Lightweight Crowd Counting Network based on Depthwise Separable Convolution

Dengguo Yao¹, Yuanping Xu*¹, Chaolong Zhang²,³, Zhijie Xu², Jian Huang¹, and Benjun Guo¹

¹ School of Software Engineering, Chengdu University of Information Technology, Chengdu, Sichuan, 610225, China
² School of Computing and Engineering, University of Huddersfield, Huddersfield, West Yorkshire, HD1 3DH, UK
³ Chengdu CHG Robots & Intelligent Equipment Institute, Chengdu, Sichuan, 610042, China

*Corresponding author’s e-mail: ghxpy@hotmail.com

Abstract. Crowd counting on the image is a challenging problem. Many neural network-based methods usually use two-branch and multi-branch networks to extract high-level features of different scales or densities, and then merge these features by a fusion operation. Although these methods can reduce the error of crowd counting, it makes the amount of parameters enormous, so that the efficiency of training and optimization of the model is low, and the calculation resource consumption is high. To this end, a residual network based on depthwise separable convolution is proposed for image crowd counting. The network can not only reduce the amount of calculation through depthwise separable convolution, but also deepen the network depth through the residual structure to extract more effective high-level features. The experiment proves that, compared with the start-of-the-art methods, the method in this paper dramatically reduces the parameter amount to 1.91 Million when the accuracy is comparable.

1. Introduction
The Corona Virus Disease that broke out in early 2020 swept through China, but the epidemic was controlled within a short period. This is closely related to the national prevention and control measures: preventing crowds from gathering and limiting the number of personnel in some necessary places, such as: Limiting the number of people and the distance between people in supermarkets, vegetable markets and student cafeterias. The spread of the disease caused by the gathering of people has caused irreparable damage to people’s lives and safety, and it has increased the difficulty of national prevention and control. To prevent and reduce the spread of this virus caused by crowd gathering events, surveillance equipment were installed in different places to monitor people’s activities[1]. Therefore, crowd counting has aroused great interest, and many researchers have tried to solve this problem. With the development of computer vision and deep learning, researchers have proposed many deep learning-based methods to solve the crowd counting problem. These methods have significantly improved the resolution of scale diversity and density diversity. But due to a large number of parameters and calculations, it is difficult to be applied in real-world applications. To solve
these problems, this paper studies and designs a lightweight crowd counting network based on depthwise separable convolution to calculate the number of people.

2. Related Works

Whether using regression or density maps, the Convolutional Neural Networks (CNN) methods can achieve excellent results[2-4]. Zhang et al.[5] proposed a deep convolutional neural network for crowd counting, confirming that CNN has higher efficiency and accuracy than traditional methods. Zhang et al.[6] proposed a three-column CNN structure. Sam et al.[7] proposed a switched CNN model by using different networks of scale information division and counting. But the distinct disadvantage is that the number of network columns limits the diversity of the crowd scale, and the number of network parameters in multi-columns is enormous. Sindagi et al.[8] proposed a cascaded CNN model to learn global correlation features and crowd density estimates. Liu et al.[9] proposed a method that combines detection and regression. Li et al.[10] proposed a network model divided into a front-end and a back-end. Ranjan et al.[11] proposed a two-branch CNN structure. However, these methods have many branch networks with a large number of parameters, which is inefficient during training and optimization.

3. Proposed method

The current crowd counting still has the problem of scale diversity of people in the picture. When extracting features through various convolutional networks to generate density maps, the parameters are enormous, and the efficiency is low, so that excellent computing resources are required. Although using a shallow layer of a network can improve efficiency, it cannot get more useful features. Although using a deeper layer of a network can get more features, the training and optimization of the model are arduous and only used in an experimental environment. To tackle the above problems, this paper proposes a residual structure network based on depthwise separable convolution. Compared with the straight structure of VGG network, the residual structure reuses the image features better and eases with the network deepening. The introduction of depthwise separable convolutions dramatically reduces the number of network parameters.

3.1. Depthwise separable convolution

Depth separable convolution is to divide the convolution operation of the image feature matrix into two dimensions: space and channel. It is divided into two parts: depthwise convolution in spatial dimension and pointwise convolution in channel dimension. The depthwise separable convolution operation is shown in figure 1.

![Depthwise separable convolution](image1)

Figure 1. Depthwise separable convolution

For depthwise separable convolution and ordinary convolution, when the stride=1, the calculation amount is shown as the follows:

\[ C_D = H_k \times H_k \times M \times H_I \times H_I + M \times N \times H_I \times H_I \]  

(1)
\[ C_o = H_k \times H_k \times M \times N \times H_i \times H_i \]  
(2)

where \( C_D \) is the calculation amount of the depthwise separable convolution, \( C_O \) is the calculation amount of the ordinary convolution, \( M \) is the number of channels of the input feature matrix, \( N \) is the number of output feature matrix, \( H_k \) is the width of the convolution kernel, and \( H_i \) is the width of the input feature matrix.

\[
\frac{C_D}{C_O} = \frac{H_k \times H_k \times M \times N \times H_i \times H_i + M \times N \times H_i \times H_i}{H_k \times H_k \times M \times N \times H_i \times H_i} = \frac{1}{N} + \frac{1}{H_k^2}
\]  
(3)

Formula (3) can be obtained from formula (1) and formula (2). It can be obtained from the formula (3): \( C_D/C_O = (1/N)+1/9 \). When the number of convolution kernels becomes larger, \( C_D/C_O \approx 1/9 \). In theory, the calculation of ordinary convolution is about nine times the depthwise separable convolution.

