ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0040

Crowd Counting Algorithm based on Fusion of Channel Spatial Attention Mechanism

Peilong Yang, Shuyue Chen*, Jiahong Wang, Shangyu Yang

School of Microelectronics and Control Engineering, Changzhou University, Changzhou, Jiangsu 213164, China

Abstract

A new crowd counting algorithm, DF-CSAM, is proposed to solve the problems of scale variation and background interference in crowd counting tasks. The algorithm consists of two parts: a front-end network that fuses sampling layers using a channel-spatial attention mechanism and a back-end network that uses dilated convolutional module. The sampling layer fusion method used by the front-end network is to redistribute the channel space weights of the feature map through the channel space attention mechanism, to solve the problems of background interference and scale change. The dilated convolutional layers used in the back-end network, whose outputs contain rich spatial and global information, are beneficial for generating high-quality crowd density maps. Comparative experiments are carried out on the ShanghaiTech dataset, UCF_CC_50 dataset and UCF_QNRF dataset, respectively, showing the effectiveness of the algorithm.

Keywords

Crowd Counting; Crowd Density Map; Channel Space Attention Mechanism; Dilated Convolution.

1. Introduction

Crowd counting[1] is a branch of computer vision that counts the number of people in an image from a given crowd image. On April 30, 2021, when a bonfire festival was held in northern Israel, a severe stampede occurred, which eventually resulted in the death of 45 people and hundreds of injuries. Therefore, in order to avoid the recurrence of such incidents, it is very necessary to carry out accurate crowd counting when holding large-scale events. Accurate crowd counting in such crowded scenes is very challenging due to the mutual occlusion of crowds, scale changes, light projection distortions, and spatial occlusions[2].

Before 2015, the research direction of crowd counting was dominated by traditional image processing methods, which were divided into the following two types: detection-based methods[3-5] and regression-based methods[6,7]. Both detection-based and regression-based methods have their drawbacks, and the counting results are not very good in crowded scenes. Crowded crowds cause pedestrians to occlude each other, and detection-based methods cannot effectively identify each pedestrian. At the same time, in a crowded crowd, the scale of each pedestrian varies greatly, and some key information cannot be extracted manually, so regression-based methods cannot solve this problem.

In recent years, the strong performance of convolutional neural networks in the field of computer vision has made accurate crowd counting possible[8,9]. The current mainstream research method is no longer to count the number of people directly, but to transform the counting problem into the summation problem of the density map. Zhang et al. first proposed a method of using convolutional networks to regress crowd density maps to deal with the crowd counting problem[1]. In order to

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0040

extract feature information at different scales, a multi-column convolutional network model MCNN was proposed. However, after experimental verification by Li et al. , the results show that the features extracted by different branches of MCNN are almost the same[10]. It can be seen that using a multi-column convolutional neural network model cannot effectively solve the problems of scale change and background interference. In addition, with the deepening of multi-column convolutional network layers, the amount of computation and computational complexity will increase exponentially.

In response to the above problems, a new crowd counting algorithm DF-CSAM (Deep Fusion network through Channel-Spatial Attention Module) is proposed. The algorithm consists of two parts: the front-end network based on the first 13 layers of VGG-16[11] and the backend network consists of 5 dilated convolutional layers and one convolutional layer. On the one hand, when the front-end network extracts features, in order to ensure that the output feature map is the same size as the input image. In this case, the background interference and scale changes cannot be effectively solved if only using the inverse sampling operation to improve the resolution. The CBAM[12] module proposed by Woo et al. is a lightweight general-purpose module that can be fused with any convolutional network to improve its performance, while adding a small amount of computation. Based on the above advantages, this paper draws on the method of Woo et al. and transforms it into a channel spatial attention module to fuse the down-sampling layer and the up-sampling layer[12]. Such a fusion sampling layer can extract richer spatial information and pedestrians. feature information, which helps to address background disturbance and scale variation. On the other hand, the dilated convolutional layer of the back-end network can obtain a larger receptive field without increasing parameters, and its output contains richer spatial and global information[10]. Therefore, the use of dilated convolutional layers in the back-end network is beneficial to generate higher-quality crowd density maps.

Finally, we choose to use Huber loss[13] as the loss function of our method. The choice of this algorithm as the loss function is based on the following two aspects: First, this algorithm can speed up the model convergence speed and reduce the model training time. Second, compared to MSE loss, Huber loss[13] is more robust to outliers. Therefore, using this algorithm can effectively save training time and improve the robustness of the algorithm.

