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

a r t i c l e i n f o

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

Communicated by Wang QI

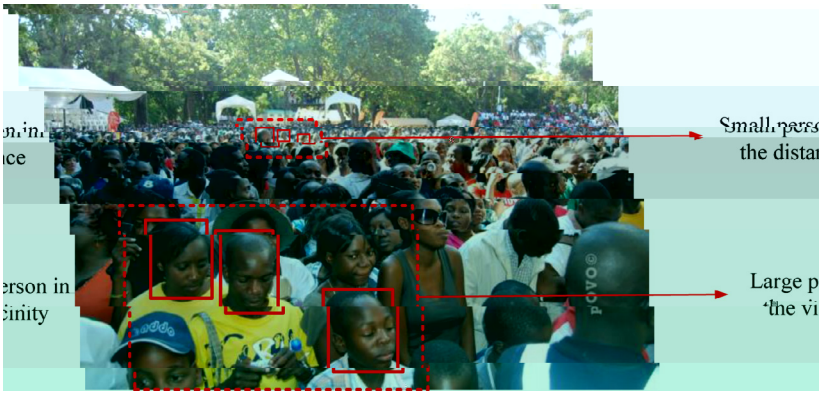


Fig. 1. Different sizes of person's head at the different distances from a camera.

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

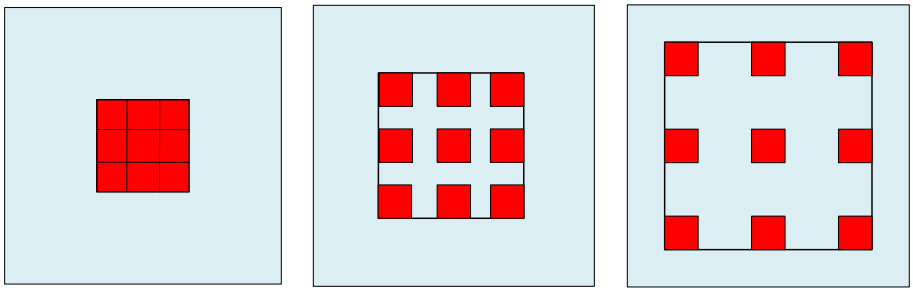


Fig. 2. Different dilation rate on feature maps.

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.

