Scale-Aware Multi-stage Fusion Network for Crowd Counting

Qi Liu[1,2], Jun Sang[1,2](✉), Fusen Wang[1,2], Li Yang[1,2], Xiaofeng Xia[1,2], and Nong Sang[3]

[1] Key Laboratory of Dependable Service Computing in Cyber Physical Society of Ministry of Education, Chongqing University, Chongqing 400044, China
[2] School of Big Data and Software Engineering, Chongqing University, Chongqing 401331, China
jsang@cqu.edu.cn
[3] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China

Abstract. Crowd counting has been widely researched and many hopeful results have been obtained recently. Due to the large-scale variation and complex background noise, accurate crowd counting is still very difficult. In this paper, we raise a simple but efficient network named SMFNet, which focuses on dealing with the above two problems of highly congested noisy scenes. SMFNet consists of two main components: multi-scale dilated convolution block (MDCB) for multi-scale features extraction and U-shape fusion structure (UFS) for multi-stage features fusion. MDCB can address the challenge of scale variation via capturing multi-scale features. UFS provides an effective structure that continuously combines outputs of different stages to achieve the capability of optimizing multi-scale features and increasing resistance to background noise. Compared with the existing methods, SMFNet achieves better performance in capturing effective and richer multi-scale features through progressively multi-stage fusion. To evaluate our method, we have demonstrated it on three popular crowd counting datasets (ShanghaiTech, UCF_CC_50, UCF-QNRF). Experimental results indicate that SMFNet can achieve state-of-the-art results on highly congested scenes datasets.

Keywords: Crowd counting · Multi-scale features · Multi-stage fusion

1 Introduction

Crowd counting has receiving considerable attention from researchers recently. With the development of deep learning on computer vision, many researchers have leveraged convolutional neural network based method to generate high-quality density map and perform accurate crowd counting [1–8]. Although these methods have achieved remarkable progress, the problem of accuracy degradation will still occur when applied to highly congested scene. And scale variation

Supported by National Natural Science Foundation of China (No. 61971073).

© Springer Nature Switzerland AG 2021

T. Mantoro et al. (Eds.): ICONIP 2021, CCIS 1517, pp. 3–11, 2021.
https://doi.org/10.1007/978-3-030-92310-5_1

is the major issue that seriously affects the quality of the estimated density maps. Recently, many CNN-based methods [1,2,9,10] with multi-column structure, multi-branch structure and multi-scale module like Inception [11] have been proposed to deal with scale variation problem. These modules adopt the filters of different sizes to capture variation in head sizes and show good improvements. However, there still exist some blatant shortcomings. The larger target in crowd scene is mainly extracted through kernels with large receptive fields (5×5, 7×7). And they can enhance the ability to capture large heads by adding more big kernels, but each additional large filter could significantly increase the number of training parameters.

In addition, these methods [1,2,9,10] mainly focus on catching the final scale diversity by stacking multi-scale features blocks directly, while ignoring different stages integration. As for crowd counting, shallow layer contains more low-level details [6] that guarantee the precise position of the people head, deep layer involves more high-level context that preserves more contextual information and overall looking of crowd region. Hence, it can be noticed that fusion of different layers may play an important role for better accuracy. Therefore, our method employs a novel multi-stage fusion for more robust crowd count.

Based on the above two points, we propose a novel Scale-aware Multi-stage Fusion Network called SMFNet, which better integrates the output of multiple stages in backend network, and achieves state-of-the-art crowd counting performance in highly congested noisy scenes. The structure of the proposed SMFNet is shown in Fig. 1. SMFNet is built upon the multi-scale dilated convolution block (MDCB) and U-shape fusion structure (UFS). MDCB is applied for multi-scale features extraction, which is constructed by different dilated convolutional layers stacked upon each other to strengthen the robustness for different heads sizes. To further enhance scale diversity captured by MDCB, we design the UFS to progressively optimize and enrich the extracted multi-scale features through merging different stages features together, while filtering low-level background noise that causes interference to density maps.

To summarize, the major contributions of our paper are as follows:

- We propose a multi-scale dilated convolution block (MDCB) to extract features at different scales to overcome scale diversity without involving extra computations.
- We propose an U-shape fusion structure (UFS) to progressively optimize and enrich extracted multi-scale features by concatenating and fusing features from multiple stages.
- We propose a novel SMFNet by integrating MDCB and UFS that achieves state-of-the-art performance in highly crowded scenes.