In summary, the contributions of this paper in this matter are as follows:

(1) A new single-column convolutional network model is proposed, which consists of a front-end feature extraction network and a back-end feature fusion network. The fusion sampling layer in the front-end feature extraction network can effectively solve the problems of scale change and background interference. The dilated convolution layer used in the back-end feature fusion network expands the receptive field of convolution, and its output contains richer spatial and global information, which is beneficial to generate higher-quality crowd density maps.

(2) A channel spatial attention module is designed to fuse the sampling layers in the front-end network. It avoids the loss caused by the direct fusion of the sampling layer. The channel spatial attention mechanism captures the connection between the feature channel information and the spatial information by reassigning the channel and spatial weights of the feature map, so that the final generated crowd density map is in the pixel neighborhood. Smooth transitions between domains.

(3) On the ShanghaiTech dataset[1], UCF_CC_50 dataset[14] and UCF_QNRF dataset[15], a large number of comparative experiments are conducted, which proves that the algorithm proposed in this paper has better counting performance.

2. Counting Algorithm based on the Fusion of Channel Spatial Attention Mechanism

As the number of layers of the multi-column convolutional network model deepens, the number of parameters and computational complexity will increase exponentially. Therefore, this paper proposes a new single-column convolutional network model, which improves the feature extraction ability of

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0040

the model by adding a channel spatial attention module to the sampling layer of the front-end network, while maintaining the simplicity of the model structure.

The structure diagram of the DF-CASM network model is shown in Figure 1. In this section, the proposed algorithm will be explained from the downsampling-upsampling fusion module, the channel spatial attention module, and the dilated convolution module.

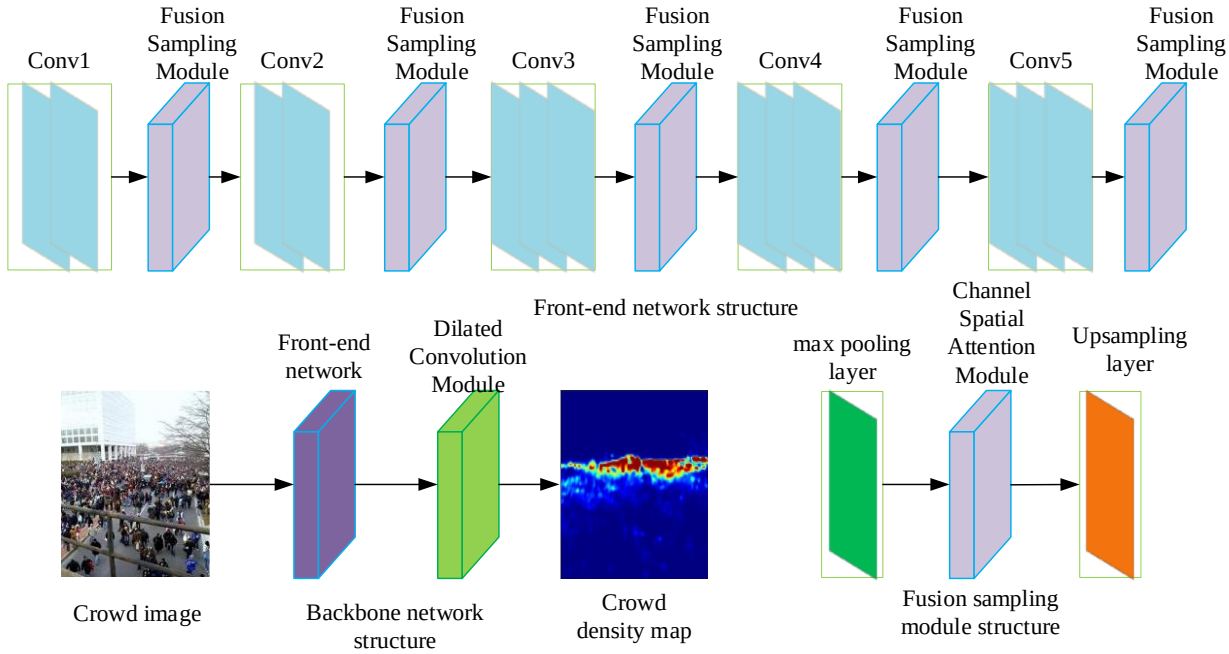


Figure 1. DF-CSAM network structure diagram

2.1 Downsampling-Upsampling Fusion Module

When collecting crowd counting images, due to the different positions of pedestrians from the camera, the size of the collected pedestrians is different, that is, there is a problem of scale diversity. In order to extract pedestrian features of different scales and also to solve the problem of background interference, a downsampling-upsampling fusion method is proposed.

