A high-robustness and low resource-consumption crowd counting model

Han Jia, Xuecheng Zou
School of Optical and Electronic Information, Huazhong University of Science and Technology
Luoyu Road 1037, Wuhan, Hubei 430074
China

Received: December 30, 2020. Revised: January 18, 2021. Accepted: January 21, 2021.
Published: January 29, 2021.

Abstract—A major problem of counting high-density crowded scenes is the lack of flexibility and robustness exhibited by existing methods, and almost all recent state-of-the-art methods only show good performance in estimation errors and density map quality for select datasets. The biggest challenge faced by these methods is the analysis of similar features between the crowd and background, as well as overlaps between individuals. Hence, we propose a light and easy-to-train network for congestion cognition based on dilated convolution, which can exponentially enlarge the receptive field, preserve original resolution, and generate a high-quality density map. With the dilated convolutional layers, the counting accuracy can be enhanced as the feature map keeps its original resolution. By removing fully-connected layers, the network architecture becomes more concise, thereby reducing resource consumption significantly. The flexibility and robustness improvements of the proposed network compared to previous methods were validated using the variance of data size and different overlap levels of existing open source datasets. Experimental results showed that the proposed network is suitable for transfer learning on different datasets and enhances crowd counting in highly congested scenes. Therefore, the network is expected to have broader applications, for example in Internet of Things and portable devices.

Keywords—Systems Theory, Signal Processing, neural Networks, Crowd counting, deep neural networks, convolutional neural networks, dilated convolution.

I. INTRODUCTION

The real-time, automatic analysis of crowded scenes has been broadly applied in crowd management, traffic control, and surveillance [1]. Thus, crowd counting has attracted considerable research and application interests. With development of crowd counting researches, analysis methods have been developed from simply counting to understanding saliency information [2]-[6] (such as concentration location, movement direction, etc.). Thus, traditional counting methods have faced considerable limitations due to lack of spatial coherence. Deep neural networks (DNNs) have been introduced to solve these with output of density map, and shown desirable performance in image semantic segmentation [7] and visual saliency detection [8]. Based on these researches, works based on hardware implementation including field programmable gate array (FPGA) [9]-[10] and application specific integrated circuits (ASICs) [11] enhance the feasibility of mapping DNN to surveillance devices, and validates effectiveness of DNN to deal with high-density crowd counting problems and other image processing applications [12]-[14].

For the existing crowd counting algorithms, networks based on multi-scale architectures [15]-[17] have achieved state-of-the-art performances. Among these, multi-column neural network (MCNN) [15] proposed by researchers from Shanghai Tech University is the most generally used method. However, these architectures still have several significant drawbacks. First, training multi-scale architectures is complicated and time-consuming. Second, multi-column networks are more time- and space-consuming than sequential networks without commensurate performance improvement. Moreover, the complicated architecture of MCNN networks limited its applications in hardware implementation. Except for MCNN structures, various CNN models have been proposed. Researchers from Visual Geometry Group of Oxford and Google DeepMind Co. put forward visual geometry group-net (VGG-Net). Switching convolutional neural network (Switch-CNN) is developed by researchers from Indian Institute of Science. These typical CNN architectures consists of input layers, convolutional layers, pooling layers, and fully connected layers. Such structures face various defects such as information loss; moreover, the pooling and up-sampling layers

are deterministic. Dilated convolution can be taken as a proper solution for the information loss, whose reception field of convolutional filters can be exponentially expanded.

In this study, we propose a network with a deeper CNN architecture and fewer parameters. Small convolution filters are adopted in all layers to minimize the network size. In addition, we use dilated convolution layers as back ends [18]-[19]. Thus, without losing resolution, the perception field is enlarged as contextual information is aggregated. The contributions of this study are as follows: 1) We propose a network that achieves considerable improvements in counting accuracy and density map quality compared to conventional methods; 2) we show that much better performance can be achieved with minimal resource consumption and easier end-to-end training methods, which is crucial in applications; 3) and we show that certain architectures can be robust enough to handle different datasets, making our network applicable in various scenes and aspects.

II. Related Work

Over the years, several algorithms have been proposed for crowd counting. Most early studies focused on detection-based methods that involve a sliding window detector over two consecutive frames of a video sequence to estimate the number of pedestrians [20]. In these methods, detection is realized by training a classifier to extract low-level features (such as Haar wavelets [21], histogram of oriented gradients [22], edgelet [23], and shapelet [24]) from the whole body. However, occlusions in highly-crowded scenes significantly affect the detection performance, limiting the estimation accuracy of the methods. To resolve this problem, detection methods have been proposed to train classifiers for specific body parts, which detect and count persons in particular regions and scenes [25].

Regression-based methods are also proposed solutions. These methods learn the direct mapping between the low-level features extracted from the local image blocks and head count. The low-level features are formed by various features such as foreground, edge, texture, and gradient features [26]. Idrees et al. [27] proposed a model based on these methods to extract features with multiple sources, namely Fourier analysis with head detections and SIFT interest point-based counting in local neighborhoods.

