ATTENTION GUIDED REGION DIVISION FOR CROWD COUNTING

Xiaoqi Pan, Hong Mo, Zhong Zhou, Wei Wu*

State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
School of Computer Science and Engineering, Beihang University, China
{panxiaoqi,mandymo,zz,wuwei}@buaa.edu.cn

ABSTRACT

Crowd counting has drawn more and more attention in computer vision. There are two mainstream approaches to deal with crowd counting tasks, regression and detection. Regression-based methods usually overestimate the count in sparse areas, while detection-based methods tend to underestimation in dense areas. In this paper, we propose a two-branch network combining regression and detection. We introduce the attention mechanism to make the network adaptively divide dense and sparse areas and employ appropriate methods on them respectively. The regression branch predicts density map in extremely dense areas. An improved detection network is applied to detect multi-scale heads in relatively sparse areas. Our method is able to obtain precise head bounding boxes in sparse areas with ensuring counting accuracy in dense areas. Experimental results show that our method achieves state-of-the-art on challenging public crowd counting datasets.

Index Terms— Crowd counting, Head detection, Density regression

1. INTRODUCTION

Crowd counting is a task aiming to estimate the number of people in images. With the urban expansion and population growth, crowded people bring a lot of problems, such as traffic inconvenience and accident risk. Crowd counting has attracted widespread attention due to its application in city management and public security.

Existing methods can be divided into two categories, regression and detection. Regression-based methods predict the approximate distribution of crowds. These methods employ a Gaussian filter [1] on the point map of heads to generate the density map as ground truth. Numerous research has been done on regression-based methods. [2] proposed MCNN, a multi-column network with different convolution kernel sizes. Different kernel sizes can learn different scales of features, which make the model more robust. [3] proposed Switching-CNN, which had three branches with different kernel sizes, and a selector was designed to decide which branch was responsible for prediction. [4] proposed CSRNet, which used dilated convolution to capture large scale information instead of using a multi-branch structure. Different from regression, detection-based methods count people by object detector. Recent studies have greatly improved the ability of detection network. Object detectors such as [5, 6, 7] achieve good performance for common objects. [8] proposed a feature pyramid network to deal with the low resolution of small objects. [9] proposed focal loss to mitigate the influence of imbalance between positive and negative samples in the one-stage detection network. [10] proposed an end-to-end people detector for crowded scenes.

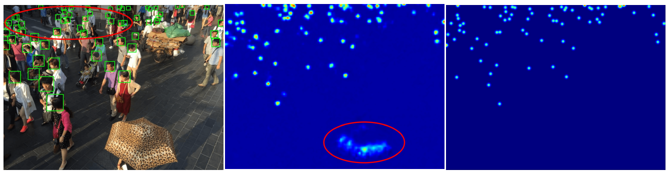
(a) Detection:63 (b) Regression:84 (c) Ground Truth:71

Fig. 1. Result comparison of two methods on ShanghaiTech dataset part B. The main source of error is in the red circle. (a) is the result of YOLOv3. (b) is the result of CSRNet. (c) is ground truth.

The detection-based methods have the ability to predict the precise location and head size of each person, which can be used for further analysis, such as pedestrian tracking and behavior analysis. However, detection-based methods suffer from low resolution and occlusion, resulting in poor performance on the datasets with high density. Regression-based methods, on the contrary, only predict the approximate distribution of people without distinction between each individual and have a good performance in congested situations, but in sparse situations, these methods are easily affected by background texture. Fig 1 shows a typical situation on ShanghaiTech dataset [2] part B. Image on the left shows the result of detection method YOLOv3 [7], the middle one shows the

[*] Corresponding author.

This work is supported by the National Key R&D Program of China under Grant No.2018YFB2100603 and Natural Science Foundation of China under Grant No.61872024.

result of regression method CSRNet [4] and the right one is the ground truth density map. We can find that the result of detection method is unreliable in the dense area, the number is underestimated. Instead, regression method overestimates in the sparse area because of the background such as the umbrella and clothes.