As shown in Figure 1, the backbone of the front-end network is based on VGG-16[11] and retains the first 13 convolutional layers, which has strong feature extraction ability and simple structure. In order to extract feature information of different scales, it is necessary to use large-sized convolution kernels to expand the receptive field of convolution. However, such operations will bring a huge amount of calculation, and the use of large-sized convolution kernels is not conducive to building deeper networks[11]. In order to solve this problem, the network selected in this paper only uses 3×3 sized convolutions (except that the last convolutional layer uses 1×1 instead of a fully-connected layer), and the superposition of multiple 3×3 sized convolutions can obtain the same size as large-sized convolutions. For example, one 5×5 convolution can be replaced by two 3×3 convolutions, and one 7×7 convolution can be replaced by three 3×3 convolutions. By analogy, using this method can not only extract features of different scales, but also reduce the amount of computation to build a deeper network.

The functions of each layer in the network are different. The convolutional layer is used to extract features, and the sampling layer is used to filter features. If a network uses only convolutional layers without sampling layers, it is difficult to maintain a balance between two convolutional layers. The sampling layer in VGG-16[11] uses max pooling, which can improve the spatial invariance of features. However, this also leads to discontinuities in the information of features and reduces the resolution of feature maps, which is not conducive to pixel-level regression tasks[16]. In order for the model to output a feature map of the same size as the input image, upsampling is required to keep the resolution of the feature map unchanged. Therefore, more feature information can be extracted from the network

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0040

by fusing the downsampling layer with the upsampling layer using the channel spatial attention module.

As shown in Figure 1, based on the VGG-16[11] backbone network, a fused sampling layer is added between each convolutional layer. The feature map generated by the convolutional layer is input into the channel spatial attention module after maximum pooling. At the same time, the attention module reassigns the channel and spatial weights of the feature map to enhance the feature extraction ability of the network. Then through the upsampling operation (linear interpolation), the size of the feature map is restored to the size of the original input. The feature map output by the last fusion sampling layer is passed to the back-end network composed of dilated convolutional layer modules to generate a crowd density map. The following will introduce the channel spatial attention module and the dilated convolution module of the back-end network in detail.

2.2 Channel Spatial Attention Module

In recent years, with the deepening of research, the attention mechanism in neural networks has become more and more important, and the attention mechanism can be used in many research fields. For example, the CBAM module of Woo et al. can not only model the relationship between the feature map channels, but also model the interdependence on the spatial information of the feature map, so as to improve the expressive ability of the network.[12] On the one hand, many previous works do not consider the weights of channel and spatial features, but directly combine the feature maps generated by different convolutional layers; on the other hand, due to the lack of channel and spatial information, the feature map is Channel and spatial information are often ignored. This module redistributes the weight of the channel space information of the feature map, which can selectively strengthen the useful information, while suppressing the unimportant information, and help the network to extract more comprehensive feature information. Finally, the module itself is ingenious in design. Although there are so many benefits, adding this module to the model will not increase the complexity of the network and the consumption of computing resources.

In addition, it has been experimentally proved that the CBAM module can improve the performance of the model and can be embedded in any network to perform cumulative optimization on the entire network[12]. Therefore, this module is converted into the channel spatial attention module of this paper. The channel spatial attention module is divided into two sub-modules: the channel attention module and the spatial attention module. The following will introduce the channel attention module and the spatial attention module respectively. The structure diagrams are shown in Figures 2 and 3.