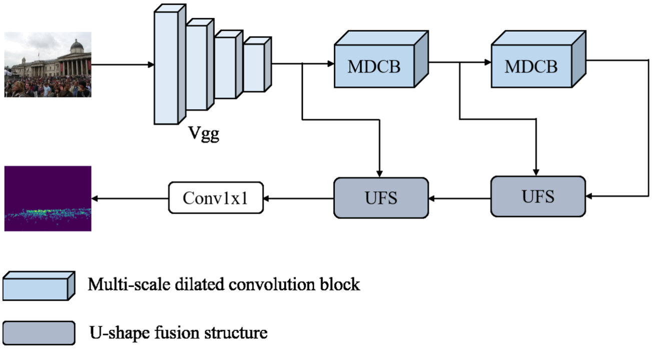

Fig. 1. The architecture of the Scare-aware Multi-stage Fusion Network (SMFNet).

2 Our Proposed Method

2.1 Multi-scale Feature Extraction

Following the previous methods [3,4,7,8], we adopt the first ten layers of VGG-16 [12] as our backbone network, which can effectively capture the features of an input crowd image.

In CNN, different kernel sizes can capture feature at different scales [9]. Many previous CNN-based works [2,9,10] employ multiple filter sizes to adapt the changes in head sizes. However, these methods usually tend to utilize large kernels to extract large head, which could lead to an increase in training parameters when more big kernels are added. Inspired by [3,12], dilated convolution kernel can expand the receptive field size without adding more parameters while keeping details of input image, and the stacking of more small kernels has better representation than directly adopting fewer large kernels. Hence, to tackle the challenge of scale diversity better, we all employ small kernels (1×1, 3×3) with different increasing dilation rates to capture features in various scales. Motived by the success of Inception structure [10] in image recognition domain, we design the multi-scale dilated convolution block (MDCB) as multi-scale feature extraction of SMFNet. An overview of MDCB is illustrated in Fig. 2.

To be specific, MDCB consists of a simple convolutional layer with 1×1 kernel size and dilation rate 1, and three convolutional layers with kernel sizes 3×3 and increasing dilation rates 1, 2, 3, which are equivalent to the filters of 1×1, 3×3, 5×5 and 7×7 respectively. The branch with filter size 1×1 is utilized to keep the feature scale from front layer to cover little targets, while others gradually increase receptive field sizes to capture large targets without involving extra parameters. This setting ensures that a smaller receptive field with different dilation rates can extract more information. Then, features from different branches in parallel are subsequently integrated by an Element-wise add operation together. Through the ablation experiments, we finally choose two cascaded multi-scale dilated convolution blocks to extract multi-scale features.

2.2 Multi-stage Feature Fusion

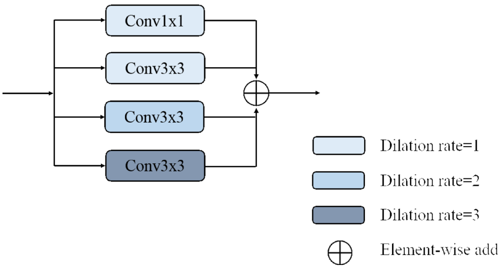

Fig. 2. Overview of multi-scale dilated convolution block (MDCB).

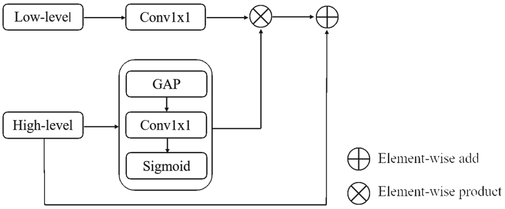

Fig. 3. Overview of the U-shape fusion structure (UFS).

