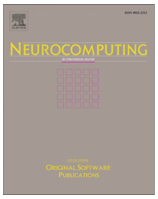
Neurocomputing 451 (2021) 12–24

Contents lists available at ScienceDirect

Neurocomputing

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



Crowd counting based on attention-guided multi-scale fusion networks

Bo Zhang [a], Naiyao Wang [a], Zheng Zhao [a], Ajith Abraham [b], Hongbo Liu [a,*]

[a] College of Information Science and Technology, Dalian Maritime University, 116026, China
[b] Machine Intelligence Research Labs, Scientific Network for Innovation and Research Excellence, Auburn, WA 98071, USA

ARTICLE INFO

Article history:
Received 20 June 2020
Revised 25 March 2021
Accepted 12 April 2021
Available online 18 April 2021
Communicated by Zidong Wang

Keywords:
Crowd counting
Density estimation
Attention mechanism
Dilated convolution

ABSTRACT

In this paper, we propose an attention-guided multi-scale fusion network (named as AMS-Net) for crowd counting in dense scenarios. The overall model is mainly comprised by the density and the attention networks. The density network is able to provide a coarse prediction of the crowd distribution (density map), while the attention network helps to distinguish crowded regions from backgrounds. The output of the attention network serves as a mask of the coarse density map. The number of persons in the scene is finally estimated by applying integration on the refined density map. In order to deal with persons of varied resolutions, we introduce a multi-scale fusion strategy which is built upon dilated convolution. Experiments are carried out on the standard benchmark datasets, covering varied over-crowded scenarios. Experimental results demonstrate the effectiveness of the proposed approach.

1. Introduction

Crowd counting is an emerging topic in the field of computer vision in recent years, which can be applied in a wide range of applications in visual surveillance, such as crowd simulation [1,2], crowd dynamics modeling [3], crowd clustering [4], abnormal behavior detection [5], pedestrian identification and facial recognition [6,7], and group behavior analysis [8,9], to name a few.

However, crowd counting in real environments is very challenging due to the intrinsic characteristics of the problem: (1) people in dense crowd always have lower resolutions with varied scales; (2) frequent occlusions make it impossible to observe the whole body of a person in the scene; (3) background clutters always exert negative influences on the counting accuracy.

In this work, we propose an attention-guided multi-scale fusion network to address the problems mentioned above, which is named as AMS-Net. Particularly, head regions are exploited to distinguish persons from each other, which are more suitable for dense environments comparing to the whole body representation, even in a low quality and low resolution scenario. In order to deal with persons of varied resolutions, we propose a multi-scale fusion strategy, which is built upon dilated convolution. The dilated convolution can expand the receptive field without introducing extra computational cost, thus allowing a wider range of perceptual scopes. In addition, it should be noticed that environmental cues will be mistaken into account occasionally in some off-the-shelf methods, due to background clutters. For example, in Fig. 1, trees are mis-labeled as persons, thus leading to the decrease in the counting accuracy. In order to tackle this problem, we introduce an attention mechanism in the framework, which allows the model to concentrate on crowds.

The main components in the overall framework are the density network and the attention network. The former can provide a coarse distribution of the crowd, while the later aims at differentiating crowded regions from backgrounds. The output of the attention network can be viewed as a mask of the coarse density map. The number of persons in the scene is finally estimated by applying integration on the refined density map.

To sum up, the main novelties and contributions of this work are presented as follows: (1) we propose an attention-guided framework to deal with the problem of crowd counting in realistic scenarios; (2) in order to perceive persons in varied resolutions, we adopt a multi-scale fusion strategy, which is built upon dilated convolution; (3) we introduce an attention mechanism in the framework, which allows the model to concentrate on crowd regions; (4) we use a channel-wise weighted strategy, which can further promote the counting accuracy.

The rest of the paper is organized as follows: In Section 2, we briefly review the recent progress in crowd counting and density estimation. In Section 3, we present the whole framework, including the multi-scale fusion strategy, the attention network, and the density network. Experimental results are demonstrated in Section 4, where we evaluate the performances of crowd counting

[*] Corresponding author.
E-mail addresses: bzhang@dlmu.edu.cn (B. Zhang), wny@dlmu.edu.cn (N. Wang), zhaozheng@dlmu.edu.cn (Z. Zhao), ajith.abraham@ieee.org (A. Abraham), lhb@dlmu.edu.cn (H. Liu).

https://doi.org/10.1016/j.neucom.2021.04.045
0925-2312/© 2021 Elsevier B.V. All rights reserved.

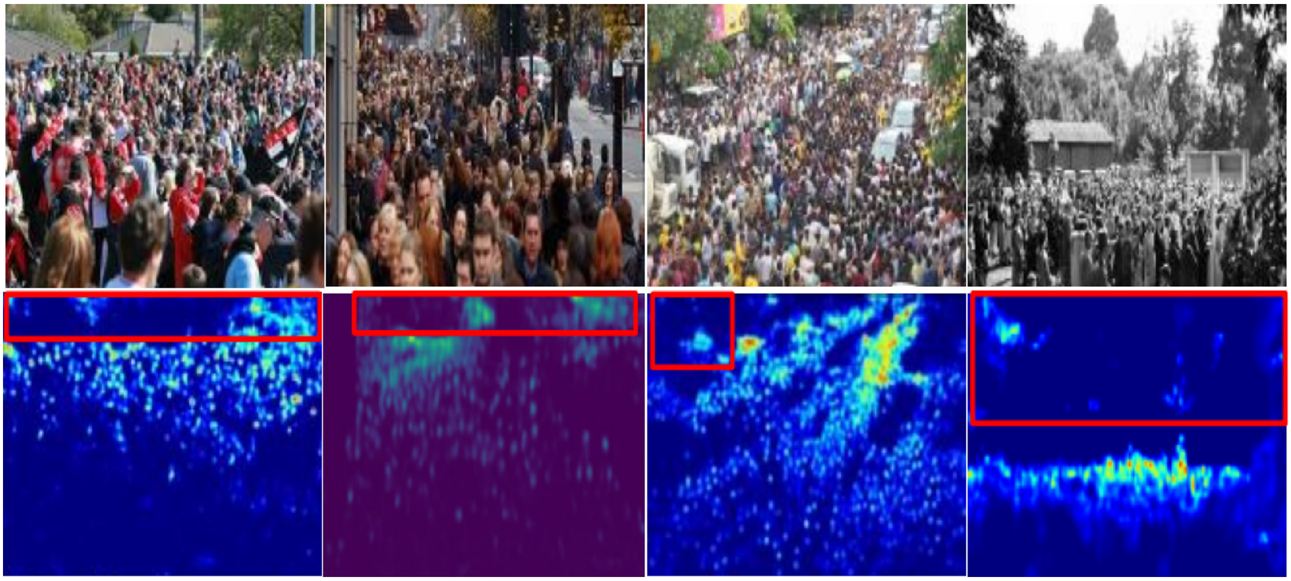


Fig. 1. Background clutters. Top: original scenarios; Bottom: density maps estimated using the method presented in [10]. The red bounding boxes indicate the background regions that being mistaken into consideration in the crowd density maps.

comprehensively and visualize the results. We conclude our work in Section 5.

2. Related work

In the last decade, a lot of effort has been spent by the research community in the area of crowd counting, mostly considering a bird-eye viewpoint, with applications to density estimation, abnormal event detection, crowd evacuation, and early warning, etc. Early works exploited the low-level visual features (i.e., HOG, wavelet) for pedestrian detection in relatively sparse scenarios. However, these features are hand-crafted, which are not suitable for dense areas, where background clutters, occlusions, non-uniform crowd distribution, and illumination changes always exist. More recently, with the rapid development of modern deep learning techniques, the performances of crowd counting in more complicated and realistic scenarios have been promoted significantly.

At the beginning, the convolutional neural network (CNN) and its variants were widely adopted. Boominathan et al. [11] presented the so-called CrowdNet, where the shallow and the deep fully convolutional neural networks are combined to capture both the high-level semantic information and the low-level visual features. Shang et al. [12] proposed an end-to-end convolutional neural network which exploits contextual information for both local and global count estimation. Sindagi et al. [13] provided an end-to-end CNN-based cascaded network, where the coarse count in an image serves as the high-level prior in the training procedure. In [14], Ranjan et al. presented a two-branch convolutional neural network, where one branch is used to generate a low resolution density map, and the other exploits the obtained low resolution density map to predict the corresponding high resolution density map.

Since scale variations are widely existed in dynamic scenarios due to varied camera viewpoints, it is essential to take multi-scale perceptions into consideration. In the recent years, varied multi-scale feature fusion strategies have been proposed, such as [15–17]. In the filed of crowd counting, Zhang et al. [10] proposed a multi-column convolutional neural network for crowd density estimation, which exploits filters with varied kernel sizes. This approach is able to deal with arbitrary sizes of input images and variations in people's resolution. In [18], Sam et al. presented a switching convolutional neural network for crowd counting. Each independent CNN is equipped with a specific receptive field. A switch classifier is exploited to relay crowd patches of varied scales to the proper CNNs. Li et al. [19] adopted the dilated convolutional neural network to expand the receptive field, in order to generate high-quality density maps. Cao et al. [20] first adopted multi-scale convolution kernels for spatial feature extraction, and then exploited deconvolution to generate the crowd density map. Yan et al. [21] proposed a perspective-guided convolutional neural network for crowd counting, which is able to deal with scale variations of people caused by the perspective effect. Yang et al. [22] proposed a reverse perspective network for perspective-aware object counting, which can solve the scale variations in an unsupervised manner. Bai et al. [23] proposed an adaptive dilated convolution network, which learns a specific continuous dilation rate to effectively match the scale variations at different locations.

