Multimedia Tools and Applications (2021) 80:6661–6674
https://doi.org/10.1007/s11042-020-10002-8

Multi-scale and multi-column convolutional neural network for crowd density estimation

Lei Chen[1] · Guodong Wang[1] · Guojia Hou[1]

Received: 7 January 2020 / Revised: 21 August 2020 / Accepted: 29 September 2020 /
Published online: 20 October 2020

Abstract
In order to accurately identify objects of different sizes, we propose an efficient Multi-Scale and Multi-Column Convolutional Neural Network (MSMC) to estimate the crowd density. On the one hand, the ground truth is generated based on the existed label information. On the other hand, the image is fed into our model to find the relationship between the ground truth and the predicted density map. The network is composed of three components: feature extraction, feature fusion and feature regression. First, VGG16 is utilized for faster feature extraction. Second, different sizes layers from VGG16 are fused, which helps the detection of objects with different sizes. Third, we apply multi-channel convolution to further solve the issue of multi-sizes. After the fusion block, the dilated convolution is employed to strengthen the receptive field without increasing the amount of parameters. In the crowd density estimation, the combination of multiple sizes and multiple channels enhances the ability of receiving information, improves the mapping ability of the original image and the density map, and promotes the accuracy of crowd density estimation. In this paper, the test results of the ShanghaiTech Dataset and UCF_CC_50 Dataset are provided in the Experiment section, which shows that the proposed method makes an excellent performance in both accuracy and robustness.

Keywords Convolutional neural network · Density map · Multi-channels · Dilated convolution

1 Introduction

In some special festivals and special occasions, overcrowded people will encounter unexpected losses. Whether the stampede of the Lantern Festival in Beijing fifteen years ago or

✉ Guodong Wang
doctorwgd@gmail.com

[1] College of Computer Science and Technology, Qingdao University, Qingdao, People's Republic of China 266071

the stampede of religious activities in Ningxia five years ago, the stampede has been existing all the time. Therefore, real-time prediction of crowd density is of great significance in preventing accidents. If the crowd density can be predicted in time, some measures can be taken in advance to avoid the terrible situation. For example, in schools, shopping malls and railway stations, relevant departments evacuate the crowds in advance. And at tourist attractions, visitors can enter the spots in batches under the guidance. Of course, timely prediction of crowd density not only efficiently save human from risks. In the mall, real-time understanding of passenger flow and solve the existing problems can identify potential customers and increase economic benefits. At present, the popularity of monitoring systems provides a large amount of data in the direction of crowd density estimation. At the same time, it also brought many problems, such as small objects detection at long distances, light changes, occlusion, etc. Therefore, it is of great importance to improve the accuracy of detection.

In our model, the VGG16 is used as the backbone to extract feature information of image, while applying the same size kernel in the whole model. On the basis of VGG16, a wider range of fused feature maps is obtained through different sizes of layers. During image feature extraction, the use of max-pooling could reduce image resolution and lose. Without max-pooling, it will be difficult to learn global information. So max-pooling will be the key to global information. And a convolution kernel with a larger size in the middle layer of the network layer could increase the receptive field while the calculation burden increases dramatically [36]. Therefore, we choose the dilated convolution [44] to make the original close kernels "fluffy" while increase no calculation burden. The point is the convolution kernels need to calculate are stable, which means, the positions of the fluffy kernels are filled with 0 but calculated according to the original calculation method of the convolution kernels.

The MCNN [49] model divides the whole feature learning into three channels to estimate various objects, corresponding to the targets of large, medium, and small. Then, the combined features obtained from the three channels are dimensional reduced to gain a density map. However, the image feature processing of MCNN is too shallow and the feature information is not mature which make it impossible to include all targets in multi-channel estimation. Due to the less parameter number and shallow depth, MCNN cannot fully extract feature details and consequently barely able to reflect the real crowd density. In contrast, the proposed method based on the mature features extracted from VGG16, and then the feature maps is concluded with the use of multi-layer semantic fusion, finally the multi-channel estimation comes into effect. CSRNet choose VGG16 as the backbone network and incorporate dilated convolution to obtain deeper feature information. However, it ignores the information in the hidden layers while keep going deeper. In crowd density estimation, there always exists targets which takes only a few pixels. They are tend to lose during the three times maxpooling operation in VGG16. Based on these consideration, we propose our MSMC network.

The contributions of this paper are listed as follow:

- Fuse feature maps of multiple size of VGG16 and perform regression in different levels of fusion.
- Three channels are used to process targets of different sizes in the regression.
- Employed dilated convolution to change the receptive field without increasing the calculate burden.

2 Related work

There have been a large number of studies on crowd density estimation [3, 17, 30, 35]. According to the existed related algorithms, traditional crowd density estimation methods are mainly divided into the following categories: video-based methods, detection-based methods, regression-based methods and density-map-based methods [26], and the Convolutional Neural Networks (CNN) [8], which is the popular method [40, 46, 47] recently (Fig. 1).