Although the proposed MDCB provides scale variation, the hierarchical features between different blocks are not fully utilized. As discussion in first section, features from shallow layers can accurately locate the head position but preserve much low-level noise, while the features from the deep layers contain more contextual information. Hence, we improve the architecture by the proposed U-shape fusion structure (UFS), which could progressively optimize the extracted multi-scale features by concatenating and fusing adjacent stages features, while filtering low-level noise features that lead to error estimation. Compared with previous crowd counting methods based on multi-column, multi-branch or multi-scale module [1,2,9,10] that ignore taking advantage of the features from different layers, SMFNet achieves a more robust crowd counting performance through further multi-stage feature fusion.

The structure of U-shape fusion structure is presented in Fig. 3, which is used to combine features from adjacent layers in the network backend. Firstly, the feature maps from low-level layer could reduce the channels through a 1×1 convolution. The high-level feature maps through GAP (global average pooling) compute large-range context along the channel dimension, which captures the important channels while suppressing unnecessary channels. Then generated global contextual information is through a 1×1 convolution with Sigmoid activation function to calculate weight information, the value of which ranges from 0 to 1 that indicates regions the network focus on. Secondly, this computed weight information is adopted as guidance to weight low-level feature maps through

Table 1. Estimation errors on ShanghaiTech PartA, ShanghaiTech PartB, UCF_CC_50 and UCF-QNRF respectively.

Methods	PartA		PartB		UCF_CC_50		UCF-QNRF	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MCNN [9]	110.2	173.2	26.4	41.3	337.6	509.1	227.0	426.0
SwitchCNN [10]	90.4	135.0	21.6	33.4	318.0	439.2	228.0	445.0
CSRNet [3]	68.2	115.0	10.6	16.0	266.1	397.5	–	–
SANet [2]	67.0	104.5	8.4	13.6	258.4	334.9	–	–
TEDNet [7]	64.2	109.1	8.2	12.8	257.1	363.5	113.0	188.0
DSPNet [15]	68.2	107.8	8.9	14.0	243.3	307.6	107.5	182.7
SCLNet [16]	67.9	103.0	9.1	14.2	258.9	326.2	110.0	182.5
ADCrowdNet [5]	63.2	98.9	**7.7**	12.9	266.4	358.0	–	–
HACNN [17]	62.9	**96.1**	8.1	13.4	256.2	348.4	118.1	180.4
CAN [8]	62.3	100.0	7.8	**12.2**	**212.2**	**243.7**	107.0	183.0
(SMFNet) Ours	**57.7**	**96.1**	8.8	14.7	239.6	337.4	**95.5**	**166.7**

Element-wise product operation to extract effective multi-scale features while filtering low-level features with noise. Finally, when the weighted low-level feature maps involve more multi-scale noise-free features, final feature maps can be generated by adding with high-level features maps to further boost multi-scale features. The computed final feature maps are adopted as the high-level feature maps for the next U-shaped fusion process.

3 Implementation Details

3.1 Ground Truth Generation

Similar to previous methods [3,4,7,10] of generating density map for ground truth, we employ a normalized Gaussian (which is normalized to 1) kernel to blur each head annotation. For ShanghaiTech [9] and UCF_CC_50 [13], the fixed kernel is adopted to generate density maps. Whereas the adaptive-geometry kernel is utilized for UCF-QNRF [14], because of large variation in crowd scene.

3.2 Loss Function

To reinforce the consistency of density levels in the global and local areas, we adopt Euclidean loss and multi-scale density level consistency loss proposed in [4]. Let L_e denote Euclidean loss, which can be defined as follows:

$$L_e = \frac{1}{N} \left\| G(X_i;\theta) - D_i^{GT} \right\|_2^2 , \tag{1}$$

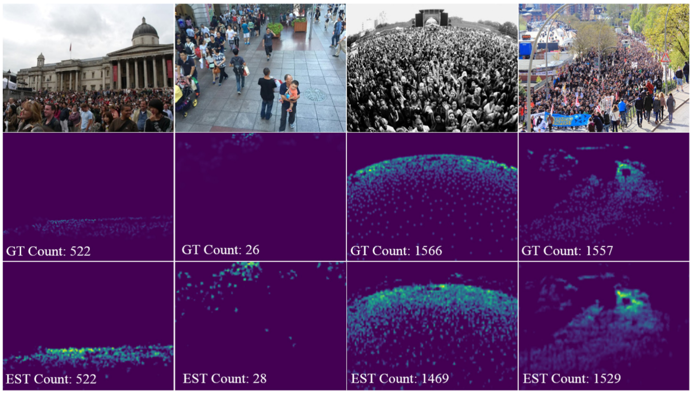

Fig. 4. Visualization of the estimated density maps generated by the proposed SMFNet and description of the counting results. From the first row to the third row represent the crowd images, corresponding ground truth maps and estimated density maps from ShanghaiTech PartA [9], ShanghaiTech PartB [9], UCF_CC_50 [13] and UCF-QNRF [14] datasets respectively.

