Journal of Information and Communication Engineering (JICE), 6 (1): 339-346, June 1, 2020
e-ISSN: 2432-5465, p-ISSN: 2186-9162
Applied Science and Computer Science Publications

An attention mechanism-based multi-scale network crowd density estimation algorithm

Yaoyao Li
School of Electrical and Electronic Engineering
Shanghai Institute of Technology
Shanghai, China
18721533191@163.com

Huailin Zhao
School of Electrical and Electronic Engineering
Shanghai Institute of Technology
Shanghai, China
zhao_huailin@yahoo.com

Li Wang
School of Electrical and Electronic Engineering
Shanghai Institute of Technology
Shanghai, China
736036396@qq.com

***Abstract*—It is becoming more and more important to calculate the people number in terms of the requirement for the safety management, because that the crowd gathering scenes are common whether or not it is daily urban traffic or some special gatherings. Calculating the people number in high-density crowd is a very difficult challenge due to the diversity of ways people appear in crowded scenes. This paper proposes a multi-branch network which combines the dilated convolution and attention mechanism. By combining dilated convolution, the context information of different scales of the crowd image are extracted. The attention mechanism is introduced to make the network pay more attention to the position of the head of the crowd and suppress the background noise, so as to obtain a higher quality density map. Then add all the pixels in the density map to get the total people number. Through a large number of experiments, this network can better provide effective crowd density estimation features and improve the dissimilarity of density map distribution, which has stronger robustness.**

Keywords*—Attention mechanism; Crowd density estimation; Dilated convolutional

I. BACKGROUND

The scenes of large-scale crowd gathering can be seen everywhere with the population growing and the modern transportation facilitating. Overcrowding in crowds can easily lead to dangerous situations such as panic and trampling. Therefore, it is necessary to calculate the number of people in the video surveillance area, and take effective measures according to actual conditions.

The current crowd counting algorithm mainly includes two parts: the crowd counting algorithm based on traditional machine learning and the crowd counting algorithm based on deep learning. The traditional crowd counting algorithm mainly includes crowd counting algorithm based on detection model, crowd counting algorithm based on regression model and crowd counting algorithm based on density map estimation.

The crowd counting algorithm based on deep learning mainly uses the deep neural network to directly establish the relationship between the input image and the crowd density map. By optimizing the Euclidean distance between the output density graph and the true density graph, the network can learn the nonlinear mapping relationship between the input image and its corresponding density graph.

However, due to the influence of the loss function, the density map learned by the network is blurred, and the density map corresponding to the real one is quite different. At the same time, it may cause overestimation problems, which makes it difficult to accurately estimate the crowd density. Moreover, accurate crowd counting has always been a challenging problem in computer vision due to problems such as occlusion, perspective distortion, scale changes, and diversity of population distribution.

In order to solve the problem of scale change, this paper designs a network structure that combines the dilated convolution, so that the network branches with different receptive field scope can extract the crowd features of different scales, thereby expanding the context information and assisting the crowd counting.

In recent years, attention models have achieved great success in various computer vision tasks. The attention mechanism does not require extracting features from the entire image, but rather allows the model to focus on the most relevant features as needed. Inspired by the attention module SE Block [1], we added attention mechanism to the network. Introduce the attention mechanism to the crowd counting problem, guide the network to pay more attention to the head position of the crowd through the attention module, and suppress the background noise, so as to provide more effective crowd density map estimation information and help to improve the problem of dissimilar density distribution.

II. RELATED RESEARCH

The traditional crowd counting algorithm can be divided into the following three categories: crowd counting algorithm based on detection model, crowd counting algorithm based on

Corresponding Author: Yaoyao Li. School of Electrical and Electronic Engineering, Shanghai Institute of Technology, China. E-mail: 18721533191@163.com

regression model and crowd counting algorithm based on density estimation [2].

Most of the initial research focused on the detection style framework, using a sliding window detector to detect pedestrians in the scene, and then using the detected information for counting the number of people. This approach has achieved great success in low-density scenarios. However, for high-density scenarios, many parts of pedestrians are blocked, so the test results are greatly affected.

