Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

A novel convolutional neural network method for crowd counting*

Jie-hao HUANG†, Xiao-guang DI†‡, Jun-de WU, Ai-yue CHEN

Control and Simulation Center, Harbin Institute of Technology, Harbin 150080, China

†E-mail: 18s004055@hit.edu.cn; dixiaoguang@hit.edu.cn

Received June 5, 2019; Revision accepted Feb. 6, 2020; Crosschecked June 11, 2020

Abstract: Crowd density estimation, in general, is a challenging task due to the large variation of head sizes in the crowds. Existing methods always use a multi-column convolutional neural network (MCNN) to adapt to this variation, which results in an average effect in areas with different densities and brings a lot of noise to the density map. To address this problem, we propose a new method called the segmentation-aware prior network (SAPNet), which generates a high-quality density map without noise based on a coarse head-segmentation map. SAPNet is composed of two networks, i.e., a foreground-segmentation convolutional neural network (FS-CNN) as the front end and a crowd-regression convolutional neural network (CR-CNN) as the back end. With only the single dot annotation, we generate the ground truth of segmentation masks in heads. Then, based on the ground truth, FS-CNN outputs a coarse head-segmentation map, which helps eliminate the noise in regions without people in the density map. By inputting the head-segmentation map generated by the front end, CR-CNN performs accurate crowd counting estimation and generates a high-quality density map. We demonstrate SAPNet on four datasets (i.e., ShanghaiTech, UCF-CC-50, WorldExpo'10, and UCSD), and show the state-of-the-art performances on ShanghaiTech part *B* and UCF-CC-50 datasets.

Key words: Crowd counting; Density estimation; Segmentation prior map; Uniform function
https://doi.org/10.1631/FITEE.1900282 **CLC number:** TP391.4

1 Introduction

A dense crowd scene is common in cities with rapidly developing economy and growing population. Generally, there is a high risk of unexpected accidents, such as stampede, in highly dense crowd scenes.

Crowd analysis is crucial in crowd flow control, safety management, anomaly monitoring, and other safety services. However, crowd analysis usually faces many challenges, such as low-resolution crowd images, severe occlusion between persons, perspective distortions, and unfixed camera viewpoints (Zhan et al., 2008; Li T et al., 2015), which make many crowd counting methods fail, like detection-based methods (Dai et al., 2016) and regression-based methods (Chan et al., 2008; Idrees et al., 2013). Currently, convolutional neural network (CNN) based methods are widely used in many fields of computer vision, including crowd counting (Zhang C et al., 2015; Zhang YY et al., 2016; Sam et al., 2017). These CNN-based methods learn to map the crowd image to its density map, i.e., crowd count per unit area (Lempitsky and Zisserman, 2010). The density map provides us not only the crowd counting but also spatial information like crowd distribution. However, it is not easy to obtain a high-quality density map due to the large variation of the head/body

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 61775048) and the Fundamental Research Funds for the Central Universities, China (No. ZDXMPY20180103)

ORCID: Jie-hao HUANG, https://orcid.org/0000-0003-1412-1324; Xiao-guang DI, https://orcid.org/0000-0002-5709-6862

sizes (huge diversity of densities) in a dense crowd scene. Various kinds of approaches have been proposed to enable CNN to learn features with different scales. For instance, the multi-column convolutional neural network (MCNN) proposed by Zhang YY et al. (2016) and contextual pyramid CNN (CP-CNN) proposed by Sindagi and Patel (2017) adopt multiple adaptive columns, aiming at solving the problem of the variation of head sizes. Furthermore, incrementally growing CNN (IG-CNN) proposed by Sam et al. (2018) forms a regressor tree to identify crowds with different density levels. The advantage of the above methods is that they can perform well in huge diversity of densities by training the network with corresponding density-level datasets (Liu et al., 2018; Sam et al., 2018). They are trained to handle different density levels, but they predict the density level of the whole image that has a large range of crowd density. Therefore, they always present complementary effect in predicting a density map (Liu et al., 2018). The complementary effect brings a lot of noise to regions without person, which reduces the quality of the density map and gives unreliable crowd counting because the crowd counting is obtained by integrating the whole density map, including the noisy parts. Moreover, a complex training procedure and a large parameter number require more training time. Hence, an improved method is required to generate a high-quality density map.

Lempitsky and Zisserman (2010) presented a density-based method for counting tasks, which shows that a two-dimensional (2D) point can be blurred into an isotropic covariance matrix through the Gaussian function. The Gaussian point in the matrix indicates the percentage of a person presenting in that pixel. To give the crowd counting, the integration value of the Gaussian matrix should be equal to one, indicating that each Gaussian point in the matrix is a small value (1e-4 to 1e-2). However, if CNN uses the activation function ReLU (i.e., rectified linear unit), which is a commonly used activation function, such a small-value density map presents a fitting problem. Since the small value (1e-4 to 1e-2) is nearly 0, it is possible to make the output of the activation function less than 0 by backpropagation. Then, the gradient cannot pass through the dying ReLU (Li HH et al., 2018) in this case. Therefore, it is hard for the network to converge to the optimal result. Hence, we should find a new function to generate a large-value map like a segmentation map.

