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

Aggregated context network for crowd counting*

Si-yue YU[1], Jian PU‡[1,2]

[1]*School of Computer Science and Technology, East China Normal University, Shanghai 200062, China*

[2]*Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China*

E-mail: 51174500148@stu.ecnu.edu.cn; jianpu@fudan.edu.cn

Received Sept. 8, 2019; Revision accepted Jan. 5, 2020; Crosschecked May 18, 2020; Published online Aug. 5, 2020

Abstract: Crowd counting has been applied to a variety of applications such as video surveillance, traffic monitoring, assembly control, and other public safety applications. Context information, such as perspective distortion and background interference, is a crucial factor in achieving high performance for crowd counting. While traditional methods focus merely on solving one specific factor, we aggregate sufficient context information into the crowd counting network to tackle these problems simultaneously in this study. We build a fully convolutional network with two tasks, i.e., main density map estimation and auxiliary semantic segmentation. The main task is to extract the multi-scale and spatial context information to learn the density map. The auxiliary semantic segmentation task gives a comprehensive view of the background and foreground information, and the extracted information is finally incorporated into the main task by late fusion. We demonstrate that our network has better accuracy of estimation and higher robustness on three challenging datasets compared with state-of-the-art methods.

Key words: Crowd counting; Convolutional neural network; Density estimation; Semantic segmentation; Multi-task learning

https://doi.org/10.1631/FITEE.1900481 **CLC number:** TP391

1 Introduction

Crowd counting aims to predict the accurate number of people in an image or a video. Over the last few years, it has been applied to various applications such as video surveillance, traffic monitoring, assembly control, and other public safety applications (Li YH et al., 2018). The major challenges of crowd counting are perspective distortion and background interference. The perspective distortion causes dramatic changes in each image (Shi et al., 2019), and a complex background always leads to overestimation of the number of people (Wang LY et al., 2019).

Researchers have tried numerous methods to address the challenges mentioned above. Representative studies include detection-based (Dollar et al., 2012), clustering-based (Rabaud and Belongie, 2006), and regression-based (Idrees et al., 2013) counting. Recently, convolutional neural network (CNN) based approaches have demonstrated significant improvements for crowd counting and density estimation (Cao et al., 2018; Li YH et al., 2018; Shen et al., 2018). Multiple methods rely on multi-column-based architectures (Zhang YY et al., 2016; Deb and Ventura, 2018) to tackle the perspective and scale variation issue. However, some researchers verified that the multi-column architecture with different kernel sizes is inefficient to form a multi-scale representation (Li YH et al., 2018). In addition, some researchers pointed out that the multi-column architecture limits scale diversity due to saturated performance. Also, the different receptive fields cause a large amount of training time with more parameters

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 61702186, 61672236, and 61602459)

ORCID: Si-yue YU, https://orcid.org/0000-0001-9569-8541; Jian PU, https://orcid.org/0000-0002-2949-4273

(Cao et al., 2018). To eliminate the interference factors due to highly similar visual effects as pedestrians, we can use the contextual information in images by simulating the human visual system. The network can generate top-down feedback to suppress and correct false density prediction (Sam and Babu, 2018).

Inspired by recent advances in multi-task learning for computer vision (Pu et al., 2014; Ruder, 2017), we introduce an auxiliary task of semantic segmentation for crowd counting, which is used for salient person segmentation (Cong et al., 2019b; Li CY et al., 2019). The purpose of this auxiliary task is to obtain more global information to compensate for the contextual information for the main task. To this end, we use the dilated convolutional layers to ensure a large receptive field, and thus to generate high-quality density maps. We also apply global average-pooling to capture the global contextual information (Chen LC et al., 2017). Finally, the semantic segmentation mask is incorporated into the main task by late fusion. Distinct from previous crowd segmentation, we use a two-branch fully convolutional network instead of hand-crafted representations (Chen K et al., 2012) without additional perspective maps.

We propose a novel multi-task aggregated context counting network (ACCNet) to learn density map estimation and foreground-background segmentation simultaneously. ACCNet is an end-to-end fully convolutional network supporting arbitrary input size. For the main crowd counting task, we use nine convolutional layers from conv1_2 to conv4_3 of a fine-tuned VGG-16. We propose a multi-column block (MCB) in ACCNet to resolve the scale variation problem with stride = 2 of MCB but with the same kernel size for each column. This alternation is a crucial factor in obtaining rich spatial information to tackle the perspective distortion issue. We add a skip connection between the processed output of fine-tuned VGG-16 and the features learned by MCB to provide more scale information. For the auxiliary task, we design a two-branch fully convolutional network to capture global context information for efficient semantic segmentation.

The main contributions of this work are summarized as follows:

1. We propose a novel fully convolutional multi-task network to learn density map estimation and semantic segmentation jointly and efficiently. In the main task, to solve the perspective issue more efficiently, the MCB with four columns and altered stride length is crucial in obtaining rich spatial information. Also, there is a skip connection to tackle scale variations. In the auxiliary task, the designs of the two-branch network and global average-pooling are the key points for capturing the global contextual information.

2. In the experiments on three benchmark datasets, we demonstrate that the proposed ACCNet is better than or comparable to the state-of-the-art methods.

3. Ablation studies on benchmark datasets further verify the effectiveness of the auxiliary semantic segmentation network.

2 Related work

There have been a variety of approaches addressing the problem of crowd density estimation or crowd counting. Existing methods for such applications can be roughly categorized into the following three classes (Loy et al., 2013): detection-based, regression-based, and density estimation based methods. Due to the rapid progress in deep learning (Cheng et al., 2018), we also review the recently developed deep neural network based methods for crowd counting.

2.1 Detection-based methods

Detection-based methods are usually intuitive by applying a moving window to detect pedestrians or body parts (Dollar et al., 2012; Sindagi and Patel, 2018). Training classifiers for the whole body is one straightforward method using low-level features, such as histograms of oriented gradient (HOG) (Dalal and Triggs, 2005) and Haar wavelets (Viola and Jones, 2004). Though successful for low-density crowds' images, most of these methods fail to deal with high-density crowds' images, since targeted pedestrians are severely obscured. As a remedy, researchers try to estimate the number of body parts using head- or shoulder-like detectors for crowd scenes analysis (Li M et al., 2008). However, these methods are designed mostly for some specific scenes, and cannot be easily applied to generalized datasets.

