Multimedia Tools and Applications (2020) 79:29145–29159
https://doi.org/10.1007/s11042-020-09541-x

Two stages double attention convolutional neural network for crowd counting

Zhao Zou[1] · Chaofeng Li[1] · Yuhui Zheng[2] · Shoukun Xu[3]

Received: 12 April 2020 / Revised: 9 July 2020 / Accepted: 4 August 2020 /
Published online: 8 August 2020

Abstract
Crowd counting has captured wide attention in computer vision, which aims to accurately count the number of people in still images or video scenes. However, it's still a challenging task due to the scale variation and cluttered background in crowd scenes. In this paper, we propose a 2-stage Double Attention convolutional neural network for crowd counting, and call it 2-DA-CNN, which could deal with scale variation and cluttered background in crowd counting. The proposed 2-DA-CNN includes three parts. The first part is the front-end module which consists of a set of convolution operations, whose function is to extract abundant feature of crowd. The second part is the first double attention module, which contains trunk branch and mask branch. The former is mainly composed by multi-column CNN module, which is to deal with scale variation in crowd scenes. The latter can generate two masks, which aims to assign interesting regions reasonably in cluttered situation. The third part is the second double attention module, similar to the first double attention module, which can enhance the performance of multi-column CNN module further. In addition, we propose progressive training method to improve the drawback of using geometry-adaptive kernels to generate ground truth. The experimental results on three mainstream datasets (ShanghaiTech part B, ShanghaiTech part A and UCF_CC_50) suggest that the proposed 2-DA-CNN is competitive with the state-of-the-art methods.

Keywords Crowd counting · Convolutional neural network · Double attention · Progressive training

✉ Chaofeng Li
wxlichaofeng@126.com

[1] Institute of Logistics Science and Engineering, Shanghai Maritime University, Shanghai 201306, China

[2] College of Computer and Software, Nanjing University of Information Science & Technology, Nanjing 210044, China

[3] School of Information Science and Engineering, Changzhou University, Changzhou 213164, China

1 Introduction

Crowd counting question aims to estimate the number of people in still images or dynamic videos. It can be used to analyze abnormality, alleviate serious occlusions and reduce some security issues in dense crowd scenes. With the development of artificial intelligence, many intelligent researches [1, 18] have a great influence on our daily life. Crowd counting has also been widely used in crowd monitoring [4], scene understanding [22, 34] and safety management [3]. However, crowd counting remains to be a challenge task due to scale variation and cluttered background in crowd scenes.

The methods of crowd counting can be classified into three categories: the detection-based methods [7, 8, 14, 20], the regression-based methods [5, 19, 29] and the density map estimation-based methods [10, 35, 36]. The detection-based methods use a sliding window detector to estimate the number of people, but they have poor performance in dense crowd scenes. The regression-based methods count the number of people by learning a mapping between the extracted features of crowd images and the number of people, but they couldn't express crowd distribution. The density map estimation-based methods can not only effectively estimate the number of people, but also get the crowd distribution, so it has already become hot topic.

At present, with the development of deep learning, it has widely used for convolutional neural network in the density map estimation-based methods. Fu *et.al* [9] first applied CNNs on crowd counting. However, it only estimated density level of images and didn't give specific number of people in images. Zhang *et.al* [31] found the existing methods dropped significantly in an unseen scene, which is usually caused by varied crowd distribution, different number of people and kinds of background in crowd scenes. To overcome this issue, they proposed a crowd CNN model and designed a nonparametric fine-tuning to improve the performance of this model in cross-scene crowd counting. To deal with scale variation in crowd counting, Zhang *et.al* [33] designed a multi-column CNN called MCNN. Intuitively, each column of MCNN has different receptive filed to extract different scale information. Despite this model could extract multi-scale information in crowd scenes, it also leads to the redundancy of information. To solve this problem, Sam *et.al* [21] proposed Switch-CNN, which includes one classifier and three regressors. Firstly, the classifier estimates the most appropriate regressor according to the density level of image. Then, the image is transferred to the most appropriate regressor to generate the estimated density map. Using the classifier to select the suitable regressor, this method alleviates the redundancy of multi-scale CNN, but the training process becomes complicated. Sindagi *et.al* [24] presented a Contextual Pyramid CNN, which consists of four parts: Global Context Estimator, Local Context Estimator, Density Estimator and Fusion CNN. The first two estimators aim to capture the contextual information on the patches, and Density Estimator transforms the input image into a set of high-dimensional feature maps. At last, the output of the first three modules will be transferred to Fusion CNN to generate the final density map. Li *et.al* [15] designed CSRNet, which uses a modified VGG [23] structure as the front-end module and a set of dilated convolution as the back-end module. Although the CSRNet has great progress in crowd counting, it is time-consuming. Ranjan *et.al* [26] introduced an iterative crowd counting framework, which combines low resolution feature maps with high resolution feature maps to generate high-resolution density map. Liu *et.al* [17] incorporated the multi-scale contextual information into an end-to-end trainable pipeline CAN, which is beneficial to exploit the right context at each image location. Cao *et.al* [2] proposed an encoder-decoder network, in which the encoder is to extract multi-scale features and the

decoder can generate high-resolution density maps by using a set of transposed convolutions. Wang *et.al* [27] combined dilated convolution with normal convolution to construct a Multi-scale-CNN, which could aggregate various context information systematically in crowd counting.