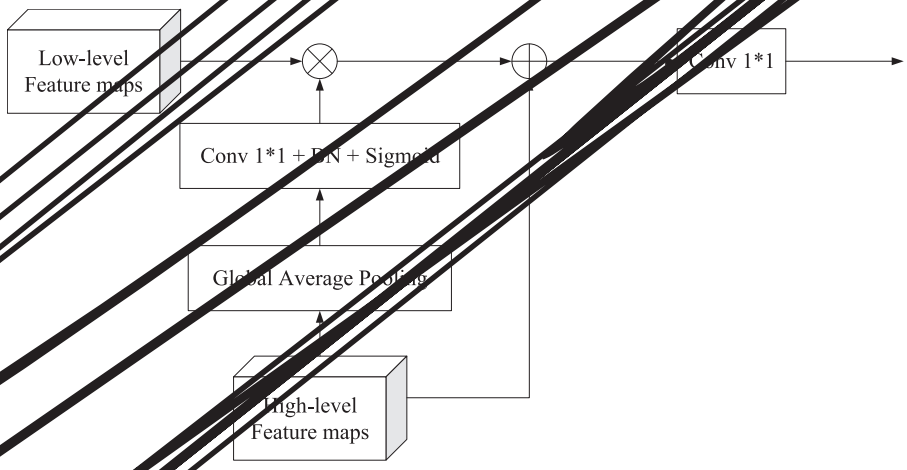
3. Proposed method

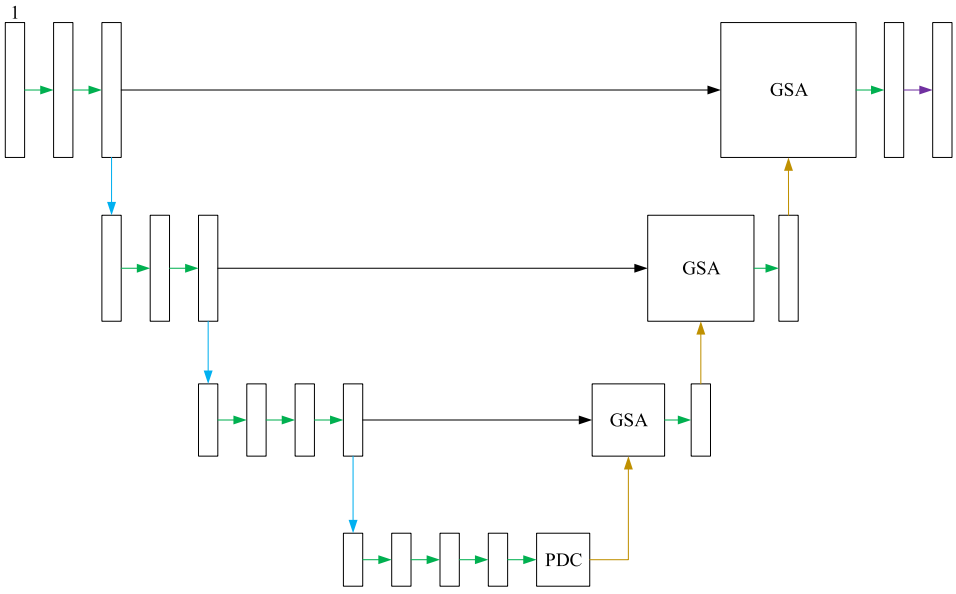
In this section, we first introduce the proposed Global Self-Attention (GSA) Module and Pyramid Dilated Convolution (PDC) Module. Then we describe the complete encoder–decoder network architecture DCGSA, designed for the joint task of predicting crowd density map and crowd counting.

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


Low-level
Feature maps
Conv 1*1 + BN + Sigmoid
Global Average Pooling
High-level
Feature maps
Conv 1*1


1
GSA
GSA
GSA
PDC

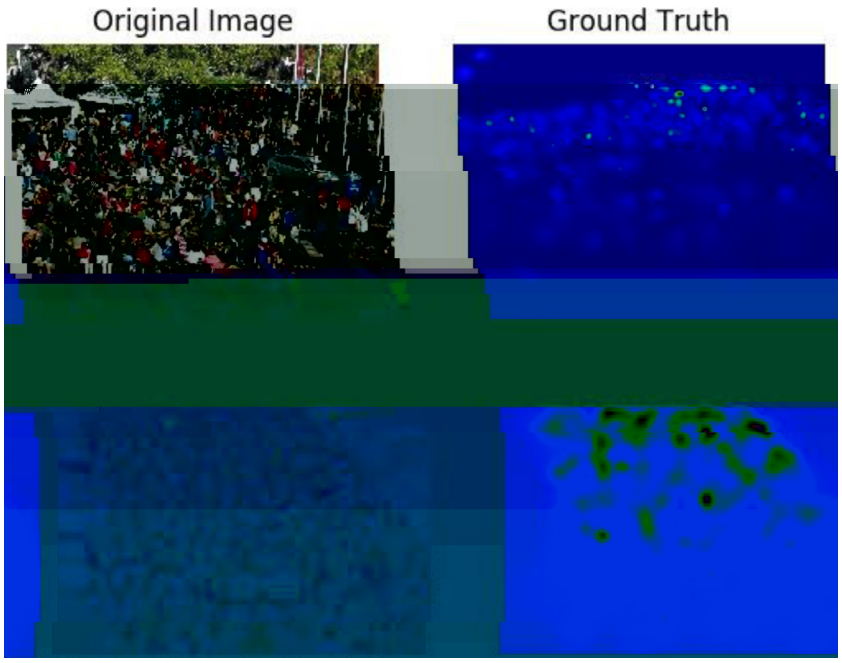


Fig. 7. Generated density maps by crowd location markers.

