Neurocomputing 378 (2020) 455–466

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

DCGSA: A global self-attention network with dilated convolution for crowd density map generating

Liping Zhu[a,b], Chengyang Li[a,b,*], Bing Wang[a,b], Kun Yuan[c], Zhongguo Yang[d,e]

[a] College of Information Science and Engineering, China University of Petroleum (Beijing), Beijing, China
[b] Key Lab of Petroleum Data Mining, China University of Petroleum (Beijing), Beijing, China
[c] School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada
[d] Beijing Key Laboratory on Integration and Analysis of Large-Scale Stream Data, North China University of Technology, Beijing, China
[e] School of Computer Science, North China University of Technology, Beijing, China

ARTICLE INFO

Article history:
Received 2 April 2019
Revised 2 September 2019
Accepted 20 October 2019
Available online 31 October 2019

Communicated by Wang QI

Keywords:
Crowd density map
Convolutional neural networks
Self-attention
Global context
Dilated convolution

ABSTRACT

Due to non-uniform density and variations in scale and perspective, estimating crowd count in crowded scenes in different degree is an extremely challenging task. The deep learning models mostly use pooling operation so that the density map of original resolution is obtained through the last upsampling. This paper aims to solve the problem of losing local spatial information by pooling in density map estimation. Therefore, we propose a dilated convolution neural network with global self-attention, named DCGSA. Especially, we introduce a Global Self-Attention module (GSA) to provide global context as guidance of low-level features to select person location details and a Pyramid Dilated Convolution module (PDC) that extracts channel-wise and pixel-wise features more precisely. Extensive experiments on several crowd datasets show that our method achieves lower crowd counting error and better density maps compared to the recent state-of-the-art methods. In particular, our method also performs well on the sparse dataset UCSD.

1. Introduction

Massive stampede events around the world have caused great injuries and deaths. Accurate population density estimation can bring convenience to timely crowded warnings. Due to the rapid development of deep learning, crowd counting and crowd density estimation [1–5] have obtained great improvement in recent years. Crowd counting outputs a total person number of persons in a crowd image while crowd density estimation presents a heatmap of crowd distribution. In this paper, we aim to solve the joint task of estimating both people number and density map from a single image with the arbitrary perspective angle.

This task faces two main challenges now: (1) Since the predicted density map needs to follow pixel-by-pixel prediction, the output density map must approximate the spatial structure of the original image so that they can render the smooth transition between each pixel and its nearest neighbors. However, some spatial structure information of the original image is missing during the pooling process, which has an adverse impact on the smooth transition. (2) Pixels of each person's head in one image range widely due to the distance to the monitoring camera. A phenomenon of "Small person in the distance, large person in the vicinity" is formed as shown in Fig. 1. Due to the advantages of low MAE (mean average error) and MSE (mean squared error), deep convolutional neural networks have become the mainstream method instead of traditional methods. Most methods extract features through several convolution layers. Pooling layers are both used and finally the feature maps are restored to the original image size by bilinear interpolation. They do not adopt the idea of multi-scale like SSD [6]. For example, CSRNet [7] deploys the first 10 layers from VGG-16 [8] as the front-end and dilated convolution layers as the back-end. Only deep features are used for final prediction, yet shallow features are discarded. The loss of spatial information caused by pooling layers is not compensated. Thus, the generated density map may lose some vital information about person location. Besides, the enlargement by bilinear interpolation in the final stage is not conducive to estimating such a pixel-level task in crowd density map prediction.

In this paper, we propose a new method called DCGSA (Dilated Convolution with Global Self-Attention). It aims at solving

* Corresponding author at: College of Information Science and Engineering, China University of Petroleum (Beijing), Beijing, China.
E-mail addresses: zhuliping@cup.edu.cn (L. Zhu), 2017215536@student.cup.edu.cn (C. Li), 2018215554@student.cup.edu.cn (B. Wang), kyuan033@uottawa.ca (K. Yuan), yangzhongguo@ncut.edu.cn (Z. Yang).

https://doi.org/10.1016/j.neucom.2019.10.081
0925-2312/

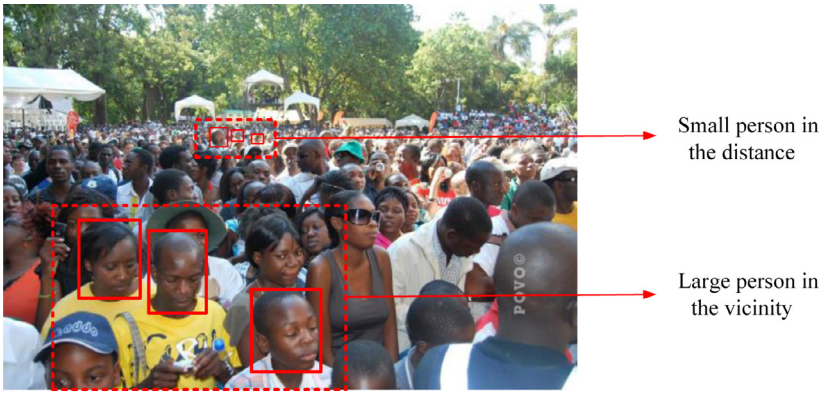


Fig. 1. Different sizes of person's head at the different distances from a camera.

the issue that high-level features are skilled in making density prediction, while weak in restructuring original resolution binary prediction. U-shape networks [9–12] which use the low-level information to help high-level features recover spatial details. Referring to U-shape networks and SENet [13], we design a decoder module named Global Self-Attention Module (GSA) which can extract the global context of high-level features as guidance to weight low-level features. Besides, the existence of persons at multiple scales brings difficulty in density prediction. Thus, inspired by Dilated Convolution [14] and SPPNet [15], we design an effective feature exactor module called PDC (Pyramid Dilated Convolution). It is used to extract both pixel-wise and channel-wise context for high-level features extracted from deeper layer of convolutional neural network.

In summary, there are two contributions in this paper. Firstly, we design Global Self-Attention Module to make up for lost information during the pooling process. Secondly, we propose Pyramid Dilated Convolution Module to embed extracted person features at different scales. By combining Global Self-Attention Module and Pyramid Dilated Convolution Module, the total structure DCGSA has achieved better performance.

The rest of the paper is structured as follows. Section 2 presents previous works of crowd density estimation, CNN attention mechanism, and dilated convolution. Section 3 introduces the details of the proposed method while Section 4 presents the experimental results on different datasets. In Section 5, we make a conclusion of the paper.

2. Related work

2.1. Crowd density estimation

Crowd Density estimation aims to map an input crowd image to its corresponding density map. The density map indicates the number of people per pixel presented in the crowd image. Over recent years, researchers have tended to use density regression-based methods for crowd counting. Especially, the features extracted by CNN are more robust than previous hand-crafted features.

Multi-column CNN fuses features through several CNN columns to regress the crowd density map. Zhang et al. [2] proposed a multi-column based architecture (MCNN). The network includes three columns corresponding to filters with receptive fields of different sizes (large, medium, small). These different columns are designed to cater to different person scales present in the images. Boominathan et al. [3] combined deep and shallow fully convolutional networks to predict the density map. The combination of two networks aims at solving non-uniform scaling of crowd and variations in perspective. Unlike the methods above, CSRNet [7] use VGG-16 as a backbone for feature exacting. It uses dilated convolution at the end of the network for understanding highly congested scenes. To embed local structural information, Wang et al. [16] proposed a deep network with metric learning. The learning of better representations and distance measurement are simultaneous. It proves that the metric learning can guide the training process of deep networks with high-level semantic features. Another research [17] by Wang proposes a Multiview-based Parameter Free framework (MPF) for group detection. A novel Structural Context descriptor is put forward to profile the structural properties of feature points. Two versions of the Self-weighted Multiview Clustering method are designed to integrate the points' correlations from both the orientation and context views. They also propose a tightness-based merging strategy for combining the coherent local groups reasonably.