Considering the advantages of regression and detection, some exploration has been done to combine them. [11] designed a multi-branch network, employing regression and detection methods respectively and another branch was trained as pixel-wise weights. Without multi-scale receptive field, the network has weakness in adaptability. Its performances on dense datasets were not reported. [12] made use of the results of regression network, concatenate them with the features from detection network and predict the bounding box, but the method relies on depth map to some extent and also difficult to handle the extremely dense situation. [13] horizontally divided images into nearby and distant regions and employed detection and regression respectively, but this kind of division is too rough to achieve better performance. With regard to the shortcomings of existing combination methods, we propose an attention guided division network to combine regression and detection in this paper.

The main contributions of this paper can be summarized as follows. A two-branch network is proposed to combine regression and detection methods with the attention mechanism. The network takes advantage of both methods, improves count accuracy and obtains head bounding box at the same time. Meanwhile, we design a head detection network and propose a low-cost method to generate ground truth bounding box. Experimental results reveal that our method achieves state-of-the-art performance on public datasets.

2. PROPOSED METHOD

The architecture of our network contains two branches, regression branch and detection branch. The regression branch is responsible for dividing the image into dense and sparse areas and predict density map in the dense area. The detection branch is responsible for detecting heads in the sparse area. The sum of two branches is the predicted people count. The overall architecture is shown as Fig 2.

2.1. Division of Dense and Sparse People

A key issue of our method is how to decide which parts of image are predicted by regression branch and which parts by detection branch. We define the local crowd density as the number of people in a small area. The crowds are divided into dense and sparse according to local density. Specifically, if there are more than C heads within $K \times K$ area, we consider them as dense people, otherwise they are sparse people. Besides, we also consider the person with a head radius less than 5 as dense people, because we are unable to detect these heads due to the setting of anchor boxes. We set $C = 5, K = 40$ empirically and no more adjustments in experiments.

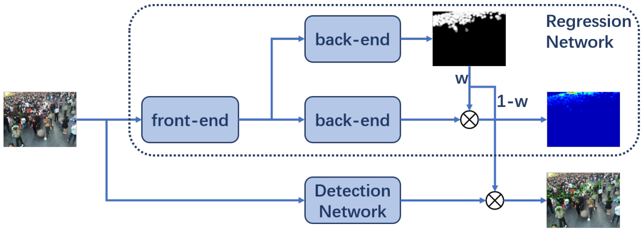

Fig. 2. Architecture of the proposed network. $\otimes$ denotes using attention map to filter density map and detection results.

2.2. Regression Branch

The regression branch is inspired by CSRNet [4]. We use the first ten layers of VGG-16 as front-end. Following the front-end are two parallel back-ends. The structure of front-end and back-end is shown in Fig 3. We introduce attention map to make the network focus on dense areas. The first back-end is followed by a convolution layer with sigmoid activation function to predict the attention map. The output of another back-end is multiplied by attention map and followed by a convolution layer without activation function to predict the density map within the dense area.

We generate ground truth of density map by the method in [2]. The density map can be illustrated with formula:

$$D^{GT} = \sum_{i=1}^{N} \delta\left(x - x_i\right) \times G_{\sigma}\left(x\right) \tag{1}$$

We only generate density map for dense people, thus we use a small fixed Gaussian kernel and set $\sigma = 5$. Attention map can be simply generate by binarization of density map as following formula:

$$\forall x \in D^{GT}, A^{GT}\left(x\right) = \begin{cases} 0, & x \leq t \\ 1, & x > t \end{cases} \tag{2}$$

We set the threshold t as $1e - 4$.

The loss of regression branch consists of Euclidean loss for density map and binary cross entropy loss for attention map. It can be illustrated with formula:

$$L_{den} = \frac{1}{N} \sum_{i=1}^{N} \left\| D_i - D_i^{GT} \right\|_2^2 \tag{3}$$

$$L_{att} = -\frac{1}{N} \sum_{i=1}^{N} \left(A_i^{GT} log A_i + \left(1 - A_i^{GT}\right) log \left(1 - A_i\right)\right) \tag{4}$$

$$L_{reg} = L_{den} + \lambda L_{att} \tag{5}$$

Where D_i and A_i denote density map and attention map. We set λ as 0.1 to balance the loss of two parts.