2.2 Regression-based methods

To tackle the scenes with severe occlusions, some methods try to count by regression. Features such as edge features, foreground features, gradient features, and texture are first independently extracted from images or image patches. Afterward, the relationship between these features and the crowd count is learned by regression (Chan and Vasconcelos, 2012). It was suggested that a single feature or detection method is not reliable in obtaining a precise number for highly congested scenes, especially for severe occlusion, perspective, and low-resolution scenes (Idrees et al., 2013). Therefore, several kinds of features are extracted using different approaches and then fused to form a comprehensive feature for crowd counting (Idrees et al., 2013). Recently, fully convolutional networks are also used to extract the foreground features (Long et al., 2015).

2.3 Density estimation based methods

Density estimation is associated with crowd counting, where each pixel of the corresponding density map indicates the number of people. Compared with direct counting of the number, the density estimation based method can preserve more spatial information of crowd scenes. To leverage more density information, Lempitsky and Zisserman (2010) suggested learning a linear mapping between features in the local patches and corresponding object density maps. By estimating image density, the number of objects in the images is obtained by the integral over the density map without explicitly detecting or localizing the target objects. Following similar approaches, Rodriguez et al. (2011) improved the performance of head detection. Furthermore, since linear mapping is difficult to learn, Pham et al. (2015) proposed to learn a non-linear mapping using random forest regression between the local patch and density maps.

2.4 CNN-based methods

Deep convolutional neural networks (CNNs) are currently the state-of-the-art approaches (Cao et al., 2018; Li YH et al., 2018; Shen et al., 2018), and many CNN-based approaches have been introduced to predict the density map for crowd counting. Wang C et al. (2015) adopted a basic convolutional structure, which is an end-to-end deep CNN regression model for counting people from images in extremely dense crowds based on the AlexNet framework. Walach and Wolf (2016) proposed the basic CNNs with layered boosting and selective sampling.

Particularly, to overcome perspective distortions and scale variations, researchers have developed various kinds of models with different network structures. The counting method combines a deep network with a shallow network to tackle scale variations across images (Boominathan et al., 2016). Multi-column counting networks are found to be useful in solving variations in the people scale originating from the perspective effect (Zhang YY et al., 2016). Recently, using a similar multi-column structure, Switch-CNN trains a classifier to relay the particular input patches to the optimal regressor from multi-column regressors (Sam et al., 2017). The aggregated multi-column dilated convolution network (AMDCN) uses dilated filters to construct an aggregation module in a multi-column CNN (Deb and Ventura, 2018). In this work, unlike the common significance of multi-column, we use the multi-column architecture with the same kernel size for each column. Furthermore, we change the stride length from one to two. Such alternation in the multi-column block can encode richer spatial information, which is beneficial to solving the problem of perspective distortion.

Contextual information has been proved useful in predicting both the local and global counts. Shang et al. (2016) proposed to learn the counting task by incorporating local regions and the overall images. Recently, a CNN structure using high-level prior information is presented to improve the crowd counting performance while learning to classify crowds into several groups (Sindagi and Patel, 2017a). Instead of learning a sophisticated density level classifier for the auxiliary task, we design a two-branch network that uses dilated convolutional layers and global average pooling. For the main crowd counting task, the network can provide a sufficient receptive field with global context information to compensate for the contextual information.

3 Proposed aggregated context counting network

In this section, we present the proposed ACCNet in detail. ACCNet is a multi-task network

with parameters shared by some of the convolutional layers. The main task is to generate high-quality density maps without losing resolution, and thus to predict crowd counting accurately. The auxiliary task is foreground-background segmentation to acquire global contextual information. It can reduce false predictions caused by background interference and further improve the performance of density map estimation and crowd counting.

The structure of ACCNet is shown in Fig. 1. A fine-tuned visual geometry group (VGG) network is used first to extract low-level image features, followed by two tasks, one for density map estimation and the other for foreground-background segmentation. Finally, the semantic segmentation mask is integrated into the main pathway by late fusion.

3.1 Shared convolutional layers

Inspired by Zhang YY et al. (2016) and Li YH et al. (2018), we build the model using part of VGG-16. Instead of using a pre-trained model, we prefer learning from scratch, and adopt nine convolutional layers (from conv1_2 to conv4_3) of the VGG-16 network. For three max-pooling layers of the original VGG-16 network, we remove the first and third pooling layers and keep only the second max-pooling layer to maintain sufficient invariance and generate high-quality density maps. As a remedy, we use a 1×1 convolutional layer after conv4_3. We also adopt a dilated convolutional layer instead of max-pooling to enlarge the receptive field and extract deeper information.

3.2 Density map estimation

One major difficulty in density map estimation is perspective distortion (Sindagi and Patel, 2018), which generates scale variations even in a single image. The perspective problem is caused by the geometry and depth information (Cong et al., 2019a) in the image and the spatial context information. In this study, the four columns of a convolutional layer with stride = 2 in MCB are enough to extract rich spatial information for density estimation. The appropriate stride enables to capture more receptive fields, and the mode of multi-column enables to capture more context information. Various kinds of methods have been proposed to tackle the perspective and scale issues (Fig. 2). MCB in Fig. 2a can provide flexible receptive fields across multiple columns but result in a large amount of training time with more parameters. Some research (Li YH et al., 2018; Shen et al., 2018) found that the effect of multi-column design on multi-scale information extraction is weak. The Skip-Net in Fig. 2b can extract multi-scale features by receptive fields of different sizes. It can connect low- and high-level features and obtain more accurate scale information. However, the joint features should be carefully chosen; inappropriate