2.2. CNN attention mechanism

A neural network with attention mechanisms can focus more on relevant elements of the input than on irrelevant parts. It is first studied in Natural Language Processing (NLP). Encoder–decoder models with attention modules are designed to facilitate neural machine translation [18–20]. In computing the output for a given query element, certain key elements are prioritized according to the query. Self-attention modules were then presented for modeling intra-sentence relations [21–24]. Especially, the Transformer attention module [24] has achieved state-of-the-art performance. The success of attention mechanisms in NLP has motivated itself to computer vision. Thus, different kinds of attention module are applied to both object detection and semantic segmentation [25–28]. Here, the query and key are visual elements such as image pixels or regions of interest in computer vision.

Channel-wise feature attention [13,25,29,30] is the representative of spatial self-attention. As different feature channels encode different semantic concepts, these works aim at capturing the correlations among these concepts. This can be achieved by activation/deactivation of certain channels. Meanwhile, relationships among elements at different spatial positions are modeled. Different attention weights are assigned to corresponding feature channels, as shown in Formula 1.

$$F_{out} = F^{C}_{(x,y)} \times W^{C} \tag{1}$$

Here, $F^{C}_{(x,y)}$ represents the pixel value of position (x, y) on the channel C of the input feature maps. W^{C} represents the attention weight corresponding to the channel C. The attention weight is generated by the network itself.

The encoder-decoder model is to encode the input sequence into an intermediate context. This context is a specific length of

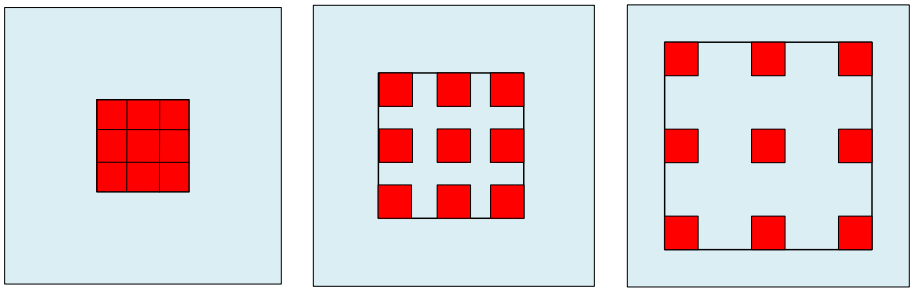


Fig. 2. Different dilation rate on feature maps.

encoding (which can be considered as a vector) and then restored to an output target sequence through this context. Many works [31,32] suggest that attention mechanism in encoder–decoder model plays a role similar to word alignment in traditional approaches [33–35]. The input elements accorded high attention weights are responsible for the model outputs. In encoder–decoder attention, the key and the query are from two different sets of elements. The two sets of elements need to be properly aligned mostly. For example, in the encoder–decoder attention of neural machine translation, the key and the query elements correspond to the words in the input and the output sentences, respectively. Similarly, in semantic segmentation, the key and the query elements can correspond to the concepts in the low-level and high-level features. The low-level features can be prioritized according to the high-level features.

2.3. Dilated convolution

Dilated convolution [14] comes from the field of image semantic segmentation. When an image enters a neural network, convolutional filters are used to extract features and pooling is used to reduce the image size and increase the receptive field. Since image segmentation is a pixel-wise prediction, it is necessary to restore the smaller image to its original size by upsampling. Although the input is finally resized by the upsampling operation, many details (pixel missing) are lost by pooling forever. Thus, dilated convolution turned out which increases the receptive field without reducing the size of feature maps.

There is an important parameter in dilated convolution called dilation rate. It represents the size of dilation as shown in Fig. 2. From the perspective of feature maps, dilation is just sampling on the feature maps. The sampling frequency is set according to the parameter. When rate is 1, the feature maps will not lose any information after sampling. At this time, dilated convolution is considered as a standard convolution. When rate $>$ 1, it is sampled every rate-1 pixels. The feature maps after sampling are finally convolved with the kernel, which actually increases the receptive field in disguise. On the other hand, dilation can enlarge the kernel size. (Rate-1) zeros are inserted between adjacent points. From Formula 2 and Formula 3, changes of the receptive field can be observed. Formula 4 and Formula 5 show the size changes of feature maps after ordinary convolution and dilated convolution.

$$Field = k * k \tag{2}$$

$$Field_d = k + (k-1)*(rate-1) \tag{3}$$

$$W = \frac{W_{in} - k + 2p}{s} + 1 \tag{4}$$

$$W_d = \frac{W_{in} - k - (k-1)*(rate-1) + 2p}{s} + 1 \tag{5}$$

Here, k is the kernel size, *rate* is the dilation rate, p is the padding size and s is the stride size.

Yet, there are two potential problems with a structure based entirely on Dilated Convolution: (a) The Gridding Effect. If the 3×3 kernel of 2 dilation rate is superimposed multiple times, not all pixels will be used for calculation. This may lose the continuity of the image information, leading to worse pixel-wise dense prediction. (b) Long-ranged information might be not relevant. Large dilation rate may only be effective for segmentation of large objects, while it may be disadvantageous for small objects. Therefore, we design a new module called Pyramid Dilated Convolution module. It referred to SPPNet [15], but replaced pooling with dilated convolution. Experiments prove that it can use all information of feature maps and have a better prediction.

3. Proposed method

In this section, we first introduce the proposed Global Self-Attention (GSA) Module and Pyramid Dilated Convolution (PDC) Module. Then we describe the complete encoder–decoder network architecture DCGSA, designed for the joint task of predicting crowd density map and crowd counting.

3.1. Global self-attention

Crowd density prediction is to generate corresponding density values for each pixel. To some extent, it is similar to the idea of semantic segmentation. Therefore, decoder architectures which perform well in the semantic segmentation task can be migrated to the crowd density prediction task. For example, PSPNet [36] or Deeplab [37] uses bilinearly upsample directly while DUC [38] uses large channel convolution combined with reshaping. Both of them lack different scales of low-level feature map information. This may be harmful to recover spatial localization to origin resolution. Deep Network in [7] has already obtained considerable performance and capability to obtain person information. However, they all ignore to repair person pixel location. Therefore, we consider to fully use high-level features with abundant person information for weighting low-level context to select precise resolution details.

SENet [13] assigns the vector obtained by global average pooling as the weight to each pixel on the feature map of each channel. It adaptively recalibrates channel-wise feature responses by explicitly modeling inter-dependency between channels. However, global context just has high semantic information, which is not helpful for recovering the spatial information. It can be observed that the network encodes finer spatial information in the lower stage, but it has poor semantic consistency due to small receptive view. While in the high stage, it has strong semantic consistency due to large receptive view, but the prediction is spatially coarse. Overall, the lower stage makes more accurate spatial predictions, while the higher stage gives more accurate semantic

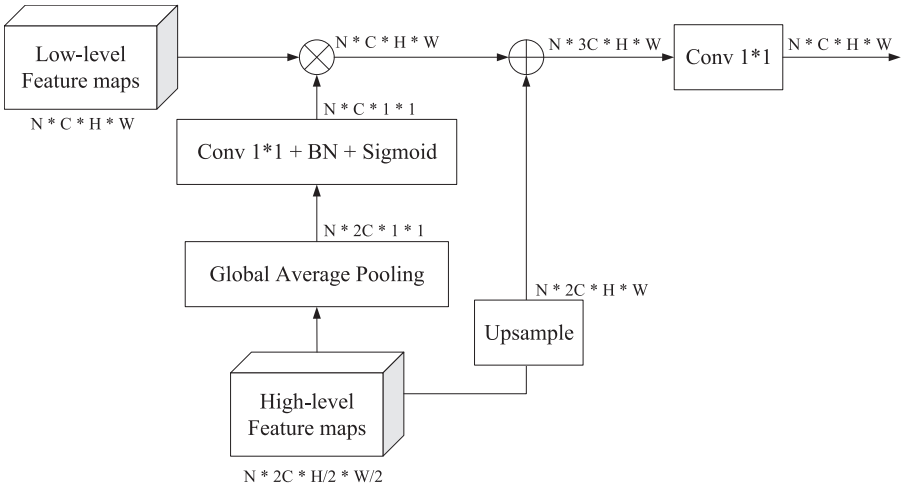


Fig. 3. The structure of Global Self-Attention Module (GSA).

predictions. Inspired by this, we assume whether the channel-wise context generated by global average pooling of deep layers can guide shallow layers to learning person localization details. Thus, we design Global Self-Attention Module in order to provide global context to low-level features. The destination is to use high-level features to guide the low-level features for the optimal prediction.