The regression-based crowd counting algorithm mainly establishes the nonlinear mapping relationship between the features in the image and their corresponding numbers. This type of algorithm focuses on the extraction of low-order features and the choice of regression models. By using regression counting, the dependence on the learning detector is avoided. However, due to external factors such as low resolution, severe occlusion, and perspective, these methods do not provide enough information to accurately count in high-density crowd.

The crowd estimation algorithm based on density estimation focuses on mining the spatial position information of the crowd in the scene, so as to improve the calculation accuracy of the crowd counting algorithm. At the same time, it avoids the difficult work of accurately detecting and locating pedestrians in the crowd, and estimates the number of people in the area by summing all the pixels of the density map.

Crowd counting algorithm based on deep learning is mainly realized by convolutional neural network. Add multi-column and multi-resolution information to the basic convolutional neural network structure, or use context information to assist crowd density estimation, so that the crowd output density map and the true value density map are more similar, achieving a more accurate count.

In reference [3], for the first time, the convolutional neural network was used to directly establish the nonlinear mapping relationship between density map and input image. A multi-column convolutional neural network MCNN was proposed. The MCNN network uses a multi-scale convolution kernel to adapt to the size of the human head at different scales, and then combines the three columns of convolutional neural networks to obtain the final crowd density map.

After that, Sam et al. proposed a new convolutional neural network (Switch-CNN) [4], which sends different blocks to the corresponding convolutional neural network regressions according to different crowd sizes through the switch layer, so that each column of regression was processed only for a crowd of a specific size. A similar approach was used in reference [5] to first pre-classify images according to different population densities.

At present, the results obtained by many crowd counting methods are only close to the truth value in the counting result, but the visualization of the density map output by the network shows that there is a big difference from the truth value density map. Therefore, the reference [6] proposed a context pyramid convolutional neural network (CP-CNN), which outputs feature maps through three different modules (global context estimation module, local context estimation module and density map estimation module). The fusion convolutional neural network is merged to obtain a high-quality crowd density map containing context information.

Reference [7] proposed a dilated convolutional neural network (CSRNet), which uses a convolution kernel with holes (extended convolution) instead of the pooling layer and the convolution layer. The receptive field is enlarged without increasing the number of parameters or the amount of calculation, thereby obtaining a density map which is closer to the true value.

In the last two years, attention models have been widely used in various types of deep learning tasks. In reference [8], decidenet first adopted adaptive counting estimation based on detection and regression to estimate the crowd number under the guidance of the attention mechanism. By dynamically evaluating the quality of each pixel to capture the different importance weights of density maps of sparse and crowded scenes, the final crowd size can be obtained.

Reference [9] proposed a new scale perceptive attention network, in which extracted features are re-weighted according to global and local attention scores, so that the network automatically focuses on some global and local scales suitable for images. Compared with reference [8], the model is simpler.

Based on the above research, this paper aims to solve the similarity problem of density map and designs the network structure integrating dilated convolution and attention mechanism. The dilated convolution is used to expand the receiving field to extract deeper information, and the attention mechanism is introduced to highlight the crowd position. In this way, the estimated density map of network output can be improved and be more similar to the truth density map.

III. Our Proposed Method

A. Multi-scale Fusion Network

Dilated convolution. In this paper, the CSRNet in reference [7] is used as the basic network to return to the crowd density map. The first ten layers of vgg-16 were selected as the front-end of the network. Because of its strong learning ability and flexible architecture, it is easy to generate density map for connecting the back-end. The network replaces the classification portion of the original VGG-16 fully connected layer with dilated convolution. Through the previous convolution and pooling layers, the output size of this front-end network is 1/64 of the original input size. If continue to use more convolutional and pooling layers, the output size will be further reduced, making it difficult to generate high-quality density maps. Therefore, the back end of CSRNet adopts dilated convolutional layer to avoid the resolution loss of density map caused by oversampling and to increase the sensing field and maintain the resolution by cavitation convolution. Although pooling layers are widely used to maintain invariance and control overfitting, they also significantly reduce the spatial resolution of feature graphs,

which means that spatial information is lost.

The dilated convolution uses a convolution kernel with holes instead of the pooling layer and the convolution layer, which expands the receptive field range. Although adding more convolution layers can also generate a larger receiving field, the operation will become Cumbersome. In the dilated convolution, the small-size kernel with $k \times k$ convolution kernel can be extended to k+ (k-1) (r-1) and has an extended stride r. Therefore, it is flexible to aggregate multi-scale context information while maintaining the original resolution.