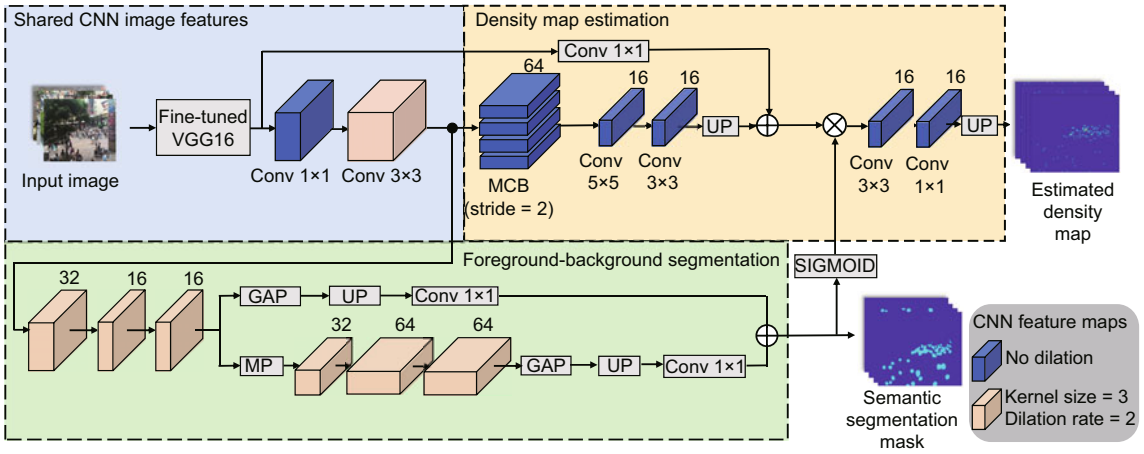

Fig. 1 Overview of ACCNet for density map estimation and semantic segmentation

Cube represents convolutional neural network (CNN) feature maps. The blue cube represents a traditional convolutional layer, and the khaki cube represents a dilated convolutional layer with dilation rate = 2. The numbers above cubes denote the corresponding numbers of channels. For the semantic segmentation task, batch normalization layers are applied after convolutional layers. ReLU as a nonlinear activation function is used after convolution layers. MP, GAP, and UP indicate max-pooling, global average-pooling, and upsampling, respectively. References to color refer to the online version of this figure

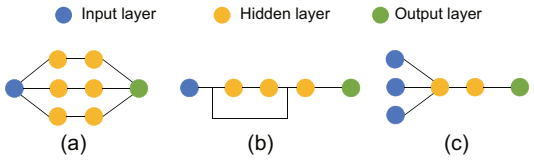

Fig. 2 Different multi-scale architectures: (a) multi-column block; (b) Skip-Net; (c) multi-scale input

connection may lead to performance degradation. The multi-scale network in Fig. 2c accepts several scales of input images to deal with scale variation. However, for images with low resolution, the quality of scaling image patches cannot be guaranteed. Scaling image patches might corrupt the distribution of input images (Wang LY et al., 2018).

In this study, we combine the merits of MCB and Skip-Net to tackle perspective distortion in crowd scenes. We first adopt MCB with the same kernel size of each column. We set stride = 2 for each convolutional layer in MCB to gain rich spatial context information. Kernels of non-uniform size after MCB are applied to achieve robust performance. To capture more scale information, we add a skip connection between the interpolated features and the final output of the fine-tuned VGG-16. We map the connected feature to a preliminary density map by subsequent convolutional layers. To compensate for the downsampling effect, we use bilinear interpolation to generate the full-resolution density map.

3.3 Foreground-background segmentation

The primary goal of the auxiliary task is to provide global context information to boost the main task of crowd counting. Similar to crowd counting, semantic segmentation also requires rich spatial details and semantic information. Instead of obtaining a fine-grained segmentation (He et al., 2018) to distinguish buildings, trees, and person, our task focuses on person segmentation (Huang JH et al., 2020) and enhancement of person saliency (Cong et al., 2018, 2019c). To this end, we regard the crowd as foreground and the rest of the region as background, to segment the foreground from the background to extract enough information about the person.

Unlike traditional crowd segmentation (Xie et al., 2014; Zhu and Peng, 2016), instead of using hand-crafted representations and methods by graph or the boosting algorithm, we propose a two-branch fully convolutional network. Different from the well-known fully convolutional network (Long et al., 2015) for semantic segmentation, however, we use dilated convolutional layers to increase the network receptive field without an exponential increase of the number of parameters. The replacement of convolutional layers can aggregate diverse contextual information and keep the spatial resolution of the feature map (Chen LC et al., 2018). The batch normalization layer is applied with a dilated convolutional layer to avoid the vanishing of the gradient.

Inspired by the importance of contextual effects in semantic segmentation (Chen LC et al., 2017) and related works about multi-scale methods (Peng et al., 2018), we adopt global average-pooling (GAP) to acquire global contextual information. Then by the use of bilinear interpolation, we upsample the feature to the desired resolution. After that, we reduce the dimension of the feature map by a 1×1 convolutional layer. The above is related to the single branch of our proposed segmentation network. The other branch uses max-pooling (MP) to encode rich spatial details. The final segmentation map obtained by the sigmoid function is applied to the main pathway for further fusion. A series of convolutional layers are used to transform the feature map, and the estimated density map and crowd count are obtained.

3.4 Objective function

In terms of the density estimation task, we measure the difference between the generated density map and the ground truth using the traditional Euclidean distance. The objective function is

$$L_D(\theta) = \frac{1}{2N}\sum_{i=1}^{n}\left\|F_d(X_i;\theta) - F_i^{\mathrm{gt}}\right\|_2^2, \qquad (1)$$

where N is the number of training samples, and θ is the network parameter of ACCNet. $F_d(X_i;\theta)$ indicates the generated density map learned from ACCNet with parameters θ. F_i^{gt} represents the ground truth density map of the input image X_i.

The auxiliary task for foreground-background segmentation is a pixel-level classification problem. We use binary cross-entropy for this task. The probability of the foreground $S_{\mathrm{Seg}}^i(\mathrm{foreground})$ corresponding to the input image X_i is given by

$$S_{\mathrm{Seg}}^i(\mathrm{foreground}) = \hat{y}_i = \frac{1}{1+\mathrm{e}^{-p_i}}, \qquad (2)$$

where p_i is the pixel value of segmentation map S_{Seg}^i. Therefore, the complementary probability of

the background $S^i_{\text{Seg}}(\text{background})$ is given by

$$S^i_{\text{Seg}}(\text{background}) = 1 - \hat{y}_i. \tag{3}$$

Denote the ground truth segmentation map for the input image X_i by gt_i, and then the segmentation loss is formed as

$$L_S = -\frac{1}{N}\sum_{i=1}^{n}\left[\text{gt}_i \log \hat{y}_i + (1-\text{gt}_i)\log(1-\hat{y}_i)\right]. \tag{4}$$

We combine density map estimation loss L_D and segmentation loss L_S by a proportionality coefficient of λ. Therefore, the final loss is given by

$$L = L_D + \lambda L_S. \tag{5}$$

The weight of density map estimation loss is set to be higher than that of the segmentation loss, since we require the former have more influence during training. Besides, in our experiments, we found that L_D is a few orders of magnitude smaller than L_S. To achieve a more accurate density map estimation, λ is set to 0.01 in the final loss for all experiments.

