MSPNET: MULTI-SUPERVISED PARALLEL NETWORK FOR CROWD COUNTING

*Bo Wei, Yuan Yuan, Qi Wang**

School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China

ABSTRACT

Crowd counting has a wide range of applications such as video surveillance and public safety. Many existing methods only focus on improving the accuracy of counting but ignore the importance of density maps. It's no doubt that a high-quality density map contains more information such as localization and movement of the crowd. In this paper, we propose a multi-supervised parallel network (MSPNet) to achieve high accuracy of crowd counting and generate high-quality density maps. We conduct multiple supervisions in the training process, which can supplement the details lost in pooling and up-sampling operations to improve the quality of density maps. In addition, to reduce the impact of background noise, the attention mechanism is employed to help the network focus on the crowd. Extensive experiments on two mainstream benchmarks show that MSPNet achieves significantly improvement over the state-of-the-art in terms of counting accuracy and the quality of density maps.

***Index Terms*—** Crowd counting, high-quality density map, multiple supervisions, attention mechanism

1. INTRODUCTION

With the rapid growth of the global population, crowd counting has received more and more attention because of its wide applications such as video surveillance, traffic control and public safety. But crowd counting is still a challenging problem due to scale variation, occlusions, background noise and so on. According to[1][2], the traditional approaches of crowd counting can be divided into detection-based methods and regression-based methods. Detection-based methods usually use sliding windows to detect the individuals in images[3][4]. But these methods are computationally expensive and hardly work when pedestrians are heavily occluded or densely spread. Regression-based methods focus on learning a mapping between features to counts[5][6], which are more robust in high-density crowd scenes, but hand-crafted representations make the results far from optimal.

*Qi Wang is the corresponding author. This work was supported by the National Natural Science Foundation of China under Grant U1864204, 61773316, 61632018 and 61825603.

In recent years, with the rapid development of deep learning, many CNN-based methods have achieved great improvement over the traditional methods[7][8][9][10]. However, few of them focus on improving the quality of density maps. In fact, density maps can improve the accuracy of crowd counting and are more useful in practice because they can show the localization and movement of the crowd intuitively. There are two challenging problems in generating high-quality density maps: 1) the details of feature maps are easily lost in pooling and up-sampling operations; 2) background noise such as trees, stone and buildings are often recognized falsely in the density maps, as shown in Fig 1.

To solve the problems mentioned above, we propose an end-to-end CNN-based method called multi-supervised parallel network (MSPNet). On one hand, we conduct multiple supervisions in training process with groundtruth maps of different sizes, which supplement the details lost in pooling and up-sampling operations. They also provide features in different scales, which make the network have a better robustness to the scale variation. On the other hand, to reduce the influence of background noise, attention mechanism is employed in our network to separate the crowd from the background.

The main contributions made in this paper is summarized as follows:

- We use groundtruth of different sizes to conduct multiple supervisions, which improve the quality of proposed density maps;
- We use attention mechanism to make the network focus on pedestrians, which help the network handle the diverse crowd distribution and the complex background;
- We achieve the best performance on two commonly-used benchmark datasets and significantly improvement over several recent state-of-the-art approaches.

2. RELATED WORK

Many CNN-based crowd counting methods have been proposed in recent years[12][13][14][15]. Though almost all of them obtained crowd counting by generating density maps, only few can generate high-quality density maps. Zhang et

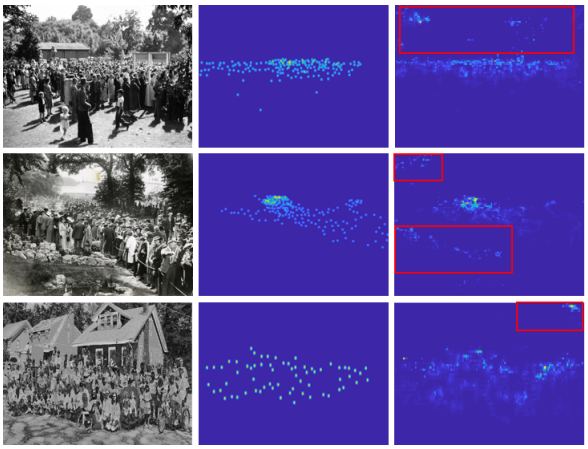


Fig. 1. The result of CSRNet[11] on ShanghaiTech dataset[9]. First column: Images; Second column: Groundtruth; Third column: CSRNet[11]. The red boxes clearly show the effect of background noise on the CSRNet[11].