In recent years, many researchers introduced attention mechanism to deep convolutional neural network to tackle crowd counting. Liu *et.al* [16] proposed the DecideNet which combines the detection-based with the regression-based method together. To extract information effectively, they utilized attention module to assess the reliability of the two types of estimation adaptively. Gao *et.al* [11] designed the Spatial-/Channel-wise Attention Regression Network to estimate the density map, which uses the Spatial-wise Attention Model to encode pixel-level information for the entire image and the Channel-wise Attention Model to extract discriminative features among different channels. Zhang *et.al* [34] introduced a MRA-CNN which makes use of attention mechanism to automatically focus on head regions. In order to deal with highly congested scenes in crowd counting, Sindagi *et.al* [25] constructed Hierarchical Attention-based Crowd Counting Network, which combines a spatial attention module with a set of global attention modules. The spatial attention module can select interesting regions in the feature maps, which is beneficial to dynamically enhance the feature responses. The global attention mechanism is similar to the channel-wise attention mechanism, whose module calculates attention along the channel dimension. Hossain *et.al* [12] designed a scale-attention mechanism to copy with the scale variation in crowd scenes. Chen *et.al* [6] proposed a novel end-to-end model called CAT-CNN, which utilizes attention mechanism to assess the importance of a head at each pixel location.

According to current research status, there still be a series of challenges for crowd counting question as following.

(1) Various scales of the people in the images. Due to the different distance between camera and people, the scale of people might be significant variation in the images. Thus, we need to get multi-scale features to reduce the error in crowd counting.
(2) Cluttered background in crowd images. There are many buildings, trees and various objects in crowd images, which often be discriminated as head of people in the estimated density maps. In the other word, we should eliminate as much complex background information as possible to count people accurately.
(3) Nonuniform crowd distribution in images. Usually, we use fixed kernel to generate ground truth for sparse crowd images and geometry-adaptive kernels for dense crowd images. However, the crowd density is nonuniform in different regions for an image. The sparse crowd regions and the dense crowd regions should be taken into account for a crowd image.

In order to estimate the number of crowd accurately, we propose a 2-stage double attention convolutional neural network (2-DA-CNN) to deal with these problems, which multi-column CNN is used to construct multi-scale network, and the double attention module in two stages is designed for generating two masks to reduce the impact of cluttered background. Then, for dealing with the nonuniform crowd density in images, we propose progressive training method by combining the advantages of geometry-adaptive kernels with fixed kernel. Finally, experimental results on three mainstream datasets demonstrate the advantages of our proposed 2-DA-CNN. In summary, our main contributions are as below:

(1) We analyze the drawbacks of popular multi-scale CNN, and propose a 2-DA-CNN for crowd counting, which can effectively deal with scale variation and cluttered background in crowd scenes.
(2) We construct a novel double attention model, which could generate two masks to assign weight reasonably for the regions of interest in feature maps. It is beneficial to extract more effective features, and generate high-quality density maps for crowd counting.
(3) During training, we design a progressive training strategic, which improves the drawback of using geometry-adaptive kernels to generate ground truth.

The remainder of the paper is organized as follows. Section 2 presents the proposed methodology. The results and detailed analysis are introduced in Section 3. Finally, we make a conclusion in Section 4.

2 Methodology

2.1 Overview

The structure of proposed 2-DA-CNN is shown in Fig. 1, which consists of three parts: the front-end module, the first double attention module and the second double attention module. The numbers of feature map are given at the top of models in Fig. 1, which the first number in the bracket is the number of input feature maps, and the second number is the number of output feature maps. In our method, only the stride of convolution kernels in the front-end module is set to 2, the stride of other convolution kernels is 1, and all convolution kernels are initialized with uniform distribution.

The front-end module contains 10 convolutional layers, whose structure is "Conv(3, 64)-Conv(3, 64)-MP-Conv(3, 128)-Conv(3, 128)-MP-Conv(3, 256)-Conv(3, 256)-Conv(3, 256)-MP-Conv(3, 512)-Conv(3, 512)-Conv(3, 512)". "Conv(n, m)" represents the convolutional layer with m filters whose size is $n \times n$, and "MP" denotes a 2×2 max-pooling layers with a

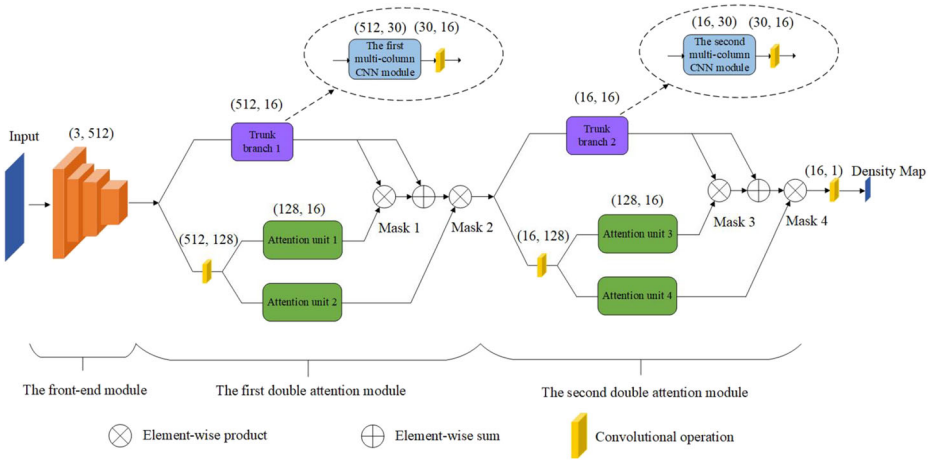

Fig. 1 Architecture of the proposed 2-DA-CNN

stride length of 2. The front-end module is used to filter complex background information and extract effective features in crowd images.