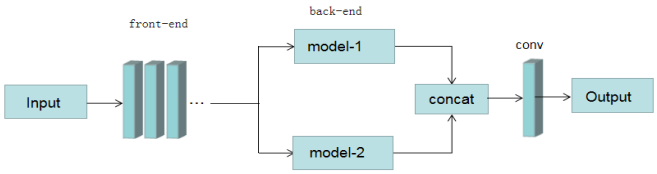


Fig.1. Multi-scale Fusion Network

Model-1: In this design, the network module 1 selects the convolution kernel with the size of 3*3 with the above expansion rate of 1, namely the conventional convolution layer, which has 6 layers. The number of convolution kernels in each layer is {512,512,512,256,128,64}, so that the number of picture channels finally output by the module 1 is 64.

Model-2: The network module 2 selects the convolution kernel with the size of 3*3 with the above expansion rate of 2, which has six layers. The number of convolution kernels of each layer is the same as that of module 1, and the number of picture channels finally output by module 2 is also 64.

Calculate the receptive field (i.e. the receptive field of the first feature in the feature graph). First, the first layer is the input image, whose feature number n is the size of the input image, the size of the sensing field r=1, and the distance between two adjacent features j=1. Set the size of the convolution kernel to k, the padding size to p, and the step size to s.

The feature interval j of the output feature graph is:

$$j_{\text{out}} = j_{\text{in}} \times s \tag{1}$$

The receptive field size r of the calculated output feature graph is:

$$r_{\text{out}} = r_{in} + (k-1) \times j_{in} \tag{2}$$

Through the above two formulas, the perceptive field size of each layer of convolution or pooling is calculated successively, and it can be obtained that the perceptive field size of module 1 is 188, and that of module 2 is 284. The modules with different sensory field sizes were fused to extract the features of different scales on the crowd images and obtain the context information. Its network structure is shown in the figure 1. The characteristics output by module 1 and module 2 were fused, and the number of image channels output was 128. Then the final crowd density map was output through a convolution of 1*1.

B. Attention Mechanisms Fuse Networks

Attention mechanisms have been used extensively in deep neural networks in recent years, and its performance has been greatly improved in many tasks. It is commonly used in conjunction with threshold functions such as Softmax and Sigmoid. For crowd counting, the attention model can be used as an effective tool to guide the network to focus on the head position, which is the most important clue of the network.

Therefore, this paper introduces an attention module to determine the degree of attention to different location features. Specifically, this paper uses attention blocks to pay more attention to the head area while suppressing the background area and body part in the image. Our goal is to direct the network to selectively focus on the head region when estimating the density map. Get more accurate density mapping for crowd counting, no matter how complex the background is, how complex the distribution is.

Model-3: Module three reference network SENet, select SE block as the attention model used in this design. The SE block is a lightweight threshold mechanism designed to model the correlation between channels. First, use a global average pooling layer to compress global space information and generate statistics for each channel. Then use two fully connected layers to form a bottleneck form that enhances the generalization capabilities of the model. The normalized weight between 0 and 1 is obtained by adding the sigmoid activation function to enhance the distinguishability of the feature. Finally, the normalized weights are weighted to the characteristics of each channel by a scale operation, and the characteristics of each channel are adjusted to enhance the representation capability of the network.

By combining SE blocks, we start from the channel dimension and model the dependencies between channels to adaptively adjust the characteristic response values of each channel. In this way, the representation ability of the network is improved, so that the network can learn global information, and the network can selectively enhance the characteristics of the crowd head. This makes it easier for the model to learn deep features to improve model performance and improve the network learning process.

For the attention model, we consider two fusion modes, namely series and parallel.

Series connection: in the method of series connection, we connect module 3 after the first 10 layers of vgg-16, and refer to the design of CSRNet, replace the full connection layer of vgg-16 back-end with SE Block. Its network structure is shown in Figure 2.