4 Implementation details

4.1 Ground truth generation

Since we have merely the head number and the labeled coordinate of each head for most datasets, we first introduce the process of ground truth generation for the crowd counting task and semantic segmentation task.

Numerous methods have been proposed to generate the ground truth density map according to the labeled coordinate of the heads. The Gaussian kernel density centered on the locations of the provided points is the most famous method (Lempitsky and Zisserman, 2010). In recent years, researchers found that perspective normalization is of great importance in creating the ground truth density map. Zhang C et al. (2015) proposed a human-shaped Gaussian kernel with two parts, i.e., the head and the body. Density maps via geometry-adaptive kernels were proposed to tackle highly crowded scenes (Zhang YY et al., 2016; Li YH et al., 2018; Wang LY et al., 2018).

We use the geometry-adaptive kernels proposed by Zhang YY et al. (2016) to generate the ground truth density map $F(x)$ as follows:

$$F(x) = \sum_{i=1}^{N}\delta(x-x_i)\cdot G_{\sigma_i}(x), \tag{6}$$

where $\sigma_i = \beta\overline{d_i}$, β is a constant term coefficient. For each annotated head at pixel x_i, we represent the pixel as a delta function $\delta(x-x_i)$. To generate the density map, we convolve $\delta(x-x_i)$ with a filter template G_{σ_i}. The parameter σ_i denotes the standard deviation of the Gaussian kernel, chosen according to the average distance $\bar{d}_i$ of k nearest neighbors to each targeted object x_i. Following Zhang YY et al. (2016), we set $k = 3$ and $\beta = 0.3$. For sparse crowd scenes, i.e., the Shanghaitech Part_B dataset (Zhang YY et al., 2016) and UCSD dataset (Chan et al., 2008), we generate the ground truth density maps by annotating every person with a fixed parameter $\sigma_i = 4$.

For the semantic segmentation task, we generate the ground truth label according to the generated ground truth density map. As mentioned before, each pixel in the corresponding density estimation map represents the number of people. In the part of the input image with a person's presence, the pixel value varies between zero and one, and all values sum to one. Unlike the property of density values, the benefit of semantic segmentation is binary classification. The pixel intensity of the ground truth density map greater than zero is set as the foreground, and the rest of the pixels are considered as background.

4.2 Training and evaluation details

Considering a dataset with various resolutions and downsampling operations in ACCNet, we need to ensure that the x and y coordinates of images are both multiples of four. For instance, the original image resolutions of the UCSD dataset are all 158×238, and we resize them to 160×240.

Note that we do not employ any data augmentation method (i.e., patch cropping or mirroring) to approximate the original distribution. Since only one max-pooling layer is applied in ACCNet, there is very little loss of context information. Therefore, the performance of our network can hardly be improved by traditional data augmentation methods. In Section 6.1, we compare the performance of ACCNet with and without data augmentation. For the UCF_CC_50 dataset, we randomly select 40 images

for training, and the remaining 10 images for testing. The datasets are divided into a training set and a testing set. If without specific descriptions, 80 percent of images in the training set is used for training, and the remaining images are used as the validation set for model selection. We report the performance of the UCSD dataset using two frequently used splitting methods.

The network parameters are initialized by Gaussian distribution with zero mean and 0.01 standard deviation. We start from scratch and set the initial learning rate at 10^{-5}, and decrease it by half every 60 epochs. ACCNet is trained end to end by an Adam optimizer using Pytorch (Paszke et al., 2017). We report the performance at the 500^{th} epoch. The experiments are implemented on one GeForce® GTX 1080 Ti.

In the testing stage, we compare the performance of our model with that of state-of-the-art methods by the mean absolute error (MAE) and mean squared error (MSE). For N test images, the metrics are defined as

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|C_i - C_i^{\text{gt}}\right|, \tag{7}$$

$$\text{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|C_i - C_i^{\text{gt}}\right|^2}, \tag{8}$$

where C_i^{gt} is the ground truth crowd number of image X_i, and C_i denotes the estimated counting number. Generally speaking, MAE represents the accuracy of estimation, and MSE represents the robustness of the model for crowd density estimation.

5 Experimental results

In this section, we report the performance comparison between ACCNet and state-of-the-art methods on three benchmark datasets (Chan et al., 2008; Idrees et al., 2013; Zhang YY et al., 2016). The image samples of these datasets are illustrated in Fig. 3. Note that the scene, number of crowds, and camera perspective vary dramatically from dataset to dataset.

5.1 ShanghaiTech dataset

We first compare the prediction performance on the ShanghaiTech dataset (Zhang YY et al., 2016).

Fig. 3 Sample images from three benchmark datasets: (a) ShanghaiTech Part_A; (b) ShanghaiTech Part_B; (c) UCSD; (d) UCF_CC_50

This dataset contains 1198 annotated images with a total of 330 165 persons with head center annotations. It involves two subsets: Part_A contains 482 images at various resolutions and Part_B contains 716 images with a fixed resolution 768×1024. Part_A has more massive density scenes with an average number of 503 persons per image as compared to Part_B with an average of 123. Consistent with Zhang YY et al. (2016), we divide the images into training and testing sets. We use the geometry-adaptive Gaussian method for Part_A and fix the spread parameter for Part_B. As shown in Table 1, for the Part_A dataset, ACCNet has increased by 15.06% compared with ACSCP (Shen et al., 2018) in terms of MAE; for the Part_B dataset, our approach achieves the best MAE and MSE.