Video-based detection [37] with a sequence of consecutive frames in the video estimates the number of pedestrians based on the movement of the crowd and the characteristics of the human body, in the meanwhile, they also isolate the background and foreground from a sequence of consecutive frames and compare the foreground to the features of a person [12, 21]. But the detection is lack of the ability to calculate the still people and images.

The earliest method based on the detection is sliding window detection [11], through the specified size, the whole image is traversed to get the crowd of the corresponding window size, the sliding window increased in order to traverse and obtain the data information of people with different sizes. Such learning mainly refers to the adoption of image processing methods such as histogram oriented gradients (HOG) [9], edge extraction, and erosion expansion for SVM classifiers, boosting, and random forests. It is suitable for sparse crowds but not suitable for the calculation of dense crowds because of the occlusion and space changes, and the calculation burden is heavy. This disadvantage will be harmful to the accuracy and robustness [7, 20].

It is acknowledged that learning low-level characteristic responses and crowd numbers is a kind of regression-based approach [6]. The basic features are first extracted, such as background separation, edge extraction and texture processing. And linear regression, ridge regression and Gaussian process regression are utilized to get the number of people. However,

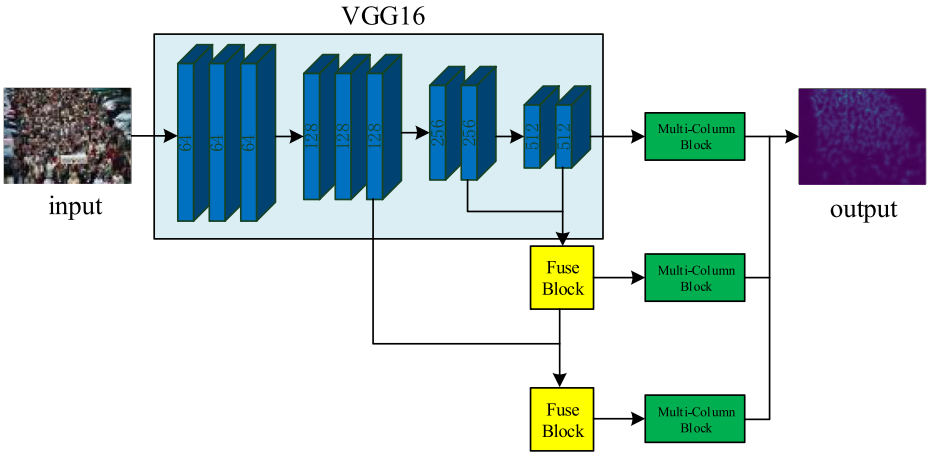

Fig. 1 The structure of the proposed Multi-Scale and Multi-Column convolutional neural network (MSMC) for crowd density map estimation. Feature maps of different sizes are extracted from VGG16, and fused to obtain the feature map contains shallow detail information and deep semantic information. Finally, a more accurate density map is gained via the Multi-Column Block

the disadvantage of the method cannot be ignored, and the spatial information of the crowd deserve more attention.

Lempitsky et al. [16] proposed density map with the guidance of the mapping of low-level features and crowd numbers. The image and the density map are linearly mapped through studying the characteristics of the crowd. The distribution of the crowd can be seen from the density map and the density of the crowd can be calculated as well.

The methods based on CNN [18, 23, 27, 43] make great progress [4, 25, 41, 42] compared with the traditional feature extraction. It can be concluded that the traditional method uses a filter to operate the image, such as smooth the image by mean filtering, extract the edges of the image by bilateral filtering, and implement shape detection and texture analysis by morphological filtering. Under these circumstances, the convolution kernel could obtain image features through self-learning.

LeNet [15] structure including convolutional layer, pooling layer, fully connected layer, the basic components of modern CNN networks are relatively complete. AlexNet designed by Hinton and his student Alex in 2012 [14] won the ImageNet competition and refreshed the record of image classification. Besides, they also made the position of deep learning in computer vision established in one fell swoop. AlexNet employed the non-linear activation function ReLU to prevent over fitting. VGG-Net is proposed by Professor Andrew Zisserman's group (Oxford) and won the first and second place on the two issues of ILSVRC localization and classification in 2014 [34] respectively. VGG-Net is different from AlexNet: the former used 16 or 19 layers while latter only has 8 layers. Replace several larger convolution kernels (11*11, 5*5) in AlexNet with several consecutive 3*3 convolution kernels. A small convolution kernel rather than a large one is reasonable for a given receptive field (the local size of the input picture related to the output), because multiple non-linear layers increase the network depth and ensure complex learning Mode with the fact that the cost burden is light and has less parameter. For example, three 3*3 convolution kernels with a stride size of 1 continuously act on a receptive field of size 7 with a total parameter of $3*(9C^2)$. If a 7*7 convolution kernel used directly, the total parameter is $49*C^2$, where C refers to the number of input and output channels. And the 3*3 convolution kernels maintain the image properties better. Many crowd density estimation methods have the VGG16 as the backbone, such as Switching CNN [32], L2R [24], CSRNet [19]. In the field of crowd density estimation, in order to identify small targets, MCNN introduces multi-channel convolution. MCNN makes multi-channel convolution popular in the field of crowd density estimation. Scale Aggregation Network (SaNet) [5] achieved the multi-scale feature aggregation effect through multiple multi-channel convolution fusion. PACNN [22] used a combination of four channels to make the density map more informative. MVMS [45] utilized multiple channels to extract feature maps with different angles to improve accuracy. AMDCN [10] employed dilated convolutions to increase receptive fields in the network. SaCNN [48] combined shallow features with deep features to reduce information loss.