However, regression-based methods ignore important spatial information due to the regression on the global count. Therefore, Lempitsky et al. [28] suggest learning the linear mapping between the local block features and their corresponding target density maps, which may incorporate the spatial information into the learning process. Nevertheless, there has been considerable difficulty in learning a linear mapping. Pham et al. [29] proposed a method to learn a non-linear mapping using random forest regression. Based on these methods, further related crowd counting approaches have been proposed and form density estimation-based systems.

Convolutional neural networks (CNN) have achieved substantial success in computer vision tasks due to their superiority in classification and recognition fields [30]-[32]. Thus, several researchers applied CNN to learn non-linear functions from crowd images to their corresponding density maps or counts. Moreover, dilated convolution can expand the reception field without resolution loss and remedy the information loss in the pooling-layers. Thus, dilated convolution has been increasingly employed in recent CNN architectures [33]-[35]. Relevant studies can be found in [36] and [37].

Recent state-of-the-art methods have achieved significant improvements in crowd counting by using multi-column architecture and density level classifier. However, these methods still have several defects, which are stated as follows. First, training difficulty is a crucial stinker for multi-column CNN-based methods [15]-[17]. Additionally, multi-column CNNs have redundancies in their architecture, which increase resource consumption. Further, the manual classification of the density map level is inflexible; i.e., differences between the testing and training datasets will weaken the performance of the density classification network. Lastly, large numbers of parameters are occupied by density-level prediction [16]-[17]. To address these problems, we propose a single-column, deep convolutional network with dilated convolution layers.

III. Materials and Methods

Datasets are integral in the research for architectures aimed at crowd counting. Recently, the density and diversity of datasets have grown as crowd counting networks have been applied in more complicated scenes. This study was mainly based on ShanghaiTech dataset, which is a large dataset containing a total of 1,198 images and 330,165 annotated heads. This dataset consists of two parts: Parts_A and B. In Part A, there are 482 images of high-density scenes randomly chosen from the Internet. Part_B contains 716 images taken from busy streets in Shanghai. The images in Part B are sparser than those in Part_A. The crowd density significantly varies between the two subsets, making crowd estimation a challenging task. Fig. 1 contains three images from the dataset with 118 persons in each image, whereas the image scenes, camera angles, and spatial distributions are all different. Thus, it is difficult to create a model to interpret the same information presented in several ways. The density map of these three images are included in Fig. 1.

Fig. 1 Images from ShanghaiTech Part B dataset. The three images contain 118 persons each, but they also have totally different spatial distributions. The second row shows their density maps.

Other datasets we used include UCF_CC_50 and WorldExpo'10 datasets. UCF_CC_50 is a challenging dataset containing 50 images with different perspectives and resolutions. The difficulty with the dataset not only comes from its limited number of images, but also from the stark differences in the crowd count of the images. The head count ranges from

94 to 4,543 per image. The WorldExpo'10 dataset consists of 3,980 annotated frames from 1,132 video sequences captured by 108 surveillance cameras, all from Shanghai 2010 WorldExpo. The dataset was divided into two parts: training set and testing set. The training and test sets contained 3,380 and 600 frames, respectively. The images were taken from five different scenes with 120 frames per scene. All the frames of the whole dataset were masked with region of interest as preprocessing.

Based on the good performance of dilated convolution in segmentation [18]-[19], our proposed method utilizes dilated convolution in learning multi-scale contextual information. In the following subsections, we detail the architecture and training method of our proposed model.

A. Dilated convolutions for deep convolutional networks

Please submit your manuscript electronically for review as e-mail attachments.

1) Dilated convolution

We propose dilated convolution as an attractive alternative to convolution layers; this can exponentially expand the receptive field while maintaining the original resolution of the input image. Recent applications involve scenes with different perspectives; thus, perspective-free counting has been an essential task in crowd counting. An effective method of enlarging the convolutional kernel size through the networks has been employed in recent studies. This method enables the extraction of multiple-scale features. However, a defect of the method is the increase of the kernel size while extracting large-scale features: the number of parameters significantly increases, which affects further application of the network. Moreover, pooling can result in information loss. Therefore, we propose the use of dilated convolution rather than larger kernels. Fig. 2 is an example of convolutional kernels of different dilation rates. As shown in the Fig. 2, the reception field exponentially expands as the dilation rate increases.