al.[9] proposed a multi-column architecture to extract features in different scales and fuse the features to generate density maps. But the resolutions of them were only 1/8 of the original images. The same problem also existed in Switch-CNN[16], Hydra-CNN[17] and DecideNet[7]. Though CP-CNN[8] combined the global and the local context to make the density maps contain more details, the network structure was cumbersome and the the features extracted by different braches were similar. CSRNet[11] used dilated convolutional layers instead of pooling layers to maintain the spatial information in feature maps and achieved state-of-the-art performance. But the results were easily effected by background noisy.

Compared with the above approaches, we conduct multiple supervisions in our network to increase the resolutions of density maps without losing details, and employ attention mechanism to reduce the influence of background noise. The two methods achieve significant improvement on the quality of generated density maps.

3. PROPOSED METHOD

In this section, the proposed MSPNet is introduced. First, the overall architecture of the network is proposed in Section 3.1. Then, Section 3.2 shows how to generate the groundtruth. Finally, Section 3.3 gives the loss function.

3.1. Network Architecture

As discussed in Section 2, though many methods have been proposed to improve the accuracy of crowd counting, the density maps generated by them are usually low-resolution and susceptible to background noise. To solve these problems, we employ multiple supervisions and attention mechanism in our network. Fig.2 illustrates the architecture of our MSPNet, which is comprised of four parts, i.e. a shared feature extractor, a density map estimator (DME), an attention map estimator (AME) and some supervision modules. Same as most CNN-based methods in crowd counting[18][11], we adopt the first ten convolutional layers of VGG-16 model[19] pre-trained from the ImageNet[20] as the feature extractor because they have strong learning ability. DME and AME are two branches with the same structure. DME generates density maps of original resolution and AME aims to produce attention maps, which help the network emphasize head regions to tackle background noise. For simplicity, we take the DME as an example to introduce the structure. The DME contains three similar modules called map generator (MG), which increase the resolutions of feature maps gradually. The design of the module is shown in Fig.3. The module consists of two convolutional layers and one deconvolutional layer and the kernel sizes in convolutional layers are all 3×3. In order to enlarge the receptive field, we employ dilated convolutional layers instead of traditional convoltional layers in Module II and III. The deconvolutional layers are used to up-sampling the feature maps instead of interpolation. The output feature maps are fused by a 1×1 convolutional layer to generate the density maps.

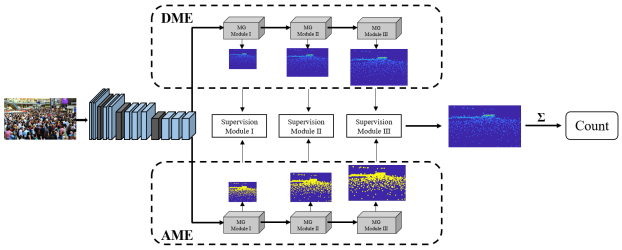


Fig. 2. The network architecture of MSPNet

Some supervision modules are employed to conduct multiple supervisions in the network. The density maps and the attention maps that generated respectively by MGs in DME and AME are input to this module. The attention map compares to the attention map groundtruth and produces the first training loss. Then a pixel-wise multiplication is conducted to fuse the two maps and generates the final density map. The second training loss is produced by the comparison between the final density map and density map groundtruth. Because the supervisions are conducted following up-sampling operations, the lost details will be supplemented in time without affecting subsequent estimations. In addition, the groundtruth used in supervisions have different sizes, which make the network learn features in various scales. The density map generated by the last supervision module is the final output of our network, which has original resolution and has been processed by the attention mechanism. The total number of crowd is obtained by integrating the density map.