In [24], Wang et al. proposed a large-scale crowd dataset for the task of counting and localization, which facilitates training the CNN-based models in a supervised manner. In [25,26], Wang et al. utilized synthetic crowd data with enough annotations to tackle the counting problem. In [27–29], the attention mechanisms [30] were adopted. In [31], Liu et al. exploited contextual information to predict crowd density. In [32], Ma et al. leveraged on the Bayesian loss for crowd counting with point supervision. In [33–35], multi-level feature fusion strategies were utilized for crowd counting and density estimation. A detailed overview on the recent literature can be found in the survey papers [36,37].

3. Methodology

3.1. Framework

The overall framework of the proposed AMS-Net is presented in Fig. 2. Raw input images are processed by the attention network and the density network directly. In order to perceive people with varied resolutions, we embed a multi-scale fusion module into the attention network, whose output is used as the mask of the density map that obtained through the density network. Finally, the

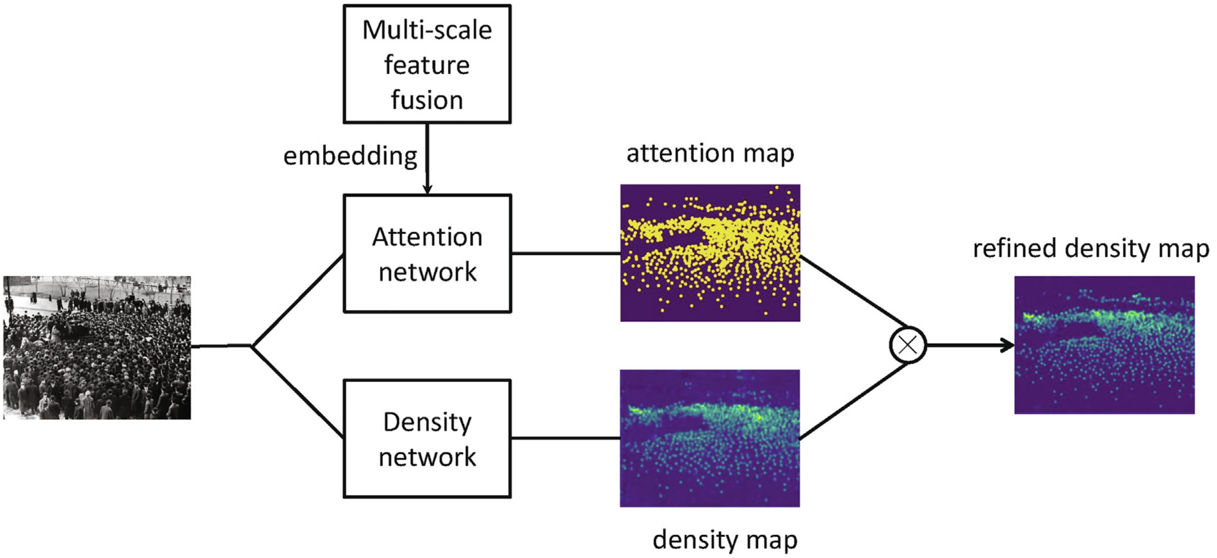


Fig. 2. The proposed framework. The attention map corresponds to the regions where crowds are more likely to locate, and the density map is a coarse estimation of the crowd distribution in the scene. ⊗ represents the dot production between the attention map and the density map.

number of persons in the crowd is estimated by applying integration on the refined density map.

3.2. Multi-scale feature fusion module

In highly dense scenarios, people always demonstrate multiple scales observing from bird-eye view cameras. Especially in the very low-resolution images, where people in the distance look rather small. When applying the standard convolution network for crowd feature extraction, the down-sampling operation will further discard spatial information, thus being not suitable to deal with people in small scales.

Therefore, we propose to use the so-called dilated convolution [38] for spatial feature extraction in the crowd. The dilated convolution is defined as in Eq. (1):

$$y(m,n)=\sum_{i=1}^{M}\sum_{j=1}^{N}x(m+r\times i,n+r\times j)w(i,j) \tag{1}$$

where $x(m,n)$ represents the original image; $w(i,j)$ represents the filter, with the size of $M\times N; r$ corresponds to the dilation rate. The dilated convolution can expand the receptive field from $k\times k$ to $k+(k-1)\times(r-1)$. When r = 1, the dilated convolution degenerates to the standard convolution operation. The advantages of dilated convolution lie in the following aspects: (1) it does not adopt the down-sampling operation, which is suitable to handle people in small scales; (2) it can expand the receptive field to a larger scope; (3) it can maintain the original kernel size, without introducing extra computational cost.

Due to the above reasons, we propose a multi-scale fusion strategy which is built upon dilated convolution with varied dilation rates. The fundamental structure is shown in Fig. 3, where the kernel size is set to $3\times 3, r$ is set to 1, 3, 6, 9, accordingly (actually more different dilation rates can be used). The whole procedure is presented as in Eq. (2). For any input I_{input}, $\mathrm{Conv}_{r_i}^{M\times M}(\cdot)$ indicates the application of a $M\times M$ kernel with the dilation rate r_i; $\mathrm{Concate}(\cdot)$ represents the concatenation of features extracted at different scales. The obtained feature f_{multi} will be further processed by the standard 1-d convolution as presented in Eq. (3). This multi-scale feature fusion module will be further integrated into the attention network in the following paragraphs. When stacking multiple dilated convolution layers sequentially, it can perceive spatial features with various scales.

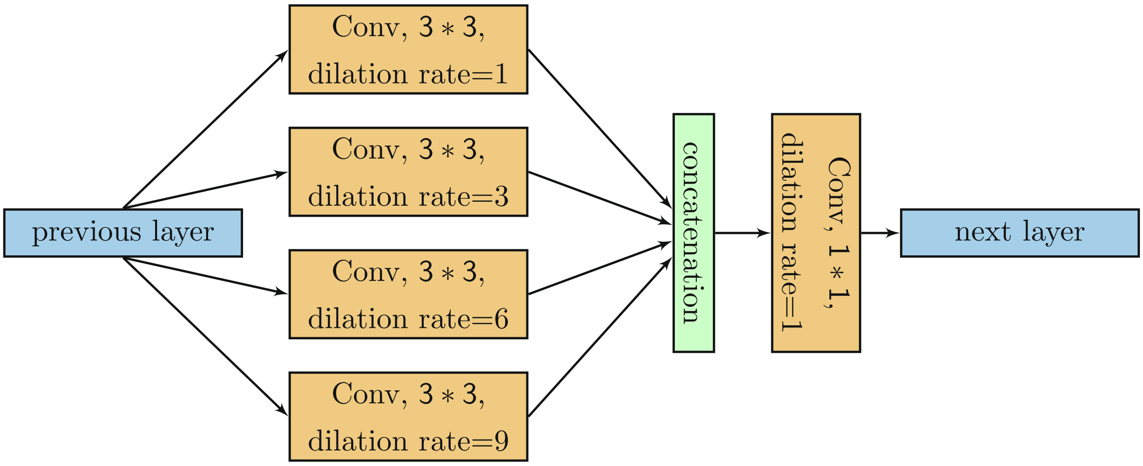

Conv, 3 ∗ 3, dilation rate=1
Conv, 3 ∗ 3, dilation rate=3
previous layer
concatenation
Conv, 1 ∗ 1, dilation rate=1
next layer
Conv, 3 ∗ 3, dilation rate=6
Conv, 3 ∗ 3, dilation rate=9

Fig. 3. The fundamental structure of multi-scale fusion network.