Here, we use *HF* for high-level features in higher stage and *LF* for low-level features in the lower stage. As default, the number of *HF*'s channels is twice that of *LF* while the feature map's width and height of *LF* are twice of that of *HF*. In detail, the global feature vector *Vector* is obtained from *HF* by Global Average Pooling (GAP). Each number in the *Vector* directly gives the abstract representation of each corresponding channel's feature map. GAP not only removes the limitation on the input size, but also retains the spatial information extracted from each channel of the previous convolution layer. *Vector* enters a 1×1 convolution with BatchNormalization (BN) [39] and Sigmoid activation function. The channels decrease to the channels of low-level feature maps while going through 1×1 convolution. In BN, the mean and standard-deviation are calculated per-dimension over the mini-batches. Parameters γ and β are learnable vectors of the input feature map's size as shown in Formula 6. BN forces such a data distribution to be scaled to a standard normal distribution with a mean of 0 variance of 1. Thus, it can lead the input value of the nonlinear transformation function to fall into the area sensitive to the input, thereby avoiding the problem of gradient disappearance. Mathematically, the estimates of its computed mean and variance are kept with a momentum of 0.1. The update rule for running statistics is shown as Formula 7:

$$y = \frac{x - E(x)}{\sqrt{Var(x) + \epsilon}} * \gamma + \beta \tag{6}$$

$$\hat{x}_{new} = (1 - momentum) \times \hat{x} + momentum \times x_t \tag{7}$$

Here, $\hat{x}$ is the estimated statistic and x_t is the new observed value. The sigmoid function scales the values in *Vector* to [−1,1]. After the data is mapped by the Sigmoid function, it will gradually move closer to the limit saturation region of the value range. Then *Vector* is multiplied by features map obtained by low-level features through the convolution. Useful features in low-level features are strengthened meanwhile useless features are weakened. Besides, upsampled high-level features are added with the weighted low-level features. This can aggregate multi-scale density prediction maps. 1×1 convolution can integrate the information of each feature channel and reduce computation. So, the fused feature maps are through 1×1 convolution to reduce channels finally. It is designed due to the success of U-Net. The whole data flow calculation is shown as follows:

$$P_{N \times C_{HF} \times 1 \times 1} = GAP(HF_{N \times C_{HF} \times H_{HF} \times W_{HF}}) \tag{8}$$

$$Vector_{N \times C_{LF} \times 1 \times 1} = ReLU(BN(Conv_{1*1}(P))) \tag{9}$$

$$Output = Conv_{1*1}(LF_{N \times C_{LF} \times H_{LF} \times W_{LF}} \times Vector + HF_{N \times C_{HF} \times H_{LF} \times W_{LF}}) \tag{10}$$

Here, N is the batch size and $N \times C \times H \times W$ is the matrix representation of data in a network. This module deploys different scale feature maps more effectively and uses high-level features provide guidance information to low-level feature maps in a simple way. The structure of Global Self-Attention Module (GSA) is shown in Fig. 3.

3.2. Pyramid dilated convolution

Inspired by Attention Mechanism and spatial pyramid structure, it is taken into account how to provide precise pixel-level information for high-level features extracted from deep convolutional layers. The pyramid module fuses feature maps under different pooling scales. Due to image scale-non-deformation and small computational burdens, pooling is widely used in various prediction tasks. However, high-level feature maps are small in size and each pixel value is critical to some extent. Spatial information loss caused by pooling may result in a less detailed density map. Also, it lacks global context prior attention to select the features channel-wisely as in SENet and EncNet [40].

Above all, we design Pyramid Dilated Convolution (PDC) module refer to PSPNet [15]. The dilated convolution architecture is based on the fact that dilated convolutions support the exponential expansion of the receptive field without loss of resolution or coverage. Thus, the pooling layers in original pyramid structures are removed to keep feature maps' resolution unchanged. 3×3 kernels of convolution with padding are used to maintain the feature map size. According to VGG-16, using more convolutional layers with small kernels is more efficient than using fewer layers with larger kernels when targeting the same size of receptive field. 3×3 convolution also brings less computation burden. To better extract context from different pyramid scales, convolutions of different dilation rate are adopted in pyramid structure respectively. Then, the

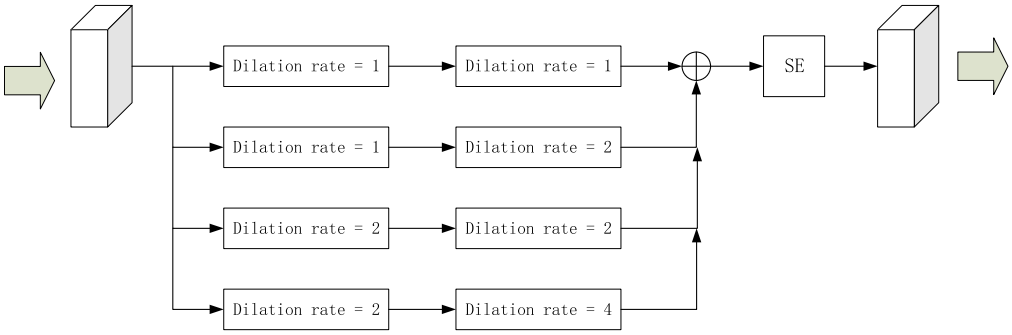


Fig. 4. The structure of Pyramid Dilated Convolution Module (PDC).

pyramid structure integrates information of different scales, which can incorporate neighbor scales of context features more precisely. Instead of pooling and upsampling, this dilated module can learn more spatial information without loss. Benefit from spatial pyramid structure, PDC module can fuse different scale context information and produce better pixel-level attention for high-level feature maps in the meantime. CSRNet proves that each column in such branch structure learns nearly identical features. However, multi-column features can be selected, rather than abandoned directly. Due to overlapping of the fused features, they are finally put into a SE-block Module to enhance spatial encoding. Useful density features are enhanced and invalid features are discarded. This will adaptively recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels. The structure of PDC module is shown in Fig. 4. Pyramid pooling module in PSPNet has four-level pooling kernels with sizes of 1×1, 2×2, 3×3 and 6×6 respectively. Similarly, we also use four-level features by dilated convolution, rather than pooling. Using this 4-level pyramid, the dilated convolution kernels cover the whole, 1/2, 1/4 and 1/8 of the image. The fused multi-scale features are weighted by SE block for more robust and accurate representation.