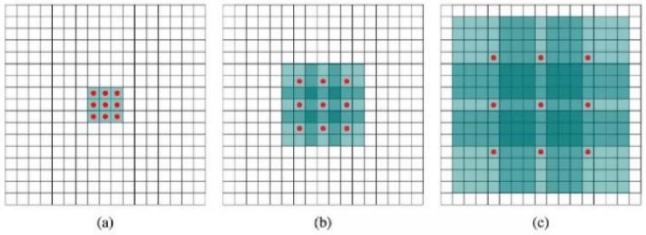

Fig. 2 A 3 × 3 convolution kernel with different dilation rates. Dilated convolution with systematic dilation rate arrangement helps expand the reception field exponentially while maintaining the resolution. (a) With a receptive field of 3 × 3, the dilation rate of normal convolution is 1. (b) Through dilated convolution with dilation rate r = 2, each convolutional kernel has a receptive field of 7 × 7. (c) Through dilated convolution with dilation rate r = 4, each convolutional kernel has a receptive field of 15 × 15.

In the dilated convolution method, filters with holes are exploited instead of pooling and convolving layers. As shown in Fig. 3, the original image goes through max-pooling, convolution (which generates a feature map that is only a quarter of its original size), and up-sampling to the original size. This process results in resolution loss. With the dilated convolution method, a feature map is generated at the original image size; hence, the resolution and spatial information can be maintained.

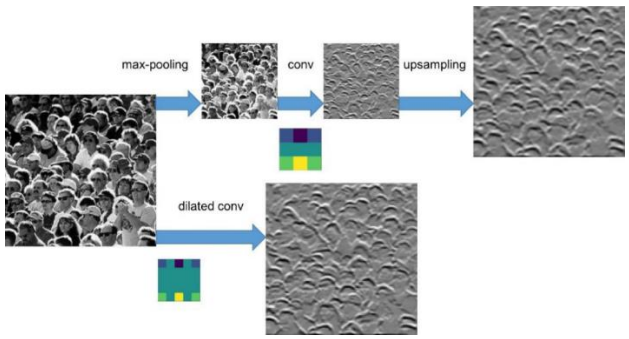

Fig. 3 Comparison between two methods: methods with dilated convolution can generate a feature map at the original image size, thus maintaining the resolution and spatial information.

2) Network architecture

As shown in Fig. 4 below, the front-end of our model consists of the fine-tuned first 10 layers from VGG-16 without dilation (we remove vestiges of the networks). The backend contains dilated layers. The frontend outputs the density map at 1/64 of its original input size, and multi-scale aggregation begins after the last max-pooling layer. As is shown in Fig.4, backend of the proposed network is consisted of 5 convolutional layers with dilation operation involved. To explore the best configuration of dilation rate in the backend, a multi-scale aggregation with a different dilation rate is implemented. The comparison between the configurations is shown in rest of the paper. The most significant performance can be reached with a dilation rate of 2 through the backend of the networks.

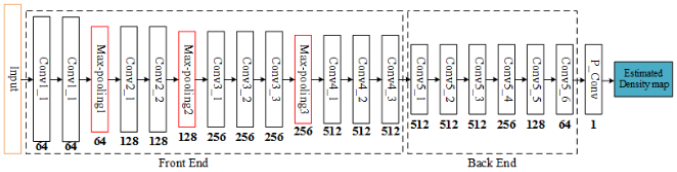

Fig. 4 Architecture of proposed network. The proposed network mainly consists of the first 10 layers of VGG-16 and the dilated layers with the dilation rate to be configured.

Considering structural similarity in Image (SSIM) [36] and peak signal-to-noise ratio (PSNR) between the ground truth and generated density map, we apply bilinear interpolation to recover feature maps at the original image resolution. The density map is resized by applying bilinear interpolation with a factor of 8.

With the proposed network architecture above, the parameter setting during training is shown in Table 1. In the process of training, stochastic gradient descent is employed with a fixed learning rate of 1e-6. Due to the limitation of testing environment, batch size is set to 1, and momentum is set to 0.95 to accelerate convergence. Weight decay is 5*1e-4 to avoid overfitting. And the number of epoch is 400 in the training. Details of training will be narrated in the next section.

Table 1. Network parameters during training.

Parameters

Learning rate	1e-6
Batch size	1
Momentum	0.95
Weight decay	5*1e-4
Epoch	400

B. Training methods

In this section, we describe our proposed easy-to-implement and fast training method. In our proposed architecture, we employed fine-tuned VGG-16 and dilated convolutional layers as frontend and backend, respectively. The network was trained with the datasets stated in Table 2, using geometry-adaptive and Gaussian kernels. Three quarters of each dataset were used as training set and the rest as testing set. Data augmentation was also employed in the training process.

Table 2. Density map generation methods for different datasets.

Dataset	Methods
ShanghaiTech Part A [15]	Geometry-adaptive
ShanghaiTech Part B [15]	kernels
The WorldExpo'10 [1]	Gaussian kernel
UCF CC 50 [28]	

1) Density map via geometry-adaptive kernels

The quality of density map given in a training process determines the performance of the CNNs. Therefore, it is important to consider the distortion caused by the homography between the ground and image planes to accurately estimate the crowd density. Thus, we propose geometry-adaptive kernels [15] or the density map of highly-crowded scenes, which can adaptively determine the spread parameter for each person based on its average distance to its neighbors.