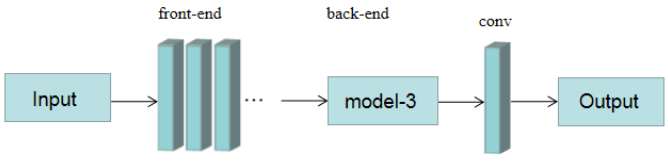


Fig.2. Module three series

Parallel connection: In this paper, module 3 is used in parallel with module 1 mentioned above. Based on the CSRNet as a whole, the empty convolution at the back-end of the network is replaced by the conventional convolution as one branch, and the attention module is added as another branch. The number of picture channels output by module 1 is 64, the number of picture channels output by module 3 is 512, the number of picture channels output by the network after fusion is 576, and then the final crowd density map is output through a 1×1 convolution.

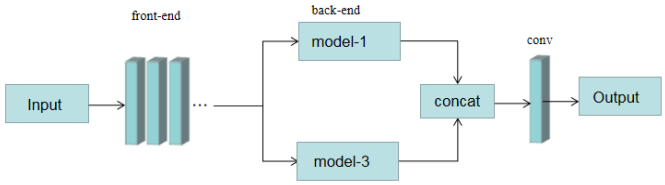


Fig.3. Module three parallel

C. Multi-branch Network

The above designs respectively integrate the dilated convolution and attention module into the network, and we consider using both the dilated convolution and attention model in a network. Therefore, the multi-branch network structure as shown in the figure is designed. The network consists of three branches.

First, the network input images go through the basic convolution and pooling layers in the first 10 layers of VGG16, and then the generated feature images are sent into three branches respectively. Where, branch 1 is the convolution of the above module 1, and the number of convolution kernels is {512,512,512,256,128,64} respectively. Branch 2 is the dilated convolution of the above mentioned module 2 with an expansion rate of 2. The number of convolution kernels is the same as that of branch 1. The range of receptive field was expanded without adding additional parameters, so as to extract population characteristics at different scales. In branch 3 (module 3 above), attention module is added to emphasize the head position of the crowd, so as to reduce background noise and enable the network to output more accurate crowd density map.

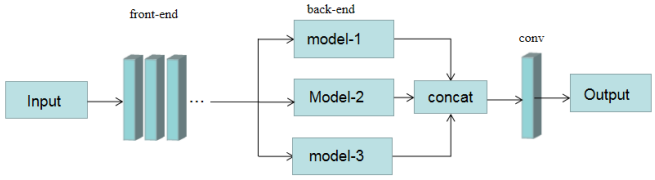


Fig.4. Multi-branch network structure

IV. EXPERIMENTS

A. Density Map Generation

The original data given in the data set is the original crowd image and its corresponding crowd location coordinates, while the training network needs to give its crowd density map, so the location coordinates should be converted into the corresponding crowd density map first. The calibration quality of crowd density in the training data determines the performance of the convolutional neural network. The generation method of crowd density map based on adaptive gaussian kernel is selected. The formula is as follows:

$$F(x)=\sum_{i}^{N}\delta\left(x-x_i\right)\times G_{\sigma_i}(x),\sigma_i=\beta\bar{d}^{l} \quad (3)$$

Among them, $G_{\sigma_i}(x)$ represents the Gauss convolution core. x_i represents the position coordinate of the human head in the image. $\delta(x-x_i)$ is the Dirac delta function for the head. N represents the total number of people included in the image. $\bar{d}^{l}=\frac{1}{m}\sum_{j=1}^{m}d^{i}{}_{j}$ stands for the average distance of M heads nearest to the human head. Usually, the size of the human head is related to the distance between two adjacent people in the center of a crowded scene. $\bar{d}^{l}$ is approximately equal to the size of the human head in a crowded situation, and β is a hyperparameter.

B. Evaluation Index

At present, there are two evaluation indexes to evaluate the crowd counting algorithm: mean absolute value error(MAE) and mean square value error(MSE). The calculation formulas are as follows:

$$MAE=\frac{1}{N}\sum_{1}^{N}\left|z_i-\hat{z}_l\right| \quad (4)$$

$$MSE=\sqrt{\frac{1}{N}\sum_{1}^{N}\left(z_i-\hat{z}_l\right)^2} \quad (5)$$

Where, N represents the total number of test sequence images. z_i represents the actual number of people in the picture. And $\hat{z}_l$ is the estimated number of people in the picture.