5.2 UCSD dataset

The UCSD dataset (Chan et al., 2008) is recorded from a video sequence at 10 frames/s in a campus scene. It has a total of 2000 images with a resolution of 158×238. The crowd count ranges from 11 to 46. The ground-truth annotation, region-of-interest (ROI), and perspective map of the scene are provided. The perspective information provided by the dataset can be used to normalize the features based on the weight scaling corresponding to reality. Since people closer to the camera appear larger than those far away, the features of closer people in perspective information account for a smaller portion. To incorporate perspective information, we divide the two-dimensional (2D) Gaussian of each person by the perspective value to generate the ground truth density map. The use of perspective information is described in Section 6.2. Here we fix the spread

Table 1 Performance comparison with state-of-the-art methods on the ShanghaiTech dataset

Method	MAE		MSE	
	Part_A	Part_B	Part_A	Part_B
Zhang C et al. (2015)'s	181.8	32.0	277.7	49.8
MCNN (Zhang YY et al., 2016)	110.2	26.4	173.2	41.3
Switching-CNN (Sam et al., 2017)	90.4	21.6	135.0	33.4
CP-CNN (Sindagi and Patel, 2017b)	73.6	20.1	106.4	30.1
Cascaded-MTL (Sindagi and Patel, 2017a)	101.3	20.0	152.4	31.1
Sam and Babu (2018)'s	97.5	20.7	145.1	32.8
IG-CNN (Sam et al., 2018)	72.5	13.6	118.2	21.1
ACSCP (Shen et al., 2018)	75.7	17.2	**102.7**	27.4
CSRNet (Li YH et al., 2018)	68.2	10.6	115.0	16.0
Huang SY et al. (2018)'s	–	20.2	–	35.6
ACCNet (ours)	**64.3**	**8.7**	104.1	**13.6**

MCNN: multi-column convolutional neural network; CP-CNN: contextual pyramid CNN; MTL: multi-task learning; IG-CNN: incrementally growing CNN; ACSCP: adversarial cross-scale consistency pursuit; CSRNet: congested scene recognition network; ACCNet: aggregated context counting network. Best performance is shown in bold

parameter to generate ground truth density maps and use ROI to constrain the testing area without the use of perspective information.

There are two methods to split the data into training and testing sets. First, following the setting in Chan et al. (2008), we select 800 frames from 600 to 1399 as the training set and the remaining 1200 frames as the testing set. Such a setting is relatively easy since all images are very similar, and the average number of persons is small (i.e., 25). As shown in Table 2, the performance is very close for all methods being compared, and ACCNet achieves comparative results.

To verify the performance of our network, we split the data in another way (Lempitsky and Zisserman, 2010; Ryan et al., 2010; Zhang C et al., 2015). The second splitting method involves four training groups: (1) maximum, (2) downscale (the most crowded), (3) upscale (the least crowded), and (4) minimum. The training frames are described in MATLAB notation as (1) 600 : 5 : 1400, (2) 1205 : 5 : 1600, (3) 805 : 5 : 1100, and (4) 640 : 80 : 1360. The frames out of the four splits are used for testing. The results are shown in Table 3. Note that ACCNet achieves the best performance for the "downscale", "upscale", and "minimum" groups. The minimum group is challenging for our method since only 10 images are used for training. Though without a data augmentation procedure, our performance is still superior.

Table 2 Performance comparison with state-of-the-art methods on the UCSD dataset

Method	MAE	MSE
Zhang C et al. (2015)'s	1.60	3.31
MCNN (Zhang YY et al., 2016)	1.07	1.35
CCNN (Oñoro-Rubio and López-Sastre, 2016)	1.51	–
Switching-CNN (Sam et al., 2017)	1.62	2.10
Huang SY et al. (2018)'s	**1.00**	1.40
ACSCP (Shen et al., 2018)	1.04	1.35
CSRNet (Li YH et al., 2018)	1.16	1.47
ACCNet (ours)	**1.00**	**1.27**

MCNN: multi-column convolutional neural network; CCNN: counting CNN; ACSCP: adversarial cross-scale consistency pursuit; CSRNet: congested scene recognition network; ACCNet: aggregated context counting network. Best performance is shown in bold

5.3 UCF_CC_50 dataset

The final dataset is UCF_CC_50 (Idrees et al., 2013). This dataset is the most challenging since it is the most crowded, and the number of training images

Table 3 Mean absolute error (MAE) on the UCSD pedestrian dataset

Method	MAE			
	Maximum	Downscale	Upscale	Minimum
Lempitsky and Zisserman (2010)'s	1.70	1.28	1.59	2.02
Fiaschi et al. (2012)'s	1.70	2.16	1.61	2.20
Codebook+RR (Arteta et al., 2014)	**1.24**	1.31	1.69	1.49
Pham et al. (2015)'s	1.43	1.30	1.59	1.62
Zhang C et al. (2015)'s	1.70	1.26	1.59	1.52
ACCNet (ours)	1.33	**1.21**	**1.10**	**1.31**

RR: ridge regression; ACCNet: aggregated context counting network. Best performance is shown in bold

is tiny. It contains only a total of 50 images with varying resolutions and diverse scenes. However, the average person count for each image is up to 1280. For this particular dataset, we use the geometry-adaptive Gaussian method to generate the ground truth density map (Zhang YY et al., 2016). Similar to Idrees et al. (2013), we report the five-fold cross-validation performance in Table 4. Note that ACCNet achieves the best performance for both MAE and MSE. Compared with ACSCP (Shen et al., 2018), a 30.7% improvement is achieved in terms of MAE. Compared with SANet (Cao et al., 2018), a 22.0% improvement is achieved in terms of MAE. For such a difficult dataset, the improvements demonstrate the effectiveness of our method in dealing with highly congested scenes.

5.4 Comparison of computation performance

The computation performance (number of parameters and runtime) is compared on the ShanghaiTech Part_A dataset. The number of network parameters is shown in Table 5. Cascaded-MTL (Sindagi and Patel, 2017a) contains the least number of parameters compared with Switching-CNN (Sam et al., 2017), CP-CNN (Sindagi and Patel, 2017b), PCC Net (Gao et al., 2019), and the proposed ACCNet. We focus on these five methods. ACCNet is very competitive in terms of these metrics of model performance. The parameter number of ACCNet is 7.86 MB, and the evaluation results are better than those of the other four methods. In terms of the runtime, ACCNet is the lowest.

Table 4 Performance comparison with state-of-the-art methods on the UCF_CC_50 dataset

Method	MAE	MSE
Zhang C et al. (2015)'s	467.0	498.5
MCNN (Zhang YY et al., 2016)	377.6	509.1
Switching-CNN (Sam et al., 2017)	318.1	439.2
CP-CNN (Sindagi and Patel, 2017b)	295.8	320.9
Cascaded-MTL (Sindagi and Patel, 2017a)	322.8	397.9
Sam and Babu (2018)'s	354.7	491.4
IG-CNN (Sam et al., 2018)	291.4	349.4
ACSCP (Shen et al., 2018)	291.0	404.6
CSRNet (Li YH et al., 2018)	266.1	397.5
SANet (Cao et al., 2018)	258.4	334.9
Huang SY et al. (2018)'s	409.5	563.7
ACCNet (ours)	**201.6**	**282.1**

MCNN: multi-column convolutional neural network; CP-CNN: contextual pyramid CNN; MTL: multi-task learning; IG-CNN: incrementally growing CNN; ACSCP: adversarial cross-scale consistency pursuit; CSRNet: congested scene recognition network; SANet: scale aggregation network; ACCNet: aggregated context counting network. Best performance is shown in bold

6 Ablation studies

In this section, ablation studies are conducted to verify the effectiveness and robustness of ACCNet. Data augmentation, additional perspective information, effectiveness of the multi-column block, skip connection, and auxiliary semantic segmentation are investigated on benchmark datasets.