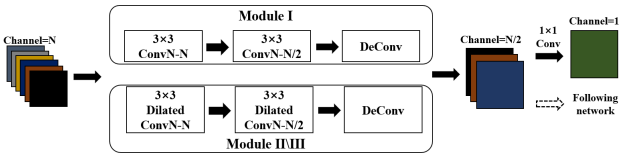


Fig. 3. The design of MG modules in DME and AME. Dilated convolutional layers are used instead of traditional convolutional layers in model II and III

3.2. Groundtruth Generation

In this paper, we use two kinds of groundtruth. The density map groundtruth is generated according to [21]. The formula is defined as:

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

where $\delta(x - x_i)$ is groundtruth and G_σ is the Gaussian kernel whose standard deviation is σ. For ShanghaiTech Part A[9] and UCF_CC_50[22], we set the σ of Gaussian kernel to 5. For ShanghaiTech Part B[9], which is sparser, σ is set to 15. The attention map groundtruth can be obtained directly based on the density map groundtruth. The formula is defined as:

$$A^{GT}(x_i) = \begin{cases} 1 & G^{GT}(x_i) > th \\ 0 & G^{GT}(x_i) \leq th \end{cases} \tag{2}$$

where th is the threshold set as 0.001. $G^{GT}(x_i)$ is the value of ith pixel in groundtruth. A binary attention map groundtruth is generated through the formula above. In addition, to match the different sizes of maps in our network, we resize the groundtruth to 1/4 and 1/2 of its original size.

3.3. Objective Function

As mentioned in section3.1, the supervision modules produce two training losses, which are MSE loss and cross entropy loss. Besides density map estimation, we employ a L1 loss to improve the accuracy of the total number of people estimated. The formulas of the three kinds of loss function is defined as:

$$L_d = \frac{1}{N} \sum_{i=1}^{N} \left\| O(X_i, \Theta) - G_i^{GT} \right\|^2 \tag{3}$$

$$\begin{aligned} L_a = -\frac{1}{N} \sum_{i=1}^{N} & \left(A_i^{GT} \log(P_i) \right. \\ & \left. + \left(1 - A_i^{GT}\right) \log(1 - P_i) \right) \end{aligned} \tag{4}$$

$$L_n = \frac{1}{N} \sum_{i=1}^{N} \left| E - n^{GT} \right| \tag{5}$$

where N is the number of samples in batch. $O(X_i, \Theta)$ is the output of the network with parameters. P_i is the predicted value of ith pixel in attention maps. E is the people counting estimated and n^{GT} is the true number.

The entire network is trained using the following defined loss:

$$L_{final} = \alpha L_n + \sum_{i=1}^{M} \frac{1}{2^{M-i}} \left(L_d^i + \beta L_a^i \right) \tag{6}$$

We use a size-dependent method to set the weights of losses in different supervision modules. M is the total number of supervision modules and i represents the ith supervision module. The weight of a supervision module is half of the next module, and the maximum weight is 1. α and β are weighting weight that are set to 0.01 and 0.0001 in our experiments.

4. EXPERIMENTAL RESULTS

In this section, we conduct the experiments on the ShanghaiTech dataset[9] and the UCF_CC_50 dataset[22] to evalute the accuracy and the quality of density maps generated by our MSPNet. The ShanghaiTech dataset[9] is a large-scale crowd counting dataset, which contains 1198 annotated images with 330,165 pedestrians. This dataset is divided into Part A and Part B. Part A contains 482 images and has high density. Part B contains 716 images and has relatively sparse density. The UCF_CC_50 dataset[22] is another challenging dataset which has extremely dense crowd density. It contains 50 images with an average of 1280 individuals per image. We employ mean absolute error(MAE) and mean squared error(MSE) for evaluting the accuracy of density maps, while PSNR and SSIM[23] are used to evalute the quality of density maps.