The first double attention module consists of trunk branch 1 and mask branch 1. The trunk branch 1 includes the first multi-column CNN module and 1 × 1 convolution. The former is used to extract multi-scale feature, whose structure will be introduced in section 2.2. The latter is used to adjust 30 multi-scale feature maps to 16 feature maps. As for mask branch 1, we firstly use 1 × 1 convolution filter to adjust 512 feature maps to 128 feature maps which is the input of attention units. Then, attention units of the first double attention module will generate masks (Mask 1, Mask 2) which could assign weight reasonably for different interesting regions. Finally, the masks are integrated with the output of trunk branch 1 to guide feature extraction. Their detailed introduction is given in section 2.3.

In order to generate high-quality density map, we use two double attention modules to guide crowd counting. The second double attention module has the same structure with the first, and only the number of input feature maps is different. At last, we use 1 × 1 convolution filters to generate the estimated density map.

2.2 Multi-column CNN module

To deal with scale variation in crowd counting, we design a multi-column CNN module shown in Fig. 2, where "Conv(m, n, k)" denotes that the size of the convolution kernel is $k \times k$, and the first two digits in the bracket denote the number of input feature maps and the number of output feature maps separately, and "C" denotes the operation of concatenation. Considering the importance of resolution in crowd counting, we don't use the pooling layers in multi-column CNN module. We find that the value of feature varies around zero in multi-column CNN module, thus we also remove ReLu which could inactivate neurons less than zero.

Despite the unequal number of input feature maps for different stages, the second multi-column CNN module has the same structure with the first multi-column CNN module in Fig. 1. The input of the first multi-column CNN module is 512 feature maps, and the input of the second multi-column CNN module is 16 feature maps.

2.3 Double attention module

The model of attention mechanism consists of trunk branch and mask branch. All weights form the mask to identify the interesting regions for images in mask branch. The output of attention model can be calculated by dot production which is between mask and the output of trunk branch. The mask branch is used to improve the performance of the whole model further.

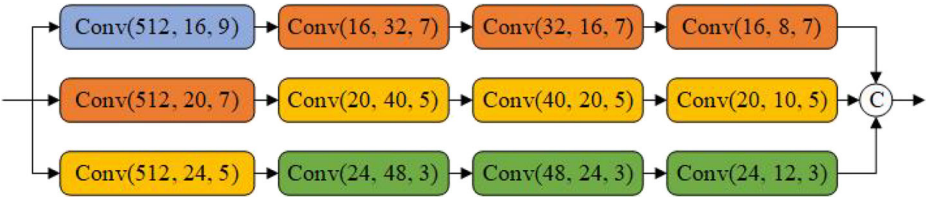

Fig. 2 Architecture of Multi-column CNN module

When we sequentially use attention mechanism to improve the performance, dot productions will be repeatedly used between the output of trunk branch and mask whose range is [0, 1]. In this case, to maintain the desired outcome of the whole model, the output of trunk branch in deep layers would continue to increase. It means that the ideal weight distribution in trunk branch could be broken, which would lead to the poor performance.

Inspired by Wang [28], we design a double attention mechanism to solve this problem, whose formulation is described as follows:

$$H_{i,c}(x_{i,c}) = (1 + M_{i,c}(x_{i,c})) \times N_{i,c}(x_{i,c}) \times F_{i,c}(x_{i,c}) \tag{1}$$

$$M_{i,c}(x_{i,c}) = \frac{1}{1 + exp(-s_1(x_{i,c}))} \tag{2}$$

$$N_{i,c}(x_{i,c}) = \frac{1}{1 + exp(-s_2(x_{i,c}))} \tag{3}$$

where $H_{i,c}(x_{i,c})$ is the output of double attention module for i*th* pixel in c*th* channel. $M_{i,c}(x_{i,c})$ and $N_{i,c}(x_{i,c})$ denote the output of corresponding attention unit, whose range of variation is [0, 1]. $F_{i,c}(x_{i,c})$ indicates the output of trunk branch. And s(·) is the scoring function of corresponding attention unit, whose result can be used to compute the importance of the pixel in crowd images.

From Eq. (1), we can find the range of variation for $H_{i,c}(x_{i,c})$ is [0, 2F], which is beneficial to keep the good property of the trunk branch, and improve the performance of model.

The structure of double attention module is shown in Fig. 3 and the structure of attention units is "C(128,64,3)-ReLu-C(64,16,3)-Sigmoid", where the numbers in the brackets separately denote the amount of input feature maps, the amount of output feature maps and the size of convolution kernels.