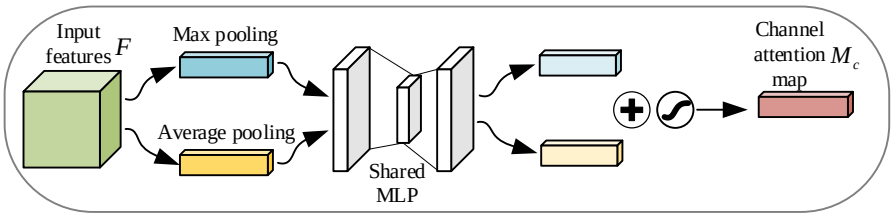


Figure 2. Structure diagram of channel attention module

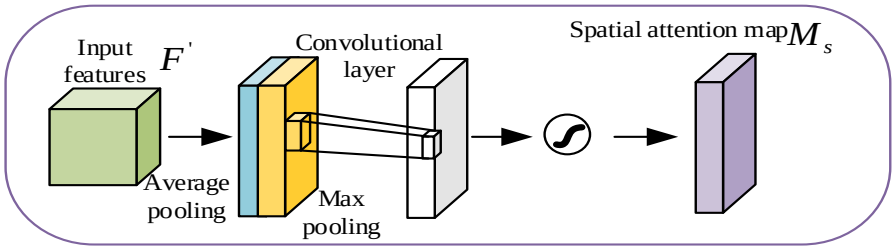


Figure 3. Structure diagram of spatial attention module

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0040

First, input the feature map generated by the convolutional layer, use average pooling and max pooling to extract the spatial information of the feature map, and generate the average pooling feature F_{avg}^{c} and the maximum pooling feature $F_{\max}^{c}$. Second, the average pooling features and max-pooled features are fed into a multilayer perceptron (MLP) to generate a channel attention map $\mathrm{M_c} \in \mathrm{R}^{\mathrm{C}\times 1\times 1}$. In order to reduce the amount of computation, the hidden layer of the multilayer perceptron is set to $\mathrm{R}^{\mathrm{C/r}\times 1\times 1}$, where r is the reduction ratio. Finally, element-wise summation is used to combine F_{avg}^{c} and F_{avg}^{c} processed by the multilayer perceptron. To sum up, the calculation method of channel attention is as follows:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F_{avg}^{c})) + W_1(W_0(F_{\max}^{c}))), \end{aligned} \tag{1}$$

where σ represents the sigmoid function, W_0 and W_1 represent the common parameters of the MLP. As a supplement to the channel attention module, the spatial attention module emphasizes the spatial information of the feature map more. Average pooling and max pooling are performed along the channel dimension of the input feature map, and pooling along the channel dimension can extract spatial information more effectively. The spatial attention map $M_s(F) \in R^{H\times W}$ is obtained by convolution operation on the pooled features, which encodes spatial information and suppresses irrelevant regions. The specific operation is as follows. First, the input feature map is pooled in the channel dimension to generate two two-dimensional feature maps: the average pooling feature map $F_{avg}^{s} \in R^{1\times H\times W}$ and the maximum pooling feature map $F_{\max}^{s} \in R^{1\times H\times W}$, and then they are connected in series. Finally, the convolution operation is performed on the concatenated values, and a two-dimensional feature map generated is the spatial attention map. To sum up, the calculation formula of spatial attention is as follows:

$$\begin{aligned} M_s(F) &= \sigma(f^{7\times 7}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(f^{7\times 7}([F_{avg}^{s}; F_{\max}^{s}])), \end{aligned} \tag{2}$$

where σ represents the sigmoid function, and $f^{7\times 7}$ represents the convolution operation with the convolution kernel size 7×7.

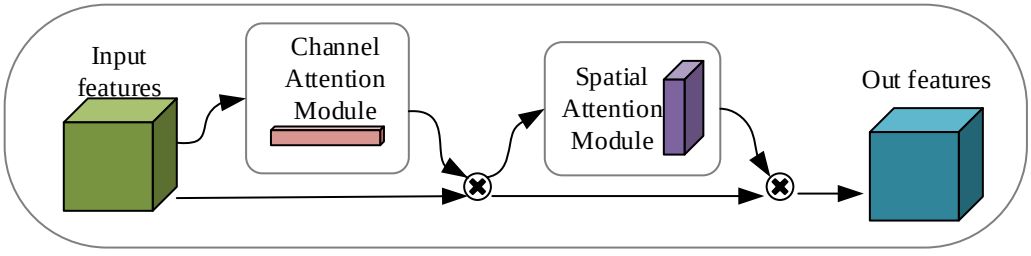


Figure 4. Structure diagram of channel spatial attention module

The channel attention module and the spatial attention module act as two complementary modules to extract "channel" and "spatial" information, respectively. Therefore, the two modules can be placed side by side or sequentially. Referring to the practice of Woo et al.[12] this paper places the two modules in sequence and places the channel attention module before the spatial attention module. Inputting the feature map $F \in R^{C\times H\times W}$ generated by the convolutional layer, the channel spatial