3 Method

In this section, we will introduce our Convolutional Neural Network from two parts. First, we will describe the backbone and dilated convolution. Secondly, we will introduce the architecture of the MSMC.

3.1 Backbone and dilated convolution

3.1.1 Backbone

Image features can be extracted based on VGG16 since the structure of VGG16 is very simple. Compared to AlexNet, the entire network uses the same size convolution kernel and the same size max-pooling. It contains 16 convolutions and 3 max-pooling. Reduce calculations by smaller convolution kernels and expand receptive field with continuous convolution. The required feature map can be received as soon as possible through the VGG16's flexible architecture and strong learning ability. In CSRNet [19], the author uses the last layer of VGG16 for density estimation, which leads to the neglect of many details. In order to make the final density map more accurate, this paper fuses multiple feature layers of VGG16 on the basis of CSRNet.

3.1.2 Dilated convolution

In order to further process the extracted feature map, dilated convolution is adopted here. Dilated convolution [38] can increase the receptive field without increasing the amount of parameters and loss of feature information. As shown in the Fig. 2, (a) corresponds to the dilated convolution with dilated rate of 1, and a normal 3*3 convolution kernel is a special dilated convolution with dilated rate of 1. Figure 2 (b) represents to the dilated convolution with dilated rate of 2, and the receptive field is 5*5. Its parameters are 36% (9/25) of the normal convolution kernels with 5*5 size. Figure 2 (c) shows the dilated convolution with dilated rate of 3, and the receptive field is 7*7. Its parameters are 18% (9/49) of the normal convolution kernel with 7*7 size. The n*n dilated convolution kernel correspond to (2n-1)*(2n-1) receptive field. Overall, dilated convolution can effectively reduce the amount of parameters, optimization computing resources.

3.2 Structure

Affected by distance, angle and perceptivity, various sizes of human heads appeared in the dataset. It is difficult to distinguish the large target at a shallow layer, since the details acquired from small one can lose information in deep layer due to the max-pooling operation, which

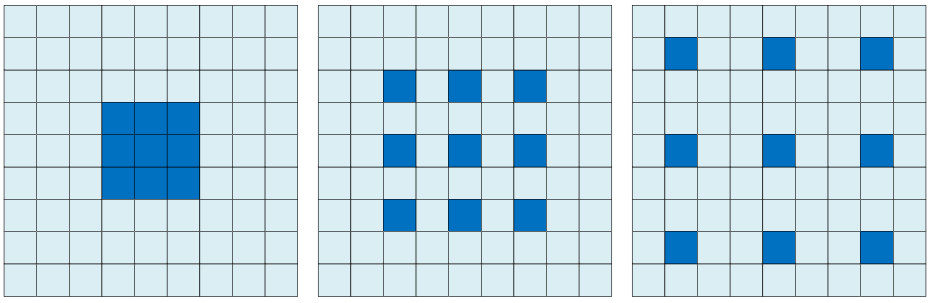

Fig. 2 3 × 3 convolution kernels with different dilated rate as 1(**a**), 2(**b**) and 3(**c**)

makes a contradiction. It is essential to analyze with more parameters and get more semantic information from deeper feature map. Moreover, different images have different sizes so detailed features from shallow layers is needed through the occlusion of people. Dilated convolution can be incorporated to identify spatial variants and the CNN cannot discriminate large-scale spatial variants simultaneously. In summary, a convolution neural network is proposed based on VGG16 without fully-connected layers, as shown in Fig. 1. The receptive field of each feature point in the output layer reached the maximum, and the semantic information is much larger than the characteristic information which both lead to the suitable details missing from results obtained from the estimated density map. Therefore, layers of different sizes from VGG16 can be combined. The VGG16 is divided into three parts and the last output layer of VGG16 as the first part. The combination of the final output layer up-sampling and the feature map before max-pooling is the second part. Finally, the third part consists of the combination of the former two parts and the feature map before max-pooling.

In the Fuse Block, in order to optimize the amount of parameters, the dimensions of feature map are reduce to 128. Then, we defined as follows:

$$fuse = 2*up + main \tag{1}$$

where *up* mean the upper feature map and *main* is the layer before maxpooling. The *up* is multiplied by 2 to align with the *main*' dimension. Detailed semantic information can be combined to strengthen features through the addition.