3.3. Auxiliary multi-scale loss

InceptionNet (GoogleNet) [41–43] shows that the network with the auxiliary branches starts to overtake the accuracy of the network without any auxiliary branch and reaches a slightly higher plateau near the end of training. The architecture of VGG-16 is treated as four stages divided by pooling layers. Therefore, feature maps at different stages can be regarded as feature maps generated by different convolution kernels (MCNN), thus avoiding feature redundancy. In a single forward pass of network training, multi-scale density predictions are obtained from different layers of the neural network. This corresponds to the problem that person's heads have different scales, depending on their distance to the camera. Therefore, four losses are adopted, leading to more detailed crowd density map. Loss1, Loss2, and Loss3 are L2 Loss while Loss4 is L1 Loss. L2 Loss (Formula 11) makes the feature maps in the network as close as possible to the density distribution. At the end of the network, L1 Loss (Formula 12) is used to estimate the number of density map in the final fitting real number, leading to more precise prediction result. Total loss (Formula 13) added from L1, L2, L3, and L4 are propagated. Then, the gradient is calculated and the network parameters are updated. Four multi-scale losses for total architecture's training are shown in Fig. 5.

$$L_{D2}(\Theta)=\frac{1}{2N}\sum_{i=1}^{N}\sum_{p=1}^{p}\|F(X_i(p);\Theta)-F_i(p)\|_2^2 \tag{11}$$

$$L_{D1}(\Theta)=\frac{1}{N}\sum_{i=1}^{N}|C(X_i;\Theta)-C_i| \tag{12}$$

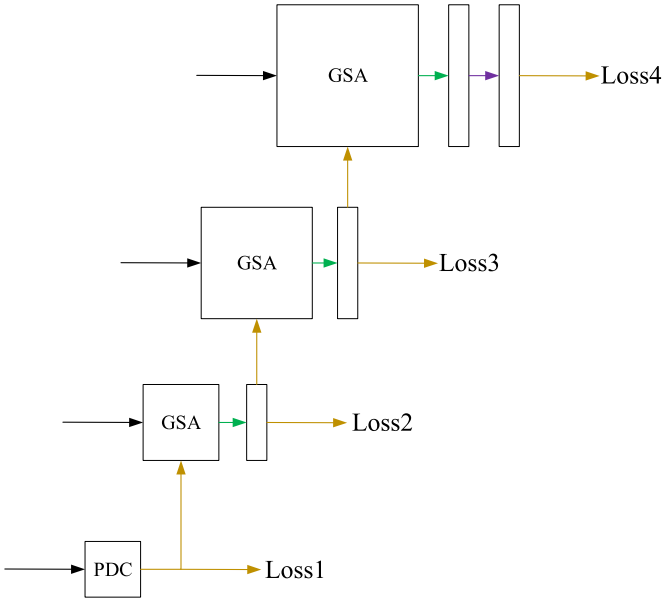


Fig. 5. Four multi-scale Loss for total architecture's training.

$$L_{total}=L_1+L_2+L_3+L_4 \tag{13}$$

3.4. Network architecture

With proposed Global Self-Attention (GSA) and Pyramid Dilated Convolution (PDC), we propose the total architecture Dilated Convolution with Global Self-Attention (DCGSA), as shown in Fig. 6. According to CSRNet [7], it shows that a best tradeoff can be achieved when keeping the first ten layers of VGG-16 with only three pooling layers instead of five. This is to suppress the detrimental effects on output accuracy caused by the pooling operation. Thus, the first 10 convolutional layers of VGG-16 are used as a backbone. The model can be divided into four stages according to the size of the feature maps. During the increase of stage, the channels of feature maps are doubled while the width and height is the half. In the final stage, the output size of feature maps is 1/8 of the original input image.

For the deepest features, the PDC module is added to gather pixel-level and channel-level information from the output of the backbone. Combined with the global context, the output is followed by three GSA module to generate multi-scale density prediction maps for predicting loss. The feature maps from the higher stage are upsampled by bilinear interpolation to be equal to the size of lower stage. Above all, the first VGG extracts fine high-level features, so it can be treated as an accurate encoder structure. The

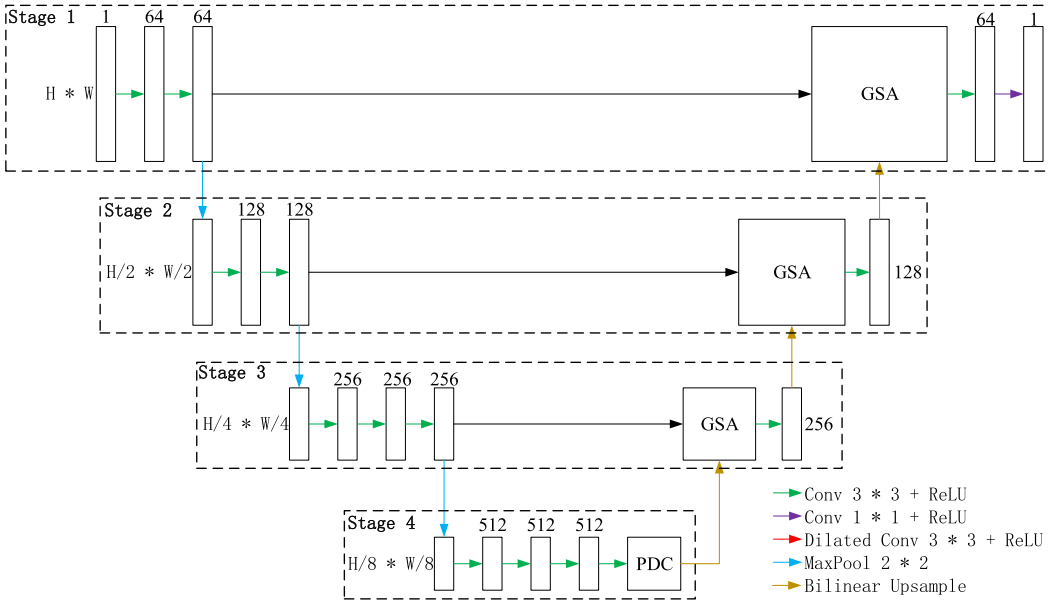


Fig. 6. The architecture of DCGSA for crowd density estimation.

GSA module decodes the high-level features to the final density map of the original size, so it can be treated as an effective decoder module. The PDC module between them can gain another round of performance boost. Finally, one 3×3 convolution kernel is used for one-channel prediction. On this basis, auxiliary multi-scale loss is also used for making more detailed prediction. Some paper like [7] only uses the final feature map which is bilinear upsample to the input size. The predicted density map is less accurate than one which is all predicted by CNN. Therefore, our method makes full use of the advantages of CNN to obtain more detailed prediction.

4. Experiment

In this section, we first introduce the ground truth generating of density maps. Then we introduce data preprocessing and evaluation metrics. Finally, we compare the proposed method to recent state-of-the-art methods on the four datasets. There are ShanghaiTech Part/A and Part/B dataset [2], UCF_CC_50 dataset [4] and UCSD dataset [44].

The batch size of the training process is set to 1. The model is trained using mini-batch adaptive moment estimation (Adam) with initial rate 1e-5. When training loss converges, the learning rate reduces to 1e-7. We used the PyTorch with one NVIDIA Tesla K80 to train and test the model.

4.1. Ground truth generating

A method for generating crowd density maps [45] is proposed to take into account perspective distortion. It generates the ground truth by estimating spread parameter of the Gaussian kernel. It is based on the size of each person's head in the image. However, it is impractical to estimate head sizes and their underlying relationship with density maps. The head size is related to distance between the centers of two neighboring persons in dense images. The spread parameter for each person is data-adaptively determined based on its average distance to its neighbors. Therefore, the average distance of each head to k nearest neighbors (other heads) can stand for geometric distortion. The method of generating density maps is followed in [2]. For each person's head x_i, the distances between it and other k nearest neighbors are calculated to be a set $\{d_1^i, d_2^i, \ldots, d_k^i\}$. According to the distance set, the mean distance of each head $\overline{d^i} = \frac{1}{k}\sum_{j=1}^{k} d_j^i$ is obtained. A delta function $\delta(x - x_i)$ is convolved with a Gaussian kernel with variance σ_i proportional to $\overline{d^i}$. An image with N heads labeled can be represented as $H(x) = \sum_{i=1}^{N} \delta(x - x_i)$. Density map $F(x)$ is produced by using Formula 14.