K-nearest neighbors is a simple machine learning algorithm used for classification and regression. The algorithm is nonparametric and will not make assumptions on the data distribution. For each head xi in a given image, we denote distances to its k-nearest neighbors as $\{ d_1^i, d_2^i, \dots, d_m^i \}$. Therefore, the average distance is

$$d_i = \frac{1}{m}\sum_{j=1}^{m} d_j^i \quad (1)$$

To estimate the crowd density around pixel xi, we convolve δ(x-xi) using a Gaussian kernel with parameter σ_i proportional to di. Thus, the density can be defined as follows:

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) \times G_{\sigma_i}(x), with\ \sigma_i = \beta d_i \quad (2)$$

We implemented geometry-adaptive kernels for better adaptation with crowd density variation among images. The labels were convolved with density kernels, which are adaptive to the local geometry of each data point. In the experiments, we set β = 0.3 and k = 3 to retrieve the best performance. For the scenes with comparatively sparse crowds, it is unnecessary to exploit geometry-adaptive kernels. Hence, we used Gaussian kernel with σ = 3. This method can be adapted to the different datasets shown in Table 2.

2) Data augmentation

Data augmentation improves the generalization ability and robustness of a model. Thus, we applied augmentation to the open-source datasets used in our experiment. Each image was cropped into 9 patches, each of which was 1/4 of the original size. Among these 9 patches, 4 contained no overlapping, and the rest 5 patches were taken randomly from the original image. To double the number of images contained in the training set, we mirrored these patches and trained the network with all patches after the augmentation.

3) Loss function

Our method involves training the network end-to-end. The first 10 convolutional layers are fine-tuned with VGG-16 weights, whereas Gaussian initialization with a standard deviation of 0.01 is applied to the others. To train the network, we employed stochastic gradient descent with a fixed learning rate at 1e-6. For the input image X_i (i = 1,…,N), the ground truth is Z_i^{GT}, and Z (X_i; θ) is the estimation density map with parameter θ. The following is the objective function:

$$L(\theta) = \frac{1}{2N}\left|\left|Z(X_i;\ \theta) - Z_i^{GT}\right|\right|_2^2 \quad (3)$$

IV. Validation and Analysis

Our model was evaluated using three different open-source datasets, and the implementation of the proposed model was based on the Caffe framework [39]. Compared to previous state-of-the-art methods, our model has a smaller size, is more robust, and is easier to train.

A. Evaluation metrics

We used mean absolute error (MAE) and mean square error (MSE) as evaluation criteria for the different methods. Generally, MAE indicates the accuracy of the estimates, whereas MSE indicates the robustness of the estimates. MAE and MSE can be defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | C_i - C_i^{GT} | \quad (4)$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} | C_i - C_i^{GT} |^2} \quad (5)$$

where N is the number of test images, C_i^{GT} is the ground truth of counting in image i, and C_i is the estimated count in image i, which is defined as

$$C_i = \sum_{l=1}^{L}\sum_{w=1}^{W} z_{l,w} \quad (6)$$

L and W are the length and width of the density map, respectively; $z_{l,w}$ is the pixel at (l, w) of the generated density map.

We also applied PSNR and SSIM to evaluate the quality of the generated density maps.

PSNR is most widely used to evaluate image quality, and it can be defined as

$$PSNR = 10\log_{10}\left(\frac{(2^n-1)^2}{MSE_1}\right) \quad (7)$$

where MSE_1 is the mean square error of the tested image X and the reference image Y; it can be defined as:

$$MSE_1 = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}(X(i,j) - Y(i,j))^2 \quad (8)$$

where H and W are the height and width of the image, respectively; n is the number of bits per pixel and is usually taken as 8. With a greater PSNR, the image is less distorted.

SSIM is also an effective way to evaluate the image quality as it measures the similarity of images in brightness, contrast ratio, and structure. The equations below describe its function:

$$SSIM(X,Y) = l(X,Y) \cdot c(X,Y) \cdot s(X,Y) \quad (9)$$

$$l(X,Y) = \frac{2\mu_X\mu_Y+C_1}{{\mu_X}^2+{\mu_Y}^2+C_1}\, c(X,Y) = \frac{2\sigma_X\sigma_Y+C_2}{{\sigma_X}^2+{\sigma_Y}^2+C_2}\, s(X,Y) = \frac{\sigma_{XY}+C_3}{\sigma_X\sigma_Y+C_3} \quad (10)$$

where μ_X and μ_Y are the means of images *X* and *Y*, respectively; σ_X and σ_Y are the variances of images *X* and *Y*, respectively; σ_{XY} is the covariance of images *X* and *Y*. The values are obtained as shown below:

$$\mu_X = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X(i,j)$$

$${\sigma_X}^2 = \frac{1}{H\times W-1}\sum_{i=1}^{H}\sum_{j=1}^{W}(X(i,j)-\mu_X)^2$$

$$\sigma_{XY} = \frac{1}{H\times W-1}\sum_{i=1}^{H}\sum_{j=1}^{W}((X(i,j)-\mu_X)(Y(i,j)-\mu_Y)) \quad (11)$$

SSIM ranges from 0 to 1, and greater values indicate lesser distortion in the image.