C. DataSet

The data set ShanghaiTech Part_A was selected in this paper. ShanghaiTech data set is a crowd counting algorithm performance evaluation data set [3]. The data set was designed by Goldman sachs research group of Shanghai university of science and technology, and was first proposed and made public in CVPR2016.

This data set contains 1198 labeled crowd counting images, among which 330,165 human heads were manually labeled. This data set is the data set with the largest crowd labeling volume so far. The data set can be divided into two parts: Part_A and Part_B. The Part_A section is mainly composed of dense group samples, and contains 482 images of crowd gathering, which are randomly selected from the network. The images are of different sizes, including 300 training images and 182 test images. Among them, the largest number of people is 3,139, while the smallest number is only 33. The average number of people per picture is about 500. Large variations in crowd density between images make the

counting task more challenging than other data sets, making it more difficult for crowd counting algorithms to accurately count people. This experiment is mainly aimed at counting the images of dense crowd, so Part_A part of this data set is selected for training and observation and analysis of test results.

D. Experimental Results and Analysis

The experimental results of the above four different networks in ShanghaiTech Part_A dataset are shown in the Table 1.

TABLE I. The experimental results

Mode	MAE	MSE
MCNN	110.2	173.2
Switch-CNN	90.4	135.0
IG-CNN	72.5	118.2
MSCNN	83.8	127.4
Model-1（CSRNet）	69.7	116.0
Model-2（CSRNet）	68.2	115.0
Model-1+2（Multi-scale Fusion）	69.1	110.0
Model-3（series）	74.9	110.8
Model-1+3（parallel）	74.7	113.1
Model-1+2+3（Multi-branch network）	71.2	110.8

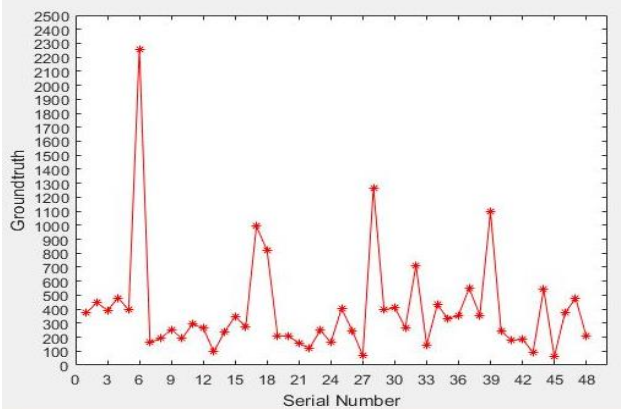
The crowd counting results obtained by using the network module 1 and module 2 alone, we refer to the test results of the proposed hollow convolutional neural network using different expansion rates in the literature CSRNet. It can be seen from the experimental results that the test results obtained by using conventional convolutional neural network in module 1 are better than those obtained by using dilated convolution with an expansion rate of 2 in module 2. It shows that the dilated convolution kernel expands the range of the sensing field to a certain extent, enables the network to gather multi-scale context information flexibly, and at the same time maintains the original resolution to prevent the loss of spatial information.By learning the research results of other scholars, it can be found that the test results of conventional convolutional neural network for crowd counting are not bad, so this paper combines the two cases for research (i.e., module 1 and module 2 are fused together).

The results show that MAE and MSE are better than module 1 alone. Although MAE is 0.9 worse than module 2 alone, MSE has improved obviously. It is shown that the network can extract more scale context information after combining with the dilated convolution, and the branches with different sensing fields are combined to improve the features extracted by the network, and the final estimated density map distribution is closer to the truth density map.