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_{\sigma_i}(x), \quad \sigma_i = \beta \overline{d^i} \tag{14}$$

Here, the empirical value β is set to be 0.3, which can produce the best result. The density map ground truth is generated based on the whole image. Some of the results are shown in Fig. 7.

4.2. Data preprocessing

A picture is processed to generate 15 new images. These new pictures are entered into the model along with the original picture.

(a) Reduce the image by 2 times and 4 times to generate 2 new pictures.
(b) Flip the image horizontally and vertically to generate 2 new pictures.
(c) Divide the image into 4 new pictures without overlapping.
(d) Randomly cut 7 new pictures on the image. The size must be bigger than 128×128.

Different scale spaces are constructed by using the specified scale factor to filter the image, thus changing the size or ambiguity of the image content. Therefore, (a) is used to make the model more adaptable to different sizes of person heads. (b)–(d) are to reduce the model overfitting on fixed spatial information.

4.3. Crowd density dataset

4.3.1. ShanghaiTech dataset

The ShanghaiTech dataset [2] includes 1198 images, whose ground truth are head-center annotations. This dataset is divided into two parts: Part/A and Part/B. The crowd distribution of images in Part/A is more congested than Part/B. In Part/A, 300 images are used for training and 182 images are used for testing. Meanwhile,

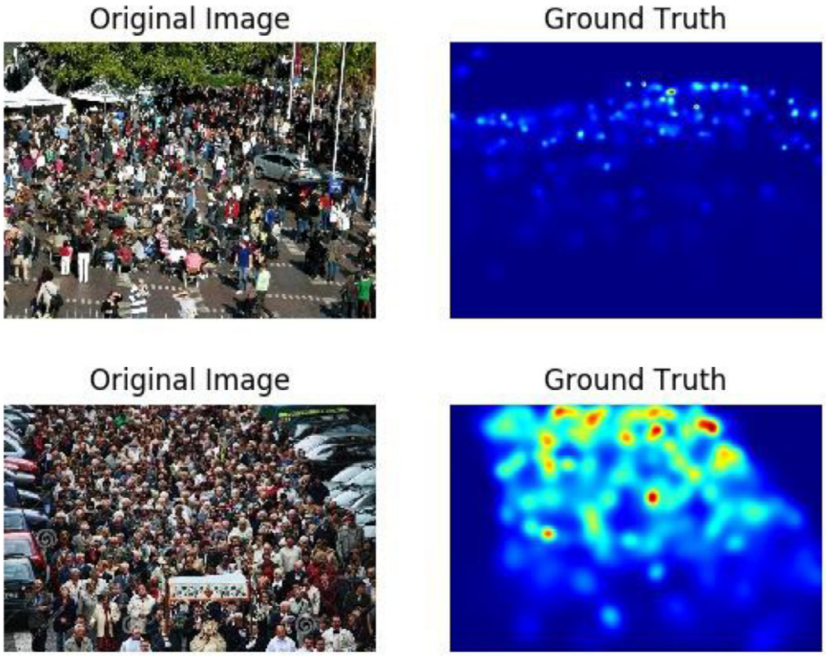


Fig. 7. Generated density maps by crowd location markers.

Fig. 8. ShanghaiTech dataset (PartA and PartB).

Fig. 9. UCF_CC_50 dataset.

400 images are used for training and 316 images are used for testing in Part/B. Some samples of ShanghaiTech dataset are shown in Fig. 8.

4.3.2. UCF_CC_50 dataset

UCF_CC_50 dataset [4] has 50 images, which has 63,974 head annotations totally. The headcounts in one image range between 94 and 4543. It is the most challenge dataset due to the small dataset size and large variance in crowd count. Here, the dataset operations are followed and predicted results are evaluated by using 5-fold cross-validation. Some samples of UCF_CC_50 dataset are shown in Fig. 9.

4.3.3. UCSD dataset

The UCSD dataset [44] has 2000 frames, which are captured by real surveillance cameras. These scenes are mostly in sparse conditions, varying from 11 to 46 persons per image. The regions of interest (ROI) are also provided. The resolution of each frame is all in low resolution (238×158). Among the 2000 frames, Frames from 601 to 1400 are used for training while the rest of them for testing [7]. Some samples of UCSD dataset are shown in Fig. 10.

4.4. Evaluation metrics

As followed existing works for crowd counting, MAE (Mean Absolute Error) and MSE (Mean Squared Error) are used for evaluation. MAE and MSE are defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y - y'\right| \tag{15}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\sqrt[2]{|y - y'|^2} \tag{16}$$

Here, y is the actual person number, while y' is the predicted number of people in the experiment. MAE can better reflect the

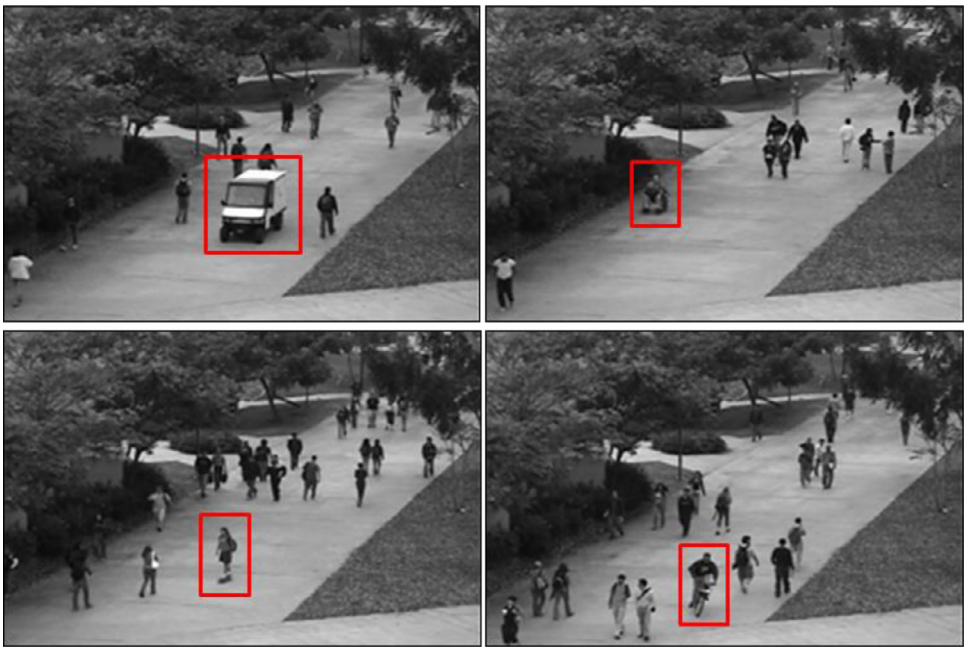


Fig. 10. UCSD dataset.