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0040

attention module calculates the channel attention map $M_c \in R^{C\times1\times1}$ and the spatial attention map $M_s \in R^{1\times H\times W}$ in turn, as shown in Figure 4. Briefly, channel spatial attention is calculated as follows:

$$\begin{aligned} F^{'} &= M_c(F)\otimes F, \\ F^{''} &= M_s(F^{'})\otimes F^{'}, \end{aligned} \tag{3}$$

where $\otimes$ represents element-wise multiplication, $F^{'}$ represents the output of the channel attention module, and $F^{''}$ represents the final output.

2.3 Channel Spatial Attention Modul

Dilated convolution is a type of convolution. A two-dimensional mathematical model of dilated convolution can be defined as:

$$H(m,n) = f(\sum_{m=1}^{M}\sum_{n=1}^{N} w_{m,n} x_{i\times r+m, j\times r+n} + b_b) \tag{4}$$

where $H(m,n)$ represents the output of the dilated convolution; $x_{i\times r+m, j\times r+n}$ represents the $i\times r$ row and $j\times r$ column elements of the input image; $w_{m,n}$ represents the parameter values of the m row and n column of the convolution kernel; b_b represents the bias term of the convolution; M, N expressed as the size of the convolution kernel; r is the dilation rate of the dilated convolution, when $r=1$, the dilated convolution is the standard convolution.

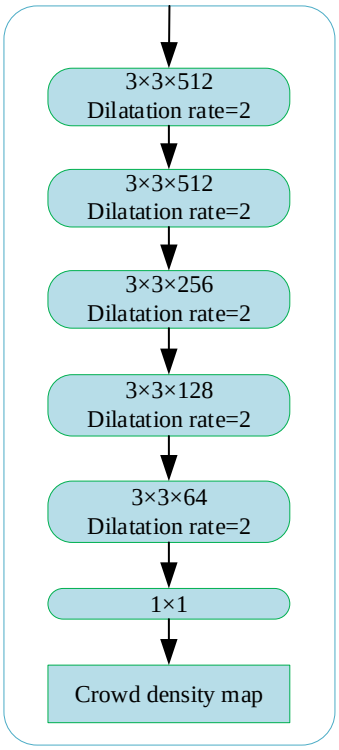


Figure 5. Schematic diagram of the structure of the dilated convolution module

The fusion sampling layer used in the front-end network can improve the feature ability of the network and solve the problems of background interference and scale change. However, using fused sampling layers still loses some original information. Therefore, directly using the features output by the front-

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0040

end network still cannot generate a high-quality crowd density map, which is not conducive to accurate crowd counting. According to Li et al. it is found that dilated convolutions can be used to solve this problem[10]. On the one hand, the dilated convolution can expand the receptive field of the convolution kernel without adding more parameters and calculations; on the other hand, the feature map can retain more detailed spatial and global information after passing through the dilated convolution layer, and at the same time, the resolution of the feature map will not be reduced, resulting in the loss of some original information.

The dilated convolution module of the back-end network is shown in Figure 5, which contains 5 dilated convolutional layers with a dilation rate of 2 and a standard convolutional layer of 1×1, which is used to regress the crowd density map.

3. Experiment and Result Analysis

First, the model in this paper uses the Pytorch framework, chooses to use the Adam optimizer to optimize the parameters of the network, and the initial learning rate is set to 0.00001. Secondly, the initial parameters of the front-end network are the pre-training parameters of VGG-16[11]. For other convolutional layers, the initial parameters are obtained by a Gaussian random function with a mean value of zero and a standard deviation of 0.01. Finally, in order to avoid vanishing or exploding gradients and improve the generalization ability of the network, bn and Relu layers are added after each convolutional layer (except the last output layer).