With reference to [28, 39, 40], in order to further improve the recognition rate of targets of various sizes. Multi-column dilated convolution is used in post processing. As shown in Fig. 3, the blue map represents dilated rate of 1 (as the Fig. 2 (a) shown) and yellow map means the rate of 2 (as the Fig. 2 (b) shown). The feature maps here are all mature feature information, it is unnecessary to do much convolution. There are only 7 convolutions, 2 of them are dilated convolutions. When the dilated rate is 2, it is equivalent to a normal convolution kernel of 5. Here we have three combinations: 3, 5*3, and 5*5*3. Finally, by connecting them, the diversified 256 dimensions feature maps can be obtained.

In a word, there are four main innovations in the proposed MSMC. Using VGG16 to extract features, the entire network uses the same size convolution kernel size (3*3) and max-pooling size (2*2), and the structure is simple. Fusion features are enhanced on feature maps of

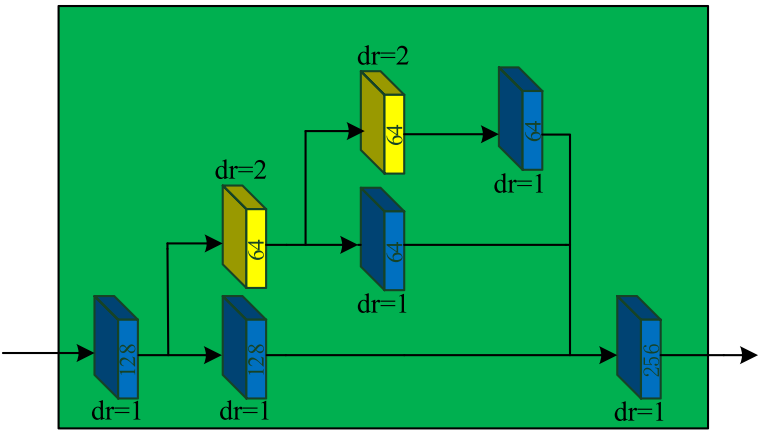

Fig. 3 The Multi-Column Block of MSMC (dr mean dilated rate)

different scales to contain semantic information without the lack of detailed features. Using multi-column to process targets of different sizes, different targets can be processed from different receptive fields. Use dilated convolution to increase the receptive field while optimizing parameters to reduce the amount of calculation.

3.3 Density map

Density maps are generated by the existed label. Each points in the tag information represents the center position of a human head. Then, a geometrically adaptive Gaussian kernel is used to generate a density map. The sum of the numbers in the density map is the number of people. It can be expressed as:

$$H(x) = \sum_{i=1}^{N} \delta(x - x_i) \tag{2}$$

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_{\sigma_i}(x), with\ \sigma_i = \beta \overline{d_i} \tag{3}$$

where x_i is the center position of a human head, N stands for the number of human. Then density map is generated by the Gaussian kernel G_{σ_i}. β represents a constant. $\overline{d_i}$ is the average sum of the Euclidean distance sum of the head from the k adjacent head in the image.

4 Experiment

4.1 Training details

In order to reduce the distance between the ground truth and the estimated density map which generated by our model, the Euclidean distance is introduced. The loss function is given as follow:

$$L(\theta) = \frac{1}{2N} \sum_{i=1}^{N} \| P(I_i; \theta) - G_i \|_2^2 \tag{4}$$

where N is the number of each batch during training. I_i is the image fed into the convolution neural network and θ represents a series of parameters obtained from the model, that is, the convolution kernel used. $P(I_i; \theta)$ is the predicted density map trained by our network and G_i stands for the ground truth of the fed image I_i. $L(\theta)$ demonstrates the loss between predicted density map and the ground truth.

4.2 Evaluation metric

The MAE and the MSE are adopted to evaluate the correctness of prediction density map, which defined as:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left| C_i - C_i^{GT} \right| \tag{5}$$

$$MSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left| C_i - C_i^{GT} \right|^2} \tag{6}$$

Table 1 Information about the dataset

Dataset		Count	Total	Average	Minimum	Maximum	Size
Shanghaitech	Part_A	300 + 182	241,677	501.4	33	3139	Variously
	Part_B	400 + 316	88,488	123.6	9	578	768 × 1024
UCF_CC_50 Dataset		50	63,075	1280	94	4543	Variously

where N is the number of the test images. C_i represents for the predicted density by our network. C_i^{GT} stands for the ground truth of the test images (Tables 1, 2 and 3).

4.3 Data augmentation

In order to get better prediction results, we augment the training dataset. First, in order to be able to contain all the content in the image, each image is divided into four quarters according to the four positions of upper left, upper right, lower left, and lower right; then, to reflect the diversity of data augmentation, four copies was randomly intercepted; finally, there are nine images including the original image. Of course, the corresponding mark information (.mat file) is also divided according to the corresponding range.