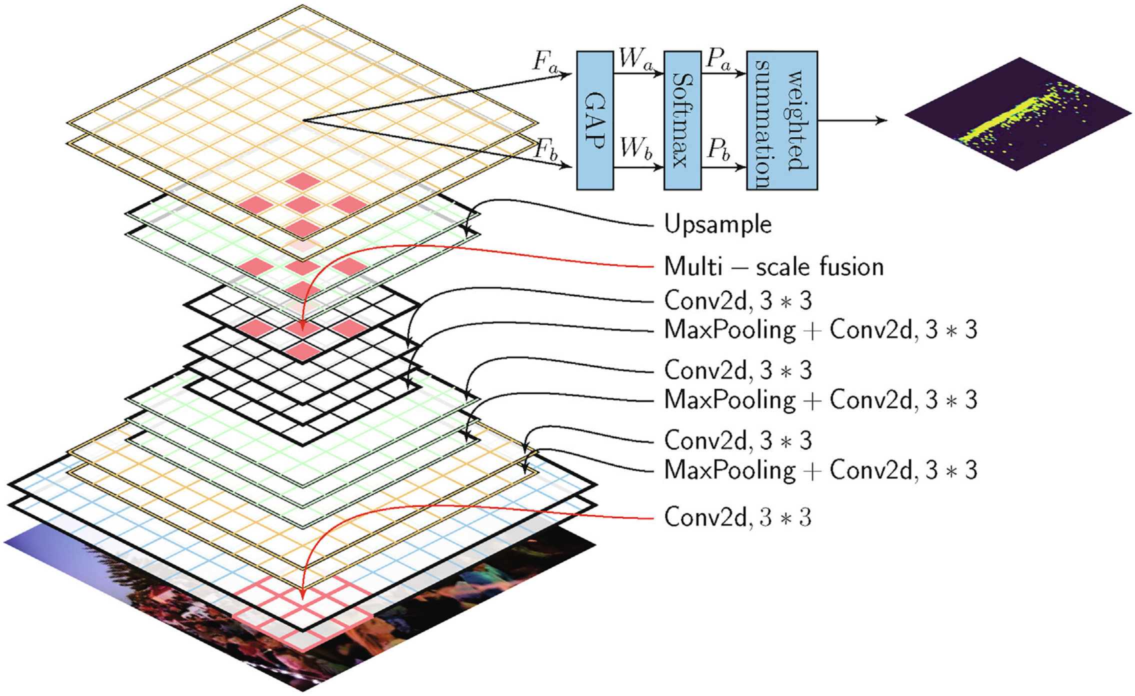


Fig. 4. The crowd attention network.

$$f_{multi} = \text{Concate}[\text{Conv}^{3\times3}_{r_i}(I_{input})]; \quad i = 1, 2, 3, \cdots; \tag{2}$$

$$f_{output} = \text{Conv}^{1\times1}(f_{\text{multi}}); \tag{3}$$

3.3. Crowd attention network

Backgrounds usually exert side effects on the performances of crowd counting and density estimation. In this work, we incorporate an attention component in the proposed framework which enables the model to focus on regions that human crowds are more likely to locate. The structure of crowd attention network is presented in Fig. 4 (the detailed implementations can be found in Table 9 in the Appendix section), where the original size of the input image is H×W×C (height, width, and channels).

The front-end of the network is built upon the VGG-16 network [39], which corresponds to Conv1 to Conv4 as presented in Table 9. The Conv5, Conv6, and Conv7 layers are implemented using the fundamental multi-scale fusion network as shown in Fig. 3, which are stacked sequentially, with up-sampling layers intersected

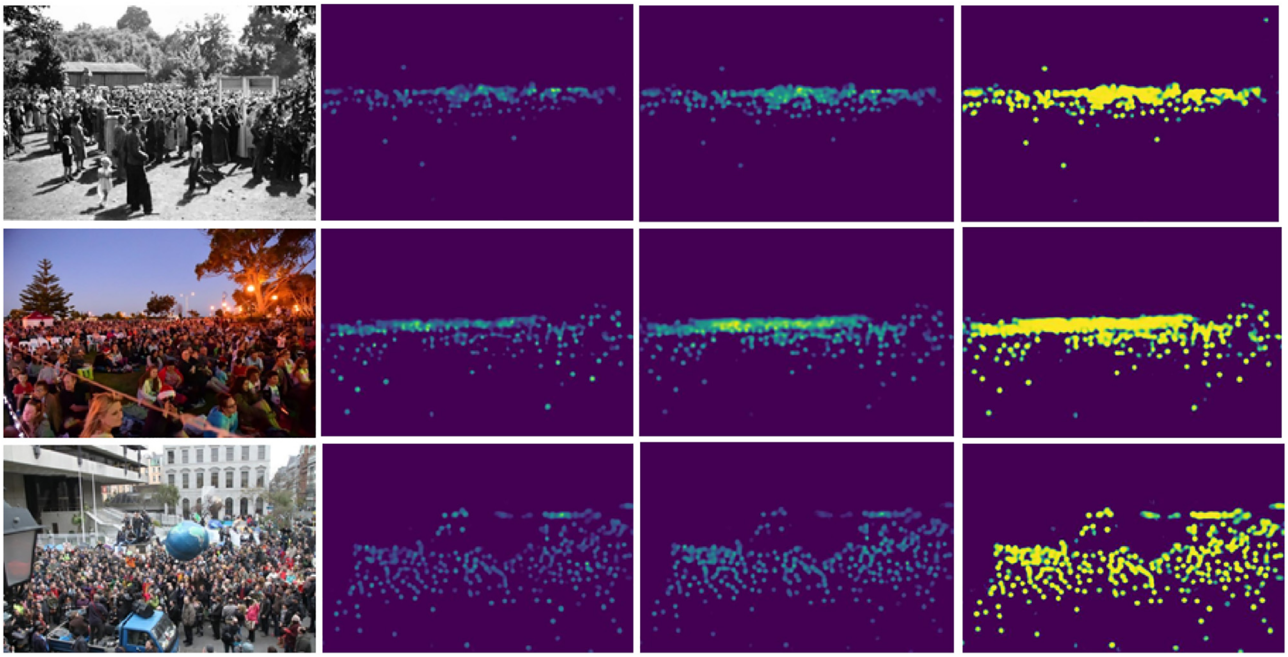


Fig. 5. Examples of attention maps. Column-1: original scenarios; Column-2: F_a; Column-3: F_b; Column-4: attention maps.

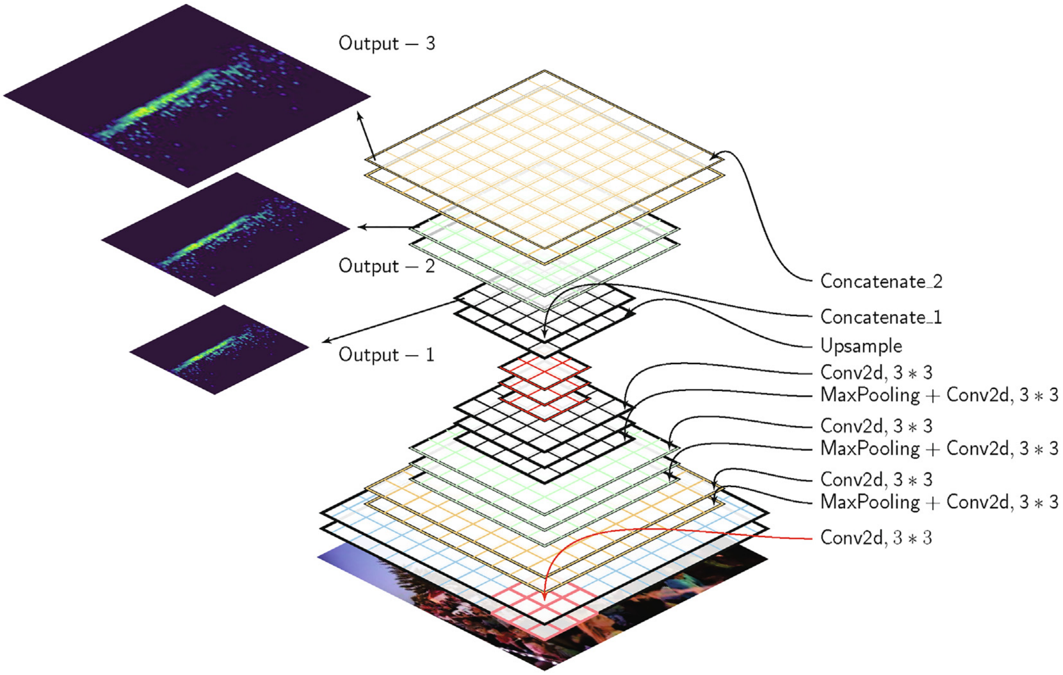


Fig. 6. The crowd density network.

alternately. The output of Conv7 consists of two feature maps, namely F_a and F_b with the size of (H/2)×(W/2). Motivated by the recent success of squeeze-and-excitation networks [40], we intend to assign channel-wise weights to F_a and F_b, respectively. We apply the global average pooling (GAP) and the softmax operation on F_a and F_b as shown in Eq. (4). The attention map is obtained using Eq. (5). This weighted operation can achieve competitive performances as compared to the adoption of the standard squeeze-and-excitation operation, which will be shown in the experimental section. Moreover, it is also computational efficient. Examples of the obtained attention maps are shown in Fig. 5.

$$(P_a, P_b) = \text{softmax}(\text{GAP}(F_a), \text{GAP}(F_b)); \tag{4}$$

$$f_{att} = \text{sigmoid}(P_a \cdot F_a + P_b \cdot F_b); \tag{5}$$

3.4. Crowd density network

The structure of crowd density network is presented in Fig. 6 (the implementation details can be found in Table 10 in the Appendix section), where the original size of the input image is H×W×C (height, width, and channels). The three output density maps (cor-