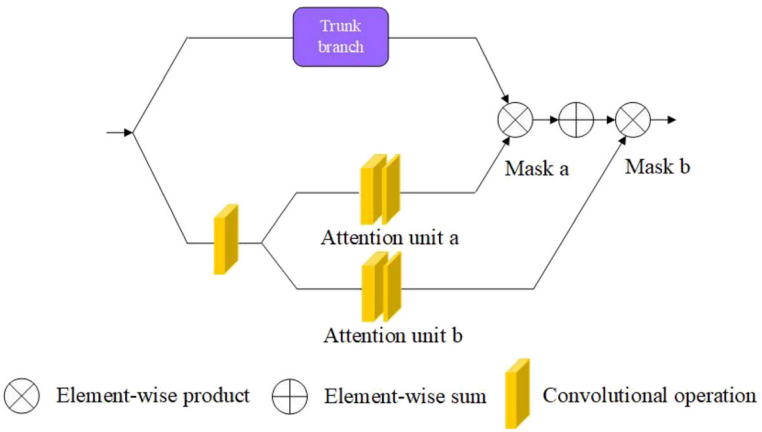

Fig. 3 The structure of double attention module

2.4 Training method

2.4.1 Progressive training

In many tasks of crowd counting, the ground truth is generated by geometry-adaptive kernels or fixed kernel. Usually, the method of geometry-adaptive kernels is used in dense crowd scenes, and the method of fixed kernel in sparse crowd scenes. In different scenes, the number of people and the distribution of crowd vary dramatically, so in this paper we use both geometry-adaptive and fixed kernels to generate ground truth. We firstly use geometry-adaptive kernels to generate ground truth to pre-train 2-DA-CNN, and then the fixed kernel is used in formal training and testing.

The geometry-adaptive kernels is defined by Eq. (4), and the fixed kernel by Eq. (5).

$$F_g(x) = \sum_{i=1}^{N} \delta(x-x_i) * N_g\left(p; P, {\sigma_i}^2\right), \sigma_i = \beta d_i, \beta = 0.3 \tag{4}$$

$$F_f(x) = \sum_{i=1}^{N} \delta(x-x_i) * N_f\left(p; P, \sigma^2\right), \sigma = 15 \tag{5}$$

where $F_g(x)$ is the density map which is generated by geometry-adaptive kernels and $F_f(x)$ by fixed kernel. x is the position of pixel in the image. $\delta(x - x_i)$ represents a head at pixel x_i. Both of $N_g(p; P, {\sigma_i}^2)$ and $N_f(p; P, \sigma^2)$ denote a normalized 2D Gaussian kernel evaluated at p with the mean at the user-placed dot P. σ_i and σ represent the standard deviation for corresponding Gaussian kernel, respectively. d_i is the average distance of 3 nearest neighbors. β is a hyper-parameter which is set to 0.3 in this paper.

2.4.2 Loss function

Like most of density map estimation-based crowd counting methods, we also use the Euclidean distance as the loss function by (6):

$$L = \frac{1}{N} \sum_{i=1}^{N} \| F(x_i, \theta) - G_i \|_2^2 \tag{6}$$

where N is the total number of samples, the $F(x_i, \theta)$ is the estimated density map for image x_i, G_i indicates the ground truth of x_i, and θ is the parameter to be learned.

3 Experimental results and analysis

3.1 Evaluation metrics

In our task, Mean Absolute Error (MAE) and Mean Squared Error (MSE) are used to evaluate 2-DA-CNN. MAE reflects the accuracy of the testing model, and MSE indicates the robustness of the testing model. They are defined by (7) and (8).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left\|c_i - c_i^{'}\right\| \tag{7}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\|c_i - c_i^{'}\right\|^2} \tag{8}$$

where N is the total number of test images, c_i stands for the estimated count for i*th* test image and $c_i^{'}$ stands for the actual count for i*th* test image.

3.2 Datasets

ShanghaiTech part B This dataset is the part of ShanghaiTech dataset, which is created by Zhang *et.al* [33], which is split by the publisher into train and test subsets consisting of 400 and 316 images respectively. All images are taken from the busy streets of Shanghai including a total of 88,488 annotated heads, whose resolution is 768 × 1024. Due to the average number of people per image is 124, the crowd density of this dataset is relatively low.

ShanghaiTech part A This dataset is also the part of ShanghaiTech dataset, which consists of 400 train images and 182 test images with total 241,677 annotated heads. Those images are crawled from Internet, whose resolution is diverse. The average number of people in each image in this dataset is 501, which have higher crowd density than those of ShanghaiTech part B.

UCF_CC_50 This dataset is a very challenging dataset, which is released by Idrees *et.al* [13]. It only contains 50 images taken from the network with a wide range of crowd densities and various resolutions. In this dataset, a total of 63,075 individuals were labeled and the average number of people in each image is 1280.

In this paper, we use 5-fold cross-validation method to train and test 2-DA-CNN, and all training process contains two steps. In the first step, only four copies of each image are used to augment the dataset (called AUG1), which is used in progressive training to pre-train proposed method. As for the second step, the augmentation method is the same as [15] (called AUG2), which is used to formally train proposed method. AUG2 is described as follows: firstly 9 patches are cropped with 1/4 size of the original image at different locations from each image, and then the patches are mirrored to further augment the dataset. It's worth noting that our all data augmentation is only used on train set. The training procedure of our 2-DA-CNN is shown as Fig. 4.