4.4 ShanghaiTech dataset

ShanghaiTech Dataset has 1198 labeled images in total. The dataset is divided into two parts, Part_A and Part_B. The images in Part_A are denser and more difficult than the images in Part_B. This dataset was first established in MCNN [49]. 300 images of Part_A was used for training and 182 images was used for testing. These images were randomly selected from the Internet, which is more universal. 400 images of Part_B was used for training and 316 images for testing. These images were taken on the streets of the Shanghai metropolis. Part_A is more challenging with different scene types, different density levels, different scales and perspective distortion.

The amount of Part_A data is relatively large, and the crowd density distribution spans 3000. In order to converge faster to the needed result, SGD is used for optimization here. SGD is to calculate the gradient of the mini-batch every iteration, and then update the parameters. It is defined as follows:

Table 2 The result on ShanghaiTech Dataset with different method

	Part_A		Part_B	
Method	MAE	MSE	MAE	MSE
TDF-CNN [31]	97.5	145.1	20.7	32.8
IG-CNN [2]	72.5	118.2	13.6	21.1
GSP [1]	70.7	103.6	9.1	15.9
L2R [24]	73.6	112	13.7	21.4
IC-CNN [29]	68.5	116.2	10.4	16.7
CSRnet [19]	68.2	115.0	10.6	16.0
MSMC-v1 (ours)	70.4	118.2	10.1	15.7
MSMC-v2 (ours)	68.1	114.1	9.2	15.0
MSMC (ours)	66.9	108.7	9.1	13.8

Table 3 The result of UCF_CC_50 Dataset

	Sec1	Sec2	Sec3	Sec4	Sec5	Avg
MAE	228.9	251.2	253.2	280.6	248.0	252.4
MSE	359.7	331.6	314.7	401.2	309.8	343.4

$$g_t = \nabla_{\theta_{t-1}} f(\theta_{t-1}) \tag{7}$$

$$\Delta\theta_t = -\eta * g_t \tag{8}$$

Here, η refers to the learning rate, and g_t is the gradient. SGD is completely dependent on the gradient of the current batch, so η can be understood as how much the gradient of the current batch is allowed to affect parameter updates. SGD easily converges to a local optimum and may be trapped in the saddle point in some cases. In the training of Part_B, we choose Adam as optimization method. Adam (Adaptive Moment Estimation) is essentially an RMSprop with a momentum term. It uses the first and second moment estimates of the gradient to dynamically adjust the learning rate of each parameter. The main advantage of Adam is that after the offset correction, the learning rate of each iteration has a certain range, which makes the parameters relatively stable. The fluctuation during training is small, and the overall trend is declining.

The experimental results are shown in Table 2. From Part_A, we can see that our model is better than TDF-CNN [31], IG-CNN [2], L2R, GSP [1], IC-CNN [29] and CSRNet, and also has better stability. The MAE we obtain is 1.9% lower than CSRNet, the accuracy rate is 30.2% higher than that of TDF-CNN. From Part_B, it can be clearly seen that this model is superior to other models in terms of prediction results and stability. Whether MAE or MSE, the results of our method are lower than the other six methods. In Table 2, MSMC-v1 refers to the network structure where Fuse Block is not used during the experiment, and MSMC-v2 refers to the network structure where Fuse Block is used once during the experiment, that is, the feature maps of the last two layers are merged. In the results of Part-A, MSMC-v2 is 2.3 higher than MSMC-v1, and MSMC is 1.2 higher than MSMC-v2. Through three comparison experiments, it can be seen that the accuracy of the density map obtained by fusing two feature maps of different sizes is higher than that of the network structure without feature map fusion or only one fusion.

Table 4 The result on UCF_CC_50 Dataset with different method

Method	MAE	MSE
TDF-CNN [31]	354.7	491.4
IG-CNN [2]	291.4	349.4
L2R [24]	279.6	388.9
CSRNet [19]	266.1	397.5
IC-CNN [29]	260.9	365.5
ADCrowdNet (DME) [33]	257.1	363.5
PACNN [22]	267.9	357.8
MSMC (ours)	252.4	343.4