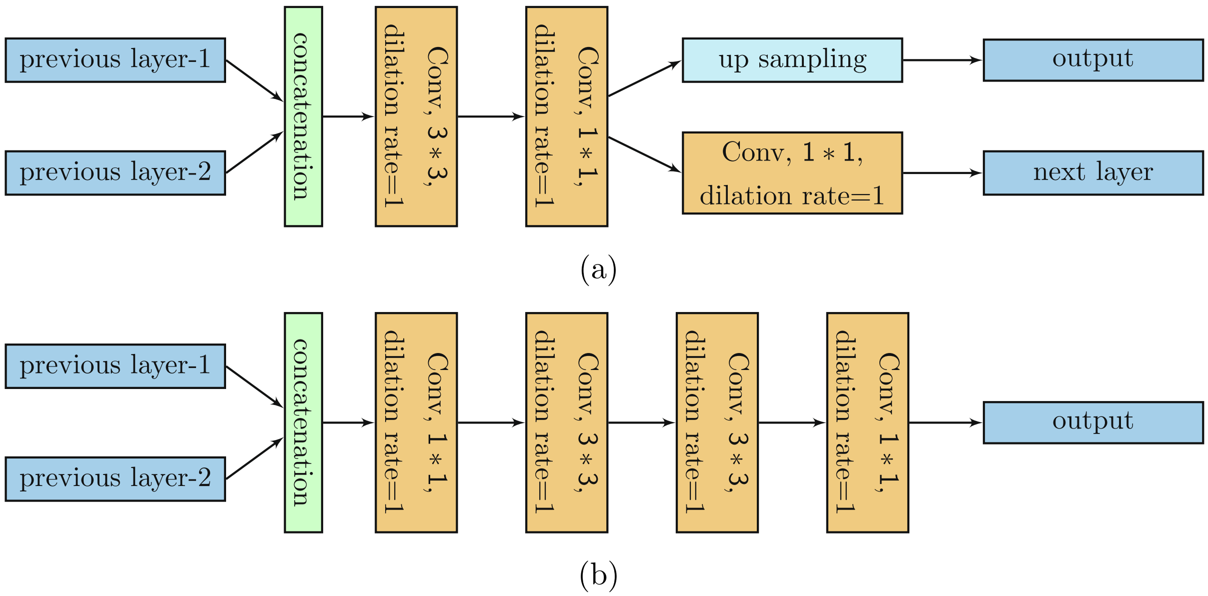

previous layer-1
previous layer-2
concatenation
Conv, 3 ∗ 3, dilation rate=1
Conv, 1 ∗ 1, dilation rate=1
up sampling
output
Conv, 1 ∗ 1, dilation rate=1
next layer
(a)
previous layer-1
previous layer-2
concatenation
Conv, 1 ∗ 1, dilation rate=1
Conv, 3 ∗ 3, dilation rate=1
Conv, 3 ∗ 3, dilation rate=1
Conv, 1 ∗ 1, dilation rate=1
output
(b)

Fig. 7. The detailed structures of the concatenation operations. (a) Concatenate_1; (b) Concatenate_2.

responding to Output1, Output2, and Output3 layers in Table 10) will be taken into account when constructing the loss function.

Particularly, the operations of Concatenate_1 and Concatenate_2 in Table 10 are shown in Fig. 7.

3.5. Model training

The overall loss is comprised by the attention loss and the density loss. In the following, we will present the details.

3.5.1. Attention loss

The attention map is a binary image, which is used to distinguish foreground regions from backgrounds. Thus, we define the attention loss using the cross-entropy form, which is formulated as in Eq. (6), where A_k represents the attention map of an input image k. We use $A_{k,i,j}$ to represent the value of the attention map A_k at a specific position (i,j). $A^{GT}_{k,i,j}$ indicates the corresponding ground truth. m is the batch size.

$$
\begin{aligned}
l_k &= \sum_{i=1}^{M}\sum_{j=1}^{N} A^{GT}_{k,i,j} log(A_{k,i,j}) + (1 - A^{GT}_{k,i,j}) log(1 - A_{k,i,j}) \\
L_{att} &= -\tfrac{1}{m}\sum_{k=1}^{m} l_k
\end{aligned}
\tag{6}
$$

3.5.2. Density loss

The density loss measures the Euclidean distance between the estimated density map and the corresponding ground truth, which is defined as in Eq. (7), where Z_i represents the density map of an input image i, and Z^{GT}_i indicates the corresponding ground truth. m is the batch size.

$$
L_{density} = \frac{1}{2m}\sum_{i=1}^{m} \|Z_i - Z^{GT}_i\|^2 \tag{7}
$$

The overall loss is defined as in Eq. (8), where λ is the weight of the attention loss, and l represents the number of output layers in the crowd density network.

$$
loss = \lambda L_{att} + \sum_{i=1}^{l} L_{density} \tag{8}
$$

The proposed model is learned in a supervised manner. Thus, we need to generate the ground truths for the density and the attention maps by leveraging on people's locations.

3.5.3. Ground truth of the density map

For a given image with N labeled persons, we use the approach presented in [10] to generate the corresponding density map. Please see Eq. (9) for details, where x_i represents the head location of a person. $G_{\sigma_i}(\cdot)$ is a Gaussian function. We compute the distances of the k-nearest persons in the neighborhood of x_i, and the average value d_i is used to initialize the variance σ_i. β is set to 0.5. '×' indicates the standard convolution operation.

$$
\begin{aligned}
F(x) &= \sum_{i=1}^{N} \delta(x - x_i) \times G_{\sigma_i}(x) \\
\sigma_i &= \beta d_i
\end{aligned}
\tag{9}
$$

Examples of the ground truths corresponding to the density maps are shown in Fig. 8.

3.5.4. Ground truth of the attention map

We binarize the density map to generate the corresponding attention map, where the threshold is set to 0.01. Examples are shown in Fig. 9.

4. Experiments

In this section, we will present the details of the experiments. First, we briefly introduce the standard benchmark datasets for validation. Next, we provide the evaluation protocols and the setting of hyper-parameters. Finally, we evaluate the performances of crowd counting comprehensively, and visualize the experimental results.

4.1. Benchmark datasets

At the early stage, the ShanghaiTech [41] and the UCF_CC_50 [42] datasets were considered as the standard benchmarks to evaluate different crowd counting algorithms. In the recent years, more large-scale crowd datasets have been released, with increasing number of sample images and more challenging scenarios, such as the UCF-QNRF [43], the JHU-CROWD++ [44], and the NWPU Crowd [24] datasets. In this section, we first use the ShanghaiTech

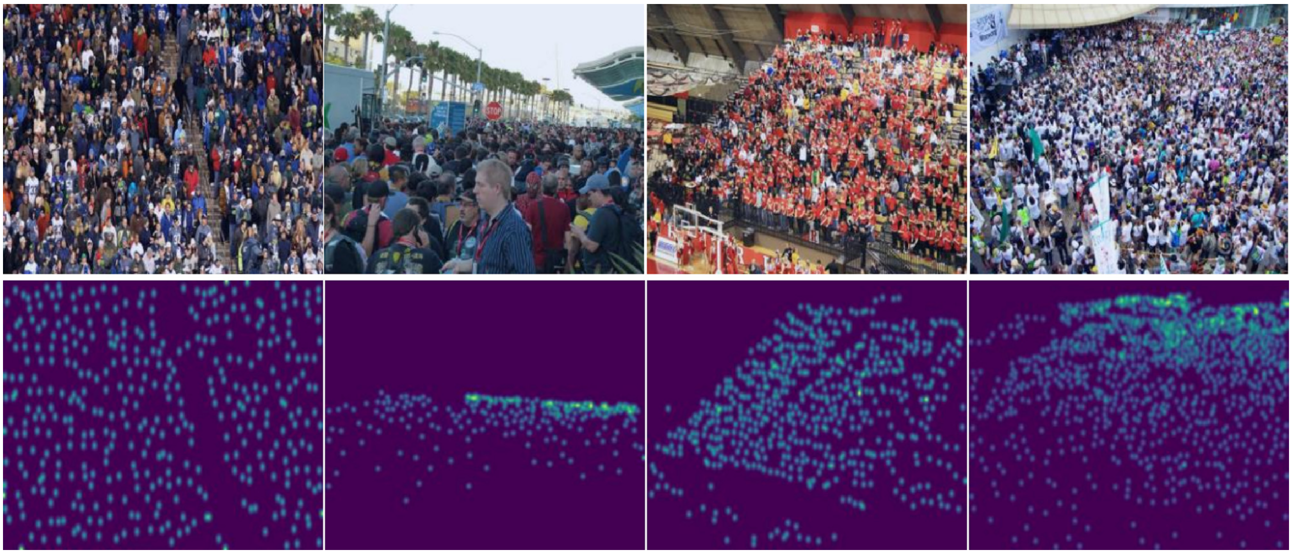


Fig. 8. Ground truths of density maps. Top: original scenarios; Bottom: ground truths of density maps obtained using Eq. (9).