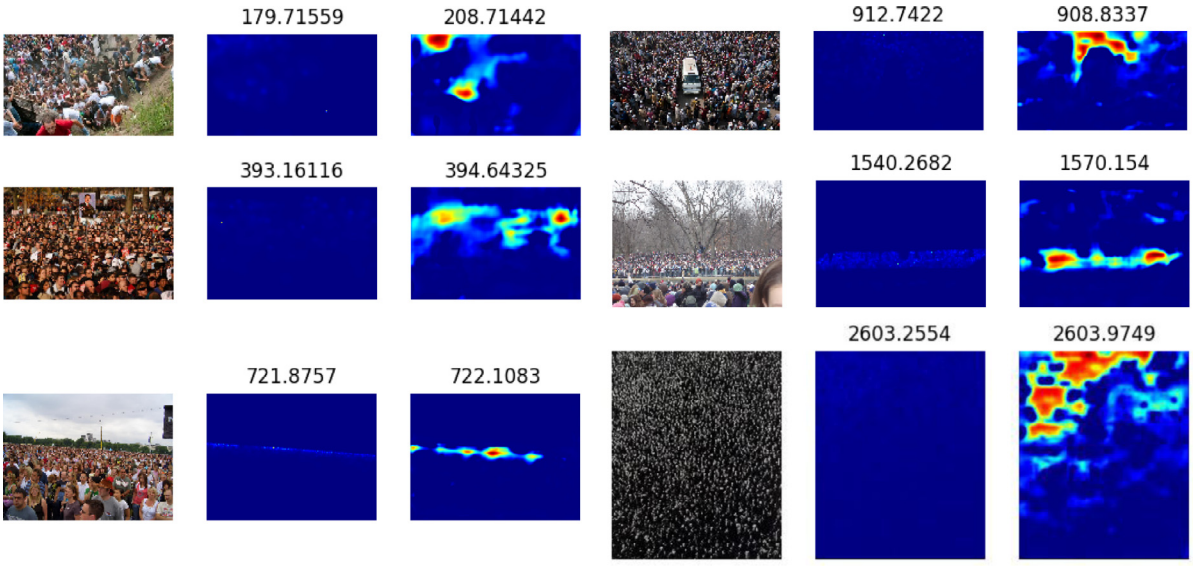


Fig. 11. Crowd density estimation results of our method.

actual situation of the prediction value error. MSE is a measure of the dispersion degree of random variables or data sets. The larger MSE is, the larger the dispersion is; the smaller MSE is, the better the accuracy and robustness of the model predicting the test data are.

PSNR [46] (Peak Signal-to-Noise Ratio) and SSIM [47] (Structural Similarity in Image) are also used to evaluate the quality of the predicted density map. The larger PSNR and SSIM are, the better quality the prediction has. To calculate the PSNR and SSIM, the preprocess is referred to [7], which follows normalization for both ground truth and predicted density map.

4.5. Performance evaluation

Results of our method on Shanghai Part/A dataset are shown in Fig. 11. From left to right, they are the original image, ground truth of density map and predicted density map, respectively. The number above the image represents the total number of people in the image. The number of predicted density maps is obtained by accumulating the values of all pixels in the image, as most paper use. The predicted density map is grayscale image. We map it into RGB space, leading to an intuitive crowd heatmap. The redder the place in the image is, the denser the crowd is. It can be observed that our method achieves good performance on dense conditions at different levels. Due to precise pixel-level prediction of PDC module, GSA module focus on using low-level features to recover pixel localization by pooling. The whole encoder-decoder can be treated as four stages and each stage has different scale features. In the encoder, feature maps of high stage are obtained by pooling from low stage. In the decoder, feature maps of low stage are generated by upsampling from high stage. This method of stage-by-stage upsampling to the original resolution of the input image can return each pixel value to an approximate truth value. All pixel values of density map prediction can be learned by neural network inference. In

Table 1
Density estimation results of different methods on four datasets.

Method	PartA		PartB		UCF_CC_50		UCSD	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN [2]	110.2	1173.2	26.4	41.3	377.6	509.1	1.07	1.35
SwitchNet [1]	90.4	135.0	21.6	33.4	318.1	439.2	1.62	2.10
SaCNN [5]	86.8	139.2	16.2	25.8	314.9	424.8	–	–
CP-CNN [48]	73.6	106.4	20.1	30.1	295.8	**320.9**	–	–
ACSCP [49]	75.7	**102.7**	17.2	27.4	291.0	404.6	**1.04**	**1.35**
M-task [50]	73.6	112.0	13.7	21.4	279.6	388.9	–	–
D-CNet [51]	73.5	112.3	18.7	26.0	288.4	404.7	–	–
IG-CNN [52]	72.5	118.2	13.6	21.1	291.4	349.4	–	–
SAANet [53]	*63.7*	*104.1*	*8.2*	*12.7*	*238.2*	*310.8*	–	–
CSRNet [7]	68.2	115.0	10.6	16.0	266.1	397.5	1.16	1.47
Our Method	**65.6**	107.2	**9.8**	**15.7**	**257.0**	343.9	1.08	1.44

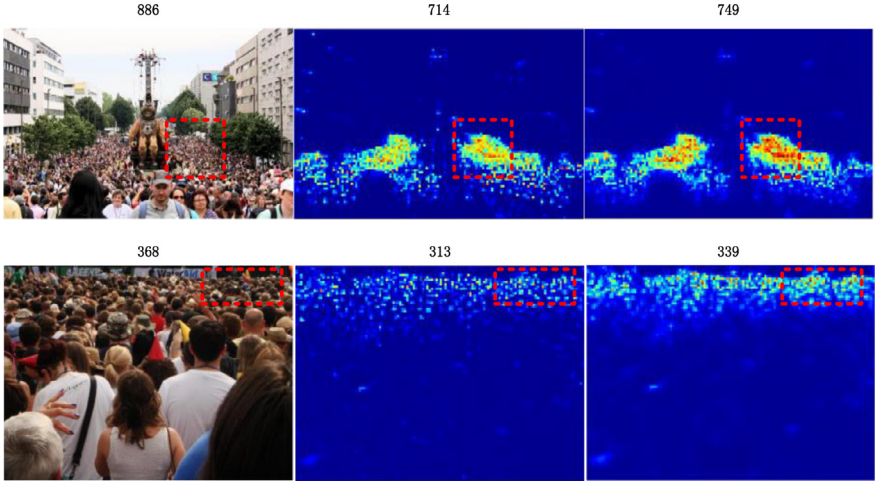


Fig. 12. The comparison between CSRNet and our method.

addition to predicting better on the number of persons in an image, our model also shows better localized predictions. Its density maps are much detailed than those output by the sub-model. The sub-model tends to over-smooth large crowd's regions, especially evident in Fig. 13. Besides, our hypothesis is validated that directly using low-level layers to help high-level layers build lost contents and compute loss is beneficial for localizing small-scale person, as these low-level feature maps have detailed spatial layout.

In this paper, it is found in the experiments that the quality of density map ground truth's generation is directly related to the prediction results of the model. The coordinates of person's heads are given in the crowd density datasets. Each coordinate corresponds to a person head, but not necessarily at the center of the head. This may have a negative impact on the head characteristics of model learning.

Our model is compared with several approaches in recent papers on the datasets introduced in Section 4.4. As shown in Table 1, our method has almost reached the state-of-the-art level on these four datasets. It can be seen that our method has been greatly improved, compared with CSRNet. Expect for our method, other methods don't make up for the loss of pixel information caused by pooling. They only generate the prediction by upsampling the final output several times. Fig. 12 shows that our method predicts better than CSRNet on the region of relatively dense crowd. Also, in Table 2, an improvement on MAE and MSE can be observed, compared with CSRNet. Even in the very sparse condition (UCSD), our method get −0.08 lower MAE and −0.03 lower MSE. SAANet [53], recently released by Amazon, uses the same scale-Aware idea as Multi-scale Loss in Section 3.3. Besides, it also designs an attention mask module and optimized loss regularization. Attention mask is very similar to the classification activation map. When training, it can guide the network to learn in a better direction and predict better. The ideas may become the future optimization direction of this paper.

4.6. Ablation study

In this subsection, an ablation study is performed to analyze the effects of different modules in the proposed method. Each module is added sequentially to the network and results for each configuration are compared on ShanghaiTech Part/A dataset. Due to large variations in crowd density and scale across images in this dataset, it is difficult to estimate density maps and crowd count with high accuracy. Thus, this dataset is chosen for a detailed analysis of the proposed method.

Following five configurations are evaluated: (1) VGG16 (Baseline); (2) Baseline + GSA: Baseline network with Global Self-Attention module in Section 3.1; (3) Baseline + PDC: Baseline network with Pyramid Dilated Convolution module in Section 3.2; (4) Baseline + GSA + PDC: Baseline network with both Global Self-Attention module and Pyramid Dilated Convolution module; (5) Baseline + GSA + PDC + Multi-scale Loss: Baseline network with GSA module, PDC module and Multi-scale Loss in Section 3.3. This is the total structure. We also add a comparison with MCNN, CP-CNN, and CSRNet. MAE, MSE, PSNR, and SSIM of each component are calculated and compared, as shown in Table 2.