where N is the number of images in the batch, $G(X_i;\theta)$ is the estimated density map for image X_i with parameter θ, D_i^{GT} is the ground truth for X_i. Let L_c denote multi-scale density level consistency loss, which can be defined as follows:

$$L_c = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{S}\frac{1}{k_j^2}\left\|P_{ave}(G(X_i;\theta),k_j) - P_{ave}(D_i^{GT},k_j)\right\|, \tag{2}$$

where S is the number of scale levels, P_{ave} is average pooling with parameter k_j that is the output size of average pooling. Following the set of [4], we employ three output levels (1, 2, 4) of average pooling to estimate consistency of the density levels in global and local areas. Final objective function can be formulated as:

$$L = L_e + \alpha L_c, \tag{3}$$

where α is the parameter that provides tradeoff between these two loss functions. For ShanghaiTech PartA [9] and UCF-QNRF [14], α is set as 1000, while 100 is applied for ShanghaiTech PartB [9] and UCF_CC_50 [13].

4 Experiments

4.1 Results and Comparison with State-of-the-Art

We perform experiments on three popular datasets and compare it with existing state-of-the-art crowd counting methods. The experimental results are listed in Table 1 for ShanghaiTech PartA [9], ShanghaiTech PartB [9], UCF_CC_50 [13] and UCF-QNRF [14] respectively.

Table 2. Estimation errors for distinct modules of our proposed SMFNet on ShanghaiTech PartA. The figure in parentheses represents the number of corresponding modules.

Combination of different modules	MAE	RMSE
Common Conv3x3(1,2)	62.9	102.9
MDCB(1)	61.3	101.4
MDCB(1,2)	59.2	98.8
MDCB(1,2,3)	60.8	101.7
Common Conv3x3(1,2) + UFS(1,2)	58.7	97.3
MDCB(1) + UFS(1)	59.5	96.5
MDCB(1,2) + UFS(1,2)	**57.7**	**96.1**
MDCB(1,2,3) + UFS(1,2,3)	60.4	97.2

For ShanghaiTech PartA, our model achieves the best result on MAE and comparable result on RMSE, and we get 7.3% MAE improvement compared with the state-of-the-art method CAN [8]. For ShanghaiTech PartB, ADCrowdNet [5] achieves the best MAE 7.7 and CAN achieves [8] the best RMSE 12.2. On UCF_CC_50, SMFNet achieves 6% and 3% improvement for MAE and RMSE respectively compared with the third best approach HACNN [17]. On UCF-QNRF, SMFNet achieves the lowest results which delivers 10.2% lower MAE and 8.9% lower RMSE than state-of-the-art method CAN [8]. As for highly congested scenes like ShanghaiTech PartA and UCF-QNRF, it can be observed that the SMFNet outperforms all existing methods. Figure 4 shows several examples from testing set in ShanghaiTech PartA, ShanghaiTech PartB, UCF_CC_50 and UCF-QNRF datasets and proves the good performance of our proposed method SMFNet in counting people.

4.2 Ablation Experiments on ShanghaiTech PartA

Our network architecture consists of backbone network [12], multi-scale dilated convolution block (MDCB) and U-shape fusion structure (UFS). To demonstrate their effectiveness and find the number of blocks, we conduct experiments by gradually adding these components one by one. In addition, to make full use of the different stage features in network backend, the number of UFS is the same as the number of blocks. The experimental results are shown in Table 2.

In Table 2, we firstly adopt two common convolutional layers with 3x3 kernel size and a 1×1 convolution after backbone network as baseline model, where MAE is 62.9. After that, by only replacing the above two common convolutional layers with the proposed MDCB incrementally, the MAE decrease to 61.3, 59.2 and 60.8 respectively. Above experiments could illustrate that the features with different receptive fields caused by MDCB are beneficial to count crowd accurately.

