNN [16]	21.6	33.4
TransCrowd [17]	9.3	16.1
SCAR [18]	9.5	15.2
DecideNet [19]	21.5	31.9
MS-GAN [20]	18.7	30.5
Ours	8.98	14.47

Table 2. Comparison with other methods on the UCF-QNRF dataset
表 2. 在 UCF-QNRF 数据集上与其他方法的对比

Methods	MAE	RMSE
MCNN [1]	277.0	426.0
CMTL [15]	252.0	514.0
Switch-CNN [16]	228.0	445.0
Ours	144.41	250.55

如图 3 为本文算法的可视化预测效果，图片来自两个数据集上的测试集。从预测图可知，网络在两个数据集上都实现了较高的预测精度，其预测密度图与真实密度图的分布相似、计数精度较高。

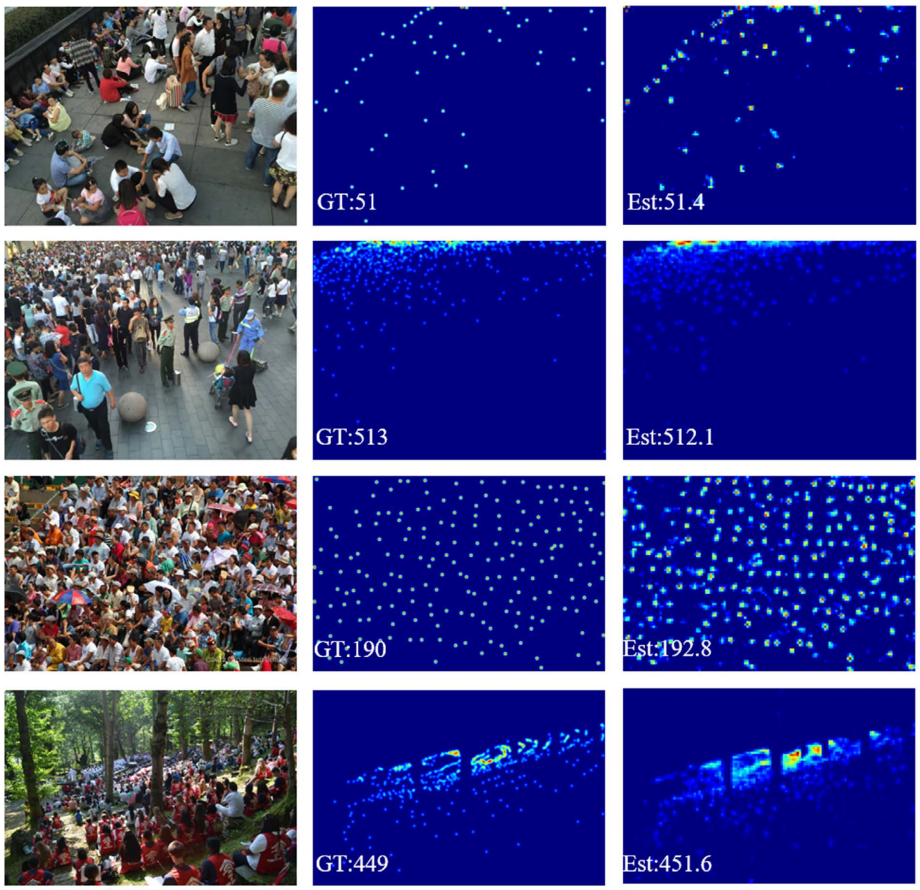

Figure 3. Partial images and density maps of the test set of the two datasets
图 3. 两个数据集的测试集部分图像及密度图

从图 3 中可以看到，对于人群密集和相对稀疏的场景，该模型均能较为准确地预测。因此，可以说明本文使用双分支自注意力模块能够有效细化行人特征，将人与背景区分开来，具有较高的实用价值。

4.5. 消融实验

为了深入验证双分支自注意力模块的有效性，设计了 2 组对比实验，2 组实验均采用 MSE 损失函数进行训练。

其中，Baseline 是基于 VGG-16 网络构建的密集人群计数网络。表 3 的结果显示，该模块比 Baseline 的 MAE 值降低了 19.3%。表 4 的结果显示，该模块比 Baseline 的 MAE 值降低了 15.9%。可以发现，在两个数据集上，该算法表现稳定，比较 RMSE 值，可以发现该算法鲁棒性较强。这表明提出的双分支自注意力模块可以有效帮助网络提取更有用的特征信息，提升模型性能。

Table 3. Ablation experimental results of dual-branch self-attention module on UCF-QNRF dataset
表 3. 双分支自注意力模块在 UCF-QNRF 数据集上的消融实验结果

Methods	MAE	RMSE
Baseline	178.96	303.81
Ours	144.41	250.55

Table 4. Ablation experimental results of dual-branch self-attention module on Shanghai Tech PART B dataset
表 4. 双分支自注意力模块在 Shanghai Tech PART B 数据集上的消融实验结果

Methods	MAE	RMSE
Baseline	10.7	16.5
Ours	8.98	14.47

5. 结束语

为了有效解决现有计数模型对人员前景特征和背景特征的关联不够的问题，本文提出基于双分支自注意力的密集人群计数算法。该网络模型使用双分支自注意力模块，促使网络关注有效的人员区域，提升了网络的特征精炼能力，增强了模型的鲁棒性。在两个公开数据集上进行了大量实验，消融实验的结果证明所提出的模块提升了人群计数的准确性。此外，实验结果表明所提出方法获得比其他经典方法更好的实验结果。