6.1 Data augmentation

We first compare the performance of our network with and without data augmentation. The data augmentation procedure is the same as in Li YH et al. (2018). The input image is cropped into 3×3 patches at different locations with $1/4$ size of the original image. After that, all image patches are doubled by mirroring. As illustrated in Table 6, ACCNet without data augmentation achieves better performance. The reason is that the employment of max-pooling is remarkably reduced and the dilated convolutional layers are applied as a remedy. The dilated convolutional layers can acquire large receptive fields and keep the input resolution with fewer parameters and little information loss. We can train the network using the distribution as close to the original data distribution as possible.

Table 5 Comparison of computation performance

Method	MAE	MSE	Parameter number (MB)	Runtime (ms)
Cascaded-MTL (Sindagi and Patel, 2017a)	101.3	152.4	0.12	3
Switching-CNN (Sam et al., 2017)	90.4	135.0	15.1	153
CP-CNN (Sindagi and Patel, 2017b)	73.6	106.4	68.4	5113
PCC Net (Gao et al., 2019)	73.5	124.0	0.55	89
ACCNet (ours)	64.3	104.1	7.86	23

MTL: multi-task learning; CNN: convolutional neural network; CP-CNN: contextual pyramid CNN; PCC Net: perspective crowd counting network; ACCNet: aggregated context counting network

Table 6 Evaluation of data augmentation on benchmark datasets

Dataset	Data augmentation	MAE	MSE
ShanghaiTech Part_A	with	67.4	108.8
	w/o	**64.3**	**104.1**
UCSD	with	**1.00**	1.33
	w/o	**1.00**	**1.27**
UCF_CC_50	with	221.6	337.4
	w/o	**201.6**	**282.1**

with: train ACCNet using a data augmentation procedure; w/o: train ACCNet without any data augmentation procedure. Best performance is shown in bold

6.2 Perspective-free counting by ACCNet

Next, we use the perspective normalization map provided by the UCSD dataset to investigate whether it can further improve the prediction performance. As mentioned in Section 5.2, the perspective map weighs pixels that originate from the objects closer to the camera less than those from the objects further away from the camera. The perspective map value $p(x, y)$ is the weight at pixel (x, y) with a range from zero to eight in the UCSD dataset. This can be used to adjust the ground truth density map. Following Lempitsky and Zisserman (2010), we divide the 2D Gaussian of each person by the perspective value to form the ground truth density map.

The performance comparison is shown in Table 7. Although the perspective map is helpful for AMDCN (Deb and Ventura, 2018) with a lower MAE, the prediction performance slightly declines when the perspective map is incorporated in ACCNet. Although without a perspective map, our method can achieve higher performance than AMDCN (Deb and Ventura, 2018) with a perspective map. We conjecture that ACCNet achieves accurate perspective-free counting using abundant context information in the image.

Table 7 Evaluation of the performance when perspective information is incorporated using the UCSD dataset

Method	MAE	MSE
AMDCN with perspective map	1.72	–
AMDCN w/o perspective map	1.74	–
ACCNet with perspective map	1.12	1.47
ACCNet w/o perspective map	**1.00**	**1.27**

AMDCN: aggregated multi-column dilated convolution network; ACCNet: aggregated context counting network. Best performance is shown in bold

6.3 Effectiveness of multi-column block and skip connection

We also verify the effectiveness of the multi-column block in our network. As mentioned in Section 3.2, the key to our proposed multi-column block is the setting of the stride and the mode of the multi-column. The appropriate stride enables to capture more receptive fields, and the mode of the multi-column enables to capture more context information.

To better verify the effectiveness of the multi-column block, we use three settings of networks on benchmark datasets as follows: (a) remove the multi-column block from the network; (b) adopt the multi-column block with stride = 1 to verify the setting of the stride; (c) adopt a convolution layer with stride = 1 instead of the mode of the multi-column.

We provide comparison analysis on ACCNet and MCNN in terms of the multi-column block. The two settings of the multi-column block are as follows: (d) a new multi-column block with stride = 2 and different kernel sizes (3, 5, 7, and 9); (e) a new multi-column block with stride = 1 and different kernel sizes (3, 5, 7, and 9).

In addition, we verify the effectiveness of skip connection in our network as follows: (f) remove the skip connection from the network.

The results are reported in Tables 8 and 9. Note that ACCNet achieves the best performance in most cases.

Table 8 Evaluation of the effectiveness of the multi-column block on benchmark datasets

Dataset	MAE						MSE					
	(a)	(b)	(c)	(d)	(e)	ACCNet	(a)	(b)	(c)	(d)	(e)	ACCNet
ShanghaiTech Part_A	80.1	74.9	82.6	72.2	79.5	**64.3**	132.4	117.1	127.5	115.9	131.0	**104.1**
ShanghaiTech Part_B	21.4	13.2	20.1	14.6	18.7	**8.7**	37.0	**12.5**	35.6	23.1	34.0	13.6
UCSD	10.1	6.8	8.1	6.0	9.0	**1.0**	9.0	7.0	8.3	7.2	11.2	**1.3**
UCF_CC_50	265.2	270.1	300.1	289.2	320.1	**201.6**	400.7	397.6	370.6	403.4	398.8	**282.1**

(a)–(e) denote the various methods described in Section 6.3; ACCNet: aggregated context counting network. Best performance is shown in bold

Table 9 Evaluation of the effectiveness of skip connection on benchmark datasets

Dataset	MAE		MSE	
	(f)	ACCNet	(f)	ACCNet
ShanghaiTech Part_A	86.4	**64.3**	137.0	**104.1**
ShanghaiTech Part_B	19.8	**8.7**	38.1	**13.6**
UCSD	10.33	**1.00**	12.07	**1.27**
UCF_CC_50	281.2	**201.6**	397.6	**282.1**

(f) denotes the method described in Section 6.3; ACCNet: aggregated context counting network. Best performance is shown in bold