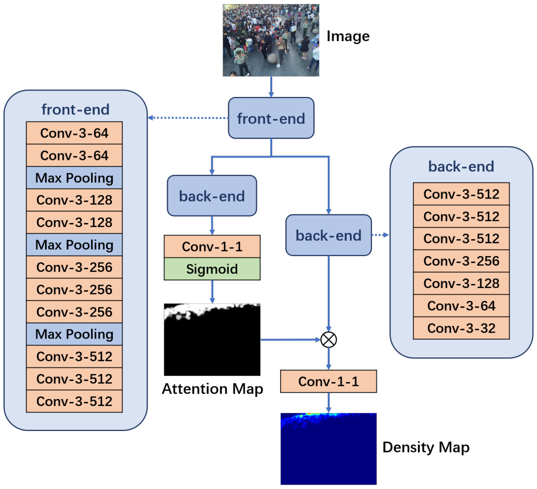

Fig. 3. Structure of regression branch in detail.

2.3. Detection Branch

Inspired by YOLOv3 [7], we design a fast and one-stage head detector. Commonly used backbone networks have too many downsampling layers, leading to low resolution of the output feature map, making the network unable to detect small heads. In the tasks of crowd counting, what we need is more spatial information instead of more semantic information. Therefore, we adjust the allocation of layers and parameters, make the output feature map have higher resolution. Meanwhile, we reduce the depth and channels for saving memory. The configuration of backbone is shown as Table 1. Let b_i denote the feature map from Block i. We can calculate f_i as follow, $f_4 = Conv\,(b_4)$, $f_i = Conv\,(Concat\,(b_i, Up\,(f_{i+1})))$, $i = 2, 3$. Where $Conv$ denotes a series of convolution layer with 3×3 kernel, $Concat$ is concatenation of channels and Up is bilinear upsampling. We set 3 kinds of anchor boxes for each detection scale. Each f_i, $i = 2, 3, 4$ is followed by a 1×1 convolution layer with 15 filters to give the predicted bounding boxes and confidence.

The cost to label head bounding box is extremely high. Thus, mainstream crowd counting datasets only provide heads location. We propose a relatively simple method to generate bounding box ground truth. For a typical pinhole camera, we can derive formula $\frac{R}{r} = \frac{d}{f}$ and $r \propto \frac{1}{d}$ based on similar triangles. Where R and r denote the head size in real world and in image. d is object distance and f is focal length of the camera. Following [14], we assume that all heads are on the same plane, in other words, the y-coordinate of the head point in the image reflects the distance from camera and $y \propto \frac{1}{d}$. Thus, $r \propto y$. Furthermore, let $r = ky + b$, where r is the head size in image and y is the y-coordinate of the head, k and b are constants to be solved. Therefore, for one scene, we only need to label the size of several heads at different location and we can estimate other bounding boxes by linear regression.

Layers	Output Size	Parameters
Conv	512×512	$3 \times 3, 32$
Conv	256×256	$3 \times 3, 64$, stride 2
Block1	256×256	$\begin{bmatrix} 1 \times 1, 32 \\ 3 \times 3, 64 \end{bmatrix} \times 2$
Conv	128×128	$3 \times 3, 128$, stride 2
Block2	128×128	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 128 \end{bmatrix} \times 8$
Conv	64×64	$3 \times 3, 256$, stride 2
Block3	64×64	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 256 \end{bmatrix} \times 8$
Conv	32×32	$3 \times 3, 512$, stride 2
Block4	32×32	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 512 \end{bmatrix} \times 4$

Table 1. The backbone of detection branch.

For detection branch, we use Euclidean loss for location and binary cross entropy loss for classification.

$$L_{loc} = \frac{1}{N} \sum_{i=1}^{N} \left\| I_p \left(B_i - B_i^{GT} \right) \right\|_2^2 \tag{6}$$

$$L_{cls} = -\frac{1}{N} \sum_{i=1}^{N} \left(I_p log P_i + I_n log \left(1 - P_i \right) \right) \tag{7}$$

$$L_{det} = L_{loc} + L_{cls} \tag{8}$$

Where B_i denotes head location and size, which is x, y, w, h, P_i denotes confidence. We set $I_p = 1$ for positive sample and $I_n = 1$ for negative sample, otherwise 0.