B. Results

As shown in Table 3, the three popular open-source datasets that were used to verify our proposed model contained highly crowded scenes with a high variance in the number of people. The results of the experiments showed that our model achieved desirable performance on all three datasets; in other words, the model exhibits high flexibility among different datasets with different dataset size and image resolution.

Table 3. Existing datasets used in the experiment

Dataset	Resolution	Num	Max	Min	Ave	Total
ShanghaiTech Part_A	Different	482	3139	33	5014	241,677
ShanghaiTech Part_B	768×1024	716	578	9	123.6	88,488
UCF_CC_50	Different	50	4543	94	1279.5	63,974
WorldExpo'10	576×720	3980	253	1	50.2	199,923

Num is the number of images, Max is the maximal crowd count, Min is the minimal crowd count, Ave is the average crowd count, Total is total number of persons in the dataset.

1) ShanghaiTech dataset

ShanghaiTech dataset [15] is consisted of 1198 images with 330,165 annotated human heads. The dataset is separated into two parts: part_A and part_B. ShanghaiTech_part_A consists of images randomly selected from internet with considerable crowd density, which contains 300 images in training set and 182 images in testing set, resolution of images are variable in this part; while ShanghaiTech_part_B includes images captured from street scenes of Shanghai with fixed resolution of 768X1024, it has 400 images in training set and 316 images in testing set. By comparing our method with six state-of-the-art architectures, we found that our method achieved the best performance in terms of MAE in Part A. In Part B, the network achieved the lowest MSE (which significantly reduced the MAE) compared to existing methods; the results can be found in Table 4. For the density map quality, the method also achieved better results on SSIM and PSNR, as shown in Table 5; moreover, we observed the results of our work on all three datasets, as shown in Table 6.

Table 4. Estimation errors on ShanghaiTech dataset.

	Part_A		Part_B	
Method	MAE	MSE	MAE	MSE
Zhang et al. [1]	181.8	277.7	32.0	49.8
Marsden et al. [40]	126.5	173.5	23.8	33.1
MCNN [15]	110.2	173.2	26.4	41.3
Cascaded-MTL	101.3	152.4	20.0	31.1
Switching-CNN [16]	90.4	135.0	21.6	33.4
CP-CNN [17]	73.6	**106.4**	20.1	30.1
DBNet (ours)	**68.2**	115.0	**10.6**	**16.0**

Table 5. Quality of density map on ShanghaiTech Part A dataset.

Method	PSNR	SSIM
MCNN [15]	21.4	0.52
CP-CNN [17]	21.72	0.72
DBNet (ours)	**23.79**	**0.76**

Table 6. Quality of density map generated by the proposed method.

Dataset	PSNR	SSIM
ShanghaiTech Part_A [15]	23.79	0.76
ShanghaiTech Part_B [15]	27.02	0.89
UCF CC 50 [28]	18.76	0.52
The WorldExpo'10 [1]	26.94	0.92

2) UCF_CC_50 dataset

UCF_CC_50 dataset is an extremely challenging dataset consists of images with different crowd density and perspective distortion, there are 63,075 annotated human heads in total. The dataset contains 50 images with different resolution, and number of person in each image varies from 94 to 4543, which makes the dataset even more challenging. Due to the limited number of images in the UCF_CC_50 dataset, we performed 5-fold cross-validation on the dataset, which is the same data augmentation approach used in ShanghaiTech dataset. By applying MAE and MSE as evaluation metrics, our method was compared with several state-of-the-art methods; it achieved the best MAE and an MSE comparable to that of the other methods. The results are listed in Table 7.

Table 7. Estimation errors on UCF_CC_50 dataset.

Method	MAE	MSE
Idrees et al. [27]	419.5	541.6
Zhang et al. [1]	467.0	498.5
MCNN [15]	377.6	509.1
Onoro et al. [13] Hydra-2s	333.7	425.2
Onoro et al. [13] Hydra-3s	465.7	371.8
Walach et al. [38]	364.4	341.4
Marsden et al. [40]	338.6	424.5
Cascaded-MTL [41]	322.8	397.9
Switching-CNN [16]	318.1	439.2
CP-CNN [17]	295.8	**320.9**
DBNet (ours)	**266.1**	397.5

3) WorldExpo'10 dataset

WorldExpo'10 dataset consists of 3980 images captured by surveillance cameras in 2010 Shanghai World Expo with various scenes, whose resolutions are fixed 576X720. The dataset has 199,923 annotated human heads in total, and number of person in each image varies from 1 to 253. WorldExpo'10 dataset contains 3380 images in the training set, and testing set is separated into 5 groups with 120 images each. The proposed network is also validated and compared with other existing networks on this dataset. According to the results

in Table 8, our method also achieved the best performance in terms of average MAE.