To illustrate the effectiveness of U-shape fusion structure, we embed two proposed UFS to enrich the baseline model, the MAE can reach 58.7 that has decreased by 4.1 compared with the baseline. Moreover, when the UFS are gradually adding to above three networks with different number of MDCB, the MAE decrease to 59.5, 57.7 and 60.4 respectively. Obviously, it also can be noticed that when UFS is embedded between the various layers of backend network, all the results are improved. Hence, we can proof the validity of the U-shape fusion organization, which helps draw rich and effective multi-scale features by multi-stage fusion. According to experimental results, the combination of two MDCB and two UFS achieves the best performance, so we choose it as our model.

5 Conclusion

In this paper, we have proposed a novel Scale-aware Multi-stage Fusion Network called SMFNet for more robust crowd counting in the highly congested noisy scenes. To deal with scale diversity problem, we design a multi-scale dilated convolution block (MDCB) to capture different scales of features. To further optimize and enrich extracted multi-scale features, we put forward a U-shape fusion structure that progressively integrates features output by multiple stages, while filtering low-level noise features. Experimental results have indicated the effectiveness and robustness of our proposed SMFNet, which achieves the advanced counting performance on highly congested scenes datasets such as ShanghaiTech PartA and UCF-QNRF datasets, and comparable results on ShanghaiTech PartB and UCF_CC_50 datasets.

References

1. Sindagi, V.A., Patel, V.M.: Generating high-quality crowd density maps using contextual pyramid CNNs. In: ICCV, pp. 1861–1870 (2017)
2. Cao, X., Wang, Z., Zhao, Y., Su, F.: Scale aggregation network for accurate and efficient crowd counting. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018, Part V. LNCS, vol. 11209, pp. 757–773. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01228-1_45
3. Li, Y., Zhang, X., Chen, D.: CSRNET: dilated convolutional neural networks for understanding the highly congested scenes. In: CVPR, pp. 1091–1100 (2018)
4. Dai, F., Liu, H., Ma, Y., Cao, J., Zhao, Q., Zhang, Y.: Dense scale network for crowd counting. arXiv preprint arXiv:1906.09707 (2019)
5. Liu, N., Long, Y., Zou, C., Niu, Q., Wu, H.: ADCrowdNet: an attention-injective deformable convolutional network for crowd understanding. In: CVPR, pp. 3225–3234 (2019)
6. Sindagi, V.A., Patel, V.M.: Multi-level bottom-top and top-bottom feature fusion for crowd counting. In: ICCV, pp. 1002–1012 (2019)
7. Jiang, X., Xiao, Z., Zhang, B., Zhen, X., Cao, X., Doermann, D.: Crowd counting and density estimation by trellis encoder-decoder networks. In: CVPR, pp. 6133–6142 (2019)
8. Liu, W., Salzmann, M., Fua, P.: Context-aware crowd counting. In: CVPR, pp. 5099–5108 (2019)
9. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: CVPR, pp. 589–597, June 2016

10. Sam, D.B., Surya, S., Babu, R.V.: Switching convolutional neural network for crowd counting. In: CVPR, pp. 4031–4039 (2017)
11. Szegedy, C., et al.: Going deeper with convolutions. In: CVPR, June 2015
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
13. Idrees, H., Saleemi, I., Seibert, C., Shah, M.: Multi-source multi-scale counting in extremely dense crowd images. In: CVPR, pp. 2547–2554 (2013)
14. Idrees, H., Saleemi, I., Seibert, C., Shah, M.: Composition loss for counting, density map estimation and localization in dense crowds. In: ICCV, pp. 532–546 (2018)
15. Zeng, X., Wu, Y., Hu, S., Wang, R., Ye, Y.: DSPNet: deep scale purifier network for dense crowd counting. Expert Syst. Appl. **141**, 112977 (2020)
16. Wang, S., Lu, Y., Zhou, T., Di, H., Lu, L., Zhang, L.: SCLNet: spatial context learning network for congested crowd counting. Neurocomputing **404**, 227–239 (2020)
17. Sindagi, V.A., Patel, V.M.: HA-CCN: hierarchical attention-based crowd counting network. TIP **29**, 323–335 (2020)