基金项目

国家自然科学基金(No. 62362003)，江西省自然科学基金(No. 20232BAB202017)。

参考文献

[1] Zhang, Y., Zhou, D., Chen, S., *et al.* (2016) Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 589-597. https://doi.org/10.1109/CVPR.2016.70

[2] Boominathan, L., Kruthiventi, S.S.S. and Babu, R.V. (2016) CrowdNet: A Deep Convolutional Network for Dense Crowd Counting. *Proceedings of the* 24*th ACM International Conference on Multimedia*, 640-644. https://doi.org/10.1145/2964284.2967300

[3] Cao, X., Wang, Z., Zhao, Y., *et al.* (2018) Scale Aggregation Network for Accurate and Efficient Crowd Counting. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018, *Lecture Notes in Computer Science*, Vol. 11209, Springer, Cham, 734-750. https://doi.org/10.1007/978-3-030-01228-1_45

[4] Zeng, L., Xu, X., Cai, B., *et al.* (2017) Multi-Scale Convolutional Neural Networks for Crowd Counting. 2017 *IEEE International Conference on Image Processing*, Beijing, 17-20 September 2017, 465-469. https://doi.org/10.1109/ICIP.2017.8296324

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

[6] Hossain, M., Hosseinzadeh, M., Chanda, O., *et al.* (2019) Crowd Counting Using Scale-Aware Attention Networks.

2019 *IEEE Winter Conference on Applications of Computer Vision*, Waikoloa, 7-11 January 2019, 1280-1288. https://doi.org/10.1109/WACV.2019.00141

[7] Vaswani, A., Shazeer, N., Parmar, N., *et al.* (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems*, **30**.

[8] Chen, L., Zhang, H., Xiao, J., *et al.* (2017) SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 21-26 July 2017, 5659-5667. https://doi.org/10.1109/CVPR.2017.667

[9] Hu, J., Shen, L. and Sun, G. (2018) Squeeze-and-Excitation Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018,7132-7141. https://doi.org/10.1109/CVPR.2018.00745

[10] Wang, Q., Wu, B., Zhu, P., *et al.* (2020) ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 11534-11542. https://doi.org/10.1109/CVPR42600.2020.01155

[11] Wang, X., Girshick, R., Gupta, A., *et al.* (2018) Non-Local Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7794-7803. https://doi.org/10.1109/CVPR.2018.00813

[12] Dai, J., Qi, H., Xiong, Y., *et al.* (2017) Deformable Convolutional Networks. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 764-773. https://doi.org/10.1109/ICCV.2017.89

[13] Srivastava, R.K., Greff, K. and Schmidhuber, J. (2015) Training Very Deep Networks. *Advances in Neural Information Processing Systems*, **28**.

[14] Chen, Y., Dai, X., Liu, M., *et al.* (2020) Dynamic Convolution: Attention over Convolution Kernels. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 11030-11039. https://doi.org/10.1109/CVPR42600.2020.01104

[15] Sindagi, V.A. and Patel, V.M. (2017) CNN-Based Cascaded Multi-Task Learning of High-Level Prior and Density Estimation for Crowd Counting. *IEEE International Conference on Advanced Video and Signal Based Surveillance*, Lecce, 29 August 2017, 1-6. https://doi.org/10.1109/AVSS.2017.8078491

[16] Sam, D.B., Surya, S. and Babu, R.V. (2017) Switching Convolutional Neural Network for Crowd Counting. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 5744-5752. https://doi.org/10.1109/CVPR.2017.429

[17] Liang, D., Chen, X., Xu, W., *et al.* (2022) Transcrowd: Weakly-Supervised Crowd Counting with Transformers. *Science China Information Sciences*, **65**, Article No. 160104. https://doi.org/10.1007/s11432-021-3445-y

[18] Gao, J., Wang, Q. and Yuan, Y. (2019) SCAR: Spatial-/Channel-Wise Attention Regression Networks for Crowd Counting. *Neurocomputing*, **363**, 1-8. https://doi.org/10.1016/j.neucom.2019.08.018

[19] Liu, J., Gao, C., Meng, D., *et al.* (2018) Decidenet: Counting Varying Density Crowds through Attention Guided Detection and Density Estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 5197-5206. https://doi.org/10.1109/CVPR.2018.00545

[20] Zhou, Y., Yang, J., Li, H., *et al.* (2020) Adversarial Learning for Multiscale Crowd Counting under Complex Scenes. *IEEE Transactions on Cybernetics*, **51**, 5423-5432. https://doi.org/10.1109/TCYB.2019.2956091

[21] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition.

[22] He, K., Zhang, X., Ren, S., *et al.* (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[23] Idrees, H., Tayyab, M., Athrey, K., *et al.* (2018) Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018, *Lecture Notes in Computer Science*, Vol. 11206, Springer, Cham, 532-546. https://doi.org/10.1007/978-3-030-01216-8_33

[24] Wang, Q., Gao, J., Lin, W., *et al.* (2019) Learning from Synthetic Data for Crowd Counting in the Wild. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 15-20 June 2019, 8198-8207. https://doi.org/10.1109/CVPR.2019.00839

[25] Gao, J., Lin, W., Zhao, B., *et al.* (2019) C^ 3 Framework: An Open-Source Pytorch Code for Crowd Counting. https://arxiv.org/abs/1907.02724