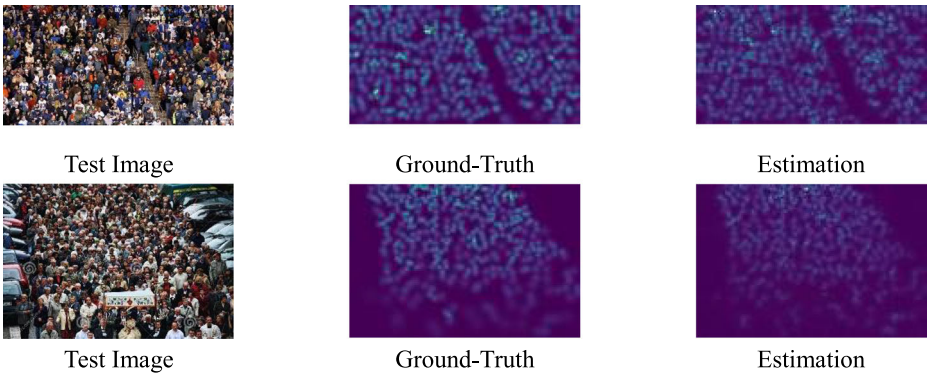

Fig. 4 The density map of Part_A

4.5 UCF_CC_50 dataset

The UCF_CC_50 Dataset contains scenes of various densities and various rallies from different perspectives such as concerts, protests, marathons, speeches, etc. and contains 50 images of different resolutions that each image contains an average of 1280 people. The number of people varies from 94 to 4543 and the density varies widely. A total of 63,075 people are tagged in the entire dataset. 50 pictures was copied into 5 groups for cross-experiment due to the lack of images. Each group takes 40 of them for training and 10 for testing, as a reminder, the 10 taken each time are different. The test set was augmented and each test set contained 360 images. Table 3 shows the MAE and MSE of our model test UCF_CC_50 Dataset. Table 4 compares the results of the model with several other models. It can be seen from Table 4 that the prediction result of the method is 4.7 higher than the best 2019 ADCrowdNet in the table, and 102.3 higher than the 2018 TDF-CNN.

Figures 4, 5 and 6 show the comparison between the estimated density map and the ground truth. The “test image” is the original image. As we can see from the pictures listed, the estimated density distribution and density of the density map are basically the same as the true

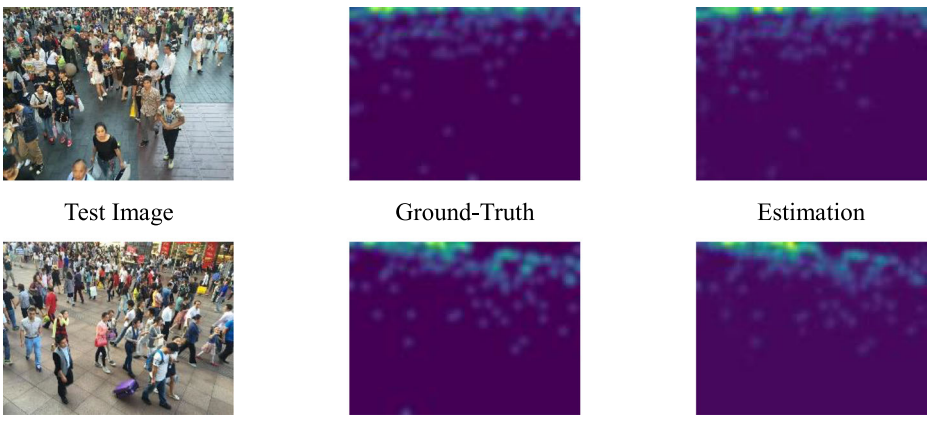

Fig. 5 The density map of Part_B

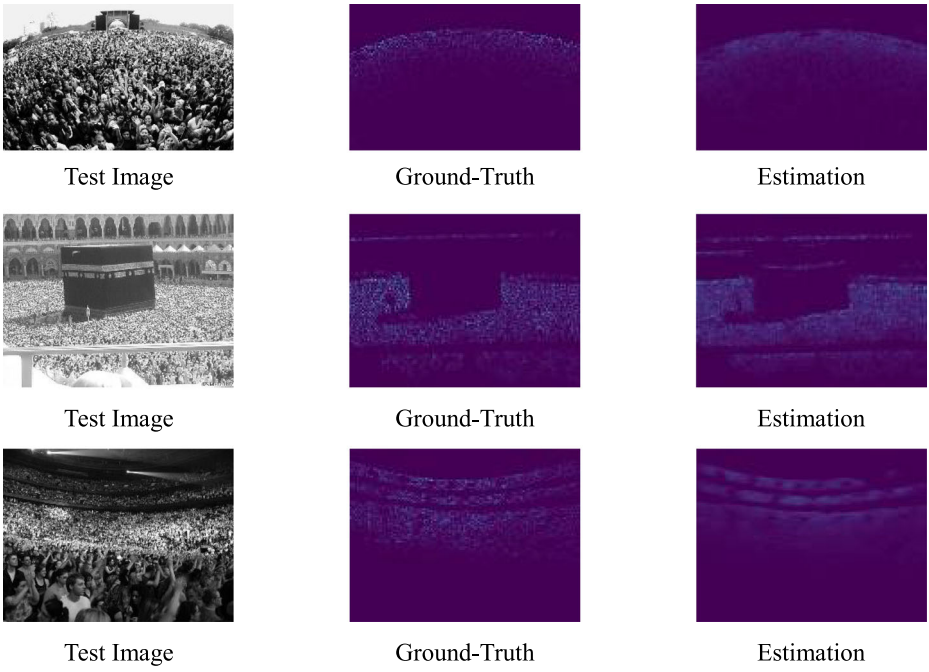

Fig. 6 The density map of UCF_CC_50 Dataset

value. In terms of distribution or density, the density map estimated by the model is consistent with the Ground-Truth.