6.4 Benefits of semantic segmentation

Finally, we verify the benefits of the semantic segmentation task in our network. We compare the performance with a single density map estimation network. For fair comparison, we use the same experimental settings and the same training and testing data. The results are reported in Table 10. The gap between the two methods is significant for both MAE and MSE. It demonstrates that the proposed semantic segmentation task is of great importance in improving the accuracy and robustness of crowd counting. To better verify the effect of semantic segmentation and context information, we have replaced the sigmoid operation by the binarization operation. Pixel values of output greater than zero are set to one, and the remaining pixels are set to zero. In Table 10, the results of ACCNet with binarization demonstrate the inadequacy of filtering out merely the background and the effect of semantic segmentation and context information. To further verify the effectiveness of our proposed architecture and the necessity of multi-task mode, we experiment with an auxiliary segmentation network alone. We train the pruning network on the UCSD dataset and compute the density map estimate generated from the segmentation network. The results (Table 11) show the superior performance of ACCNet.

Table 11 Evaluation of the performance of auxiliary segmentation network alone on the UCSD dataset

Method	MAE	MSE
With auxiliary network alone	5.79	6.30
ACCNet	**1.00**	**1.27**

Best performance is shown in bold

7 Conclusions

In this study, we have proposed a multi-task network combining density map estimation with semantic segmentation to provide a mutual promotion of crowd counting and semantic segmentation. To tackle perspective distortion and background interference, the specially designed multi-column block, the skip connection, and the foreground-background segmentation task are incorporated. The density map estimation task with the multi-column block and skip connection is to extract spatial information and multi-scale information. The global average-pooling and dilated convolutional layers are adopted to provide a comprehensive view of the foreground and background information. Although our method has fewer network parameters, it has superior performance on three benchmark crowd counting datasets. The ablation studies also verify the effectiveness and robustness of the proposed network.

Contributors

Jian PU guided the research. Jian PU and Si-yue YU designed the research. Si-yue YU drafted the manuscript. Jian PU helped organize and polish the manuscript. Si-yue YU revised and finalized the paper.

Compliance with ethics guidelines

Si-yue YU and Jian PU declare that they have no conflict of interest.

References

Arteta C, Lempitsky V, Noble JA, et al., 2014. Interactive object counting. European Conf on Computer Vision,

Table 10 Evaluation of the performance of semantic segmentation on benchmark datasets

Dataset	MAE			MSE		
	ACCNet	ACCNet w/o segmentation	ACCNet with binarization	ACCNet	ACCNet w/o segmentation	ACCNet with binarization
ShanghaiTech Part_A	**64.3**	74.9	71.3	**104.1**	118.2	112.0
ShanghaiTech Part_B	**8.7**	17.6	12.1	**13.6**	22.6	25.0
UCSD	**1.00**	1.77	1.92	**1.27**	3.29	3.01
UCF_CC_50	**201.6**	234.0	224.8	**282.1**	322.8	310.9

Best performance is shown in bold

p.504-518.
https://doi.org/10.1007/978-3-319-10578-9_33

Boominathan L, Kruthiventi SSS, Babu RV, 2016. CrowdNet: a deep convolutional network for dense crowd counting. ACM Int Conf on Multimedia, p.640-644.
https://doi.org/10.1145/2964284.2967300

Cao XK, Wang ZP, Zhao YY, et al., 2018. Scale aggregation network for accurate and efficient crowd counting. European Conf on Computer Vision, p.757-773.
https://doi.org/10.1007/978-3-030-01228-1_45

Chan AB, Vasconcelos N, 2012. Counting people with low-level features and Bayesian regression. *IEEE Trans Image Process*, 21(4):2160-2177.
https://doi.org/10.1109/TIP.2011.2172800

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

Chen K, Loy CC, Gong SG, et al., 2012. Feature mining for localised crowd counting. British Machine Vision Conf, Article 21. https://doi.org/10.5244/c.26.21

Chen LC, Papandreou G, Schroff F, et al., 2017. Rethinking atrous convolution for semantic image segmentation.
https://arxiv.org/abs/1706.05587

Chen LC, Papandreou G, Kokkinos I, et al., 2018. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans Patt Anal Mach Intell*, 40(4):834-848.
https://doi.org/10.1109/TPAMI.2017.2699184

Cheng J, Wang PS, Li G, et al., 2018. Recent advances in efficient computation of deep convolutional neural networks. *Front Inform Technol Electron Eng*, 19(1):64-77.
https://doi.org/10.1631/FITEE.1700789

Cong RM, Lei JJ, Fu HZ, et al., 2018. Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation. *IEEE Trans Image Process*, 27(2):568-579.
https://doi.org/10.1109/TIP.2017.2763819

Cong RM, Lei JJ, Fu HZ, et al., 2019a. Going from RGB to RGBD saliency: a depth-guided transformation model. *IEEE Trans Cybern*, in press.
https://doi.org/10.1109/TCYB.2019.2932005

Cong RM, Lei JJ, Fu HZ, et al., 2019b. Review of visual saliency detection with comprehensive information. *IEEE Trans Circ Syst Video Technol*, 29(10):2941-2959.
https://doi.org/10.1109/TCSVT.2018.2870832

Cong RM, Lei JJ, Fu HZ, et al., 2019c. Video saliency detection via sparsity-based reconstruction and propagation. *IEEE Trans Image Process*, 28(10):4819-4831.
https://doi.org/10.1109/TIP.2019.2910377

Dalal N, Triggs B, 2005. Histograms of oriented gradients for human detection. IEEE Conf on Computer Vision and Pattern Recognition, p.886-893.
https://doi.org/10.1109/CVPR.2005.177

Deb D, Ventura J, 2018. An aggregated multicolumn dilated convolution network for perspective-free counting. IEEE Conf on Computer Vision and Pattern Recognition Workshops, p.195-204.
https://doi.org/10.1109/CVPRW.2018.00057

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

Fiaschi L, Nair R, Koethe U, et al., 2012. Learning to count with regression forest and structured labels. Int Conf on Pattern Recognition, p.2685-2688.