3.3 Comparison with other methods on different datasets

3.3.1 Results on ShanghaiTech part B dataset

The comparative results with the state-of-art methods are shown in Table 1. In Table 1, MCNN [33] and CP-CNN [24] use multi-column structure to get multi-scale information, and ic-CNN [26] employs the multi-resolution feature map to get multi-scale information, and MRA-CNN

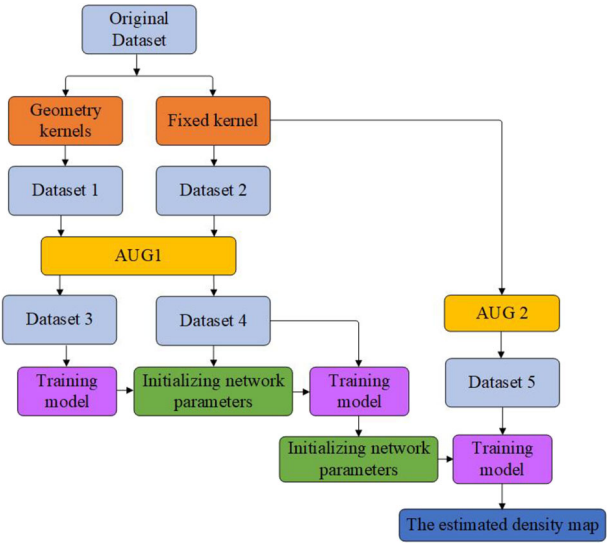

Fig. 4 Training procedure of 2-DA-CNN

[32], SCAR [11] and FDCNet [30] utilize attention mechanism to deal with crowd counting, and Switching-CNN [21] and CSRNet [15] use a single-column regressor to achieve crowd counting. From Table 1, it can be seen our 2-DA-CNN gets the lowest MAE and MSE, and outperforms the state-of-art methods. We also give the results on four images shown in Fig. 5, which intuitively demonstrate the performance of 2-DA-CNN in relatively sparse crowd scenes.

3.3.2 Results on ShanghaiTech part A dataset

In order to better validate 2-DA-CNN in dense crowd scenes, we also test 2-DA-CNN in ShanghaiTech part A, and the results are shown in Table 2. From Table 2, it can be seen that our proposed model achieves the best MAE and the second-grade MSE (very close to the best MSE). We also report 4 test images in Fig. 6, which further suggest the effectiveness of proposed 2-DA-CNN in relatively dense crowd scenes.

Table 1 Performance comparison on the ShanghaiTech part B dataset

Model	MAE	MSE
MCNN [33]	26.4	41.3
Switching-CNN [21]	21.6	33.4
CP-CNN [24]	20.1	30.1
ic-CNN [26]	10.7	16.0
CSRNet [15]	10.6	16.0
MRA-CNN [32]	11.9	21.3
FDCNet [30]	10.3	15.8
SCAR [11]	9.5	15.2
2-DA-CNN	**8.9**	**13.9**

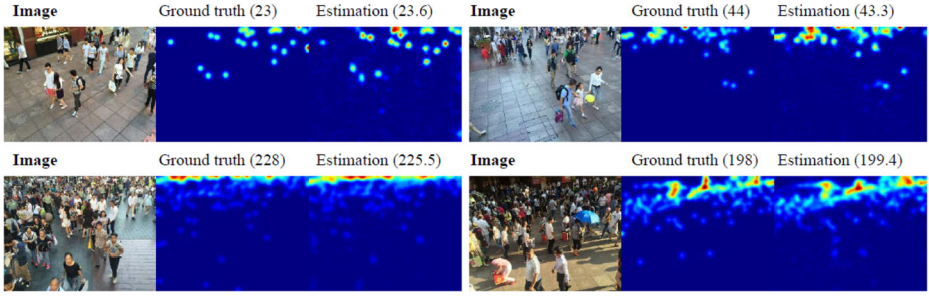

Fig. 5 Four examples from ShanghaiTech part B

3.3.3 Results on UCF_CC_50 dataset

The comparative results on UCF_CC_50 dataset are given on Table 3. Observing the results in Table 3, we find that the proposed 2-DA-CNN gets the third grade on both MAE and MSE. Four test samples are shown in Fig. 7, which also suggests the competitiveness of 2-DA-CNN with the state-of-art methods in fully challenged crowd scenes.

3.4 Ablation study

3.4.1 Analysis of multi-scale CNN network

To enhance the ability of multi-column CNN module, we combine the front-end module with multi-column CNN module to construct the multi-scale CNN network, which can filter the cluttered background information more effectively in crowd images. We give the comparative results with MCNN [33], CSRNet [15] shown in Table 4. It can be seen that the proposed multi-scale CNN network shows the best on MAE and the second grade on MSE. Moreover, it is worth noting that multi-scale CNN network needs less parameters than CSRNet.