5 Conclusion

This paper proposes a multi-scale and multi-channel dilated convolution network for crowd density estimation. The network obtains information-rich fused feature layers through the mutual fusion of feature map of different sizes. The fused feature map contains low-level feature details and high-level semantic features. Moreover, the reuse of low-level features prevents from fitting and increases the information content of feature map [13]. In order to be able to detect targets of different sizes, multi-channel training is performed on the basis of fused feature maps. Different channels use receptive fields of different sizes. Here, dilated convolution is adopted to change the size of the receptive fields. The comparison of the experimental results of ShanghaiTech Dataset and UCF_CC_50 Dataset proves that MSMC has more superior performance in estimating crowd density. However, after a lot of experiments, some problems are found. That is the data is pooled three times through the network and becomes 1/8 of its original size. If there are targets that only occupy a few pixels in the data, they will be lost in the pooling process. Future work will consider extracting deeper semantic information at the shallow feature layer.

Acknowledgements The research work is supported by the Natural Science Foundation of Shandong Province, China (No. ZR2019MF050, ZR2019BF042), National Natural Science Foundation of China (No. 61901240).

References

1. Aich S, Stavness I (2019) Global sum pooling: a generalization trick for object counting with small datasets of large images. In Proc. IEEE Conf. CVPR, pp. 73–82
2. Babu Sam D, Sajjan NN, Venkatesh Babu R, et al (2018) Divide and grow: capturing huge diversity in crowd images with incrementally growing cnn. In Proc. IEEE Conf. CVPR, pp. 3618–3626
3. Boominathan L, Kruthiventi SSS, Babu RV (2016) Crowdnet: a deep convolutional network for dense crowd counting. In Proc.of the 2016 ACM on Multimedia Conf., ACM, pp. 640–644
4. Cai W, Wei Z (2020) PiiGAN: generative adversarial networks for pluralistic image Inpainting. IEEE Access 8:48451–48463
5. Cao X, Wang Z, Zhao Y, et al (2018) Scale aggregation network for accurate and efficient crowd counting. In Proc. ECCV, pp. 734–750
6. Chan AB, Vasconcelos N (2009) Bayesian poisson regression for crowd counting. In Proc. IEEE Conf. ICCV, pp. 545–551
7. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
8. Cireşan D, Meier U, Schmidhuber J (2012) Multi-column deep neural networks for image classification. In Proc. IEEE Conf. CVPR, pp. 3642–3649
9. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In Proc. IEEE Conf. CVPR, pp. 886–893
10. Deb D, Ventura J (2018) An aggregated multicolumn dilated convolution network for perspective-free counting. In Proc. IEEE Conf. CVPR, pp. 195–204
11. Dollar P, Wojek C, Schiele B et al (2011) Pedestrian detection: an evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell 34(4):743–761
12. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell 32(9):1627–1645
13. Hu S, Wang G, Wang Y, Chen C, Pan Z (2020) Accurate image super-resolution using dense connections and dimension reduction network. Multimed Tools Appl 79:1427–1443
14. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In Proc. NIPS, pp. 1097–1105
15. LeCun Y, Bottou L, Bengio Y et al (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
16. Lempitsky V, Zisserman A (2010) Learning to count objects in images. In Proc. NIPS, pp. 1324–1332
17. Li T, Chang H, Wang M, Ni B, Hong R, Yan S (2015) Crowded scene analysis: a survey. IEEE Trans on Circuits and Syst for Video Technol 25(3):367–386
18. Li K, Ma W, Usman S et al (2020) Object detection with convolutional neural networks. Deep Learning in Computer Vision: Principles and Applications 30(31):41–62
19. Li Y, Zhang X, Chen D (2018) Csrnet: dilated convolutional neural networks for understanding the highly congested scenes. In Proc. IEEE Conf. CVPR, pp. 1091–1100
20. Li M, Zhang Z, Huang K, et al (2008) Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection. In Proc IEEE Conf CVPR, 1–4
21. Lin SF, Chen JY, Chao HX (2001) Estimation of number of people in crowded scenes using perspective transformation. IEEE trans. On Syst. Man, and Cybernetics-Part A: Systems and Humans 31(6):645–654
22. Liu N, Long Y, Zou C, et al (2019) ADCrowdNet: an Attention-injective Deformable Convolutional Network for Crowd Understanding. In Proc. IEEE Conf. CVPR, pp. 3225–3234
23. Liu L, Ouyang W, Xiaogang W et al (2020) Deep learning for generic object detection: a survey. Int J Comput Vision 128(2):261–318
24. Liu X, van de Weijer J, Bagdanov AD (2018) Leveraging unlabeled data for crowd counting by learning to rank. In Proc. IEEE Conf. CVPR, pp. 7661–7669
25. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In Proc. IEEE Conf. CVPR, pp. 3431–3440
26. Loy CC, Chen K, Gong S, et al (2013) Crowd counting and profiling: Methodology and evaluation. In Modeling, Simulation and Visual Analysis of Crowds. Springer, pp. 347–382
27. Mahmoud H and Ali IA (2020) Deep learning in computer vision: principles and applications. CRC Press
28. Onoro-Rubio D, López-Sastre RJ (2016) Towards perspective-free object counting with deep learning. In Proc. ECCV, pp. 615–629
29. Ranjan V, Le H, Hoai M (2018) Iterative crowd counting. In Proc. ECCV, pp. 270–285
30. Revathi T and Rajalaxm TM (2020) Deep Learning for People Counting Model Soft Computing for Problem Solving, https://doi.org/10.1007/978-981-15-0035-0_43