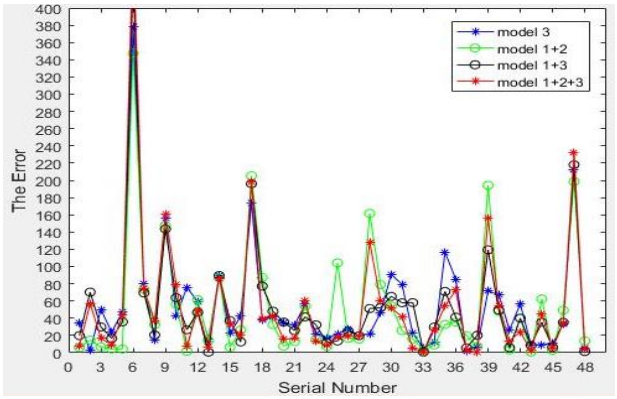
The attention mechanism is added into the network, and two fusion methods, series and parallel, are used respectively. When the attention module is used in series, the MSE test results are better. MAE is reduced by 0.2 when the attention module is used in parallel. Compared with CSRNet, the experimental results of the two methods have made significant progress in MSE. Therefore, the use of the attention module can indeed help improve the performance of the network and enhance the discrimination of features.

For the three-branch network structure, the experimental results MAE and MSE are better than the two methods adding the attention model, which shows the effectiveness of fused dilated convolution. Compared with CSRNet, its MSE is improved. However, MAE and MSE are slightly inferior to module-one and module-two fused networks. In this regard, the factors of crowd density level are considered in this paper, that is, dilated convolution and attention model have different effects on crowd images with different density levels. Therefore, 48 pictures were randomly selected from the test set for verification, and the total number of people and the error obtained by each method were statistically analyzed, as shown in Figure 5.

Figure 5 (a) is the actual number of people corresponding to each picture, and figure 5 (b) is the error obtained by testing the above four methods. It can be seen from the figure that the attention model is better added to the pictures of people with high density. For the crowd pictures with low density, it is better to add dilated convolution. It verifies our conjecture that dilated convolution and attention model have different effects on crowd images with different density levels. The following two pictures with different density levels were selected, and the test results were obtained through four different networks.



(a) .The actual number of people corresponding to each picture



(b) .The error obtained by testing the above four methods

Fig.5. Error comparison of four models

The test result graph of each picture is divided into four parts, the first is the original input picture, the second is the truth density graph, the third is the density graph of network output, and the fourth is the error between the estimated density graph of network and the truth density graph. There are two values in the true-value picture, and the top one is the sum of the original true-value density map.

Because VGG16 will reduce the input image sampling, for example, the final output density map of the network is 1/8 of the original image size, and the truth density map generates the original input image size. Then, in order to train the network, the truth density map should be reduced to 1/8 of the original size. The cv2.imresize function is used here. However, this would cause a problem. The truth map after resize lost many values (because many spatial pixels disappeared), and the sum was not equal to the number of original images. Therefore, the simple transformation is to multiply the truth density map after resize by 64 (8x8), and the value will be the same as before, which is the one at the bottom of the two Numbers in the picture. By observing the density map, the attention model is more suitable for the pictures of people with high density, while the dilated convolution is prominent in the pictures of people with low density, which further verifies the conjecture of this paper.