3.4.2 Analysis of double attention

Here we mainly compare residual attention mechanism with double attention mechanism in our work. The back-end structure of different attention mechanism in one stage is shown in Fig. 8, Fig. 8(a) and Fig. 8(c) are the structures of using residual attention mechanism, and Fig. 8(b) and Fig. 8(d) are the structures of using double attention mechanism. Two kinds of

Table 2 Performance comparison on ShanghaiTech part A dataset

Model	MAE	MSE
MCNN [33]	110.2	173.2
Switching-CNN [21]	90.4	135.0
CP-CNN [24]	73.6	**106.4**
ic-CNN [26]	68.5	116.2
CSRNet [15]	68.2	115.0
MRA-CNN [32]	74.2	112.5
FDCNet [30]	75.1	118.5
SCAR [11]	66.3	114.1
2-DA-CNN	**64.6**	**106.6**

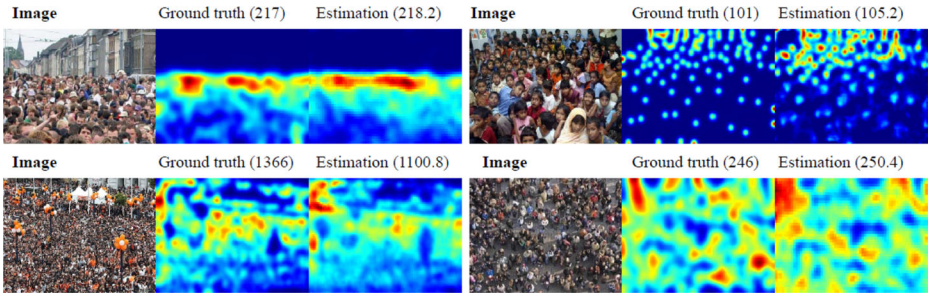

Fig. 6 Four examples on ShanghaiTech part A

attention units are used to construct different attention module, one is residual attention unit, and the other is segmented attention unit, which is consists of two convolutional layers, and its framework is "C(128,64,3)-ReLu-C(64,16,3)-Sigmoid".

In Fig. 8, the residual attention model and the residual double attention model are based on residual attention unit, and the segmented residual attention model and the segmented double attention model are based on segmented attention unit. The last convolution operation of different models makes use of 1×1 convolutional filter to generate the estimated density map.

The results in one stage are shown in Table 5. From Table 5, it can be seen that using segmented residual attention can get better results than using residual attention, and using the double attention is better than using residual attention in one stage, so we use segmented double attention on the proposed 2 -DA-CNN.

We further compare the performance of using different attention module in one stage and two stages, and the results are shown in Table 6. It can be found that those models using two stages are always better than using one stage, which suggests that the sequential use of attention module can get better accuracy and better robustness than single use, and also further indicates that double attention has better performance than residual attention.

3.4.3 Analysis of progressing training

For many crowd scenes, we observe their scale variation is large, so we design the progressing training to train 2-DA-CNN, i.e. using geometry-adaptive kernels to generate ground truth for dense crowd regions, and using fixed kernel to generate ground truth for sparse crowd regions. The comparative results of using different kernel are given in Table 7.

Table 3 Performance comparison on the UCF_CC_50 dataset

Model	MAE	MSE
MCNN [33]	377.6	509.1
Switching-CNN [21]	318.1	439.2
CP-CNN [24]	298.8	**320.9**
ic-CNN [26]	260.9	365.5
CSRNet [15]	266.1	397.5
MRA-CNN [32]	**240.8**	352.6
FDCNet [30]	246.8	322.2
SCAR [11]	259.0	374.0
2-DA-CNN	**252.0**	**340.3**

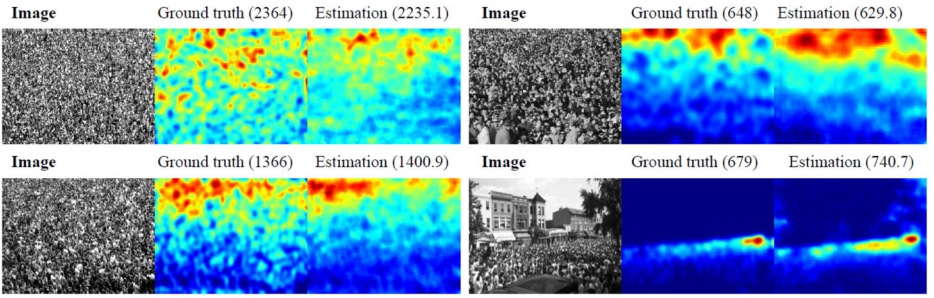

Fig. 7 Four examples on UCF_CC_50

Table 4 Performance of Multi-scale CNN network

ShanghaiTech part B	MAE	MSE	Number of parameters
MCNN	26.4	41.3	$\mathbf{1.3 \times 10^5}$
CSRNet	10.6	**16.0**	1.6×10^7
Our multi-scale CNN network	**10.54**	**16.18**	$\mathbf{9.2 \times 10^6}$

From Table 7, it can be seen that progressing training performs the best with respect to MAE of 8.94 and MSE of 13.85. This reflects that progressing training is the best plan to train 2-DA-CNN.