4.1. Performance and Comparison

We compare our MSPNet with other methods, including MCNN[9], CP-CNN[8], Switching-CNN[16], CMTL[24], CSRNet[11] and so on. The results are shown in table 1. Compared with other methods, MSPNet achieves better performance than the state-of-the-art method on both ShanghaiTech dataset[9] and UCF_CC_50 dataset[22]. The result of MSPNet is MAE of 59.8(8.5-point improvement) and MSE of 98.2(16.8-point improvement) on ShanghaiTech Part A dataset[9]. On Part B[9], MSPNet also achieves best results: MAE of 7.6(3.0-point improvement) and MSE of 14.1(1.9-point improvement). On the UCF_CC_50 datasets[22], the proposed MSPNet achieves a 59.4-point improvement in MAE and a 98.2-point improvement in MSE over the state-of-the-art(CSRNet[11]). Moreover, we compare the quality of density maps estimated by MSPNet, Switching-CNN[16], CP-CNN[8] and CSRNet[11], which also focus on generating high-quality density maps. Table 2 shows the results on ShanghaiTech Part A dataset[9]. It clearly indicates that

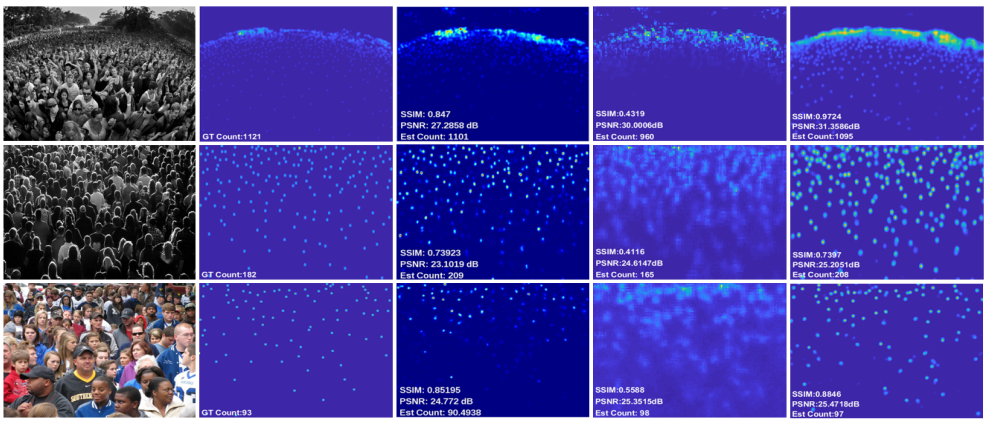


Fig. 4. Comparative visualization results with CSRNet[11] and CP-CNN[8] on ShanghaiTech dataset[9]. First column: Input Images; Second column: Groundtruth; Third column: CP-CNN[8]; Forth column: CSRNet[11]; Fifth column: MSPNet.

Table 1. The comparison between MSPNet and other methods on ShanghaiTech dataset[9] and UCF_CC_50 dataset[22]

Method	Part A		Part B		UCF_CC_50	
	MAE	MSE	MAE	MSE	MAE	MSE
MCNN[9]	110.2	173.2	26.4	41.3	377.6	509.1
Cascaded-MTL[24]	101.3	152.4	20.0	31.1	322.8	397.9
Switching-CNN[16]	90.4	135.0	21.6	33.4	318.1	439.2
CP-CNN[8]	73.6	106.4	20.1	30.1	295.8	320.9
ACSCP[25]	75.7	102.7	17.2	27.4	291.0	404.6
DecideNet[7]	-	-	20.8	29.4	-	-
SaCNN[26]	86.8	139.2	16.2	25.8	314.9	424.8
CSRNet[11]	68.2	115.0	10.6	16.0	266.1	397.5
our MSPNet	**59.8**	**98.2**	**7.5**	**14.1**	**206.7**	**299.3**

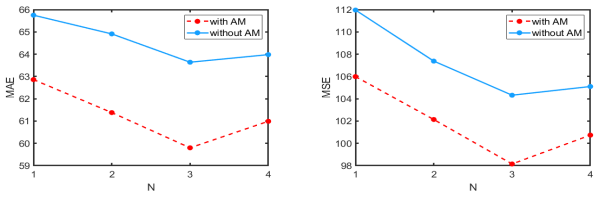
MSPNet achieves a significant improvement over the existing methods in both accuracy and quality of density maps. A visual result is shown in Fig.4.

