A Channel-aware Attention Network for Crowd Counting

Wenzhe Zhai
School of Electrical and Electronic Engineering
Shandong University of Technology
Zibo, China
wenzhezhai@163.com

Jinfeng Pan*
School of Electrical and Electronic Engineering
Shandong University of Technology
Zibo, China
pjfbysj@163.com

Qilei Li
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, United Kingdom
q.li@qmul.ac.uk

Guofeng Zou
School of Electrical and Electronic Engineering
Shandong University of Technology
Zibo, China
zgf841122@163.com

Liju Yin
School of Electrical and Electronic Engineering
Shandong University of Technology
Zibo, China
ljyin72@163.com

Mingliang Gao
School of Electrical and Electronic Engineering
Shandong University of Technology
Zibo, China
mlgao@sdut.edu.cn

2021 China Automation Congress (CAC) | 978-1-6654-2647-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/CAC53003.2021.9728649

***Abstract*— With the rapid increase of urban population, crowd counting is a popular yet difficult topic. However, the problem of scale variation in high-density scenario remains under-explored. To address this problem, we propose a channel-aware attention network in this paper. The channel attention module attempts to handle the relations between channel maps and highlight the discriminative information in specific channels. Thus, it alleviates the misestimation for background regions. Experimental results on ShanghaiTech and UCF-QNRF benchmark datasets prove that our approach achieves compelling performance compared to the state-of-the-art methods.**

***Index Terms*—Crowd counting, Density estimation, Channel attention, Convolutional neural network**

I. Introduction

The purpose of crowd counting is to estimate the population in crowd scenes. It has drawn much attention in the last few years because of its crucial role in video-based surveillance and public safety [1, 2, 3]. Crowd counting is inherently challenging due to many severe challenges, *e.g.*, perspective distortion, extreme scale variations, and non-uniform distribution.

Current methods on crowd counting algorithms can be generally categorized into detection-based methods and regression-based methods [3]. The detection-based methods identify pedestrians by detecting the body and head region of the crowd. These methods can not accurately count pedestrians in highly dense crowd scene due to the poor detection performance in such scenario. The regression-based methods devote to training regression models to directly learn a mapping from the visual features. In recent years, benefiting from the powerful learning ability of deep convolutional neural networks (CNNs), the CNN-based methods have achieved dominant performances [1, 2, 3].

In spite of the great achievements, crowd counting in high-density scenario is still a thorny issue. The main challenge in crowd counting is the scale variations caused by the camera perspective distortion, as depicted in Fig. 1. The large scale variation will decrease the quality of estimated density maps and result in the error estimation for backgrounds. To settle this problem, we put forward a channel-aware attention network.

The rest of this paper is summarized as follows. Section II reviews previous related works in this domain. Section III depicts the detailed explanation of the proposed methods. Section IV introduces the detail of experimental results and discussions. This work is concluded in Section V.

II. Related work

A. Detection-based methods

Incipiently methods count the quantity of individuals by detecting heads or bodies within the pictures. Features such as Haar wavelets [4] and HOG [5] concentrated on exploiting handcrafts features which detected head or other body parts. Dollar *et al.* [6] employed a sliding window to detect a

This work is partially supported by the National Natural Science Foundation of China (No. 61801272) and the SDUT, the National Natural Science Foundation of Shandong (No. ZR2020MF127), and Zibo District Integration Project(No. 2019ZBXC516).

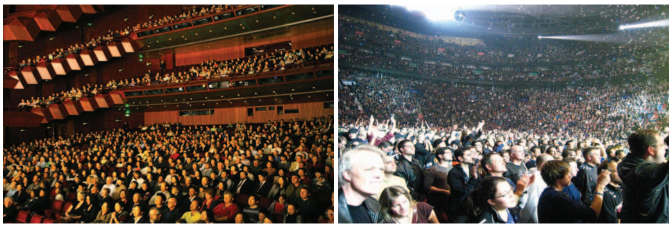

Fig. 1: Scale variations in crowd scenes.

person and counted the people by the detected bounding boxes. Li *et al.* [7] calculated the crowd density by constructing detectors of head and shoulder. Despite the progress achieved in low-density crowd scenarios, the detection-based methods underperform in crowded scenarios because of the occlusion and background noise [8].