3.1 Generation of the True Value of the Crowd Density Map

Referring to the method of Zhang et al.[1], this paper also chooses to use a geometrically adaptive Gaussian kernel to generate the ground truth of the crowd density map. Because the distribution of the crowd is random, an adaptive Gaussian kernel can be used to blur the location information of each person to get a density map. The formula for the Gaussian kernel is defined as follows:

$$F(x)=\sum_{i=1}^{N}\delta(x-x_i)\times G_{\sigma_i}(x),\sigma_i=\beta\overline{d_i} \tag{5}$$

where δ represents the ground truth, x represents any pixel in the image, $x_i, i=1,2,...,N$ represents the head position of each pedestrian in the image, $\delta(x-x_i)$ represents an adaptive Gaussian kernel with a standard deviation of σ_i, $\overline{d_i}$ represents the nearest target of i Neighbor mean. According to the relevant experimental experience of Zhang et al.[1], this paper sets β and i to 0.3 and 3, respectively.

3.2 Generation of the True Value of the Crowd Density Map

This paper uses the following two criteria: mean absolute error (MAE) and root mean square error (RMSE) to measure the accuracy and robustness of the model. The mean absolute error (MAE) is used to measure the accuracy of the model, and the root mean square error (RMSE) is used to measure the robustness of the model. MAE and RMSE can be defined by the following formulas:

$$MAE=\frac{1}{N}\sum_{i=1}^{N}\left|G_i-E_i\right| \tag{6}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}|G_i-E_i|^2} \tag{7}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0040

where N represents the number of the entire test set, G_i represents the real number of people in the i crowd image, and E_i represents the model predicted number of the i crowd image.

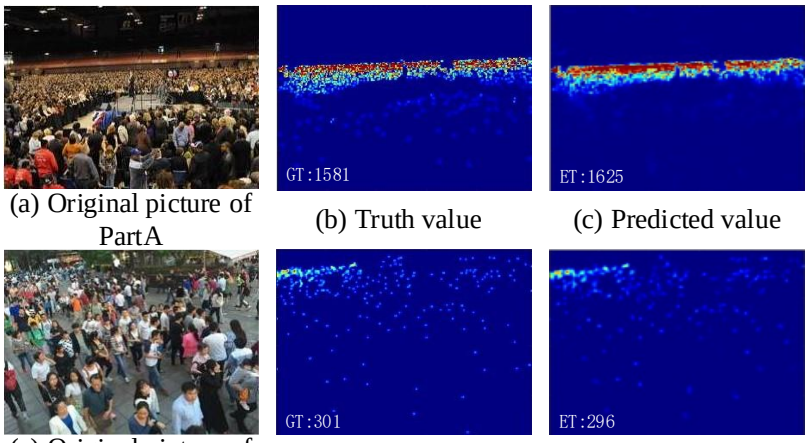
3.3 Experiments on the ShanghaiTech Dataset

The ShanghaiTech dataset was proposed by Zhang et al.[1] in 2016. The dataset consists of PartA and PartB. The crowd pictures in PartA are from the Internet and represent relatively crowded scenes; PartB is collected from busy and crowded streets in Shanghai and represents relatively sparse scenes. Referring to the experimental arrangement of Zhang et al.[1] in the PartA dataset, 300 images were selected for training, and the remaining 182 images were used for testing; in the PartB dataset, 400 images were selected for training, There are 316 images left for testing.

Table 1. Comparison of test results on ShanghaiTech dataset

Method	PartA		PartB	
	MAE	RMSE	MAE	RMSE
MCNN[1]	110.2	173.2	26.4	41.3
Switch-CNN[17]	90.4	135.0	21.6	33.4
SaCNN[18]	86.8	139.2	16.2	25.8
CSRNet[10]	68.2	115.0	10.6	16.0
FF-CAM[19]	71.0	109.8	10.3	15.8
Improved- CSRNet[20]	67.1	108.3	8.7	15.5
DF-CSAM(our)	65.3	105.4	8.5	15.0

Table 1 lists the comparison results of MAE and RMSE in this paper with other advanced methods. It can be seen from the table that this method is superior to other methods in both MAE and RMSE. In the test results on the PartA dataset, MAE and RMSE are improved by 2.7% and 2.6%, respectively, over the methods based on the improved-CSRNet[20]. In the test results on the PartB dataset, MAE and RMSE are improved by 2.3% and 3.2%, respectively, over the methods based on the improved-CSRNet[20]. The partial test results of the algorithm in the ShanghaiTech dataset are shown in Figure 6.