Table 8. Average estimated errors on WorldExpo'10 dataset.

Method	S1	S2	S3	S4	S5	Ave
Chen et al. [42]	**2.1**	55.9	9.6	11.3	3.4	16.5
Zhang et al. [1]	9.8	14.1	14.3	22.2	3.7	12.9
MCNN [15]	3.4	20.6	12.9	13.0	8.1	11.6
Shang et al. [39]	7.8	15.4	14.9	11.8	5.8	11.7
Switching-CNN [16]	4.4	15.7	10.0	11.0	5.9	9.4
CP-CNN [17]	2.9	14.7	10.5	**10.4**	5.8	8.86
DBNet (ours)	2.9	**11.5**	**8.6**	16.6	**3.4**	**8.6**

Due to the results on the three datasets shown above, the proposed network exhibits great performance on all three datasets. Specifically, compared to the classic multi-column network MCNN [15] and CP-CNN [17] which has similar multi-column structure and state-of-the-art performance, on ShanghaiTech dataset, the proposed network achieved best performance in Part_B and only slightly outperformed by CP-CNN on MSE in Part_A, this might be due to the various resolution of images in Part_A, the multi-column structure helps improve performance dealing with resolution variation, and more work on this issue will be done in the future. Also, the proposed network generates density map with best quality on this dataset, due to the introduction of dilated convolution and its deep structure. On UCF_CC_50 dataset, the proposed work still has best performance on MAE. Due to various resolution of images in the dataset, the proposed network is outperformed by CP-CNN on MSE, however, compared to other existing network models, DBNet still has significant improvement. On the third dataset WorldExpo'10 with fixed resolution, DBNet outperforms CP-CNN which has state-of-the-art performance by now.

In conclusion, the proposed network achieves best performance on the three challenging datasets. By employing first 10 convolutional layers of VGG-16 as frontend, 2D feature can be extracted with this deep structure. By removing the fully-connected layers of VGG-16, which brings in great amount of parameters, weight size of the network can be greatly reduced. The introduction of dilated convolution helps expand the receptive field while maintaining the original resolution and generates output with more detailed contextual information, thus generates good quality density map as the results shown in Table 6. In all, the proposed architecture effectively improved performance and exhibit high flexibility on various crowd counting datasets.

C. Transfer learning setting

A model trained on a large dataset containing different head sizes can be easily adapted to other datasets with varying crowd head sizes. The last few layers of MCNN can be fine-tuned to achieve a similar operation on our model, and the performance is compared between different settings. The cross-dataset and transfer learning experiments were used to verify the robustness of our network, and the results are shown in Table 9.

Table 9. Transfer learning across datasets.

Method	MAE	MSE
Fine-tune the last two layers of MCNN [15]	295.1	490.2
DBNet without fine-tuning	404.1	619.4
Fine-tune the whole DBNet	**279.8**	425.9
Fine-tune the back-end of DBNet	282.4	**407.5**

The above transfer learning setting train the models from source domain in the first step, which is ShanghaiTech Part_A here. Then UCF_CC_50 is chosen as target domain, and fine-tuning from training samples in target domain is conducted according to methods in Table 9. The results validate the model's ability to deal with fresh samples, that is the robustness of the model. This can be due to the introduction of deep, single-column structure of VGG-16 convolutional layers in the frontend.

D. Ablation study

We conducted an ablation study on the backend of our network using ShanghaiTech Part_A dataset. The frontend containing the first 10 VGG-16 layers was removed, and the performance of the backend was compared with different configurations. This enabled the verification of the configuration effect on the network performance. ShanghaiTech Part_A is a large-scale crowd counting dataset with varied perspectives and resolutions. Although the dataset is extremely challenging for crowd counting, its large size provides enough data for the stability of deep learning.

In the ablation experiments, we applied four different levels of dilation rates to the backend of our network and verified its performance on ShanghaiTech Part_A dataset to determine the most efficient configuration.

As shown in Table 10, we trained four different configured models on the ShanghaiTech Part_A dataset using the training method described above. With the evaluation metrics defined above, Table 11 reveals the performance of these models. As shown in the table 11, the model achieved the best performance with a dilation rate of 2.

Table 10. Configuration of the backend with different dilation rates.

Back_end of DBNet			
A	B	C	D
conv3-512-1	conv3-512-2	conv3-512-2	conv3-512-4
conv3-512-1	conv3-512-2	conv3-512-2	conv3-512-4
conv3-512-1	conv3-512-2	conv3-512-2	conv3-512-4
conv3-256-1	conv3-256-2	conv3-256-4	conv3-256-4
conv3-128-1	conv3-128-2	conv3-128-4	conv3-128-4
conv3-64-1	conv3-64-2	conv3-64-4	conv3-64-4

The parameters of the convolutional layers are denoted as "conv(kernel size) - (number of filters) - (dilation rate)."