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, are

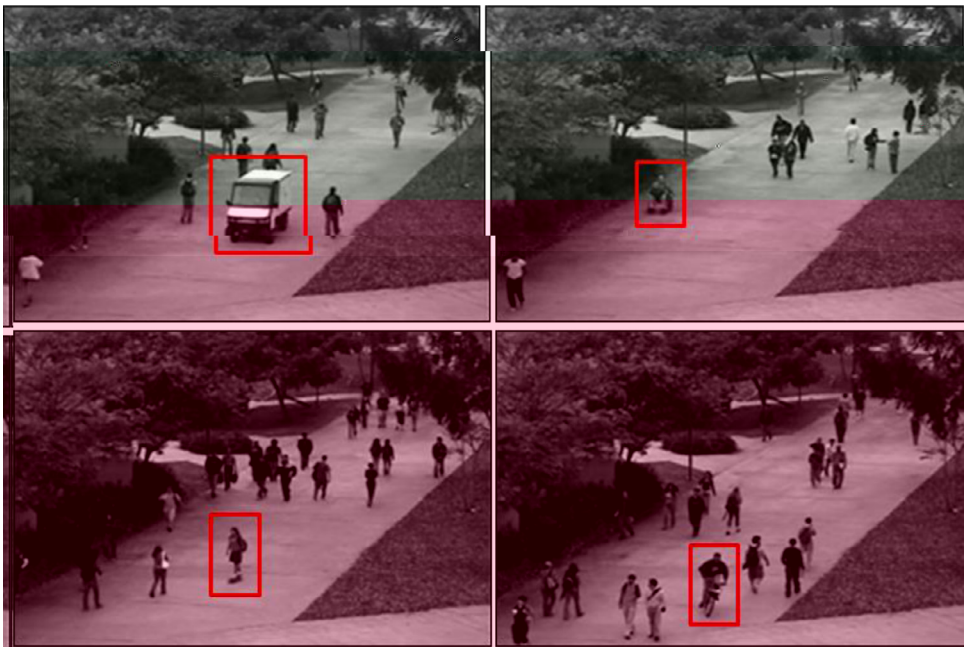


Fig. 10. UCSD dataset.

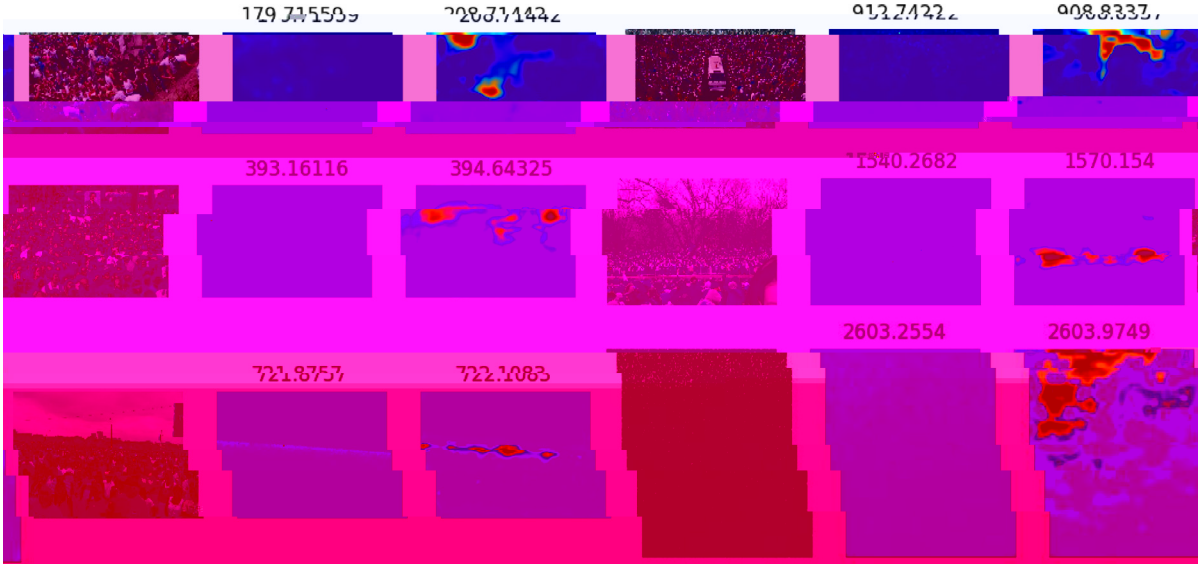


Fig. 11. Crowd density estimation results of our method.

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

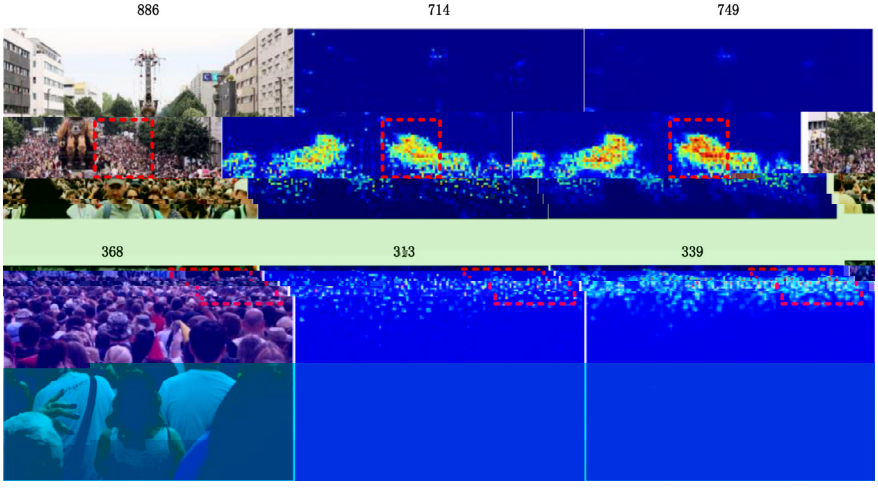


Fig. 12. The comparison between CSRNet and our method.