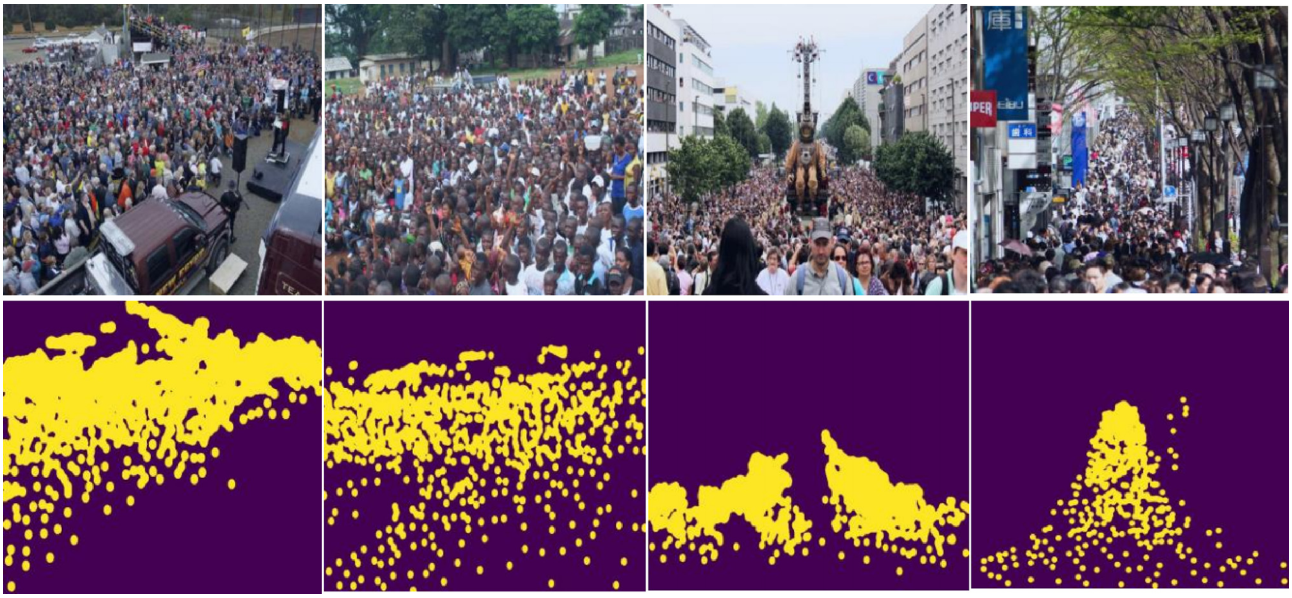


Fig. 9. Ground truths of attention maps. Top: original scenarios; Bottom: ground truths of attention maps.

and the UCF_CC_50 datasets for evaluation, in order to validate the fundamental characteristics of the proposed framework. Next, we apply our approach on the UCF-QNRF, the JHU-CROWD++, and the NWPU Crowd datasets, demonstrating the counting performances on large-scale data. The detailed descriptions of the mentioned datasets are presented as below.

ShanghaiTech: The ShanghaiTech dataset is a large-scale benchmark for crowd counting, which includes 1,198 labeled images covering 330,165 people in different scenes. This dataset can be further divided into two subsets, namely Group A and Group B. Group A contains 482 images in the dense scenarios, where 300 images are used for model training, and the rest are used for test. Group B contains 716 images in the relatively sparse scenarios, where 400 images are used for model training, and the rest are used for test.

UCF_CC_50: The UCF_CC_50 dataset contains 50 images captured from highly-crowded scenes. The number of labeled persons varies from 94 to 4,543, and the average is 1,280. The 5-fold cross-validation strategy is used for evaluation in this dataset.

UCF-QNRF: The UCF-QNRF dataset is used to evaluate crowd counting and localization approaches, which contains 1,535 images in total, where 1,201 images are used for training and the rest 334 images are used for test.

JHU-Crowd++: The JHU-Crowd++ dataset is a large-scale dataset, containing 4,372 images collected in a variety of different scenarios. Moreover, it also provides rich annotations (1.51 million in total), such as dots, bounding boxes, blur levels, etc.

NWPU Crowd: The NWPU Crowd is currently the largest dataset for crowd counting and localization, which includes 5,109 images in total with 2,133,375 head annotations. It contains a variety of crowd scenes with diverse illumination conditions and density ranges.

We present the differences of the mentioned benchmark datasets in Table 1. Examples of the corresponding crowd scenarios can be found in Fig. 10.

4.2. Evaluation protocols and hyper-parameters

We use the mean average error (MAE) and mean square error (MSE) as the criteria for evaluation, which are defined as in Eq. (10) and (11), respectively, where n is the number of images in the test set, C_i indicates the estimated number of persons in an image, and C_i^{GT} is the corresponding ground truth.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|C_i - C_i^{GT}| \tag{10}$$

$$\mathrm{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|C_i - C_i^{GT}|^2} \tag{11}$$

MAE reflects the average counting accuracy, and MSE shows the stability of the counting algorithm. Lower values of MAE and MSE imply better performances in the task of crowd counting.

Given the density map of an image, the number of persons can be approximated as in Eq. (12), where H and W represent the height and the width of a density map, $Z_{h,w}$ represents the density value at a specific position (h, w).

$$C_i = \sum_{h=1}^{H}\sum_{w=1}^{W} Z_{h,w} \tag{12}$$

In the training phase, we use the Adam algorithm to optimize the loss function, and the learning rate is set to $1e-4$. The number of epoch is set to 500, and the batch size is set to 2.

4.3. Results

4.3.1. Evaluation on the performances of crowd counting

First, we validate the role of attention mechanism in the proposed model. To this end, we compare the crowd counting performances by removing the attention network. Experimental results can be seen in Table 2, from where we can find that the attention network can promote the performances significantly in terms of MAE and MSE. The results are consistent in both dense (Group A) and sparse (Group B) scenarios in the ShanghaiTech dataset, and also in the UCF_CC_50 dataset.

In Fig. 11, we show the estimated number of persons in every image in the ShanghaiTech dataset. Group A contains 182 test images, and Group B contains 316 test images. For demonstration, we sort images according to their ground truth C_i^{GT} in the ascending order. In Fig. 12, we provide several visual examples. It can be seen clearly that without the support of attention network, some

Table 1
Characteristics of the benchmark datasets used for evaluation.

Dataset		#Total Images	Min	Max	Average	Total
ShanghaiTech	A	482	33	3,139	501	241,677
	B	716	9	578	124	88,488
UCF_CC_50		50	94	4,543	1,280	63,974
UCF-QNRF		1,535	49	12,865	815	1,251,642
JHU-Crowd++		4,372	0	25,791	346	1,515,005
NWPU Crowd		5,109	0	20,033	418	2,133,375

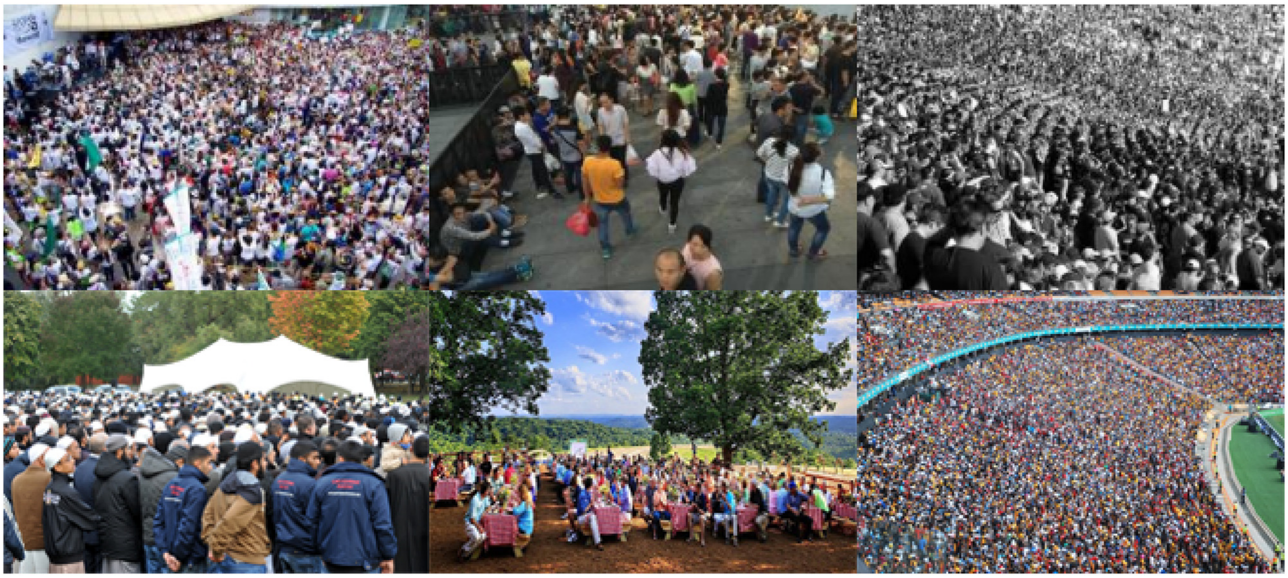


Fig. 10. Benchmark datasets. Top: ShanghaiTech Group A, ShanghaiTech Group B, and UCF_CC_50. Bottom: UCF-QNRF, JHU-Crowd++, and NWPU Crowd. It can be seen clearly that the densities and people's scales are different in varied datasets.

Table 2
Evaluation on the attention network.