The result of our total structure has lower MAE and MSE, higher PSNR and SSIM than the other three methods. PSNR and SSIM of our method have a little improvement than CSRNet. GSA module,

Table 2
Results for the different components of our architecture.

VGG16	+ GSA	+ PDC	+ Multi-scale Loss	MAE	MSE	PSNR	SSIM
√				117.2	179.6	19.61	0.48
√	√			83.9	133.8	21.57	0.66
√		√		105.7	165.3	21.44	0.55
	√	√		66.1	110.5	23.81	0.76
√	√	√	√	65.6	107.2	23.83	0.78
MCNN				110.2	173.2	21.4	0.52
CP-CNN				73.6	106.4	21.72	0.72
CSRNet				68.2	115.0	23.79	0.76

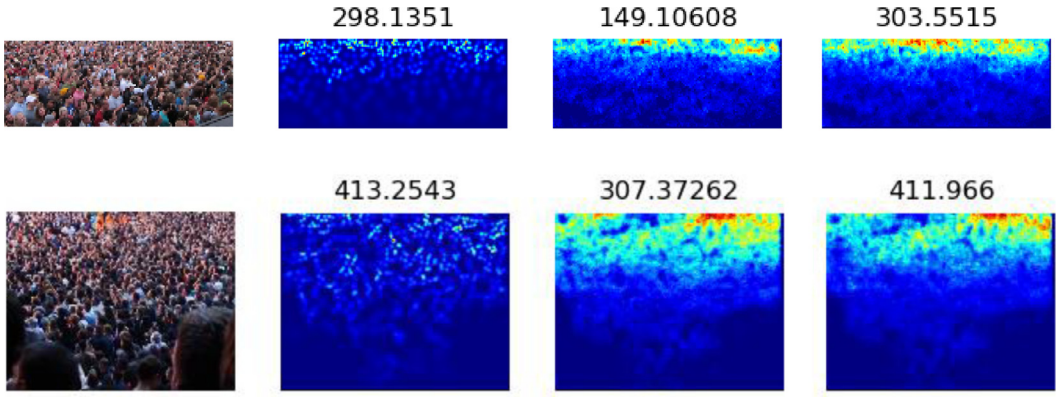


Fig. 13. Result comparison of the method with GSA module or not (from left to right: original image, ground truth, prediction by the method without GSA, prediction by the method with GSA).

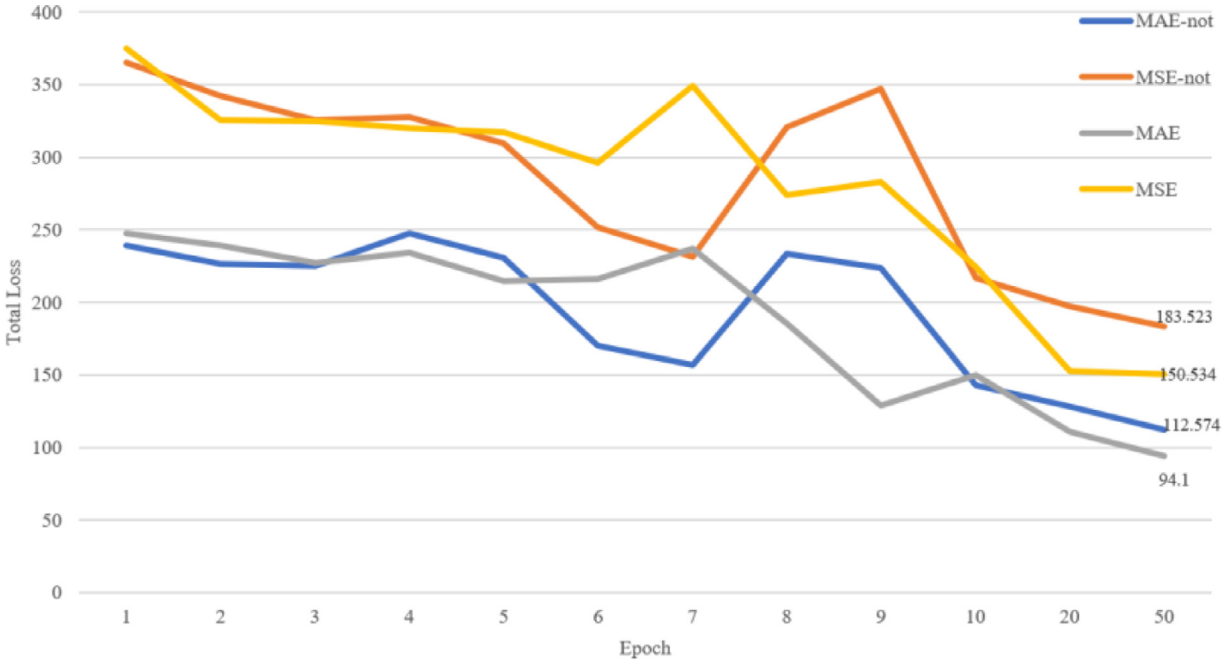


Fig. 14. The total loss changes of MAE or MSE (final stage loss) with Multi-scale Loss or not on the verification set.

PDC module, and Multi-scale Loss are proved to be beneficial and effective.

4.6.1. PDC module makes high-level features more detailed

Higher-level features with more semantic information are composed of lower-level features. Due to spatial invariance of images, it is simpler and more efficient to operate on high-level features than on the original image. ParseNet and PSPNet add a specific module to the feature map extracted by resnet50 to make the features more precise. The PDC module designed is to help clarify local confusions. Dilated convolution achieves the same functionality as pooling, but it does not lose spatial pixel information. Pyramid structure composed of dilated convolution has stronger feature extraction and expression ability. As shown in Table 2, the improvement in MAE and MSE is −11.5 and −14.3, respectively.

4.6.2. GSA module greatly increases the quality of density map

Global Self-Attention module (GSA) is aimed to recover pixel localization information by low-level features. We exploit the capability of high-level information by low-level context aggregation.

GSA module uses high-level features to guide low-level features learning. It also gradually decodes high-level features into original resolution density maps. By skillfully fusing the high-level features with the low-level features, every pixel of the image is involved in feature extraction and training. This is a simple and robust method to mix up the missing spatial information. Fig. 13 shows that it brings a moderate improvement (33.3 on MAE, 45.8 on MSE).

4.6.3. Multi-scale loss reduces the training time of networks

Using more stringent L1 Loss on the final generated density map makes the crowd count more accurate. L2 Loss is used in different scale feature maps in the process to make the model learn faster. As shown in Fig. 14, "MAE" and "MSE" presents the loss on the final stage and "not" represents whether it is with Multi-scale Loss or not. It proves that Multi-scale Loss can greatly reduce the training time of the network. Experiment confirms that Multi-scale Loss can improve our method with a little performance. We achieve lower error, 65.6 MAE and 107.2 MSE.

5. Conclusion

In this paper, we propose a new architecture called Dilated Convolution with Global Self-Attention (DCGSA). It is easy-trained and end-to-end for crowd counting and density map generating. We design two notable modules, Global Self-Attention module and Pyramid Dilated Convolution module. Global Self-Attention module exploits high-level feature maps to guide low-level features recovering pixel location. Pyramid Dilated Convolution module provides pixel-level and channel-level context and increases respective field by performing pyramid structure. Due to dilated convolution layers, our structure is capable of losing no resolution as much as possible.

In the future, there are three main research directions on crowd density estimation: (1) How to generate better density map's ground truth based on head coordinates; (2) How to find and distinguish very small heads; (3) How to use attention mechanism and generate better attention mask.