Gao JY, Wang Q, Li XL, 2019. PCC Net: perspective crowd counting via spatial convolutional network. *IEEE Trans Circ Syst Video Technol*, in press.
https://doi.org/10.1109/TCSVT.2019.2919139

He XT, Peng YX, Zhao JJ, 2018. Fast fine-grained image classification via weakly supervised discriminative localization. *IEEE Trans Circ Syst Video Technol*, 29(5):1394-1407.
https://doi.org/10.1109/TCSVT.2018.2834480

Huang JH, Di XG, Wu JD, et al., 2020. A novel convolutional neural network method for crowd counting. *Front Inform Technol Electron Eng*, 21(8).
https://doi.org/10.1631/FITEE.1900282

Huang SY, Li X, Zhang ZF, et al., 2018. Body structure aware deep crowd counting. *IEEE Trans Image Process*, 27:1049-1059.
https://doi.org/10.1109/TIP.2017.2740160

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

Lempitsky V, Zisserman A, 2010. Learning to count objects in images. Conf and Workshop on Neural Information Processing Systems, p.1324-1332.

Li CY, Cong RM, Hou JH, et al., 2019. Nested network with two-stream pyramid for salient object detection in optical remote sensing images. *IEEE Trans Geosci Remote Sens*, 57(11):9156-9166.
https://doi.org/10.1109/TGRS.2019.2925070

Li M, Zhang ZX, Huang KQ, et al., 2008. Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection. Int Conf on Pattern Recognition, p.1-4.
https://doi.org/10.1109/ICPR.2008.4761705

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

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

Loy CC, Chen K, Gong SG, et al., 2013. Crowd Counting and Profiling: Methodology and Evaluation. Springer, New York, USA.
https://doi.org/10.1007/978-1-4614-8483-7_14

Oñoro-Rubio D, López-Sastre RJ, 2016. Towards perspective-free object counting with deep learning. European Conf on Computer Vision, p.615-629.
https://doi.org/10.1007/978-3-319-46478-7_38

Paszke A, Gross S, Chintala S, et al., 2017. Automatic differentiation in PyTorch. 31st Conf on Neural Information Processing Systems, p.1-4.

Peng YX, He XT, Zhao JJ, 2018. Object-part attention model for fine-grained image classification. *IEEE Trans Image Process*, 27(3):1487-1500. https://doi.org/10.1109/TIP.2017.2774041

Pham VQ, Kozakaya T, Yamaguchi O, et al., 2015. COUNT forest: CO-voting uncertain number of targets using random forest for crowd density estimation. IEEE Int Conf on Computer Vision, p.3253-3261. https://doi.org/10.1109/ICCV.2015.372

Pu J, Jiang YG, Wang J, et al., 2014. Which looks like which: exploring inter-class relationships in fine-grained visual categorization. European Conf on Computer Vision, p.425-440. https://doi.org/10.1007/978-3-319-10578-9_28

Rabaud V, Belongie S, 2006. Counting crowded moving objects. IEEE Conf on Computer Vision and Pattern Recognition, p.705-711. https://doi.org/10.1109/CVPR.2006.92

Rodriguez M, Laptev I, Sivic J, et al., 2011. Density-aware person detection and tracking in crowds. IEEE Int Conf on Computer Vision, p.2423-2430. https://doi.org/10.1109/ICCV.2011.6126526

Ruder S, 2017. An overview of multi-task learning in deep neural networks. https://arxiv.org/abs/1706.05098

Ryan D, Denman S, Fookes CB, et al., 2010. Crowd counting using multiple local features. Proc Digital Image Computing: Techniques and Applications, p.81-88. https://doi.org/10.1109/DICTA.2009.22

Sam DB, Babu RV, 2018. Top-down feedback for crowd counting convolutional neural network. AAAI Conf on Artificial Intelligence, p.7323-7330.

Sam DB, Surya S, Babu RV, 2017. Switching convolutional neural network for crowd counting. 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. IEEE Conf on Computer Vision and Pattern Recognition, p.3618-3626. https://doi.org/10.1109/CVPR.2018.00381

Shang C, Ai HZ, Bai B, 2016. End-to-end crowd counting via joint learning local and global count. IEEE Int Conf on Image Processing, p.1215-1219. https://doi.org/10.1109/icip.2016.7532551

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

Shi MJ, Yang ZH, Xu C, et al., 2019. Revisiting perspective information for efficient crowd counting. IEEE Conf on Computer Vision and Pattern Recognition, p.7271-7280. https://doi.org/10.1109/CVPR.2019.00745

Sindagi VA, Patel VM, 2017a. CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. IEEE Int Conf on Advanced Video and Signal Based Surveillance, p.1-6. https://doi.org/10.1109/AVSS.2017.8078491

Sindagi VA, Patel VM, 2017b. Generating high-quality crowd density maps using contextual pyramid CNNs. 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

Viola P, Jones MJ, 2004. Robust real-time face detection. *Int J Comput Vis*, 57(2):137-154. https://doi.org/10.1109/ICCV.2001.937709

Walach E, Wolf L, 2016. Learning to count with CNN boosting. European Conf on Computer Vision, p.660-676. https://doi.org/10.1007/978-3-319-46475-6_41

Wang C, Zhang H, Yang L, et al., 2015. Deep people counting in extremely dense crowds. ACM Int Conf on Multimedia, p.1299-1302. https://doi.org/10.1145/2733373.2806337

Wang LY, Yin BQ, Guo AX, et al., 2018. Skip-connection convolutional neural network for still image crowd counting. *Appl Intell*, 48:3360-3371. https://doi.org/10.1007/s10489-018-1150-1

Wang LY, Yin BQ, Tang X, et al., 2019. Removing background interference for crowd counting via de-background detail convolutional network. *Neurocomputing*, 332:360-371. https://doi.org/10.1016/j.neucom.2018.12.047

Xie WX, Peng YX, Xiao JG, 2014. Weakly-supervised image parsing via constructing semantic graphs and hypergraphs. Proc 22nd ACM Int Conf on Multimedia, p.277-286. https://doi.org/10.1145/2647868.2654910

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

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

Zhu C, Peng YX, 2016. Group cost-sensitive boosting for multi-resolution pedestrian detection. 30th AAAI Conf on Artificial Intelligence, p.3676-3682.

Jian PU received his Ph.D. degree from Fudan University, China, in 2014. He was a postdoctoral researcher at the Institute of Neuroscience, Chinese Academy of Sciences from 2014 to 2016. He was an associate professor at the School of Computer Science and Technology, East China Normal University in China from 2016 to 2019. Currently, he is a professor at the Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, China. He is a corresponding expert of *Frontiers of Information Technology & Electronic Engineering*. His current research interests include machine learning, computer vision, autopilot, and medical image analysis.