3.2. Residual structure

The traditional residual structure first performs 1×1 convolution reduction operation on the input feature matrix, then using 3×3 convolution to extract the image features, and finally it outputs the feature matrix with a higher dimension. This purpose is also to reduce the amount of calculation in the residual structure. In the new residual structure, to reduce the amount of calculation, the 3×3 ordinary convolutional layer is replaced by a depthwise separable convolutional layer. In order to extract more depth features, this paper first performs 1×1 convolution to increase the dimension of the input matrix, and then using the 3×3 depthwise separable convolution to extract the features of spatial dimension and channel dimension. The branch of shortcut guarantees the multiplexing picture features, and the depthwise separable convolution guarantees the reduction of the calculation amount and the parameter amount. The residual structure is shown in figure 2. \( C \) is the number of input feature matrix channels, \( D \) is the number of feature matrix channels in the depthwise separable convolution process, and \( P \) is the number of output matrix channels.

![Residual structure](image)

Figure 2. Residual structure

3.3. Crowd counting network based on depthwise separable convolution

The lightweight crowd counting network based on depthwise separable convolution proposed in this paper is shown in figure 3. The input image passes through each layer structure of the network, and finally, the density map of the picture is output, and the number of people is obtained by regression of the density map. The network includes low-level feature extraction, high-level feature extraction and density map regression. The feature remapping block is composed of a traditional convolutional layer with a convolution kernel of 3×3 to remap the image features. Multi-Residual blob is a network.
structure consisting of a series of new residual structures similar to residual structures for extracting advanced features. In each residual structure, the parameters of the convolution kernel of depthwise separable convolution, the number of channels and the step of the feature matrix are different. Since the final output of the residual structure is output through dimensionality reduction, but the ReLU activation function has a severe loss of low-dimensional feature information, so in this residual structure, the last layer uses Liner, and every other layer is activated with ReLU. The density map regression block is mainly composed of a traditional convolutional layer with a convolution kernel of 1×1 to generate a density map. Since the eigenvalues of the density map are always positive, the ReLU activation function can be used after the regression layer of the density map to enhance the recovery of the density map.

![Diagram of crowd counting network based on depthwise separable convolution](image)

**Figure 3.** Crowd counting network based on depthwise separable convolution

### 4. Experiment

In this section, we first introduce datasets and experiment evaluation metrics. Then an evaluation of transferability is reported to demonstrate the transferability of the proposed method across datasets. Finally, we give the evaluation results and perform comparisons between the proposed method with recent methods.

#### 4.1. Dataset

ShanghaiTech dataset: a total of 1198 images including two parts A and B. As shown in figure 4, Part A consists of 482 pictures randomly selected from the Internet, the training set contains 300 pictures, and the test set contains 182 pictures; part B consists of 716 pictures taken in the streets of Shanghai city, the training set contains 400 pictures, the test set contains 316 pictures.

![Images of dataset](image)

**Figure 4.** (a) Training images in Part A dataset. (b) Test images in Part A dataset. (c) Training images in Part B dataset. (d) Test images in Part B dataset.

#### 4.2. Evaluation Metric

Mean Absolute Error (MAE), Mean Squared Error (MSE) and Number of Neural Network Parameters (PARAMS), unit: Million (M), were used to evaluate the performance of our method in the test dataset.
\[ MAE = \frac{1}{N} \sum_{i=1}^{N} |g_i - p_i| \]  
(4)

\[ MSE = \frac{1}{N} \sum_{i=1}^{N} (g_i - p_i)^2 \]  
(5)

In formulas (4) and (5), N is the total number of pictures in the test set, \( g_i \) is the ground truth value of the \( i \)-th image, and \( p_i \) is the predicted value of the \( i \)-th image.

4.3. Implementation Details

In order to improve the training effect of the model, data expansion processing is performed on the samples, and the original pictures are divided into nine subset with equal size. We use 90% of the training set images as training samples and the remaining 10% as verification samples. And we set the learning rate to 0.00001 and momentum to 0.9, and using Adam optimization.