To eliminate the noise caused by the network misdetecting the backgrounds (such as trees, buildings, and fences) as the dense crowd, a universal method is required to generate the foreground segmentation map.

The foreground segmentation method enables the network to focus on the crowded areas without being affected by the noisy background. It is easier to generate the density map from the crowd area than from the whole image. However, the crowd segmentation task usually requires more cues, such as spatial and temporal information. Therefore, we generate a Canny-edge map as texture information from a raw image using a Canny-edge operator. The Canny-edge map can reflect the difference between the foreground and background. Besides, the ground truth of the pixel-wise crowd segmentation is indispensable in the segmentation task, whereas labeling the pixel-wise segmentation ground truth is labor-intensive and takes much time. Although we do not have a pixel-wise segmentation map, we can generate a coarse head-segmentation map based on dot annotations of individuals. The motivation of our work is based on the observation that when two persons are close enough, the distance between two heads indeed reflects their head sizes. So, the masks of all heads in the segmentation map (Fig. 1) could be obtained from the dot annotations and the estimated head diameters.

Taking the above factors into consideration, we use a segmentation map to eliminate the noise in the background. Provided with only a single frame, we use a Canny-edge map to help generate the segmentation map. Also, coarse segmentation ground truth is generated with the cheap single dot annotations.

In summary, we address the problem of density map in the crowd counting task with the following contributions:

1. We design two CNN structures aiming at different targets, i.e., a foreground-segmentation network for coarsely segmenting the heads, which can effectively wipe out the misdetected noise in the background, and a crowd-regression network for differentiating head sizes and generating a high-quality density map.

2. Two networks take the prior maps as the input, i.e., a Canny-edge map for FS-CNN and a coarse

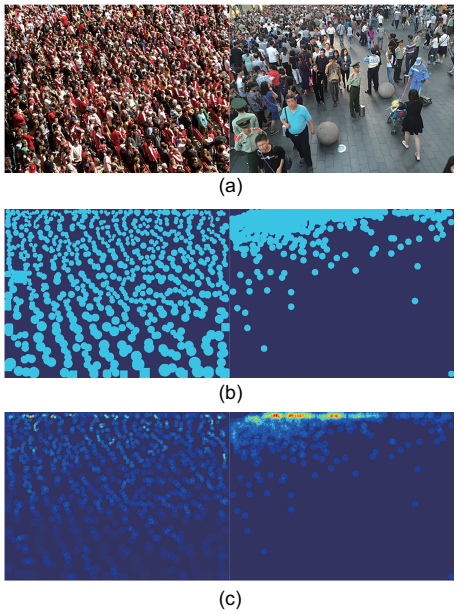

Fig. 1 Raw images (a), location-level head-segmentation ground truth map (b), and uniformly distributed ground truth density map (c)

In the left sides of (a)–(c), the geometry-adaptive method is used to differentiate head sizes. In the right sides of (a)–(c), a fixed kernel is used to adapt to the average size of heads

head-segmentation map for CR-CNN, which helps quickly recognize effective features and reduce the training complexity.

3. A novel uniform function is proposed to generate a head-segmentation map and a uniformly distributed density map with cheap single dot annotations. The ground truth of the head-segmentation map is generated for FS-CNN, whereas the uniformly distributed density map is used for counting regression in CR-CNN.

2 Related works

2.1 Crowd counting

In general, we divide the crowd counting algorithms into two main groups (Li T et al., 2015; Sindagi and Patel, 2018), i.e., traditional methods and CNN-based methods. The traditional methods include the detection-based method and regression-based method. The detection-based method (Dollar et al., 2012) uses a sliding window to detect individuals based on common features such as histogram of oriented gradient (HOG) and Haar wavelets. This method requires the target be full-body or part-body appearance with enough pixels (high resolution). In reality, however, the severe occlusion in a highly dense crowd scene causes pedestrians in the image to have few pixels. Since detection counting cannot be applied in dense crowd scenarios due to the above problem, regression-based methods (Chan and Vasconcelos, 2009; Idrees et al., 2013) were proposed. For example, Idrees et al. (2013) built a model using scale-invariant feature transform (SIFT), frequency-domain analysis, and Markov random field to estimate the crowd count.

However, regression-based methods output only the number of people but do not generate a crowd-distribution map. Since the crowd-distribution map is crucial, Lempitsky and Zisserman (2010) cast the counting task as learning a density map, which achieves huge improvements. It is also suitable for a CNN-based counting method. Thus, various CNN-based approaches (Zhang YY et al., 2016; Sam et al., 2017; Sindagi and Patel, 2017) were proposed to deal with counting problems in the dense crowd scene. Most of these approaches focus on the problem of large diversity of crowd density (different head sizes). For example, Zhang YY et al. (2016) proposed an MCNN model that uses filters of different sizes in different columns to adapt to the variation of head sizes. As the input image is sent into three columns and the density level of the input image is unknown, Sam et al. (2017) presented an improved MCNN called Switching-CNN, which uses a switch classifier to predict the density level of the image and deliver the image to the corresponding optimal CNN column. Sam et al. (2018) proposed a similar idea, i.e., IG-CNN, which builds a CNN tree. The leaf regressors in the tree are more sensitive to the crowd of different density levels. Although this kind of network achieves better results, it is more complex in both the CNN structure and training process. Moreover, it uses a large parameter number and it is hard to be applied to the real system. The density maps generated by the above methods have a severe problem; i.e., there is noise in regions without people due to the average effect. Therefore, the crowd count obtained from such a density map is unreliable because the noise in the background is added to the crowd count estimation. Considering all the above shortcomings, we propose a simple and easy-to-train model, which can output the head-segmentation map and help reduce the noise pixels to generate a high-quality density map.