Attention network	ShanghaiTech A		ShanghaiTech B		UCF_CC_50	
	MAE	MSE	MAE	MSE	MAE	MSE
No	73.5	114.8	11.4	23.3	432.4	617.8
Yes	**63.8**	**108.5**	**7.3**	**11.8**	**236.5**	**319.2**

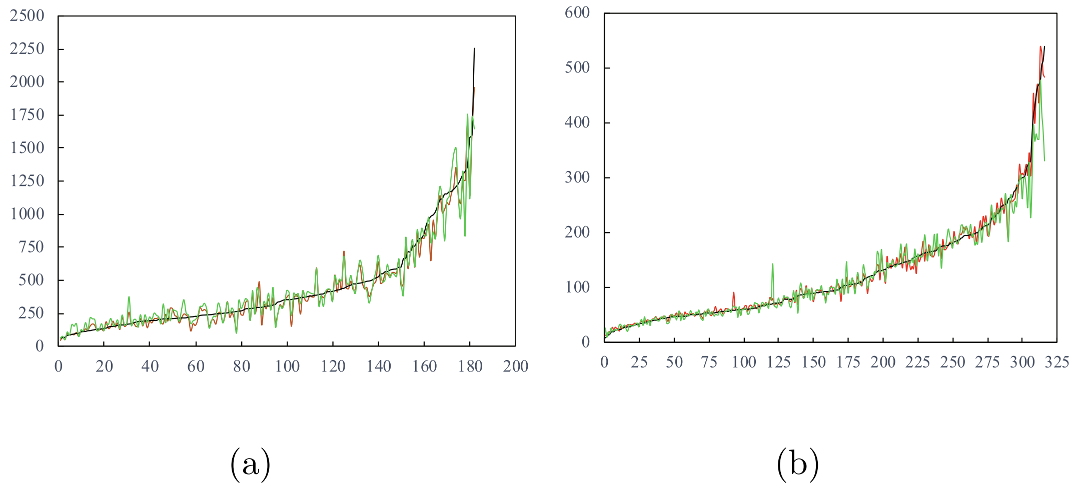


Fig. 11. The number of persons in every image in the ShanghaiTech dataset. (a) Group A; (b) Group B. The black curve indicates the ground truth; the red curve indicates the results using the attention module; the green curve indicates the results without the attention module.

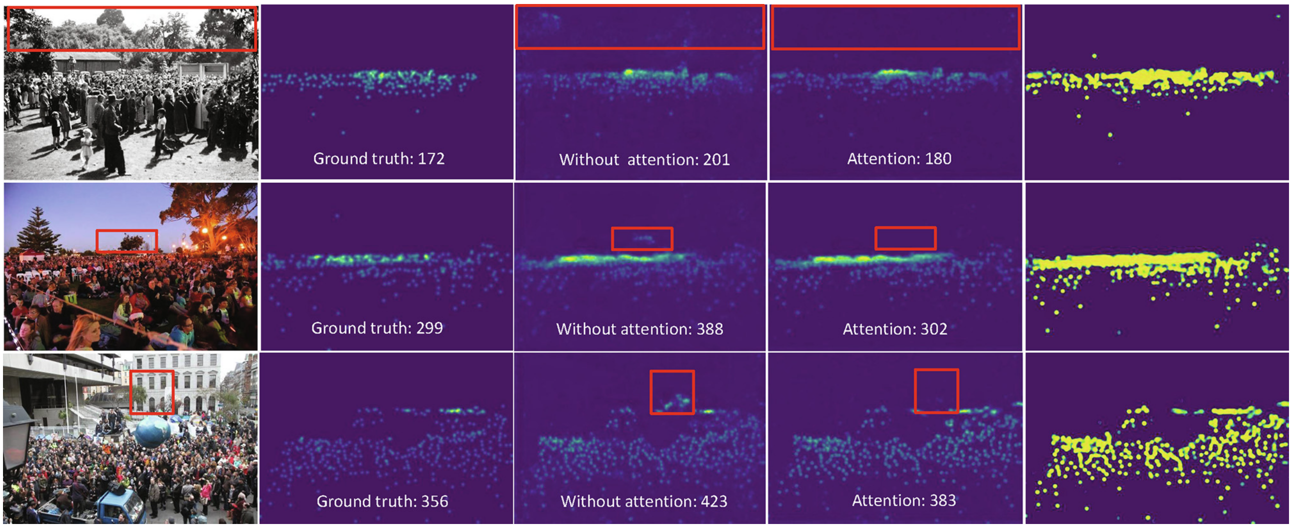


Fig. 12. Visualization. Column-1: original images; Column-2: ground truth; Column-3; estimation results without the attention network; Column-4: estimation results with the help of attention network. Red bounding boxes indicate background regions; Column-5: attention maps.

Table 3
Comparisons of different approaches on the ShanghaiTech dataset.

Method	Group A		Group B	
	MAE	MSE	MAE	MSE
MCNN [10]	110.2	173.2	26.4	41.3
Cascaded-MTL[13]	101.3	152.4	20.0	31.1
SwitchCNN [18]	90.4	135.0	21.6	33.4
CP-CNN [45]	73.6	106.4	20.1	30.1
IC–CNN [14]	68.5	116.2	10.7	16.0
CSRNet [19]	68.2	115.0	10.6	16.0
SANet [20]	67.0	104.5	8.47	13.6
Perspective-CNN [21]	**57.0**	**86.0**	8.8	13.7
RANet [46]	59.4	102.0	7.9	12.9
AMS-Net	63.8	108.5	**7.3**	**11.8**

Table 4
Comparisons of different approaches on the UCF_CC_50 dataset.

Method	MAE	MSE
MCNN [10]	377.6	509.1
Cascaded-MTL [13]	322.8	397.9
Switchin-CNN [18]	318.1	439.2
CP-CNN [45]	295.8	320.9
IC–CNN [14]	260.9	365.5
CSRNet [19]	266.1	397.5
SANet [20]	258.4	334.9
Perspective-CNN [21]	244.6	361.2
RANet [46]	239.8	319.4
AMS-Net	**236.5**	**319.2**

Table 5
Evaluation on other large-scale benchmark datasets, namely the UCF-QNRF, the JHU Crowd++, and the NWPU Crowd datasets.

Methods	UCF-QNRF		JHU Crowd++		NWPU Crowd	
	MAE	MSE	MAE	MSE	MAE	MSE
MCNN [10]	277	426	160.6	377.7	218.5	700.6
Cascaded-MTL [13]	252	514	138.1	379.5	None	None
CSRNet [19]	None	None	72.2	249.9	104.8	433.4
SANet [20]	None	None	82.1	272.6	171.1	471.51
SFCN [26]	102.0	171.4	62.9	247.5	95.4	608.3
CAN [31]	107	183	89.5	239.3	93.5	489.9
BL [32]	88.7	**154.8**	**59.3**	**229.2**	93.6	470.3
AMS-Net	**86.5**	167.2	61.3	236.1	**91.2**	**425.5**

Table 6
Evaluation on different channel-wise weighted operations on the ShanghaiTech dataset.

Weighted operations	Group A		Group B	
	MAE	MSE	MAE	MSE
without channel-wise weighted operation	65.7	114.3	8.5	16.6
squeeze-and-excitation	63.3	106.5	**7.4**	**13.6**
global average pooling	**63.8**	**108.5**	7.3	11.8

Table 7
Evaluation on the number of kernels on the ShanghaiTech dataset (Group A).

Kernel ID	ShanghaiTech Group A	
	MAE	MSE
1	66.3	112.4
1 and 2	64.3	110.2
1, 2, and 3	64.0	109.0
1, 2, 3, and 4	**63.8**	**108.5**

Table 8
Characteristics of the typical scenarios selected for demonstration.

Scene	Image resolution	Max-scale	Min-scale	Average-scale
scene-1	1024×684	30×30	5×5	17×17
scene-2	1024×680	90×90	7×7	37×37
scene-3	1024×768	85×85	8×8	40×40
scene-4	896×600	12×12	7×7	9×9

background areas in the red bounding boxes are mistaken into account, thus leading to the decrease in the counting accuracy.