Declaration of Competing Interest

We declare that all the authors of this paper have no conflict of interest.

Acknowledgment

This work was supported by National Natural Science Foundation of China (Grant No. 61672042), Models and Methodology of Data Services Facilitating Dynamic Correlation of Big Stream Data, 2017.1~2020.12.

References

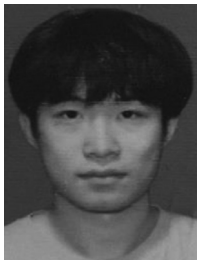
[1] D.B. Sam, S. Surya, R.V. Babu, Switching convolutional neural network for crowd counting, in: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 4031–4039.
[2] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, Single-image crowd counting via multi-column convolutional neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 589–597.
[3] L. Boominathan, S.S. Kruthiventi, R.V. Babu, Crowdnet: a deep convolutional network for dense crowd counting, in: Proceedings of the 24th ACM International Conference on Multimedia, ACM, 2016, pp. 640–644.
[4] H. Idrees, I. Saleemi, C. Seibert, M. Shah, Multi-source multi-scale counting in extremely dense crowd images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2547–2554.
[5] L. Zhang, M. Shi, Q. Chen, Crowd counting via scale-adaptive convolutional neural network, in: Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 1113–1121.
[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, Ssd: single shot multibox detector, in: Proceedings of the European Conference on Computer Vision, Springer, Cham, 2016, pp. 21–37.
[7] Y. Li, X. Zhang, D. Chen, CSRNet: dilated convolutional neural networks for understanding the highly congested scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1091–1100.
[8] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, (2014), arXiv:1409.1556.
[9] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[10] G. Lin, A. Milan, C. Shen, I. Reid, Refinenet: multi-path refinement networks for high-resolution semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1925–1934.
[11] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, Y. Bengio, The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 11–19.
[12] C. Peng, X. Zhang, G. Yu, G. Luo, J. Sun, Large kernel matters–improve semantic segmentation by global convolutional network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4353–4361.
[13] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[14] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, (2015), arXiv:1511.07122.
[15] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916.
[16] Q. Wang, J. Wan, Y. Yuan, Deep metric learning for crowdedness regression, IEEE Trans. Circuits Syst. Video Technol. 28 (10) (2018) 2633–2643.
[17] Q. Wang, M. Chen, F. Nie, X. Li, Detecting coherent groups in crowd scenes by multiview clustering, IEEE Trans. Pattern Anal. Mach. Intell. (2018).
[18] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, (2014), arXiv:1409.0473.
[19] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y.N. Dauphin, Convolutional sequence to sequence learning, in: Proceedings of the 34th International Conference on Machine Learning-Volume, 70, JMLR. org., 2017, pp. 1243–1252.
[20] M.T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, (2015), arXiv:1508.04025.
[21] Z. Lin, M. Feng, C.N.D. Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, A structured self-attentive sentence embedding, (2017), arXiv:1703.03130.
[22] A.P. Parikh, O. Täckström, D. Das, J. Uszkoreit, A decomposable attention model for natural language inference, (2016), arXiv:1606.01933.
[23] R. Paulus, C. Xiong, R. Socher, A deep reinforced model for abstractive summarization, (2017), arXiv:1705.04304.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, ..., I. Polosukhin, Attention is all you need, in: Proceedings of Advances in neural Information Processing Systems, 2017, pp. 5998–6008.
[25] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.
[26] H. Hu, J. Gu, Z. Zhang, J. Dai, Y. Wei, Relation networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3588–3597.
[27] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[28] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. Change Loy, D. Lin, J. Jia, Psanet: point-wise spatial attention network for scene parsing, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 267–283.
[29] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, ..., X. Tang, Residual attention network for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156–3164.
[30] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, A. Agrawal, Context encoding for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7151–7160.
[31] H. Ghader, C. Monz, What does attention in neural machine translation pay attention to?, (2017), arXiv:1710.03348.
[32] G. Tang, R. Sennrich, J. Nivre, An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation, (2018), arXiv:1810.07595.
[33] T. Alkhouli, G. Bretschner, J.T. Peter, M. Hethnawi, A. Guta, H. Ney, Alignment-based neural machine translation, in: Proceedings of the First Conference on Machine Translation:, 1, 2016, pp. 54–65. Research Papers.
[34] W. Chen, E. Matusov, S. Khadivi, J.T. Peter, Guided alignment training for topic-aware neural machine translation, (2016), arXiv:1607.01628.
[35] T. Cohn, C.D.V. Hoang, E. Vymolova, K. Yao, C. Dyer, G. Haffari, Incorporating structural alignment biases into an attentional neural translation model, (2016), arXiv:1601.01085.
[36] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition, 2017, pp. 2881–2890.
[37] L.C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, (2017), arXiv:1706.05587.
[38] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, G. Cottrell, Understanding convolution for semantic segmentation, in: Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 1451–1460.

[39] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, (2015), arXiv:1502.03167.
[40] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, A. Agrawal, Context encoding for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7151–7160.
[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, ..., A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[43] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[44] A.B. Chan, Z.S.J. Liang, N. Vasconcelos, Privacy preserving crowd monitoring: counting people without people models or tracking, in: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–7.
[45] V. Lempitsky, A. Zisserman, Learning to count objects in images, in: Proceedings of Advances in Neural Information Processing Systems, 2010, pp. 1324–1332.
[46] Q. Huynh-Thu, M. Ghanbari, Scope of validity of PSNR in image/video quality assessment, Electronics letters 44 (13) (2008) 800–801.
[47] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[48] V.A. Sindagi, V.M. Patel, Generating high-quality crowd density maps using contextual pyramid cnns, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1861–1870.
[49] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, X. Yang, Crowd counting via adversarial cross-scale consistency pursuit, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5245–5254.
[50] X. Liu, J. van de Weijer, A.D. Bagdanov, Leveraging unlabeled data for crowd counting by learning to rank, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7661–7669.
[51] Z. Shi, L. Zhang, Y. Liu, X. Cao, Y. Ye, M.M. Cheng, G. Zheng, Crowd counting with deep negative correlation learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5382–5390.
[52] Babu Sam, Sajjan D., N. N., Venkatesh Babu, M. Srinivasan, Divide and grow: capturing huge diversity in crowd images with incrementally growing cnn, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3618–3626.
[53] R.R. Varior, B. Shuai, J. Tighe, D. Modolo, Scale-Aware attention network for crowd counting, (2019), arXiv:1901.06026.

Liping Zhu is currently an Associate Professor and a Master Advisor with the Computer Technology Department, China University of Petroleum, Beijing. Her research interests are swarm intelligence, reservoir protection, clustering, big data and data mining. She has supervised over 40 master students. She is the Administrator of the Beijing Key Laboratory of Petroleum and Data Mining.

Chengyang Li received the degree in Computer Science and Technology from Northwest A&F University in 2017. He is currently pursuing the master degree of Computer Technology from China University of Petroleum, Beijing. His research interests include image processing, machine learning, and deep learning.

Bing Wang is completing a master's degree in computer science at China University of Petroleum (Beijing), and he graduated from China University of Petroleum (East China) with a bachelor's degree in computer science. He's broadly interested in deep learning and data analysis. His research focuses on computer vision and data modeling.



Kun Yuan is currently a master student of Artificial Intelligence program in the University of Ottawa. The focus of his research is the application of deep learning in low-level vision problem, such as super resolution, image inpainting, image deblurring and so on. In addition, his work also includes the field of medical image analysis, such as MRI super resolution, Multimodality transfering etc.

Zhongguo Yang received the Ph.D. degree in computer science from the China University of Pertroleum (Beijing), China, in 2018. He is an assist reseacher Beijing Key Laboratory on Integration and Analysis of Large-Scale Stream Data, North China University of Technology, Beijing. His research interests include service computing, the Internet of Things, and deep learning.