2.2 Crowd segmentation

Crowd segmentation, in general, is a prerequisite of many crowd analysis tasks. If the crowd segmentation map (i.e., region of interest) is supplied, crowd analysis focuses on the crowd itself and ignores noisy backgrounds such as tree and fence. However, crowd segmentation is a complex and challenging task, and needs much spatial and temporal information (Long et al., 2015; Li JJ et al., 2016). For example, it requires successive frames of a stationary camera to provide collective motion information (Kang and Wang, 2014). Nevertheless, more appearance cues should be provided because it suffers from many stationary individuals, which could not be captured based on motion information.

Although the ground truth of pixel-wise segmentation is crucial for segmentation tasks to obtain an accurate result, the cost of obtaining the pixel-wise ground truth could be high. Therefore, the performance of the crowd segmentation method is restricted to the information provided and the ground truth of pixel-wise segmentation.

3 The proposed method

3.1 Uniform function for ground truth generation

As we mentioned in Section 2, most CNN-based methods adopt the Gaussian function to generate the ground truth of the density map (Zhang YY et al., 2016). However, there are many problems to be resolved in the Gaussian-distributed density map. It is difficult to fit the Gaussian density map with a numerical range between 1e-4 and 1e-2 because the value is nearly 0. Then, the output of the convolution layer possibly becomes smaller than 0 by backpropagation and makes a dying ReLU. In this study, we propose a variant of the Gaussian kernel, i.e., a uniform kernel, to generate the head-segmentation map for foreground segmentation and the uniformly distributed density map for crowd counting regression.

Two maps can be defined by $F(\boldsymbol{x})$ as follows:

$$F(\boldsymbol{x})=\sum_{i=0}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_i)U_{\alpha_i,\gamma}(\boldsymbol{x}) \text{ with } \alpha_i=\beta\overline{e_i}. \quad (1)$$

If there is a head annotation at pixel $\boldsymbol{x}_i$, it is described as a delta function, i.e., $\delta(\boldsymbol{x}-\boldsymbol{x}_i)$. The map is generated by convolving this function with a uniform kernel $U_{\alpha_i,\gamma}$, which has two parameters: average value γ and mask size α_i. The mask size α_i for the head-segmentation map is determined by the geometry-adaptive method $\overline{e_i}=\frac{1}{m}\sum_{j=0}^{m}e_j$ to adapt to different head sizes, shown in the left side of Fig. 1. β is the empirical parameter. The average value γ is set to 1 in the head mask to generate the coarse head segmentation (1 for head and 0 for background). In this way, using only the cheap dot annotations, a coarse head-segmentation map is generated as the ground truth of the foreground segmentation task.

For uniformly distributed density maps, the method of determining the mask size α_i is the same as the method mentioned above. The parameter γ depends on the mask size α_i to normalize the distribution to 1. Since the uniformly distributed density is used for crowd regression, we call it a crowd-regression density map. The head-segmentation map is similar to the crowd-regression density map apart from the numerical difference. Therefore, if the network converges from the segmentation map to the uniformly distributed density map, it needs only to adjust the 1 value to a smaller value in each mask according to the head sizes. It is easier to regress the crowd density map from the head-segmentation map than from the raw features.

Therefore, a novel uniform function is proposed to generate two maps, i.e., the head-segmentation map and crowd-regression density map. Two structures, i.e., the foreground-segmentation network and crowd-regression network, are designed to deal with the segmentation task and regression task, respectively.

3.2 Foreground-segmentation CNN

CNN-based methods have achieved great success in foreground segmentation tasks (Long et al., 2015). Although foreground segmentation is an effective way for crowd density estimation, it is a challenging task that needs more cues (Kang and Wang, 2014) such as motion and appearance.

Since our work is provided with a single frame (i.e., appearance cues only), we combine more information to obtain a better segmentation result. Following the ideas of Kang and Wang (2014) and Li JJ et al. (2016), we use a Canny-edge detector to process the crowd images (Fig. 2). We find that in

a dense crowd scene, the Canny-edge map can provide information about what kind of structures are crowd-like and background-like.

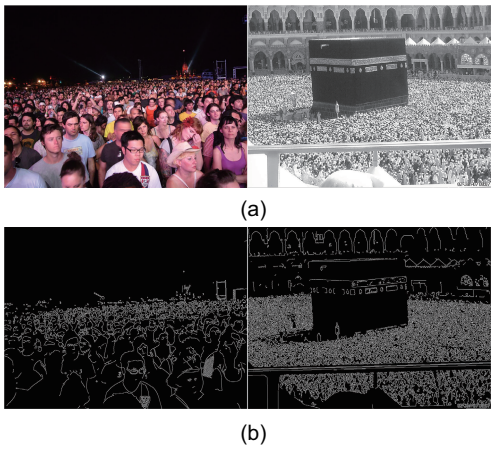

Fig. 2 Original images (a) and the edge prior maps processed by the Canny-edge detector (b)

The Canny-edge map adds prior information about the difference between the foreground and background into the network.