Next, we compare our approach with other state-of-the-art methods, and the quantitative results are shown in Table 3 (on the ShanghaiTech dataset), Table 4 (on the UCF_CC_50 dataset), and Table 5 (other newly released large-scale datasets). From the experimental results we can find that: (1) on the ShanghaiTech and the UCF_CC_50 datasets, the proposed method achieves very promising performances. Only on Group A of the ShanghaiTech dataset, the method in [21] is a better option; (2) on the UCF-

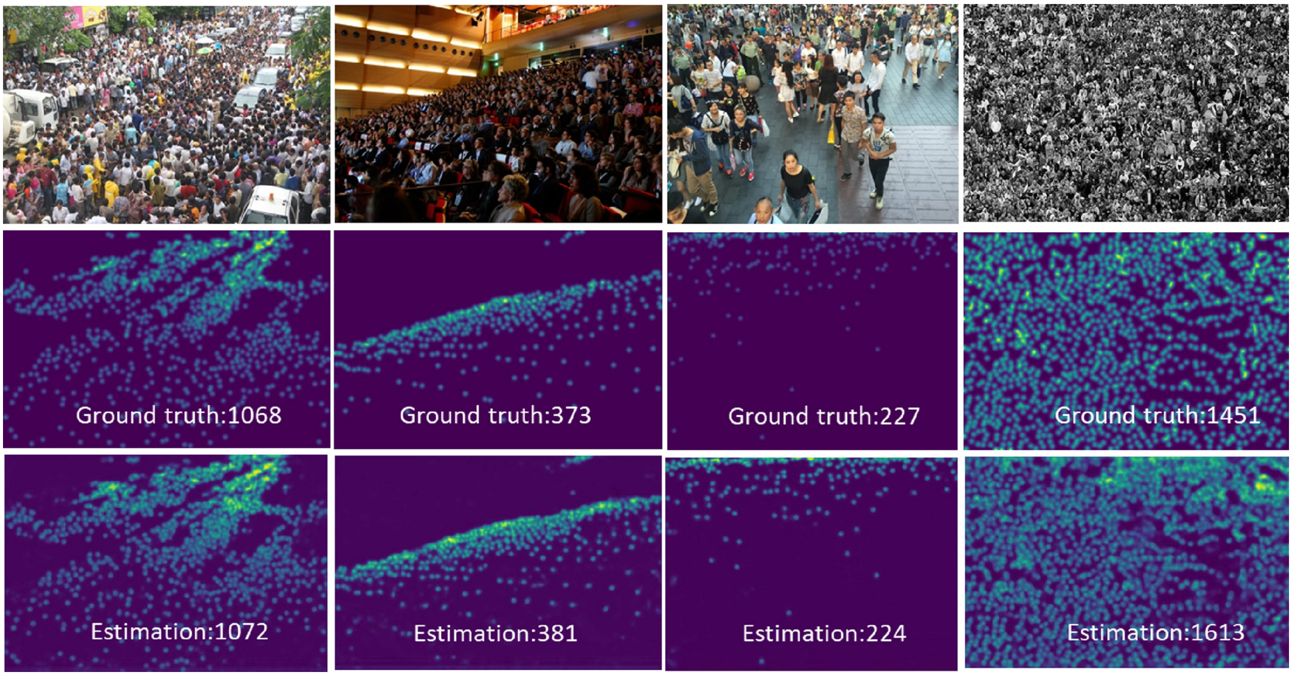


Fig. 13. Visualization: typical scenarios selected for demonstration with varied image resolutions, backgrounds, non-uniform density distribution, and scale changes. From the 1st column to the 4th column, we illustrate scene-1 to scene-4.

Table 9
The detailed structure of the attention network.

Structure	Current layer	Type	Output size	Previous layer
Conv1	Conv1_1	Conv2d	H*W*64	Input
	Conv1_2	Conv2d	H*W*64	Conv1_1
Conv2	Conv2_1	Maxpool+Conv2d	(H/2)*(W/2)*128	Conv1_2
	Conv2_2	Conv2d	(H/2)*(W/2)*128	Conv2_1
Conv3	Conv3_1	Maxpool+Conv2d	(H/4)*(W/4)*256	Conv2_2
	Conv3_2	Conv2d	(H/4)*(W/4)*256	Conv3_1
	Conv3_3	Conv2d	(H/4)*(W/4)*256	Conv3_2
Conv4	Conv4_1	Maxpool+Conv2d	(H/8)*(W/8)*512	Conv3_3
	Conv4_2	Conv2d	(H/8)*(W/8)*512	Conv4_1
	Conv4_3	Conv2d	(H/8)*(W/8)*512	Conv4_2
Conv5	Conv5_1	Multi-scale fusion	(H/8)*(W/8)*256	Conv4_3
Upsmaple1	Upsmaple1	Upsample	(H/4)*(W/4)*256	Conv5_1
Conv6	Conv6_1	Multi-scale fusion	(H/4)*(W/4)*128	Upsmaple1
Upsmaple2	Upsmaple2	Upsample	(H/2)*(W/2)*128	Conv6_1
Conv7	Conv7_1	Multi-scale fusion	(H/2)*(W/2)*2	Upsmaple2
Avgpool	Avgpool	GlobalAveragePool	2	Conv7_1
Output	Output	Pixel-wise product	(H/2)*(W/2)*1	Conv7_1,Avgpool

QNRF, the JHU-Crowd++, and the NWPU Crowd datasets, our approach can still achieve competitive performances.

Finally, we evaluate the performances of different channel-wise weighted operations. Experiments are carried out on the ShanghaiTech dataset, and the results are shown in Table 6, where we compare the performances of squeeze-and-excitation, global average pooling, and without the channel-wise weighted operation. From the results we can find that: (1) channel-wise weighted operations can promote the counting performances; (2) using the standard squeeze-and-excitation operation to weight F_a and F_b can achieve slightly better counting accuracy in the dense environments (Group A), but a little bit worse in the relative sparse scenarios (Group B); (3) Considering the complexity and computational cost, global average pooling (GAP) is still a competitive channel-wise weighted strategy in realistic applications.

4.3.2. Evaluation on multi-scale crowds

In this section, we evaluate the importance of the multi-scale feature fusion model. To this end, we first compare the performances using varied number of kernels. We take the ShanghaiTech dataset (Group A) for demonstration, and the corresponding experimental results are demonstrated in Table 7. The kernel ID is in line with Fig. 3, where ID: 1 to 4 correspond to the dilation rate r = 1, 3, 6, 9. It can be seen clearly that when increasing the number of kernels, the counting accuracy improves consistently.

Next, we validate that the proposed approach is able to handle persons with varied scales. We select several typical scenarios with varied crowd density, apparent scale changes, and different backgrounds, where the characteristics of these scenarios are presented in Table 8. The qualitative results are demonstrated in Fig. 13.

Table 10
The detailed structure of the crowd density network.

Structure	Layer	Type	Output size	Previous layer
Conv1	Conv1_1	Conv2d	H*W*64	Input
	Conv1_2	Conv2d	H*W*64	Conv1_1
Conv2	Conv2_1	Maxpool+Conv2d	(H/2)*(W/2)*128	Conv1_2
	Conv2_2	Conv2d	(H/2)*(W/2)*128	Conv2_1
Conv3	Conv3_1	Maxpool+Conv2d	(H/4)*(W/4)*256	Conv2_2
	Conv3_2	Conv2d	(H/4)*(W/4)*256	Conv3_1
	Conv3_3	Conv2d	(H/4)*(W/4)*256	Conv3_2
Conv4	Conv4_1	Maxpool+Conv2d	(H/8)*(W/8)*512	Conv3_3
	Conv4_2	Conv2d	(H/8)*(W/8)*512	Conv4_1
	Conv4_3	Conv2d	(H/8)*(W/8)*512	Conv4_2
Conv5	Conv5_1	Maxpool+Conv2d	(H/16)*(W/16)*512	Conv4_3
	Conv5_2	Conv2d	(H/16)*(W/16)*512	Conv5_1
	Conv5_3	Conv2d	(H/16)*(W/16)*512	Conv5_2
Upsmaple1	Upsmaple1	Upsample	(H/8)*(W/8)*512	Conv5_3
Output1	Output1+Concatenate1	Concatenate_1	(H/8)*(W/8)*1	Upsample1,
			(H/8)*(W/8)*256	Conv4_3
Upsmaple2	Upsmaple2	Upsample	(H/4)*(W/4)*256	Concatenate1
Output2	Output2+Concatenate2	Concatenate_1	(H/4)*(W/4)*1	Upsample2,
			(H/4)*(W/4)*128	Conv3_3
Upsmaple3	Upsmaple3	Upsample	(H/2)*(W/2)*128	Concatenate2
Output3	Output3	Concatenate_2	(H/2)*(W/2)*1	Upsample2,Conv3_3

5. Conclusions

In this paper, we propose an attention-guided framework for crowd counting in realistic environments, which is mainly comprised by an attention network and a density network, respectively. The attention network is used to detect regions where human crowds are more likely to locate, which can alleviate the impacts that background clutters exert on the counting performances. In order to perceive people of varied resolutions, we further embed a multi-scale fusion module into the attention network, which is built upon dilated convolution. The output of the attention network will serve as the mask of the crowd density map, which is obtained through the density network. We conduct extensive experiments on several popular benchmark datasets, demonstrating the effectiveness of the proposed framework. From the results we can conclude: (1) the attention network can improve the counting accuracy significantly; (2) channel-wise weighted operations can further promote the counting performances. As for the future work, we would like to embed the most recent multi-scale fusion strategies into our framework, and provide comprehensive evaluations.

CRediT authorship contribution statement

Bo Zhang: Conceptualization, Data curation, Writing - original draft. **Naiyao Wang:** Conceptualization, Methodology, Software. **Zheng Zhao:** Methodology, Software, Validation. **Ajith Abraham:** Writing - review & editing. **Hongbo Liu:** Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to thank Houtao Zhao for his scientific collaboration in this research. This work is supported by the National Natural Science Foundation of China (Grant No. 61702073, 61772102), the China Postdoctoral Science Foundation (Grant No. 2019M661079), and the Liaoning Collaborative Fund (Grant No. 2020-HYLH-17).

Appendix A

In the appendix section, we present the detailed structures of the crowd attention network (as shown in Fig. 4) and the crowd density network (as shown in Fig. 6) in Tables 9 and 10, respectively.

References

[1] J. Zhong, W. Cai, L. Luo, H. Yin, Learning behavior patterns from video: A data-driven framework for agent-based crowd modeling, in: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems, 2015, pp. 801–809.