B. *Regression-based methods*

The regression-based approaches map the visual features to the amount of people in crowds and have achieved a great success in crowd counting [3]. Idress *et al.* [9] utilized multiple sources to regress the crowd counts. Zhang *et al.* [10] adopted the trained CNN architecture to solve the cross-scence counting problems. Zhang *et al.* [11] employed MCNN model to cope with the problem of perspective distortions and scale variation. Sindagi *et al.* [12] put forward a CNN-based to classify the crowd density of pedestrian into diverse levels to promote the density estimation. Li *et al.* [13] proposed CSRNet by replacing the pooling operations with dilated kernels to aggregate the multiscale information in congested scenes. Li *et al.* [13] proposed CSRNet by utilizing dilated convolution to aggregate the multiscale information in crowd scenes.

Additionally, the attention mechanisms have been adopted in crowd counting. Liu *et al.* [14] used attention module to coordinate the model weights of crowd density. Hossain *et al.* [15] achieved the crowd counting by a global attentional network with local scale perception. Zhang *et al.* [16] presented an attention mechanism to evaluate the probability map for crowd counting in non-head area.

Despite the great advances in crowd counting, the problem of scale variations in high-density scenarios is far from settled. We demonstrate the effectiveness of introducing attention mechanisms into crowd counting. To this aim, we employed a channel-aware attention network to effectively cope with the scale variations in high-density scenario.

III. Proposed Method

The framework of the proposed model is depicted in Fig. 2. In this section, the attention module, loss function, and the implementation details are introduced.

A. *Channel attention architecture*

Given an input image I, ResNet-50 is employed to extract the feature. Considering the scale variation caused by perspective distortion in dense crowd, it is vital to introduce the attention model to restrain the side effects caused by the perturbed pattern.

As shown in Fig. 3, our model aggregates the spatial information by employing the average-pooling and max-pooling to generate two spatial context descriptors, namely M_a and M_m. Both descriptors are feed into a shared network and generate a channel attention map $D(M) \in \mathbb{R}^{1\times1\times C}$. The channel attention is formulated as,

$$\begin{aligned} D(M) =& \text{Sigmoid}[\text{FC}(\text{AvgPool}(M)) + \text{FC}(\text{MaxPool}(M))] \\ =& \text{Sigmoid}[(K \cdot M_{\text{a}}) + (K \cdot M_{\text{m}})], \end{aligned} \tag{1}$$

where $\text{FC}(\cdot)$ denotes convolution operations. The channel weight K is generated by $\text{FC}(\cdot)$ manipulation.

The proposed model produces a $1 \times 1 \times C$ channel attention map $D(M)$ to reflect the vital region of crowd. S_{o} employs the channel enhanced feature map,

$$S_{\text{o}} = D(M) \otimes M, \tag{2}$$

where $\otimes$ represents the element-wise multiplication.

B. *Loss function*

The Euclidean distance is employed as the optimization objective loss function as,

$$\text{loss} = \frac{1}{M}\sum_{i=1}^{M} \|F_\theta(I_i) - Y_i\|_2^2 \,, \tag{3}$$

where M is the batch size. $F_\theta(I_i)$ indicates the estimated density map, θ denotes the learned parameter, and Y_i is the density map of ground truth.

C. *Implementation details*

1) Ground truth of the density map: Similar to the work [13], the generation of density map $H(z)$ is formulated as,

$$\text{H}(z) = \sum_{i=1}^{N} \delta(z - z_i) * \text{G}_\sigma(z), \tag{4}$$

where N denotes the labelled person, and z_i is the annotated head location. The delta function $\delta(z - z_i)$ employs a head of pedestrian. G_σ represents Gaussian kernel with a parameter σ.

2) Training details: In these experiments, two parallel NVIDIA RTX2080S GPUs are used for training and evaluating using the PyTorch framework [17]. The default batch size is set as 4 on each GPU. We resize all images to 576x768 resolution, and generate density maps in the same size. The decay rate is 0.995. The learning rate of the whole network is initially set at 10^{-5}. The Adam [18] is adopted as the optimizer and models are trained for 400 epochs.