Furthermore, foreground segmentation always requires a pixel- or region-level ground truth, whereas we have only the coarse head-segmentation map (head masks) shown in Fig. 1. We name the head-segmentation map the location-level ground truth since it provides only the coarse location and does not cover the whole region of people with polygons (Kang and Wang, 2014).

FS-CNN has two types of inputs, i.e., a Canny-edge map and a raw image. The Canny-edge map provides global information about the difference between the foreground and background, whereas the raw image feeds more low-level features to CNN. Moreover, compared with a raw image, a Canny-edge map is an edge feature, so it cannot be sent to the convolution layers with the same number. Thus, the whole network of FS-CNN carves the first 13 layers from VGG-16 as a feature extractor for processing raw images because of the strong transfer learning ability and flexible architecture of VGG-16. The Canny-edge map is sent to the prior-bone, which has three convolution layers and three max-pooling layers. The outputs of VGG-bone and prior-bone (Table 1) are concatenated. Then, we add four dilated convolution layers with a large receptive field as the encoder network, because the dilated convolution layer significantly improves the accuracy of semantic segmentation. Finally, we add a sigmoid activation function, which predicts the probability of persons, as the output.

Table 1 Structure of SAPNet: FS-CNN and CR-CNN

Structure		
First part*	VGG-bone	(conv3-64)*2+max-pooling (conv3-128)*2+max-pooling (conv3-256)*3+max-pooling (conv3-512)*3
	Prior-bone	Conv3-64+max-pooling Conv3-128+max-pooling Conv3-256+max-pooling
Second part	FS-CNN	(dilate-conv3-64)*2 Dilate-conv1-1
	CR-CNN	(dilate-conv3-64)*4 Dilate-conv1-1

* The first part is the same for FS-CNN and CR-CNN

Taking the i^{th} raw crowd image $\boldsymbol{I}_i$ and the Canny-edge prior map $\boldsymbol{C}_i$ as inputs, the foreground-segmentation CNN D_i^{seg} generates the head-segmentation map $\boldsymbol{S}^{\text{seg}}$ with parameter $\boldsymbol{\Omega}_{\text{seg}}$, expressed as

$$\boldsymbol{S}^{\text{seg}}(\boldsymbol{I}_i|\boldsymbol{\Omega}_{\text{seg}}) = D_i^{\text{seg}}(\boldsymbol{\Omega}_{\text{seg}}, \boldsymbol{I}_i, \boldsymbol{C}_i). \tag{2}$$

3.3 Crowd-regression CNN

In dense crowd counting tasks, density classification is widely used. Some methods such as CP-CNN (Sindagi and Patel, 2017) and Switching-CNN (Sam et al., 2017) can obtain the density classification through training the network in the datasets of different density levels.

Compared with the prior information of crowd density classification, the novel prior information that we proposed, i.e., the head-segmentation map, is low level and more effective. It can be generated by FS-CNN and can provide CR-CNN with much information such as coarse crowd distribution. However, if FS-CNN has poor performance and gives an inaccurate crowd distribution map, it will have negative influence on the performance of CR-CNN. So, it is not a good choice to merge the crowd distribution map into the final convolution layer. Hence, we decide to use a similar structure (i.e., the prior-bone) as the FS-CNN to merge head-segmentation prior information into the end of the VGG-bone. The head-segmentation map from the prior-bone and the crowd feature map from the VGG-bone are concatenated.

Table 1 shows the structure of SAPNet, which

contains mainly two parts. The first part includes two parallel bones, i.e., VGG-bone and prior-bone. Then the outputs from the VGG-bone and prior-bone are concatenated. FS-CNN and CR-CNN have the same first part. The second part contains a concatenate layer and several dilated convolution layers. FS-CNN has three dilated convolution layers, while CR-CNN has five layers. Since the prior head-segmentation map is slightly different from the crowd-regression density map, CR-CNN converges quickly in the training process.

The hyper-parameters of the convolution layer are denoted as the sizes of the convolution kernel and several filters (e.g., conv3-64). The max-pooling layers are conducted over a 2×2-pixel window with stride 2, whereas the dilation rate of the dilated-convolution layer is set as 2.

Taking the i^{th} raw crowd image $\boldsymbol{I}_i$ and distribution of the head-segmentation map $\boldsymbol{S}_i$ as inputs, the CR-CNN D_i^{reg} generates the crowd-regression density map $\boldsymbol{R}^{\text{reg}}$ with parameter $\boldsymbol{\Omega}_{\text{reg}}$:

$$\boldsymbol{R}^{\text{reg}}(\boldsymbol{I}_i|\boldsymbol{\Omega}_{\text{reg}}) = D_i^{\text{reg}}(\boldsymbol{\Omega}_{\text{reg}}, \boldsymbol{I}_i, \boldsymbol{S}_i). \tag{3}$$

4 Training method

In this section, we describe the steps of training our proposed SAPNet method.