31. Sam DB, Babu RV (2018) Top-down feedback for crowd counting convolutional neural network. Thirty-Second AAAI Conf on AI
32. Sam DB, Surya S, Babu RV (2017) Switching convolutional neural network for crowd counting. In Proc. IEEE Conf. CVPR, pp. 4031–4039
33. Shi M, Yang Z, Xu C, et al (2019) Revisiting perspective information for efficient crowd counting. In Proc. IEEE Conf. CVPR, pp. 7279–7288
34. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv
35. Sindagi VA, Patel VM (2017) Generating highquality crowd density maps using contextual pyramid CNNs. In Proc. IEEE Conf. CVPR, pp. 1861–1870
36. Vedaldi A, Jia Y, Shelhamer E, et al (2014) Caffe: Convolutional architecture for fast feature embedding. In Proc.of the 22nd ACM International Conf. on Multimedia. ACM, pp. 675–678
37. Viola P, Jones MJ, Snow D (2005) Detecting pedestrians using patterns of motion and appearance. Int J Comput Vis 63(2):153–161
38. Wang Y, Hu S, Wang G et al (2020) Multi-scale dilated convolution of convolutional neural network for crowd counting. Multimed Tools Appl 78(11):1057–1073
39. Wang Y, Wang G, Chen C, Pan Z (2019) Multi-scale convolution of convolutional neural network for image denoising. Multimed Tools Appl 78:19945–19960
40. Wang Z, Zou C, Cai W (2020) Small sample classification of Hyperspectral remote sensing images based on sequential joint Deeping learning model. IEEE 8:71353–71363
41. Wei Y, Feng J, Liang X, et al (2017) Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In Proc. IEEE Conf. CVPR, pp. 1568–1576
42. Wei Y, Liang X, Chen Y, Shen X, Cheng MM, Feng J, Zhao Y, Yan S (2017) Stc: a simple to complex framework for weaklysupervised semantic segmentation. IEEE Trans Pattern Anal Mach Intell 39(11):2314–2320
43. You H, Tian S, Yu L, Lv Y (2020) Pixel-level remote sensing image recognition based on bidirectional word vectors. IEEE Trans Geosci Remote Sens 58(2):1281–1293
44. Yu F, Koltun V (2016) Multi-scale context aggregation by dilated convolutions. In Proc. ICLR
45. Zhang Q, Chan AB (2019) Wide-Area Crowd Counting via Ground-Plane Density Maps and Multi-View Fusion CNNs. In Proc. IEEE Conf. CVPR, pp. 8297–8306
46. Zhang C, Kang K, Li H, Wang X, Xie R, Yang X (2016) Data-driven crowd understanding: a baseline for a large-scale crowd dataset. IEEE Trans Multimedia 18(6):1048–1061
47. Zhang C, Li H, Wang X, et al (2015) Cross-scene crowd counting via deep convolutional neural networks. In Proc. IEEE Conf. CVPR, pp. 833–841
48. Zhang L, Shi M, Chen Q (2018) Crowd counting via scale-adaptive convolutional neural network. In Proc. IEEE Conf. WACV, pp. 1113–1121
49. Zhang Y, Zhou D, Chen S, et al (2016) Single-image crowd counting via multi-column convolutional neural network. In Proc. IEEE Conf. CVPR, pp. 589–597

Lei Chen obtained his B.Sc. degree from Qingdao University in 2018. Now he is a master degree candidate in Qingdao University. His research interests include machine learning and image processing.

Guodong Wang Male, Born in Weifang City, Shandong Province, China, in February, 1980. Now he is an associate professor in College of Computer Science and Technology, Qing University. He received bachelor degree in 2001 and master degree in 2004 in control theory and control engineer, Qingdao University of Science and Technology, and received Ph. D degree in pattern recognition and intelligent system in Huazhong University in 2008. His research Interest include: Variational Image Science, Face recognition, Intelligent video survillance, 3D reconstruction and Medical image processing and Analysis.

Guojia Hou is now an assistant professor in the College of Computer Science and Technology, Qingdao University. He received his BS degree in computer science in 2010 and his MS and PhD degrees in computer applications technology from the Ocean University of China in 2012 and 2015, respectively. He is the author of more than 20 journal and conference papers. His current research interests include image processing, image quality evaluation and pattern recognition.