IV. Experiments

A. *Evaluation metrics*

We employ evaluation metrics via the Mean Absolute Error (MAE) and Mean Square Error (MSE), which are computed in Eq. (5) and Eq. (6), respectively.

$$\text{MAE} = \frac{1}{N}\sum \left| y_i - y_i^{'} \right| , \tag{5}$$

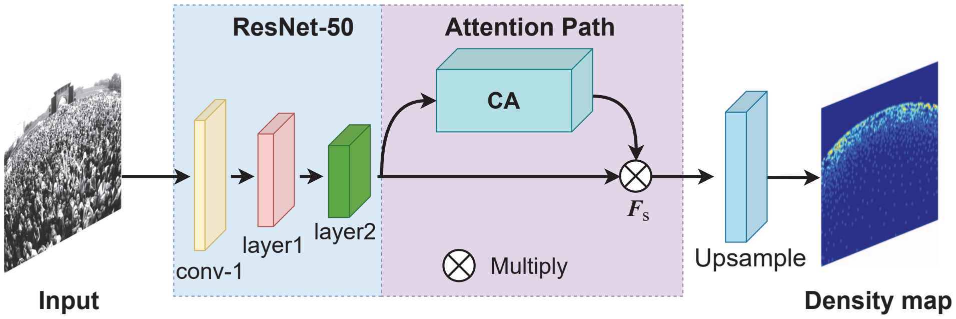

Fig. 2: Flowchart of the channel-aware attention network for crowd counting. The proposed method consists of three parts, *i.e.*, feature extractor, channel-aware attention model, and density map generator.

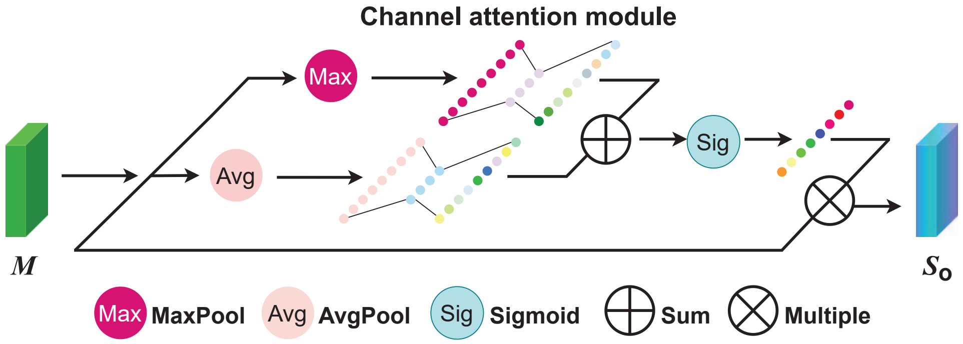

Fig. 3: Structure of the channel attention module. It takes the high level feature map M as input. The average-pooling and max-pooling are performed on two paralleled paths to squeeze a 1-dimension vector. Then the channel weight is generated by sigmoid operation. The final feature map is produced by the multiplication of weight and input M.

TABLE I: Experimental results on the ShanghaiTech dataset

Method	Part_A		Part_B	
	MAE	MSE	MAE	MSE
Zhang et al.[10]	181.8	277.7	32.0	49.8
MCNN[11]	110.2	173.2	26.4	41.3
Marsden et al.[19]	126.5	173.5	23.8	33.1
Switching-CNN[20]	90.4	135.0	21.1	30.1
CMTL[21]	101.3	152.4	20.0	31.1
ACSCP[22]	75.7	102.7	17.2	27.4
BSAD[23]	90.4	135.0	20.2	35.6
SaCNN[24]	86.8	139.2	20.7	32.8
PCC-Net[25]	73.5	124.0	19.2	31.5
DNCL[26]	**73.5**	**112.3**	18.7	26.0
Ours	74.3	127.1	**8.4**	**14.0**

$$\mathrm{MSE} = \sqrt{\frac{1}{N}\sum \left|y_i - y_i^{'}\right|^2}, \tag{6}$$

where N represents the number of test datasets, i is the number of test image. $y_i^{'}$ indicates the estimated counting results, and y_i is the groud truth number of crowd.