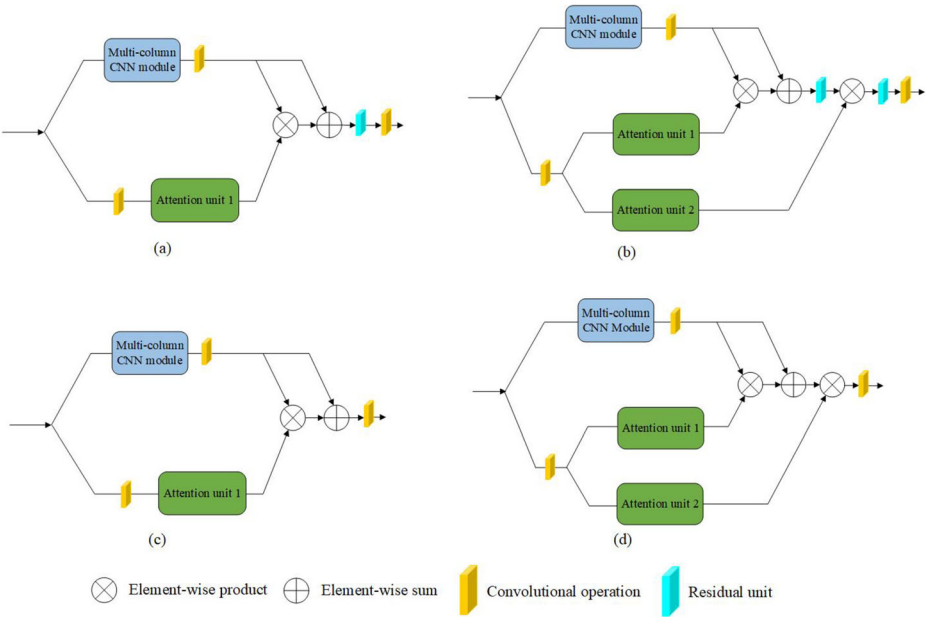

Fig. 8 one stage with different attention models. (**a**): the residual attention model; (**b**): the residual double attention model; (**c**): the segmented residual attention model; (**d**): the segmented double attention model.

Table 5 Comparative results of different attention module in one stage

Methods	MAE	MSE
One stage with the residual attention model	13.14	18.33
One stage with the residual double attention model	11.71	17.69
One stage with the segmented residual attention model	11.41	17.74
One stage with the segmented double attention model(1-DA-CNN)	**11.05**	**16.91**

Table 6 Performance comparison of different attention module

Methods	MAE	MSE
One stage with the segmented residual attention model	11.41	17.74
One stage with the segmented double attention model	11.05	16.91
Two stages with the segmented residual attention model	10.68	16.15
Two stages with the segmented double attention model(2-DA-CNN)	**10.26**	**15.85**

3.5 Comparative results on other measures

In order to assess 2-DA-CNN fully, we further analyze 2-DA-CNN in terms of parameters, model size and runtime. Considering the various resolutions of images and the dense crowd in real life, we use ShanghaiTech part A to demonstrate the performance listed as Table 8. Obviously, there is a tradeoff between parameters and estimate precision. In the future, we will develop lightweight models based on 2-DA-CNN.

4 Conclusion

In this paper, we propose a novel 2-stage double attention convolutional neural network (2-DA-CNN) to deal with scale variation and cluttered background in crowd scenes for crowd counting. 2-DA-CNN uses the double attention mechanism to learn the regions of interest in crowd scenes, which could distinguish the effective feature in cluttered scenes. Moreover, we design progressive training to improve the drawback of using geometry-adaptive kernels to generate ground truth, which enables model to deal with non-uniform crowd distribution in crowd images. The comparative results with the state-of-art methods suggest that the proposed method achieves the superior performance on three mainstream crowd counting datasets. In future work, we will explore lightweight models based on 2-DA-CNN to achieve crowd counting.

Table 7 Performance of using different training method

Methods	MAE	MSE
Fixed kernel: $\sigma = 15$	10.26	15.85
Geometry-adaptive kernels	9.95	16.15
Progressive training	**8.94**	**13.85**

Table 8 Comparative results on other measures in the ShanghaiTech part A

Model	Parameters(M)	Runtime(ms)	Model Size
MCNN [33]	**0.13**	**25**	**527.6 KB**
Switching-CNN [21]	15.1	153	57.6 MB
CP-CNN [24]	68.4	5113	>500 MB
CSRNet [15]	16.3	64	65.1 MB
2-DA-CNN	**9.8**	**224**	**74.9 MB**

Acknowledgmet This work was supported by the National Natural Science Foundation of China (No. 61771223).

Availability of data and material Not applicable.

Code availability Not applicable.

Funding information This study was funded by National Natural Science Foundation of China (No. 61771223).

Compliance with ethical standards

Conflicts of interest/competing interests The authors declare that they have no conflict of interest.

References

1. Alghamdi A, Hammad M, Ugail H, Abdel-Raheem A, Muhammad K, Khalifa HS, Abd El-Latif AA (2020). Detection of myocardial infarction based on novel deep transfer learning methods for urban healthcare in smart cities. Multimedia tools and applications:22
2. Cao X, Wang Z, Zhao Y, Su F (2018). Scale Aggregation Network for Accurate and Efficient Crowd Counting. Paper presented at the 2018 European conference on computer vision (ECCV), Munich
3. Chaker R, Aghbari ZA, Junejo NI (2017) Social network model for crowd anomaly detection and localization. Pattern Recogn 61:266–281
4. Chan AB, Liang Z-SJ, Vasconcelos N (2008). Privacy preserving crowd monitoring: Counting people without people models or tracking. Paper presented at the 2008 IEEE conference on computer vision and pattern recognition(CVPR), Anchorage, AK
5. Chan AB, Vasconcelos N (2009). Bayesian Poisson regression for crowd counting. Paper presented at the 2009 IEEE 12th international conference on computer vision (ICCV), Kyoto
6. Chen J, Su W, Wang Z (2020) Crowd counting with crowd attention convolutional neural network. Neurocomputing 382:210–220
7. Dalal N, Triggs B (2005). Histograms of oriented gradients for human detection. Paper presented at the IEEE computer society conference on computer vision and pattern recognition (CVPR), San Diego, CA
8. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell 32(9):1627–1645
9. Fu M, Xu P, Li X, Liu Q, Ye M, Zhu C (2015) Fast crowd density estimation with convolutional neural networks. Eng Appl Artif Intell 43:81–88
10. Gao J, Wang Q, Li X (2019). PCC net: perspective crowd counting via spatial convolutional network. IEEE Transactions on Circuits and Systems for Video Technology:1–1
11. Gao J, Wang Q, Yuan Y (2019) SCAR: spatial−/channel-wise attention regression networks for crowd counting. Neurocomputing 363:1–8
12. Hossain M, Hosseinzadeh M, Chanda O, Wang Y (2019). Crowd Counting Using Scale-Aware Attention Networks. Paper presented at the 2019 IEEE winter conference on applications of computer vision (WACV), Hawaii