Table 11. Comparison of architectures on ShanghaiTech Part A dataset.

Architecture	MAE	MSE
A	69.7	116.0
B	68.2	115.0

C	71.9	120.6
D	75.8	120.8

E. Result analysis

As shown by the results above, our single-column, deep neural network structure with dilated convolutional layers outperformed current architectures and exhibited high robustness. We achieved the best performance in ShanghaiTech Part_B dataset and the same performance with CP-CNN in ShanghaiTech Part_A dataset. For UCF_CC_50 dataset, our work also equaled CP-CNN in the final result. In the experiment based on WorldExpo'10 dataset, our method outperformed existing state-of-the-art architectures in most scenes, and its high robustness was verified. The best configuration of the backend was found through an ablation study. Thus, our method is a high-performing network with strong robustness. However, the results showed that our work has limited performances in scenes with extremely high crowd density, and it matched state-of-the-art architectures. The network performance should be improved in future studies.

For further researches on hardware acceleration solutions, the single column structure and simple kernel size variation makes the proposed network friendly for hardware implementation, which aims for accelerate realization with limited calculation and storage resources, light weight of the proposed architecture due to the removal of fully-connected layers further enhance this advantage. More works focusing on hardware implementation of proposed network has been done.

V. CONCLUSION

In this study, we propose a deep convolutional neural network based on dilated convolution to solve existing problems of crowd counting and density distribution estimation on high-density scenes. On one hand, the improved network contributes to aggregating the multi-scale contextual information contained in congested scenes and obtain a higher-quality density map that maintains spatial information. On the other hand, by applying the dilated convolution method, resource consumption was also reduced. Experimental results on various open-source datasets verified the state-of-the-art performance of DBNet and its robustness to scale and perspective changes in a wide range of tasks and datasets. The improvement on flexibility and robustness in various applications can be explained by the small-size-induced low-resource-consumption of DBNet. These attributes suggest that DBNet is friendly to hardware implementation based on FPGA and ASICs; thus, it can be implemented on IOT devices and portable devices. Our future studies focusing on hardware implementation improves this point.

References

[1] C. Zhang, H. Li, X. Wang, X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 2015, pp. 833–841.

[2] Xiaoheng Jiang et al., "Attention scaling for crowd counting," in Proc. 2020 IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 2020, pp. 4705-4714.

[3] Muming Zhao, Jian Zhang, Chongyang Zhang, Wenjun Zhang, "Leveraging heterogeneous auxiliary tasks to assist crowd counting," in Proc. 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 12728-12737.

[4] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, Hefeng Wu, "ADCrowdNet: an attention-injective deformable convolutional network for crowd understanding," in Proc. 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 3220-3229.

[5] Xiaolong Jiang et al, "Crowd counting and density estimation by trellis encoder-decoder networks," in Proc. 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 6126-6135.

[6] Miaojing Shi, Zhaohui Yang, Chao Xu, Qijun Chen, "Revisiting perspective information for efficient crowd counting," in Proc. 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 7271-7280.

[7] J. Long, E. Shelhamer, T. Darrell, "Fully convolutional networks for semantic segmentation," in Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA, 2015, pp. 3431–3440.

[8] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, N.E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016, pp. 598-606.

[9] J. Qiu, et al., "Going deeper with embedded fpga platform for convolutional neural network," in Proc. 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, New York, NY, USA, 2016, pp. 26–35.

[10] X. Zhang, et al., "High performance video content recognition with long-term recurrent convolutional network for fpga," in Proc. 27th International Conference on Field Programmable Logic and Applications, Ghent, 2017, pp. 1–4.

[11] R. Andri, L. Cavigelli, D. Rossi, L. Benini, "Yodann: An ultra-low power convolutional neural network accelerator based on binary weights," in Proc. 2016 IEEE Computer Society Annual Symposium on VLSI, Pittsburgh, PA, 2016, pp. 236–241.

[12] H. Fu, Y. Xu, D.W.K. Wong, J. Liu, "Retinal vessel segmentation via deep learning network and fully-connected conditional random fields," in Proc. 13th IEEE International Symposium on Biomedical Imaging, Prague, 2016, pp. 698–701.

[13] D. Onoro-Rubio, R. J. Lopez-Sastre, "Towards perspective-free object counting with deep learning," in Proc. European Conference on Computer Vision, Cham, 2016, pp. 615–629.

[14] G. French, M. Fisher, M. Mackiewicz, C. Needle. "Convolutional neural networks for counting fish in fisheries surveillance video," 2015.