B. Performance on ShanghaiTech dataset

The ShanghaiTech dataset [11] consists of two parts, among which Part_A contains 482 images and Part_B has 716 images, respectively. The results are shown in Table I. Some examples are depicted in Fig. 4. It shows that the prediction results are close to the ground truth.

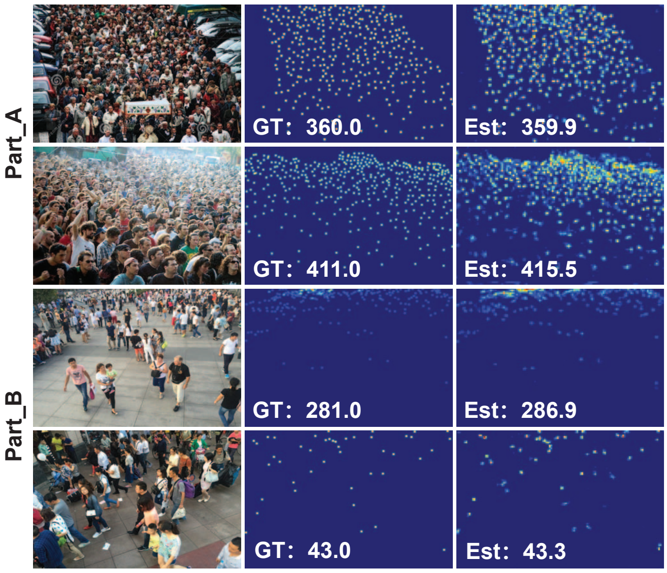

Fig. 4: The estimated density maps and counting numbers on ShanghaiTech dataset.

C. Performance on UCF-QNRF dataset

The UCF-QNRF dataset [27] contains 1535 images with huge density variance, which is extremely challenging. The training set and the test set include 1201 and 334 images,

TABLE II: Experimental results on the UCF-QNRF dataset

Methods	MAE	MSE
Zhang *et al.* [10]	467.0	498.5
MCNN[11]	277.0	509.1
CRSNet[13]	129.0	209.0
CMTL[21]	252.0	514.0
Switching-CNN[20]	228.0	445.0
PCCNet[25]	148.7	247.3
DENet[28]	121.0	205.0
LSC-CNN[29]	120.5	218.2
HA-CCN[30]	118.1	**180.4**
Ours	**113.9**	197.1

respectively. Comparative results are shown in Table II. It shows that our method has the lowest score of MAE (113.7) and second-lowest score of MSE (197.1). Fig. 5 illustrates the experimental results for sample images from the UCF-QNRF datasets.

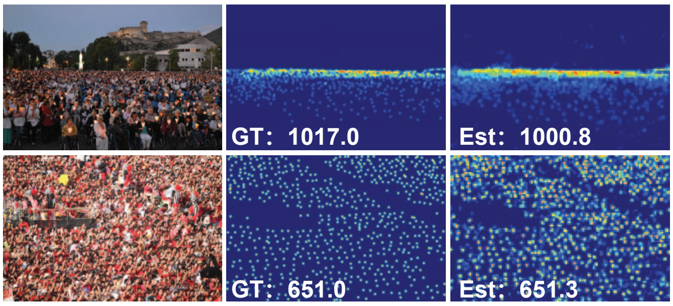

Fig. 5: The estimated density maps and counting numbers on UCF-QNRF dataset.

V. Conclusion

In this paper, we propose a channel-aware attention network to handle the crowd counting in dense crowd scene. The proposed method handles the relations between channel maps and highlights the discriminative information in specific channels. Thus, it alleviates the mistaken estimation for background regions. The experiments verify that the proposed approach accomplishes competitive performance compared with other SOTA trackers.

References

[1] Haoyue Bai and S. Chan. Cnn-based single image crowd counting: Network design, loss function and supervisory signal. *ArXiv*, abs/2012.15685, 2020.

[2] Di Kang, Z. Ma, and Antoni B. Chan. Beyond counting: Comparisons of density maps for crowd analysis tasks—counting, detection, and tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 29:1408–1422, 2019.

[3] V. Sindagi and V. Patel. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters*, 107:3–16, 2018.

[4] Paul A. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2004.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR),*, pages 886–893, 2005.

[6] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:743–761, 2012.

[7] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR),*, pages 1–4, 2008.