4.4. Comparisons with Recent Methods

We demonstrate the efficiency of our proposed method on challenging crowd counting dataset. Table 1 reports the results on ShanghaiTech.

| Methods     | Part_A | Part_B |
|-------------|--------|--------|
|             | MAE    | MSE    | MAE    | MSE    |
| LBP+RR      | 303.2  | 371.0  | 59.1   | 81.7   |
| MCNN-CCR    | 245.0  | 336.1  | 70.9   | 95.9   |
| Ours        | 235.5  | 373.3  | 64.1   | 100.9  |

Table 2. Experiment error in transfer learning

| Methods | Part_A→Part_B | Part_B→Part_A |
|---------|---------------|---------------|
|         | MAE | MSE | MAE | MSE |
| MCNN    | 85.2 | 142.3 | 221.4 | 357.8 |
| Ours    | 78.6 | 95.8 | 355.5 | 497.9 |

Table 3. Number of parameters

| Methods | MCNN | Switch-CNN | CSRNet | SANet | CP-CNN | MSCNN | Ours |
|---------|------|------------|--------|-------|--------|-------|------|
| PARAM S/(M) | 19.2 | 15.11 | 16.26 | 0.91 | 68.4 | 2.9 | 1.91 |

As shown in table 1, compared with the MCNN-CCR method in Part A and Part B of this dataset, although the MSE is slightly higher, the MAE is reduced by 9.5 and 6.8. Compared with the LBP+RR method, although the performance is slightly worse in Part B of the dataset, MAE is reduced by 67.7 in Part A. This shows that the method in this paper can count people more accurately in scenes with high crowd density.

To demonstrate the transferability of the proposed method across datasets, we conduct experiments under cross dataset settings, where the model is trained on the source domain and tested on the target domain.
The cross dataset experimental results are presented in table 2. We can observe that the proposed method generalizes well to unseen datasets. The proposed method also performs better than MCNN in transferring models trained on ShanghaiTech Part A to Part B. Yet, the improvement is not as significant as the comparison with MCNN on transferring on ShanghaiTech Part B and Part A. This is probably because MCNN is easy to extract features in scenes with low crowd density. This also confirms the generalizability of the proposed method.

As shown in table 3, the number of parameters of our proposed is the least except SANet[12]. Although the error of Switch-CNN and CSRNet is slightly lower than the method in this paper, the number of parameters is almost eight times of this paper. Compared with other methods, the method in this paper have comparable results, but the number of parameters is much reduced, which proves that the method proposed in this paper is more lightweight and has higher efficiency under the condition of accuracy guarantee.

5. Conclusions
This paper proposes a lightweight network for crowd counting. Compared with many CNN-based methods, the method in this paper reduces the number of parameters to 1.91 Million while ensuring accuracy. However, the accuracy of this method is not high enough for crowd counting. Next, we will further study how to improve the network to extract more features of different scales to improve the accuracy of crowd counting. Achieving accurate and reliable crowd counting with fewer parameters, which makes it possible to expand to real-world application scenarios, such as embedded devices and mobile devices.

Acknowledgments
This work was supported by the NSFC (61203172), the Sichuan Science and Technology Programs (2018JY0146, 2019YFH0187, and 2019135).

References
[1] Sultani, W., Chen, C., Shah, M. (2018). Real-world anomaly detection in surveillance videos. In: the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 6479-6488.
[2] Xu, C., Qiu, K., Fu, J., Bai, S., Xu, Y., & Bai, X. (2019). Learn to scale: Generating multipolar normalized density maps for crowd counting. In: the IEEE International Conference on Computer Vision(ICCV). pp. 8382-8390.
[3] Sindagi, V. A., Patel, V. M. (2018). A survey of recent advances in cnn-based single image crowd counting and density estimation. Pattern Recognition Letters., 107: 3-16.
[4] Liu, X., Van De Weijer, J., Bagdanov, A. D. (2018). Leveraging unlabeled data for crowd counting by learning to rank. In: the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 7661-7669.
[5] Zhang, C., Li, H., Wang, X., & Yang, X. (2015). Cross-scene crowd counting via deep convolutional neural networks. In: the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 833-841.
[6] Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y. (2016). Single-image crowd counting via multi-column convolutional neural network. In: the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 589-597.
[7] Sam, D. B., Surya, S., Babu, R. V. (2017). Switching convolutional neural network for crowd counting. In: the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 4031-4039.
[8] Sindagi, V. A., Patel, V. M. (2017). Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In: the IEEE International Conference on Advanced Video and Signal Based Surveillance(AVSS). pp. 1-6.
[9] Liu, J., Gao, C., Meng, D., Hauptmann, A. G. (2018). Decidenet: Counting varying density crowds through attention guided detection and density estimation. In: the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 5197-5206.

[10] Li, Y., Zhang, X., Chen, D. (2018). Csrsnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: the IEEE conference on Computer Vision and Pattern Recognition(CVPR). pp. 1091-1100.

[11] Ranjan, V., Le, H., Hoai, M. (2018). Iterative crowd counting. In: the European Conference on Computer Vision (ECCV). pp. 270-285.

[12] Cao, X., Wang, Z., Zhao, Y., Su, F. (2018). Scale aggregation network for accurate and efficient crowd counting. In: the European Conference on Computer Vision (ECCV). pp. 734-750.