13. Idrees H, Saleemi I, Seibert C, Shah M (2013). Multi-Source Multi-Scale Counting in Extremely Dense Crowd Images. Paper presented at the 2013 IEEE conference on computer vision and pattern recognition (CVPR), Portland, OR
14. Li W, Mahadevan V, Vasconcelos N (2014) Anomaly detection and localization in crowded scenes. IEEE Trans Pattern Anal Mach Intell 36(1):18–32
15. Li Y, Zhang X, Chen D (2018). CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. Paper presented at the 2018 IEEE conference on computer vision and pattern recognition(CVPR), Salt Lake City, UT
16. Liu J, Gao C, Meng D, Hauptmann AG (2018). DecideNet: Counting Varying Density Crowds Through Attention Guided Detection and Density Estimation. Paper presented at the 2018 IEEE conference on computer vision and pattern recognition(CVPR), Salt Lake City, UT
17. Liu W, Salzmann M, Fua P (2019). Context-aware Crowd Counting. Paper presented at the 2019 IEEE conference on computer vision and pattern recognition(CVPR), Long Beach, CA
18. Nestor T, De Dieu NJ, Jacques K, Yves EJ, Iliyasu AM, Abd El-Latif AA (2020) A multidimensional Hyperjerk oscillator: dynamics analysis, analogue and embedded systems implementation, and its application as a cryptosystem. Sensors 20(1):23
19. Paragios N, Ramesh V (2001). A MRF-based Approach for Real-Time Subway Monitoring. Paper presented at the 2001 IEEE conference on computer vision and pattern recognition(CVPR), Kauai, HI.
20. Sabzmeydani P, Mori G (2007). Detecting Pedestrians by Learning Shapelet Features. Paper presented at the 2007 IEEE conference on computer vision and pattern recognition (CVPR), Minneapolis, MN.
21. Sam DB, Surya S, Babu RV (2017). Switching Convolutional Neural Network for Crowd Counting. Paper presented at the 2017 IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI
22. Shao J, Kang K, Loy CC, Wang X (2015). Deeply learned attributes for crowded scene understanding. Paper presented at the 2015 IEEE conference on computer vision and pattern recognition (CVPR), Boston, MA.
23. Simonyan K, Zisserman A (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556
24. Sindagi VA, Patel VM (2017). Generating high-quality crowd density maps using contextual pyramid CNNs. Paper presented at the 2017 IEEE international conference on computer vision (ICCV), Venice
25. Sindagi VA, Patel VM (2019) HA-CCN: hierarchical attention-based crowd counting network. IEEE Trans Image Process 29:323–335
26. Viresh R, Hieu L, Minh H (2018). Iterative crowd counting. Paper presented at the European conference on computer vision (ECCV), Munich
27. Wang Y, Hu S, Wang G, Chen C, Pan Z (2019) Multi-scale dilated convolution of convolutional neural network for crowd counting. Multimed Tools Appl 79(1–2):1057–1073
28. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X (2017). Residual Attention Network for Image Classification. Paper presented at the 2017 IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI
29. Wang C, Zhang H, Yang L, Liu S, Cao X (2015). Deep people counting in extremely dense crowds. Paper presented at the the 23rd ACM international conference(ACM), New York
30. Zhang YQ, Li GH, Lei J, He JY (2019) FDCNet: frontend-backend fusion dilated network through channel-attention mechanism. Appl Sci-Basel 9(17):16
31. Zhang C, Li H, Wang X, Yang X (2015). Cross-scene crowd counting via deep convolutional neural networks. Paper presented at the 2015 IEEE conference on computer vision and pattern recognition (CVPR), Boston, MA
32. Zhang Y, Zhou C, Chang F, Kot AC (2019) Multi-resolution attention convolutional neural network for crowd counting. Neurocomputing 329:144–152
33. Zhang Y, Zhou D, Chen S, Gao S, Ma Y (2016). Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. Paper presented at the 2016 IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV
34. Zhou B, Wang X, Tang X (2012). Understanding collective crowd behaviors: learning a mixture model of dynamic pedestrian-agents. Paper presented at the IEEE conference on computer vision and pattern recognition(CVPR), Providence, RI
35. Zou Z, Cheng Y, Qu X, Ji S, Guo X, Zhou P (2019) Attend to count: crowd counting with adaptive capacity multi-scale CNNs. Neurocomputing 367:75–83
36. Zou Z, Su X, Qu X, Zhou P (2018) DA-net: learning the fine-grained density distribution with deformation aggregation network. IEEE Access 6:60745–60756