3. EXPERIMENT

We random crop the images into 512×512 patches and horizontal flip with probability of 0.5 for data augmentation. Adam optimizer with $1e-4$ learning rate is employed during training. Because of the limitation of memory, we train two branches separately, set batch size to 6, and train 500 epochs on Nvidia Titan X. Then, combine them for evaluation.

3.1. Experiment Results

We use mean absolute error (MAE) and mean squared error (MSE) as evaluation metric. The results on part A and part B of ShanghaiTech dataset [2] are shown in Table 2. On part A and part B, we achieve the best 61.4 and 7.2 MAE compared to the state-of-the-art method. The results on ShanghaiTechRGBD dataset [12] are shown in Table 3. Our method achieves 15% lower MAE than RDNet. We also did 5-fold cross validation on UCF_CC_50 dataset [15], shown as Table 4. We achieve a significantly lower MSE, owing to the strong adaptability of our method. Fig 4 shows the visualization of our method on occasions of different densities of people. The results show that our method is robust to scale variation.

Methods	Part A		Part B	
	MAE	MSE	MAE	MSE
MCNN[2]	110.2	173.2	26.4	41.3
Switching-CNN[3]	90.4	135.0	21.6	33.4
DecideNet[11]	-	-	20.7	29.4
CSRNet[4]	68.2	115.0	10.6	16.0
ASD[16]	65.6	98.0	8.5	13.7
TEDnet[17]	64.2	109.1	8.2	12.8
RDNet[12]	-	-	8.8	15.3
ADCrowdNet[18]	63.2	98.9	7.6	13.9
PACNN[14]	62.4	102.0	7.6	11.8
ours	**61.4**	**97.5**	**7.2**	**11.8**

Table 2. Results on ShanghaiTech dataset part A and part B.

Methods	MAE	MSE
MCNN[2]	7.14	9.99
Idrees et al.[19]	7.32	10.48
CSRNet[4]	4.91	7.11
RDNet[12]	4.96	7.22
ours	**4.18**	**6.75**

Table 3. Results on ShanghaiTechRGBD dataset, some results are referenced from [12].

Methods	MAE	MSE
MCNN[2]	377.6	509.1
Switching-CNN[3]	318.1	439.2
CSRNet[4]	266.1	397.5
ASD[16]	196.2	270.9
TEDnet[17]	249.4	354.5
ADCrowdNet[18]	257.1	363.5
PACNN[14]	241.7	320.7
ours	**194.7**	**246.8**

Table 4. Results on UCF_CC_50 dataset.

3.2. Ablation Study

We conduct ablation experiment on ShanghaiTech dataset part B. The results are shown in Table 5. First, we compare our detection network with YOLOv3. Results show that our method obtains significant MAE and MSE decrease. Second, we evaluate our regression network on the whole image without division. It shows that the attention mechanism improves the performance of CSRNet. Then, we combine these two branches and achieve a better result.

Methods	MAE	MSE
Only detection(YOLOv3[7])	17.3	35.8
Only detection(ours)	10.2	17.0
Only regression(CSRNet[4])	10.6	16.0
Only regression(CSRNet+attention)	7.7	12.2
Combination(ours)	**7.2**	**11.8**

Table 5. Results of ablation study on ShanghaiTech part B.