4.2. Ablation Study

In order to demonstrate the effectiveness of multiple supervision and attention mechanism, we conduct two sets of ablation experiments on ShanghaiTech Part A[9]. We compare the performance of networks with different numbers of supervisions. The results are shown in Fig.5. We can observe that the performance is not positively correlated with the number of supervisions. The network achieves a better result when the number is three. Attention mechanism can help network focus on the crowd, which improves the performance intuitively. It's easy to observe that the network with attention mechanism achieves better performance in both MAE and MSE.

Table 2. Evalution of the quality of density maps on ShanghaiTech Part A dataset[9]

Methods	MAE	MSE	PSNR	SSIM
Switching-CNN[16]	90.4	135.0	21.91	0.67
CP-CNN[8]	73.6	106.4	21.72	0.72
CSRNet[11]	68.2	115.0	23.79	0.76
our MSPNet	**59.8**	**98.2**	**23.94**	**0.78**



(a) MAE (b) MSE

Fig. 5. The performance of networks with different structures. N is the number of supervisions and with/without AM means the network with/without attention mechanism

5. CONCLUSION

To generate high-quality density maps, a novel end-to-end architecture called multi-supervised parallel network(MSPNet) has been proposed in this paper. We conduct multiple supervisions in the process of increasing the resolutions of density maps, which supplement the details lost in pooling and upsampling operations. In addition, attention mechanism is employed to reduce the influence of background noise. Extensive experiments indicate that the proposed MSPNet achieves state-of-the-art performance in two mainstream datasets and significant improvement of the quality of density maps over several recent state-of-the-art approaches.

6. REFERENCES

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

[2] Sami Abdulla Mohsen Saleh, Shahrel Azmin Suandi, and Haidi Ibrahim, "Recent survey on crowd density estimation and counting for visual surveillance," *Eng. Appl. of AI*, vol. 41, pp. 103–114, 2015.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, pp. 886–893.

[4] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *ICCV*, 2005, pp. 90–97.

[5] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *BMVC*, 2012, pp. 1–11.

[6] A. B. Chan and N. Vasconcelos, "Bayesian poisson regression for crowd counting," in *ICCV*, 2009, pp. 545–551.

[7] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, "Decidenet: counting varying density crowds through attention guided detection and density estimation," in *CVPR*, 2018, pp. 5197–5206.

[8] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid cnns," in *ICCV*, 2017, pp. 1879–1888.

[9] Y. Zhang, D. Zhou, S. Chen, S. Gao, and M. Yi, "Single-image crowd counting via multi-column convolutional neural network," in *CVPR*, 2016, pp. 589–597.

[10] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 1, pp. 46–58, 2020.

[11] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *CVPR*, 2018, pp. 1091–1100.

[12] J. Gao, Q. Wang, and X. Li, "PCC net: Perspective crowd counting via spatial convolutional network," *IEEE Transactions on Circuits and Systems for Video Technology, DOI: 10.1109/TCSVT.2019.2919139*, 2019.

[13] C. Shang, H. Ai, and B. Bai, "End-to-end crowd counting via joint learning local and global count," in *ICIP*, 2016, pp. 1215–1219.

[14] L. Boominathan, S. S. Kruthiventi, and R. V. Babu, "Crowdnet: A deep convolutional network for dense crowd counting," in *ACM Multimedia*, 2016, pp. 640–644.

[15] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8198–8207.

[16] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *CVPR*, 2017, pp. 4031–4039.

[17] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *ECCV*, 2016, pp. 615–629.

[18] X. Wu, Y. Zheng, H. Ye, W. Hu, J. Yang, and L. He, "Adaptive scenario discovery for crowd counting," in *ICASSP*, 2019, pp. 2382–2386.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[20] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[21] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *NIPS*, 2010, pp. 1324–1332.

[22] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *CVPR*, 2015, pp. 833–841.

[23] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[24] V. A. Sindagi and V. M. Patel, "Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *AVSS*, 2017, pp. 1–6.

[25] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, and X. Yang, "Crowd counting via adversarial cross-scale consistency pursuit," in *CVPR*, 2018, pp. 5245–5254.

[26] L. Zhang, M. Shi, and Q. Chen, "Crowd counting via scale-adaptive convolutional neural network," in *WACV*, 2018, pp. 1113–1121.