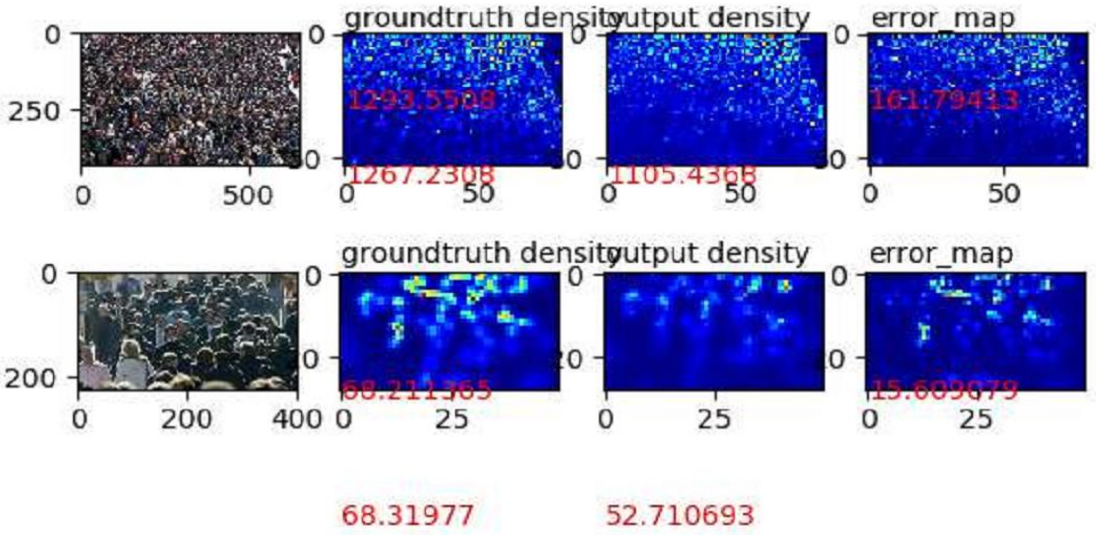


Fig.6. Model 1+2（Multi-scale fusion）

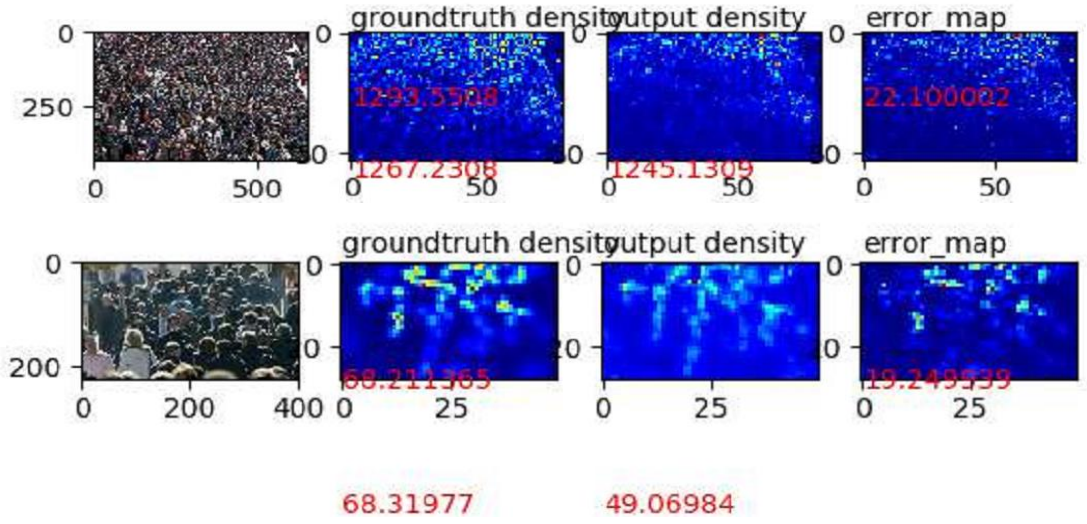


Fig.7. Model 3 (Series connection)

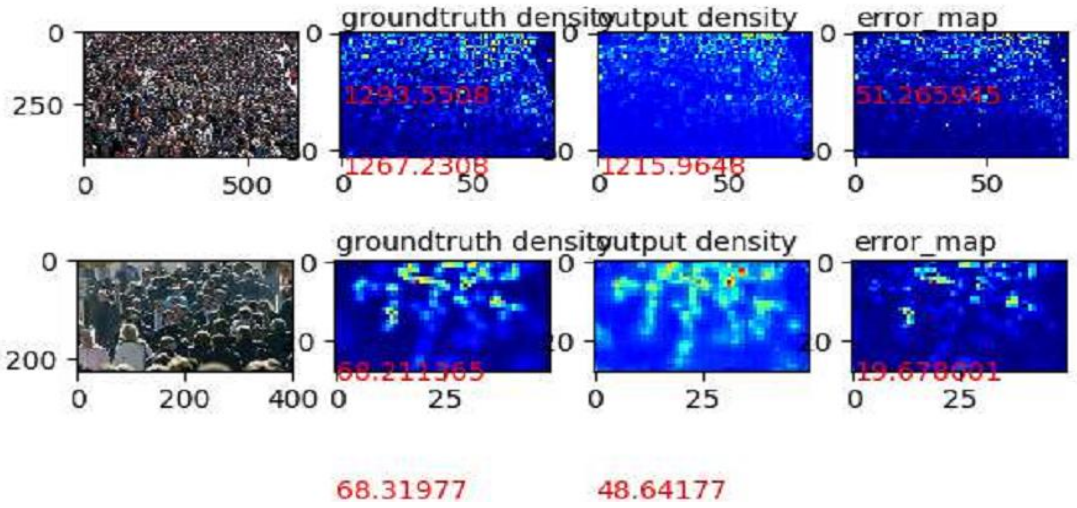


Fig.8. Model 1+3 (Module Three Parallel Connection)