[2] H. Wang, J. Ondřej, C. O'Sullivan, Path patterns: Analyzing and comparing real and simulated crowds, in: Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games, 2016, pp. 49–57.

[3] B. Zhou, X. Tang, X. Wang, Learning collective crowd behaviors with dynamic pedestrian-agents, International Journal of Computer Vision 111 (1) (2015) 50–68.

[4] I.A. Lawal, F. Poiesi, D. Anguita, A. Cavallaro, Support vector motion clustering, IEEE Transactions on Circuits and Systems for Video Technology 27 (11) (2016) 2395–2408.

[5] T. Hassner, Y. Itcher, O. Kliper-Gross, Violent flows: Real-time detection of violent crowd behavior, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 1–6.

[6] X. Wu, R. He, Z. Sun, T. Tan, A light CNN for deep face representation with noisy labels, IEEE Transactions on Information Forensics and Security 13 (11) (2018) 2884–2896.

[7] R. He, W. Zheng, B. Hu, Maximum correntropy criterion for robust face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (8) (2010) 1561–1576.

[8] J. Shao, C.C. Loy, X. Wang, Learning scene-independent group descriptors for crowd understanding, IEEE Transactions on Circuits and Systems for Video Technology 27 (6) (2016) 1290–1303.

[9] Q. Wang, M. Chen, F. Nie, X. Li, Detecting coherent groups in crowd scenes by multiview clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (1) (2018) 46–58.

[10] 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 International Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 589–597.

[11] L. Boominathan, S.S. Kruthiventi, R.V. Babu, CrowdNet: A deep convolutional network for dense crowd counting, in: Proceedings of the International Conference on Multimedia, 2016, pp. 640–644.

[12] C. Shang, H. Ai, B. Bai, End-to-end crowd counting via joint learning local and global count, in: Proceedings of the IEEE International Conference on Image Processing, IEEE, 2016, pp. 1215–1219.

[13] V.A. Sindagi, V.M. Patel, CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting, in: Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance, IEEE, 2017, pp. 1–6.

[14] V. Ranjan, H. Le, M. Hoai, Iterative crowd counting, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 270–285.

[15] Y. Li, Y. Chen, N. Wang, Z. Zhang, Scale-aware trident networks for object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6054–6063.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[17] S. Gao, M.M. Cheng, K. Zhao, X.Y. Zhang, P.H.S. Torr, Res2Net: A new multi-scale backbone architecture, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019), https://doi.org/10.1109/TPAMI.2019.2938758.

[18] D.B. Sam, S. Surya, R.V. Babu, Switching convolutional neural network for crowd counting, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 4031–4039.

[19] Y. Li, X. Zhang, D. Chen, CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 1091–1100.

[20] X. Cao, Z. Wang, Y. Zhao, F. Su, Scale aggregation network for accurate and efficient crowd counting, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 734–750.

[21] Z. Yan, Y. Yuan, W. Zuo, X. Tan, Y. Wang, S. Wen, E. Ding, Perspective-guided convolution networks for crowd counting, in: Proceedings of the IEEE International Conference on Computer Vision, IEEE, 2019, pp. 952–961.

[22] Y. Yang, G. Li, Z. Wu, L. Su, N. Sebe, Reverse perspective network for perspective-aware object counting, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2020, pp. 4373–4382.

[23] S. Bai, Z. He, Y. Qiao, H. Hu, J. Yan, Adaptive dilated network with self-correction supervision for counting, in: Proceedings of the IEEE International Conference on Computer Vision, 2020, pp. 4593–4602.

[24] Q. Wang, J. Gao, W. Lin, X. Li, NWPU-Crowd: A large-scale benchmark for crowd counting and localization, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020), https://doi.org/10.1109/TPAMI.2020.3013269.

[25] Q. Wang, J. Gao, W. Lin, Y. Yuan, Pixel-wise crowd understanding via synthetic data, International Journal of Computer Vision (2020) 1–21.

[26] Q. Wang, J. Gao, W. Lin, Y. Yuan, Learning from synthetic data for crowd counting in the wild, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2019, pp. 8198–8207.

[27] J. Chen, W. Su, Z. Wang, Crowd counting with crowd attention convolutional neural network, Neurocomputing 382 (2020) 210–220.

[28] L. Zhu, C. Li, B. Wang, K. Yuan, Z. Yang, Dcgsa: A global self-attention network with dilated convolution for crowd density map generating, Neurocomputing 378 (2020) 455–466.

[29] Y. Zhang, C. Zhou, F. Chang, A.C. Kot, Multi-resolution attention convolutional neural network for crowd counting, Neurocomputing 329 (2019) 144–152.

[30] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Muller, R. Manmatha, M. Li, A. Smola, ResNeSt: Split-attention networks, arXiv preprint arXiv:2004.08955 (2020)..

[31] W. Liu, M. Salzmann, P. Fua, Context-aware crowd counting, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2019, pp. 5099–5108.

[32] Z. Ma, X. Wei, X. Hong, Y. Gong, Bayesian loss for crowd count estimation with point supervision, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6142–6151.

[33] X. Ding, F. He, Z. Lin, Y. Wang, H. Guo, Y. Huang, Crowd density estimation using fusion of multi-layer features, IEEE Transactions on Intelligent Transportation Systems (2020), https://doi.org/10.1109/TITS.2020.2983475.

[34] Y. Fang, S. Gao, J. Li, W. Luo, L. He, B. Hu, Multi-level feature fusion based locality-constrained spatial transformer network for video crowd counting, Neurocomputing 392 (2020) 98–107.

[35] S. Vishwanath, P. Vishal, Multi-level bottom-top and top-bottom feature fusion for crowd counting, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1002–1012.

[36] V.A. Sindagi, V.M. Patel, A survey of recent advances in cnn-based single image crowd counting and density estimation, Pattern Recognition Letter (2018) 3–16.

[37] D. Kang, Z. Ma, A.B. Chan, Beyond counting: Comparisons of density maps for crowd analysis tasks-counting, detection, and tracking, IEEE Transactions on Circuits and Systems for Video Technology (2019) 1408–1422.

[38] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, arXiv preprint arXiv:1511.07122 (2015)..

[39] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014)..

[40] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[41] 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.

[42] H. Idrees, I. Saleemi, C. Seibert, M. Shah, Multi-source multi-scale counting in extremely dense crowd images, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2547–2554.

[43] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, M. Shah, Composition loss for counting, density mapestimation and localization in dense crowds, in: Proceedings of the IEEE International Conference on European Conference on Computer Vision, 2018, pp. 8–14.

[44] V.A. Sindagi, R. Yasarla, V.M. Patel, JHU-CROWD++: Large-scale crowd counting dataset and a benchmark method, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020), https://doi.org/10.1109/TPAMI.2020.3035969.

[45] 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, IEEE, 2017, pp. 1861–1870.

[46] A. Zhang, J. Shen, Z. Xiao, F. Zhu, X. Zhen, X. Cao, L. Shao, Relational attention network for crowd counting, in: Proceedings of the IEEE International Conference on Computer Vision IEEE, 2019, pp. 6788–6797.

Bo Zhang received the BSc degree in computer science and technology in 2007 and the MSc degree in computer application technology in 2010 from Jilin University, China. He received the PhD degree in telecommunications in 2015 from the University of Trento, Italy. He is currently an assistant professor in Dalian Maritime University, China. His research interests include computer vision, multimedia signal processing, and machine learning.

Naiyao Wang is currently a Ph.D. candidate in the School of Information Science and Technology, Dalian Maritime University, China. His research computer vision, self-supervised learning, and machine learning.

Zheng Zhao received his B.S. degree from Dalian University of Technology in 2010. He received his M.S. and Ph.D degrees from Zhengzhou Science and Technology Institute in 2013 and 2017. His research interests include next generation Internet and deep learing.

Ajith Abraham received Ph.D. degree in Computer Science from Monash University, Melbourne, Australia (2001) and a Master of Science Degree from Nanyang Technological University, Singapore (1998). Ajith's research and development experience includes nearly 30 years in the Academia and Industry. He works in a multidisciplinary environment involving machine (network) intelligence, cyber security, sensor networks, Web intelligence, scheduling, data mining and applied to various real world problems. He is an author/co-author of 1,200+ publications and some of the works have also won best paper awards at International conferences and also received several citations. Since 2008, Dr. Abraham is the Chair of IEEE Systems Man and Cybernetics Society Technical Committee on Soft Computing (which has over 200 + members) and served as a Distinguished Lecturer of IEEE Computer Society representing Europe (2011–2013). Currently Dr. Abraham is the editor-in-chief of Engineering Applications of Artificial Intelligence (EAAI) and serves/served the editorial board of over 15 International Journals indexed by Thomson ISI.

Hongbo Liu received his three level educations (B. Sc., M. Sc., Ph.D.) at the Dalian University of Technology, China. He is with the School of Information Science and Technology, Dalian Maritime University. Professor Liu's research interests are in cognitive computing, machine learning, big data, etc. He participates and organizes actively international conference and workshop and international journals/publications. He received the New Century Excellent Talents Award in 2010.