(a) Original picture of PartA (b) Truth value (c) Predicted value

(a) Original picture of PartB (b) Truth value (c) Predicted value

Figure 6. Partial experimental results on the ShanghaiTech dataset

3.4 Experiments on the UCF_50 Dataset

The UCF_CC_50 dataset was proposed by Idress et al.[14] in 2012. The dataset has only 50 images of different perspectives and resolutions. The number of pedestrians contained in each image is extremely uneven from 94 sparse people to 4543 crowded people. Therefore, this is an extremely challenging crowd counting dataset. In order to fully utilize each sample, 5-fold cross-validation is

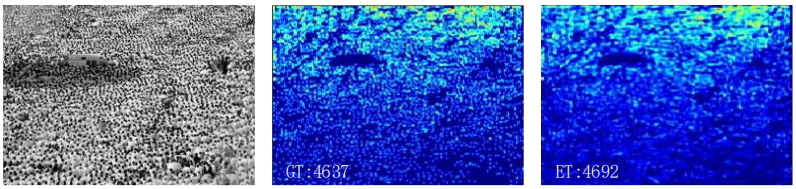
ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0040

required on the dataset. 5-fold cross-validation: The data set is randomly divided into five equal parts, four of which are taken as the training set, and the remaining one is used as the test set. A total of five training tests are performed. The experimental results are shown in Table 2. Finally, the average result of five experiments is used as the final result of our algorithm on this dataset.

Table 2. 5-fold cross-validation results on UCF_CC_50 dataset

Serial number	MAE	RMSE
1	144.4	179.3
2	190.0	239.0
3	276.9	306.2
4	274.1	316.3
5	205.3	243.3
Average value	218.1	256.8

Table 3 lists the comparison results of MAE and RMSE in this paper with other advanced methods. It can be seen from the table that the method proposed in this paper is superior to other methods in both MAE and RMSE. Compared with the improved-CSRNet[20] algorithm , the MAE and RMSE are improved by 2.0% and 2.2%, respectively. The partial test results of the algorithm in the UCF_CC_50 dataset is shown in Figure 7.



(a) Original picture of UCF_CC_50 (b) Truth value (c) Predicted value

Figure 7. Experimental results on the UCF_CC_50 dataset

Table 3. Comparison of test results on UCF_CC_50 dataset

Method	MAE	RMSE
MCNN[1]	377.6	509.1
Switch-CNN[17]	318.1	439.2
SaCNN[18]	314.9	424.8
CSRNet[10]	266.1	397.5
FF-CAM[19]	246.8	322.2
Improved-CSRNet[20]	222.5	262.5
DF-CSAM(our)	218.5	256.8

3.5 Experiments on the UCF_QNRF Dataset

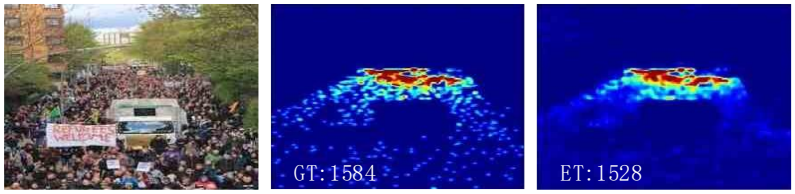
The UCF_QNRF dataset was proposed by Idress et al.[15] in 2018, and the dataset contains a total of 1535 crowd pictures. Among them, 1201 crowd images are selected as the training set, and the remaining 334 crowd images are selected as the test set. Compared with the above two datasets, the UCF_QNRF dataset contains more scenes, perspectives and crowd density changes, so it is more suitable for training neural networks.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0040

Table 4. Comparison of test results on UCF_QNRF dataset

Method	MAE	RMSE
MCNN[1]	277	426
Switch-CNN[17]	228	445
Idress et al.[15]	132	191
FF-CAM[19]	114.5	200.5
Improved-CSRNet[20]	-	-
DF-CSAM(our)	111.4	187.3

The comparison results of MAE and RMSE in this paper with other state-of-the-art methods are listed in Table 4. Compared with the FF-CAM[19] algorithm, the MAE and RMSE are improved by 2.7% and 6.6%, respectively. The partial test results of the algorithm in the UCF_QNRF dataset are shown in Figure 8.



(a) Original picture of UCF_QNRF (b) Truth value (c) Predicted value

Figure 8. Experimental results on the UCF_QNRF dataset

3.6 Ablation Experiment

In order to verify the effectiveness of the DF-CSAM algorithm proposed in this paper, an ablation experiment of the algorithm was performed on the ShanghaiTech Part A dataset[1], and the experimental results are shown in Table 5.