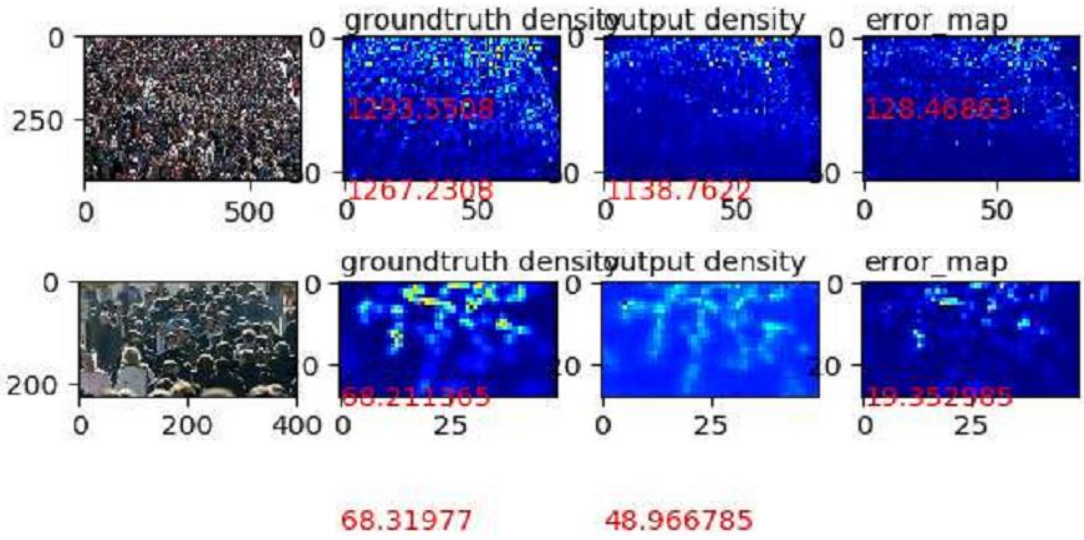


Fig.9. Model 1+2+3（Multi-branch structure)

V. Conclusion

In this paper, we design a multi-scale network structure with dilated convolution and introduce the attention mechanism into the crowd counting problem. The dilated convolution is used to extract the context information of different scales of the image, and the attention size allocated to the crowd and the background by the attention model is used to highlight the crowd position and better reflect the crowd distribution. In this way, more effective estimation characteristics of crowd density map can be provided, and the problem of dissimilar distribution of density map can be improved. The experimental results show that the characteristics extracted from the network can be improved by combining the branches with different sensing fields by combining the dilated convolution. The use of attention module is helpful to improve the performance of network and enhance the discrimination of features. For crowd images with different density levels, the effect of dilated convolution and attention model is different. For this aspect, we will carry out further research.

References

[1] Hu J, Shen L, Sun G.Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[2] Sindagi V A, Patel V M. A survey of recent advances in cnn-based single image crowd counting and density estimation[J]. Pattern Recognition Letters, 2018, 107: 3-16.

[3] Zhang Y, Zhou D, Chen S, et al. Single-image crowd counting via multi-column convolutional neural network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 589-597.

[4] Sam D B, Surya S, Babu R V. Switching convolutional neural network for crowd counting[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 4031-4039.

[5] Wang S, Zhao H, Wang W, et al. Improving Deep Crowd Density Estimation via Pre-classification of Density[C]//Proceedings of the IEEE International Conference on Neural Information Processing. Springer, Cham, 2017: 260-269.

[6] Sindagi V A, Patel V M. Generating high-quality crowd density maps using contextual pyramid cnns[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 1861-1870.

[7] Li Y, Zhang X, Chen D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1091-1100.

[8] Liu J, Gao C, Meng D, et al. Decidenet: Counting varying density crowds through attention guided detection and density estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5197-5206.

[9] Hossain M, Hosseinzadeh M, Chanda O, et al. Crowd counting using scale-aware attention networks[C]//Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019: 1280-1288.