4.1 Training procedure

Since the head-segmentation map of FS-CNN is the prerequisite input of CR-CNN, the training process is divided into two parts, i.e., FS-CNN and CR-CNN.

As for the edge prior map in FS-CNN, we choose a Canny-edge detector as the edge-detection operator (Canny, 1986). For the Canny-edge detector, we set the hyper-parameter of the Gaussian filter size as 3×3, which is small enough to detect the texture information of the dense crowd. Two empirical thresholds 150 and 250 are used to determine the potential connection of the neighbor edges. Given an edge prior map and a single image, we train FS-CNN with the head-segmentation ground truth. Then, the loss function of FS-CNN is given as follows:

$$\begin{aligned} L_{\text{seg}}(\boldsymbol{\Omega}_{\text{seg}}) =& \frac{1}{2N}\sum_{i=1}^{N}\left\{\|\boldsymbol{S}_i^{\text{seg}} - \boldsymbol{S}_i^{\text{GT}}\|_2^2\right. \\ &\left.+[1-\text{SSIM}(\boldsymbol{S}_i^{\text{seg}} - \boldsymbol{S}_i^{\text{GT}})]\right\}, \end{aligned} \tag{4}$$

where $\boldsymbol{S}_i^{\text{GT}}$ denotes the head-segmentation ground truth of the i^{th} image, $\boldsymbol{S}_i^{\text{seg}}$ the i^{th} output segmentation map of FS-CNN with training parameter $\boldsymbol{\Omega}_{\text{seg}}$, N the number of training images, and SSIM the structural similarity index.

Then, both FS-CNN and CR-CNN networks are trained using the Adam optimization algorithm until they converge on the validation set.

The loss function of CR-CNN is defined as follows:

$$L_{\text{reg}}(\boldsymbol{\Omega}_{\text{reg}}) = \frac{1}{2N}\sum_{i=1}^{N}\|\boldsymbol{R}_i^{\text{reg}} - \boldsymbol{R}_i^{\text{GT}}\|_2^2. \tag{5}$$

We preform an experiment on the ShanghaiTech (SHT) dataset and obtain two sets of parameters in parts A and B. To improve the training efficiency, parameters in the UCF-CC-50 dataset are trained based on the parameters of part A since they are datasets of dense scenes. Furthermore, parameters in the UCSD and WorldExpo'10 datasets are trained based on the parameters of part B since they are datasets of relatively sparse scenes.

4.2 Ground truth generation

The steps of generating the ground truth in this study follow the method in Zhang YY et al. (2016). The mask size in dense crowd scenes is determined by the geometry-adaptive method. In relatively sparse crowd scenes like part B, a fixed kernel is used to adapt to the average size of heads (Fig. 1). Table 2 shows the empirical parameters β and σ for different datasets.

Table 2 Ground truth generation methods for different datasets

Dataset	Generation method
SHT part A	Geometry-adaptive kernel, $\beta = 3$
UCF-CC-50	Geometry-adaptive kernel, $\beta = 3$
SHT part B	Fixed kernel, $\sigma = 15$
UCSD	Fixed kernel, $\sigma = 3$
WorldExpo'10	Fixed kernel, $\sigma = 8$

To expand the training datasets, we randomly crop eight patches with a quarter size of the original images and divide the original images into four patches without overlapping (Li YH et al., 2018).

5 Experiments

We demonstrate our SAPNet method in four different datasets (Chan et al., 2008; Idrees et al., 2013; Zhang C et al., 2016; Zhang YY et al., 2016). As we mentioned earlier, an inaccurate density map may lead to "good" crowd estimation since the crowd count is obtained by integrating the whole density map. Consequently, the crowd counting, in general, is unreliable. Therefore, we focus on not only crowd counting but also the quality of the density map (especially noise-occurrence regions).

5.1 Evaluation metric

Since the output of FS-CNN is a coarse head-segmentation map, SSIM is used as a metric to evaluate the image similarity.

For CR-CNN, we use the mean absolute error (MAE) and root mean square error (RMSE) to evaluate the performance of our method on crowd estimation, which are defined as follows:

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}|C_i-C_i^{\mathrm{GT}}|, \tag{6}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}|C_i-C_i^{\mathrm{GT}}|^2}, \tag{7}$$

where C_i is the integration of the density map outputted by CR-CNN and C_i^{GT} the actual number of people in the i^{th} crowd image.

5.2 Ablation studies

In this subsection, several ablation experiments are conducted to demonstrate the effectiveness of our method.

We preform an ablation experiment to demonstrate the effect of the Canny-edge map on the quality of the foreground segmentation map. It turns out that the Canny-edge map does help improve the quality of the crowd segmentation map (Table 3).

Then, we perform another ablation experiment to confirm the benefits of the head-segmentation prior map and uniform-kernel density map. Table 4 shows that both the head-segmentation prior map and the uniform-kernel density map improve the crowd estimation, and that the uniform-kernel density map outperforms the Gaussian-kernel density map.