As can be seen from the table, the MAE and RMSE of our method are improved by 22.9% and 18.1% respectively compared with the VGG-16 benchmark, and the results show that the proposed model structure can significantly improve the prediction accuracy and robustness. After removing the Channel Spatial Attention Module, the MAE and RMSE of the model decrease by 4.5% and 6.5%, respectively, which indicates that the channel spatial attention module can effectively improve the performance of the whole model. After removing the dilated convolution module, the MAE and RMSE of the model drop by 2.5% and 5.6%, respectively, which proves that the dilated convolution module can also effectively improve the performance of the entire model.

The above experiments prove that each module of the model in this paper can effectively improve the performance of the entire model, which also verifies the effectiveness of the algorithm.

Table 5. Comparative results of ablation experiments

Method	MAE	RMSE
VGG-16 benchmark	84.7	128.7
Remove the channel spatial attention module	68.4	112.7
Remove the dilated convolution module	67	111.6
DF-CSAM	65.3	105.4

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0040

4. Conclusion

In this paper, an end-to-end crowd counting algorithm called DF-CSAM is proposed. The front-end network uses a channel spatial attention module to fuse the sampling layers of the front-end network, which solves the problems of background interference and scale variation. The dilated convolutional layer of the back-end network aggregates the spatial and global information of the feature map output by the front-end network to generate a high-quality crowd density map. The algorithm proposed in this paper has the characteristics of simplicity and stability. It has obtained good test results on the ShanghaiTech dataset, the UCF_CC_50 dataset and the UCF-QNRF dataset. improved compared to all. In future work, we plan to use more advanced attention mechanisms to optimize the proposed algorithm so that it can adapt to more counting scenarios and strive to improve the accuracy and robustness of the algorithm.

Acknowledgments

This paper is funded by The Project: Key Research and Development Program of Jiangsu Province (BE2021012-5).

References

[1] 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.

[2] Zhang A, Yue L, Shen J, et al. Attentional neural fields for crowd counting[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 5714-5723.

[3] Enzweiler M, Gavrila D M. Monocular pedestrian detection: Survey and experiments[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 31(12): 2179-2195.

[4] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), 2005: 886-893.

[5] Li M, Zhang Z, Huang K, et al. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection[C]. 2008 19th international conference on pattern recognition, 2008: 1-4.

[6] Chen K, Loy C C, Gong S, et al. Feature mining for localised crowd counting[C]. Bmvc, 2012: 3-12.

[7] Lempitsky V, Zisserman A. Learning to count objects in images[J]. Advances in neural information processing systems, 2010, 23: 1324-1332.

[8] Liu X, Yang J, Ding W. Adaptive mixture regression network with local counting map for crowd counting[J]. arXiv preprint arXiv:2005.05776, 2020.

[9] Wan J, Kumar N S, Chan A B. Fine-grained crowd counting[J]. IEEE transactions on image processing, 2021, 30: 2114-2126.

[10] 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.

[11] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[C]. International Conference on Learning Representations (ICLR), 2014: 1-14.

[12] Woo S, Park J, Lee J-Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European conference on computer vision (ECCV), 2018: 3-19.

[13] Esmaeili A, Marvasti F. A novel approach to quantized matrix completion using huber loss measure[J]. IEEE Signal Processing Letters, 2019, 26(2): 337-341.

[14] Idrees H, Saleemi I, Seibert C, et al. Multi-source multi-scale counting in extremely dense crowd images[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2013: 2547-2554.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0040

[15] Idrees H, Tayyab M, Athrey K, et al. Composition loss for counting, density map estimation and localization in dense crowds[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 532-546.

[16] Dai F, Liu H, Ma Y, et al. Dense scale network for crowd counting[C]. Proceedings of the 2021 International Conference on Multimedia Retrieval, 2021: 64-72.

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

[18] Zhang L, Shi M, Chen Q. Crowd counting via scale-adaptive convolutional neural network[C]. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018: 1113-1121.

[19] Zhang Yuqian, Li Guohui, Lei Jun, et al. FF-CAM: Crowd Counting Based on Front-End Fusion of Channel Attention Mechanism [J]. Chinese Journal of Computers, 2021, 44(02): 304-317.

[20] Guo Haoqi, Yang Jie, Kang Zhuang. Crowd Counting Algorithm Based on Improved CSRNet [J]. Sensors and Microsystems, 2022, 41(06): 150-152+156.