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

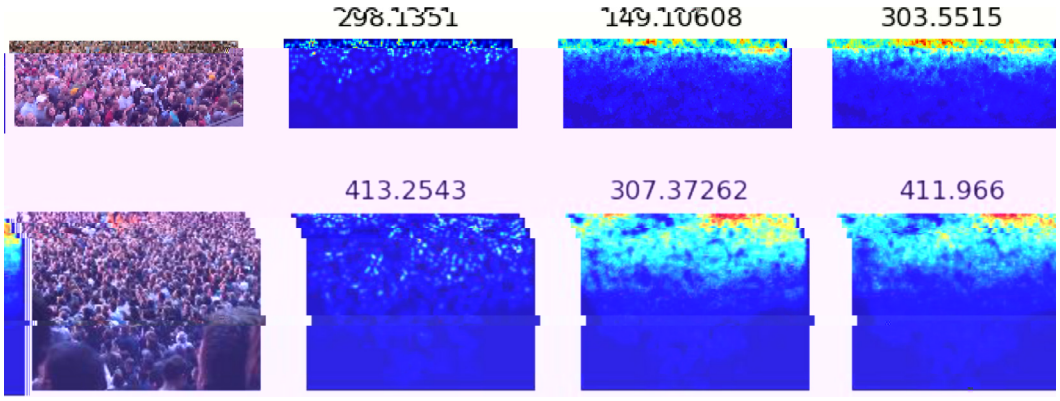


Fig. 13. Result comparison of the method with GSA module or not (from left to right: original image, ground truth, prediction by the method without GSA, prediction by the method with GSA).

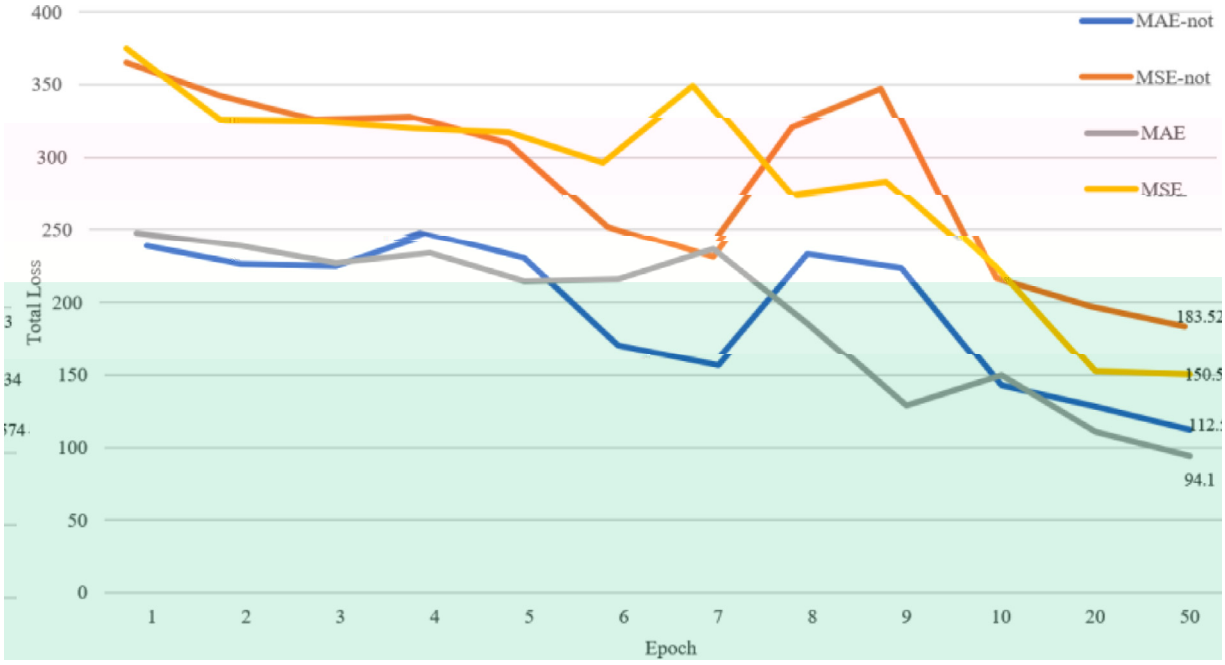


Fig. 14. The total loss changes of MAE or MSE (final stage loss) with Multi-scale Loss or not on the verification set.

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

[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 computer n

gr a5 n eepHisg and ddata