Table 4 also shows that without pre-training on other datasets, only two epochs are needed to converge to an optimal result using the head-segmentation map, whereas five epochs are required without using the head-segmentation map, thus proving that the head-segmentation map helps reduce the training difficulty. We implement all the experiments on the ShanghaiTech part B dataset.

Table 3 Ablation study results of the Canny-edge map on the ShanghaiTech part *B* dataset

Method	SSIM
VGG-bone	0.89
VGG-bone+Canny-edge map	0.93

Table 4 Ablation study results of the proposed method on the ShanghaiTech part *B* dataset

SAPNet with			MAE	Number of epochs
Gaussian kernel	Head-segmentation map	Uniform kernel		
✓	×	×	10.6	–
×	×	✓	10.1	5
✓	✓	×	9.8	–
×	✓	✓	9.4	2

We propose a new counting evaluation metric to prevent the noise in the sub-patch from being added to the crowd counting on the parent-patch (i.e., the whole image). We split the images in part B into four patches (Fig. 3). Furthermore, we evaluate the accuracy of crowd counting in sub-patches 3 and 4 to demonstrate its capability of wiping out the noise because there are fewer people and more background information in these patches. We compare our method with the congested-scene recognition network (CSRNet) (Li YH et al., 2018) and MCNN (Zhang YY et al., 2016) on the crowd counting error (Table 5).

Table 5 Comparison of the counting errors in sub-patches 3 and 4

Method	MAE	
	Sub-patch 3	Sub-patch 4
MCNN	2.02	1.86
CSRNet	0.87	0.97
SAPNet	0.60	0.51

The average numbers of sub-patches 3 and 4 are 3.00 and 3.60, respectively

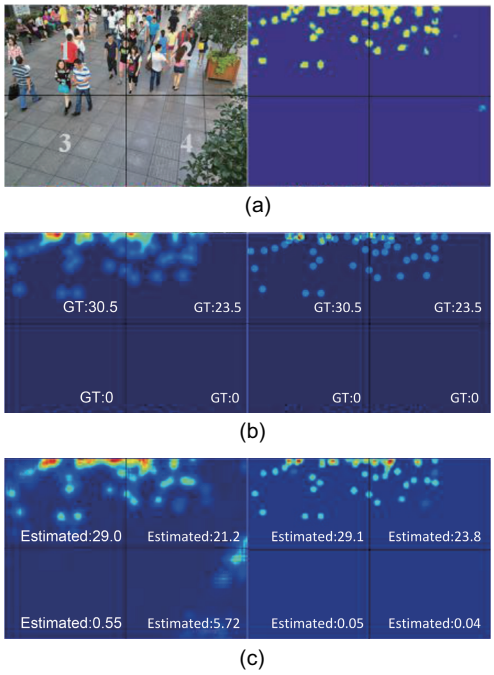

Fig. 3 Original image and head-segmentation map predicted by SAPNet (a), ground truth of two methods CSRNet and SAPNet (b), and two density maps predicted by the two methods (c)

As shown in Table 5, the counting error of our method is lower than those of MCNN and CSRNet in sub-patches 3 and 4. Since our CR-CNN is similar to CSRNet (the difference is only in the prior-bone), the noise-eliminating effect of SAPNet is attributed to the prior-bone (crowd-segmentation prior map). Fig. 3 shows the noise-eliminating effect of a common area without people (trees in the image).

5.3 ShanghaiTech dataset

The SHT dataset consists of two parts, i.e., parts A and B. Part A is a relatively dense crowd, containing 482 frames with an average of 501.4 persons per frame. Part B is a sparser crowd, including 716 frames with an average of 123.6 persons per frame.

Our method obtains the state-of-the-art results in part B and achieves a competitive performance in part A (Table 6).

Since VGG-bone uses the first 13 layers of the VGG network, the parameter number of our whole network is large and the computational complexity is high. The parameter number of SAPNet is 30.26 MB, whereas that of MCNN is 0.13 MB. However, our SAPNet significantly outperforms MCNN on all datasets. For example, our SAPNet gains huge increases of 29.7% and 64.4% on SHT parts A and B, respectively. Compared with the inference time of 0.0013 s of MCNN, the inference time of SAPNet is 0.252 s, which satisfies the real-time requirement.

As mentioned in Section 1, an inaccurate density map (which has noise in the regions without people) may lead to "good" crowd estimation since the crowd count is obtained by integrating the whole density map.

We use SSIM to evaluate the quality of the density map and achieve a better result compared with CSRNet and MCNN. To demonstrate the efficiency of our method, we compare it with CSRNet since they have similar network structures. Our proposed method has a slightly larger parameter number and higher computational complexity than CSRNet, and achieves more accurate crowd estimation and a high-quality density map (Table 7).