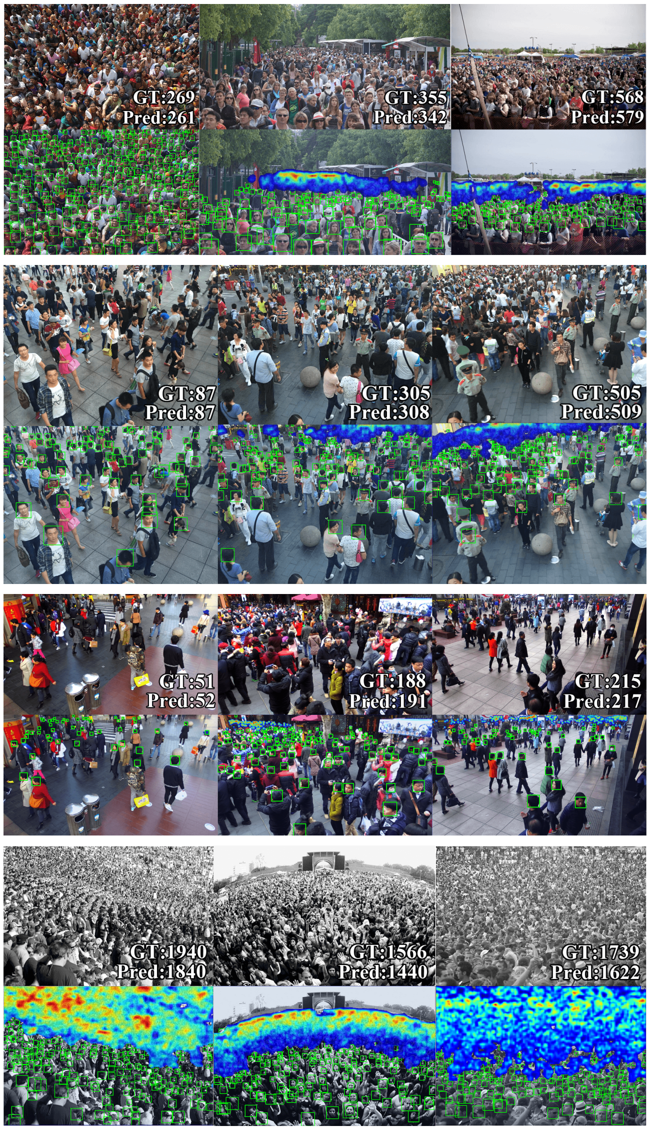

Fig. 4. The visualization results on ShanghaiTech part A, part B, ShanghatTechRGBD and UCF_CC_50. First row is original image, second row is the result of our method.

4. CONCLUSIONS

In this paper, we introduce attention mechanism to combine regression and detection, which can divide image into dense and sparse areas and choose appropriate methods on them. The combination takes full advantage of the two methods and have adaptability to scale variation. We evaluate our method on challenging public datasets with high variation in crowd densities. Experimental results show that our method achieves state-of-the-art performance.

5. REFERENCES

[1] Victor Lempitsky and Andrew Zisserman, “Learning to count objects in images,” in *Advances in neural information processing systems*, 2010, pp. 1324–1332.

[2] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma, “Single-image crowd counting via multi-column convolutional neural network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 589–597.

[3] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu, “Switching convolutional neural network for crowd counting,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4031–4039.

[4] Yuhong Li, Xiaofan Zhang, and Deming Chen, “Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1091–1100.

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, 2015, pp. 91–99.

[6] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, “Ssd: Single shot multibox detector,” in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[7] Joseph Redmon and Ali Farhadi, “Yolov3: An incremental improvement,” *arXiv preprint arXiv:1804.02767*, 2018.

[8] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[10] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng, “End-to-end people detection in crowded scenes,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2325–2333.

[11] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G Hauptmann, “Decidenet: Counting varying density crowds through attention guided detection and density estimation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5197–5206.

[12] Dongze Lian, Jing Li, Jia Zheng, Weixin Luo, and Shenghua Gao, “Density map regression guided detection network for rgb-d crowd counting and localization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1821–1830.

[13] Gaoqi He, Zhenwei Ma, Binhao Huang, Bin Sheng, and Yubo Yuan, “Dynamic region division for adaptive learning pedestrian counting,” in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 1120–1125.

[14] Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen, “Revisiting perspective information for efficient crowd counting,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7279–7288.

[15] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah, “Multi-source multi-scale counting in extremely dense crowd images,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2547–2554.

[16] Xingjiao Wu, Yingbin Zheng, Hao Ye, Wenxin Hu, Jing Yang, and Liang He, “Adaptive scenario discovery for crowd counting,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2382–2386.

[17] Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David Doermann, and Ling Shao, “Crowd counting and density estimation by trellis encoder-decoder networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6133–6142.

[18] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu, “Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3225–3234.

[19] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah, “Composition loss for counting, density map estimation and localization in dense crowds,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 532–546.