[8] Xinyue Chen, Hua Yan, Tong Li, Jialang Xu, and Fushun Zhu. Adversarial scale-adaptive neural network for crowd counting. *Neurocomputing*, 450:14–24, 2021.

[9] H. Idrees, Imran Saleemi, C. Seibert, and M. Shah. Multi-source multi-scale counting in extremely dense crowd images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR),*, pages 2547–2554, 2013.

[10] Cong Zhang, Hongsheng Li, X. Wang, and X. Yang. Cross-scene crowd counting via deep convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR),*, pages 833–841, 2015.

[11] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR),*, pages 589–597, 2016.

[12] V. Sindagi and V. Patel. Generating high-quality crowd density maps using contextual pyramid cnns. *Proceedings of the International Conference on Computer Vision*, pages 1879–1888, 2017.

[13] Yuhong Li, Xiaofan Zhang, and D. Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR),*, pages 1091–1100, 2018.

[14] J. Liu, Chenqiang Gao, Deyu Meng, and A. Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR),*, pages 5197–5206, 2018.

[15] M. Hossain, M. Hosseinzadeh, Omit Chanda, and Yang Wang. Crowd counting using scale-aware attention networks. *Proceedings of the IEEE Workshop on Applications of Computer Vision*, pages 1280–1288, 2019.

[16] Youmei Zhang, Chunluan Zhou, F. Chang, and A. Kot. Attention to head locations for crowd counting. In *Proceedings of the International Conference on Image*

and Graphics, pages 727–737, 2019.

[17] Junyu Gao, Wei Lin, Bin Zhao, Dong Wang, Chenyu Gao, and Jun Wen. c^3 framework: An open-source pytorch code for crowd counting. *ArXiv*, abs/1907.02724, 2019.

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proceedings of the International Conference on Learning Representations*, 2015.

[19] Mark Marsden, Kevin McGuinness, S. Little, and N. O'Connor. Fully convolutional crowd counting on highly congested scenes. *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 27–33, 2017.

[20] Deepak Babu Sam, Shiv Surya, and R. Venkatesh Babu. Switching convolutional neural network for crowd counting. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR),*, pages 4031–4039, 2017.

[21] V. Sindagi and V. Patel. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–6, 2017.

[22] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. Crowd counting via adversarial cross-scale consistency pursuit. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR),*, pages 5245–5254, 2018.

[23] Siyu Huang, Xi Li, Zhongfei Zhang, Fei Wu, Shenghua Gao, R. Ji, and Junwei Han. Body structure aware deep crowd counting. *IEEE Transactions on Image Processing*, 27:1049–1059, 2018.

[24] L. Zhang, Miaojing Shi, and Qiaobo Chen. Crowd counting via scale-adaptive convolutional neural network. *Proceedings of the IEEE Workshop on Applications of Computer Vision*, pages 1113–1121, 2018.

[25] Junyu Gao, Q. Wang, and Xuelong Li. Pcc net: Perspective crowd counting via spatial convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30:3486–3498, 2020.

[26] Le Zhang, Zenglin Shi, Ming-Ming Cheng, Yun Liu, Jia-Wang Bian, Joey Tianyi Zhou, Guoyan Zheng, and Zeng Zeng. Nonlinear regression via deep negative correlation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:982–998, 2021.

[27] H. Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah. Composition loss for counting, density map estimation and localization in dense crowds. *Proceedings of the European Conference on Computer Vision*, pages 532–546, 2018.

[28] Lei Liu, Jie Jiang, Wenjing Jia, Saeed Amirgholipour, Yi Wang, Michelle Zeibots, and Xiangjian He. Denet: A universal network for counting crowd with varying densities and scales. *IEEE Transactions on Multimedia*, 23:1060–1068, 2021.

[29] Deepak Babu Sam, S. Peri, Mukuntha Narayanan Sundararaman, Amogh Kamath, and R. Venkatesh Babu. Locate, size and count: Accurately resolving people in dense crowds via detection. *IEEE transactions on pattern analysis and machine intelligence*, 43:2739–2751, 2021.

[30] Vishwanath A. Sindagi and V. Patel. Ha-ccn: Hierarchical attention-based crowd counting network. *IEEE Transactions on Image Processing*, 29:323–335, 2020.