Table 6 Estimation errors on ShanghaiTech dataset

Method	MAE		RMSE	
	Part A	Part B	Part A	Part B
MCNN	110.2	26.4	173.2	41.3
Cascaded MTL	101.3	20.0	152.4	31.1
Switching-CNN	90.4	21.6	135.0	33.4
CP-CNN	73.6	20.1	106.4	30.1
IG-CNN	72.5	13.6	118.2	21.1
ACSCP	75.7	17.7	102.7	27.4
CSRNet	68.2	10.6	115.0	16.0
SAPNet	77.5	9.4	128.8	15.4

ACSCP: crowd counting via adversarial cross-scale consistency pursuit (Shen et al., 2018)

Table 7 Comparison between CSRNet and SAPNet on the ShanghaiTech part B dataset

Method	MAE	SSIM	Parameter number (MB)	Time (s)
CSRNet	10.6	0.890	16.26	0.13
SAPNet	9.4	0.955	30.26	0.25

5.4 UCF-CC-50 dataset

The UCF-CC-50 dataset contains 50 images with large variation of the crowd density. It is a challenging dataset because the resolutions of images in this dataset are low and the number of people in these images varies greatly, ranging from 94 to 4543. Five-fold cross validation is used because of the small dataset (Idrees et al., 2013). We obtain a state-of-the-art result from this challenging dataset (Table 8),

which indicates that the segmentation map is more effective for smaller datasets.

Table 8 Estimation errors on the UCF-CC-50 dataset

Method	MAE	RMSE
Zhang C et al. (2015)'s	467.0	498.5
MCNN	377.6	509.1
Switching-CNN	318.1	439.2
CP-CNN	295.8	320.9
IG-CNN	291.4	349.4
ACSCP	291.0	404.6
CSRNet	266.1	397.5
SAPNet	255.0	327.1

5.5 UCSD dataset

The UCSD dataset contains 2000 frames taken from a stationary camera. In contrast to the congested scenes in UCF-CC-50, the maximum number of pedestrians in the scenario of the UCSD dataset is only 46. Before generating a density map with a uniform kernel, we mask both the raw images and annotation with the corresponding region of interests (ROIs). Moreover, since the image size (238×158) is too small to generate a high-quality density map after propagating through multiple layers of max-pooling in the network, bilinear interpolation is used to increase the resolution of the image (952×632) (Li YH et al., 2018). To evaluate the generalization capability of our proposed method, we follow the method of partitioning the dataset in Chan et al. (2008), which takes frames 601–1400 as the training set and the remaining frames as the testing set.

From the experiment, we achieve a competitive result compared with the state-of-the-art works (Table 9). Essentially, the quality of the density map achieved using our method demonstrates a significant improvement (Li YH et al., 2018) and proves that our method can effectively remove noise from the background (Table 10).

5.6 WorldExpo'10 dataset

The WorldExpo'10 dataset (Zhang C et al., 2016) has 3980 frames taken from 108 scenes with 3380 frames and 600 frames used as the training and testing sets, respectively. The crowd scenes in this dataset are relatively sparse, and the number of people ranges from 1 to 220. The dataset provides ROIs for all the scenes, so each frame and its generated density map are masked with ROI during the preprocessing (Zhang C et al., 2015). This dataset proves to be unsuitable for our method because the raw images are of low resolution, which significantly affects the performance of the Canny-edge detector (i.e., the blurred appearance in a dense crowd significantly reduces the edge information detected by the Canny-edge detector). The network cannot generate an effective segmentation map due to the poor Canny-edge map. Therefore, a poor segmentation map causes our method poor performance on this dataset (Table 11).

Table 9 Estimation errors on the UCSD dataset

Method	MAE	RMSE
Zhang C et al. (2015)'s	1.60	3.31
Switching-CNN	1.62	2.10
CSRNet	1.16	1.47
MCNN	1.07	1.35
ACSCP	1.04	1.35
SAPNet	1.17	1.51

Table 10 Quality of density maps generated by our method

Dataset	SSIM
ShanghaiTech part *A*	0.480
ShanghaiTech part *B*	0.955
UCF-CC-50	0.390
WorldExpo'10	0.840
UCSD	0.950

6 Conclusions

In this study, we have proposed a novel network architecture, called SAPNet, which consists of a foreground-segmentation CNN (FS-CNN) and a crowd-regression CNN (CR-CNN), to achieve accurate crowd count estimation and generate a high-quality density map. This study has three major contributions. First, we proposed a novel uniform function to generate a crowd segmentation map from the data of dot annotation, which solves the common problem that the predicted density maps usually have a lot of noise in the background. Second, both networks take the prior maps as the input, i.e., a Canny-edge prior map for FS-CNN and a coarse head-segmentation prior map for CR-CNN, which helps extract effective features and reduce difficulty of training network. Finally, we demonstrated SAPNet on four benchmarks and achieved state-of-the-art performances on the ShanghaiTech part *B* and UCF-CC-50 datasets.

Table 11 Estimation errors on the WorldExpo'10 dataset

Method	MAE					Average
	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	
Zhang C et al. (2015)'s	9.80	14.1	14.3	22.2	3.7	12.82
MCNN	3.40	20.6	12.9	13.0	8.1	11.60
Switching-CNN	4.40	15.7	10.0	11.0	5.9	9.40
CP-CNN	2.90	14.7	10.5	10.4	5.8	8.86
CSRNet	2.90	11.5	8.6	16.6	3.4	8.60
IG-CNN	2.60	16.1	10.2	20.2	7.6	11.34
ACSCP	2.80	14.1	9.6	8.1	2.9	7.50
SAPNet	3.47	24.2	21.3	25.0	3.2	15.43

The possible future work includes employing network compression techniques, which have been deeply studied in recent years, to accelerate our proposed method.