[15] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma. "Single image

crowd counting via multi-column convolutional neural network," In Proc. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016, pp. 589–597.
[16] D.B. Sam, S. Surya, R.V. Babu, "Switching convolutional neural network for crowd counting," In Proc. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, 2017, pp. 4031-4039.
[17] V.A. Sindagi, V.M. Patel, "Generating high-quality crowd density maps using contextual pyramid cnns," arXiv preprint, 2017, arXiv:1708.00953.
[18] F. Yu, V. Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv preprint, 2015, arXiv:1511.07122.
[19] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834-848, April 2018.
[20] P. Dollar, C. Wojek, B. Schiele, P. Perona, "Pedestrian detection: An evaluation of the state of the art," IEEE Trans. Pattern Anal. Mach. Intell. vol. 34, no. 4, pp. 743-761, April 2012.
[21] P. Viola, M.J. Jones, "Robust real-time face detection," Int. J. Comput. Vis., Vol. 57, pp. 137–154, May 2004.
[22] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 2005, pp. 886-893 vol. 1.
[23] B. Wu, R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors," in Proc. 10th IEEE International Conference on Computer Vision, Beijing, 2005, pp. 90-97, Vol. 1.
[24] P. Sabzmeydani, G. Mori, "Detecting pedestrians by learning shapelet features," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, 2007, pp. 1-8.
[25] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE Trans. Pattern Anal. Mach. Intell. vol. 47, no. 2, pp. 6-7, Feb. 2014.
[26] A.B. Chan, N. Vasconcelos, "Bayesian Poisson regression for crowd counting," in Proc. IEEE 12th International Conference on Computer Vision, Kyoto, 2009, pp. 545-551.
[27] H. Idrees, I. Saleemi, C. Seibert, M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, 2013, pp. 2547-2554.
[28] V. Lempitsky, A. Zisserman, Learning to count objects in images, in Proc. 23rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2010, pp. 1324–1332.
[29] V.Q. Pham, T. Kozakaya, O. Yamaguchi, R. Okada, "Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation," In Proc. IEEE International Conference on Computer Vision, Santiago, 2015, pp. 3253-3261.
[30] A. Krizhevsky, I. Sutskever, G.E. Hinton, "Imagenet classification with deep convolutional neural networks," In Proc. 25th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2012, pp. 1097–1105.
[31] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint, 2014, arXiv:1409.1556.
[32] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," In Proc. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, 2017, pp. 1800-1807.
[33] F. Yu, V. Koltun, "Multi-Scale context aggregation by dilated convolutions," in Proc. International Conference on Learning Representations, San Juan, Puerto Rico, 2016.
[34] P. Wang, et al., "Understanding convolution for semantic segmentation," in Proc. 2018 IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, 2018, pp. 1451-1460.
[35] L.C. Chen, G. Papandreou, F. Schroff F, H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint, 2017, arXiv:1706.05587.
[36] Younis Ibrahim, Junyang Liu, Xuanxuan Yang, Hongwei Sha, Peng Li, Haibin Wang, "Analyzing the impact of soft errors in deep neural networks on GPUs from instruction level," WSEAS Transactions on Systems and Control, Vol. 15, Art. #70, pp. 699-708, 2020.
[37] Nivedita M., Asnath Victy Phamila Y, "A Survey on Different Deep Learning Architectures for Image Captionings," WSEAS Transactions on Systems and Control, Vol. 15, Art. #63, pp. 635-646, 2020.
[38] E. Walach, L. Wolf, "Learning to Count with CNN Boosting," in Proc. 14th European Conference on Computer Vision, Amsterdam, 2016, pp. 660-676.
[39] C. Shang, H. Ai, and B. Bai, "End-to-end crowd counting via joint learning local and global count," in Proc. IEEE International Conference on Image Processing, Phoenix, AZ, 2016, pp. 1215-1219.
[40] M. Marsden, K. McGuiness, S. Little, and N. E. O'Connor, "Fully convolutional crowd counting on highly congested scenes," arXiv preprint, 2016, arXiv:1612.00220.
[41] V. A. Sindagi and V. M. Patel, "Cnn-based cascaded multitask learning of high-level prior and density estimation for crowd counting," in Proc. 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, Lecce, 2017, pp. 1-6.
[42] K. Chen, S. Gong, T. Xiang, C. Change Loy, "Cumulative attribute space for age and crowd density estimation," in Proc. IEEE conference on computer vision and pattern recognition, Portland, OR, 2013, pp. 2467-2474.

Han Jia received the B.E. degree from Huazhong University of Science and Technology, Wuhan, China, in 2010, and is working on PhD degree in Huazhong University of Science and Technology since 2013. His main research interests include very large scale integrated circuits design, crowd counting

neural networks and AI hardware acceleration.

Xuecheng Zou received B.E. and M.S. and PhD degrees from Huazhong University of Science and Technology in 1985, 1988 and 1995, respectively. He is currently a Professor with School of Optical and Electronic Information, Huazhong University of Science and Technology. His main research interests include very large scale integrated circuits design, microelectronics and micro photonics.

Author Contributions:

Han Jia carried out main work and the composition of this paper.

Xuecheng Zou gives guidance and suggestions of research work.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0
https://creativecommons.org/licenses/by/4.0/deed.en_US