Contributors

Jie-hao HUANG and Xiao-guang DI designed the research. Jie-hao HUANG drafted the manuscript. Jie-hao HUANG, Jun-de WU, and Ai-yue CHEN processed the data. Xiao-guang DI helped organize the manuscript. Jie-hao HUANG and Xiao-guang DI revised and finalized the paper.

Compliance with ethics guidelines

Jie-hao HUANG, Xiao-guang DI, Jun-de WU, and Ai-yue CHEN declare that they have no conflict of interest.

References

Canny J, 1986. A computational approach to edge detection. *IEEE Trans Patt Anal Mach Intell*, 8(6):679-698. https://doi.org/10.1109/TPAMI.1986.4767851

Chan AB, Vasconcelos N, 2009. Bayesian Poisson regression for crowd counting. Proc IEEE 12th Int Conf on Computer Vision, p.545-551. https://doi.org/10.1109/ICCV.2009.5459191

Chan AB, Liang ZSJ, Vasconcelos N, 2008. Privacy preserving crowd monitoring: counting people without people models or tracking. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.1-7. https://doi.org/10.1109/CVPR.2008.4587569

Dai JF, Li Y, He KM, et al., 2016. R-FCN: object detection via region-based fully convolutional networks. Proc 30th Int Conf on Neural Information Processing Systems, p.379-387.

Dollar P, Wojek C, Schiele B, et al., 2012. Pedestrian detection: an evaluation of the state of the art. *IEEE Trans Patt Anal Mach Intell*, 34(4):743-761. https://doi.org/10.1109/TPAMI.2011.155

Idrees H, Saleemi I, Seibert C, et al., 2013. Multi-source multi-scale counting in extremely dense crowd images. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.2547-2554. https://doi.org/10.1109/CVPR.2013.329

Kang K, Wang XG, 2014. Fully convolutional neural networks for crowd segmentation. https://arxiv.org/abs/1411.4464

Lempitsky V, Zisserman A, 2010. Learning to count objects in images. Proc 23rd Int Conf on Neural Information Processing Systems, p.1324-1332.

Li HH, He XJ, Wu HF, et al., 2018. Structured inhomogeneous density map learning for crowd counting. https://arxiv.org/abs/1801.06642

Li JJ, Yang H, Wu S, 2016. Crowd semantic segmentation based on spatial-temporal dynamics. Proc 13th IEEE Int Conf on Advanced Video and Signal Based Surveillance, p.102-108. https://doi.org/10.1109/AVSS.2016.7738032

Li T, Chang H, Wang M, et al., 2015. Crowded scene analysis: a survey. *IEEE Trans Circ Syst Video Technol*, 25(3):367-386. https://doi.org/10.1109/TCSVT.2014.2358029

Li YH, Zhang XF, Chen DM, 2018. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1091-1100. https://doi.org/10.1109/CVPR.2018.00120

Liu J, Gao CQ, Meng DY, et al., 2018. DecideNet: counting varying density crowds through attention guided detection and density estimation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5197-5206. https://doi.org/10.1109/CVPR.2018.00545

Long J, Shelhamer E, Darrell T, 2015. Fully convolutional networks for semantic segmentation. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.3431-3440. https://doi.org/10.1109/CVPR.2015.7298965

Sam DB, Surya S, Babu RV, 2017. Switching convolutional neural network for crowd counting. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.4031-4039. https://doi.org/10.1109/CVPR.2017.429

Sam DB, Sajjan NN, Babu RV, 2018. Divide and grow: capturing huge diversity in crowd images with incrementally growing CNN. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.3618-3626. https://doi.org/10.1109/CVPR.2018.00381

Shen Z, Xu Y, Ni B, et al., 2018. Crowd counting via adversarial cross-scale consistency pursuit. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5245-5254. https://doi.org/10.1109/CVPR.2018.00550

Sindagi VA, Patel VM, 2017. Generating high-quality crowd density maps using contextual pyramid CNNs. Proc IEEE Int Conf on Computer Vision, p.1879-1888. https://doi.org/10.1109/ICCV.2017.206

Sindagi VA, Patel VM, 2018. A survey of recent advances in CNN-based single image crowd counting and density estimation. *Patt Recogn Lett*, 107:3-16. https://doi.org/10.1016/j.patrec.2017.07.007

Zhan BB, Monekosso DN, Remagnino P, et al., 2008. Crowd analysis: a survey. *Mach Vis Appl*, 19(5-6):345-357. https://doi.org/10.1007/s00138-008-0132-4

Zhang C, Li HS, Wang XG, et al., 2015. Cross-scene crowd counting via deep convolutional neural networks. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.833-841. https://doi.org/10.1109/CVPR.2015.7298684

Zhang C, Zhang K, Li HS, et al., 2016. Data-driven crowd understanding: a baseline for a large-scale crowd dataset. *IEEE Trans Multim*, 18(6):1048-1061. https://doi.org/10.1109/TMM.2016.2542585

Zhang YY, Zhou DS, Chen SQ, et al., 2016. Single-image crowd counting via multi-column convolutional neural network. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.589-597. https://doi.org/10.1109/CVPR.2016.70