LEARNING A PERSPECTIVE-EMBEDDED DECONVOLUTION NETWORK FOR CROWD COUNTING

Muming Zhao [1,2], *Jian Zhang* [2], *Fatih Porikli* [3], *Chongyang Zhang* [1*], *Wenjun Zhang* [1]

[1] Shanghai Jiao Tong University, [2] University of Technology, Sydney, [3] Australian National University
muming.zhao@student.uts.edu.au, jian.zhang@uts.edu.au,
fatih.porikli@anu.edu.au, {*sunny_zhang, zhangwenjun}@sjtu.edu.cn

ABSTRACT

We present a novel deep learning framework for crowd counting by learning a perspective-embedded deconvolution network. Perspective is an inherent property of most surveillance scenes. Unlike the traditional approaches that exploit the perspective as a separate normalization, we propose to fuse the perspective into a deconvolution network, aiming to obtain a robust, accurate and consistent crowd density map. Through layer-wise fusion, we merge perspective maps at different resolutions into the deconvolution network. With the injection of perspective, our network is driven to learn to combine the underlying scene geometric constraints adaptively, thus enabling an accurate interpretation from high-level feature maps to the pixel-wise crowd density map. In addition, our network allows generating density map for arbitrary-sized input in an end-to-end fashion. The proposed method achieves competitive result on the WorldExpo2010 crowd dataset.

Index Terms— crowd counting, deconvolution network, perspective

1. INTRODUCTION

Counting pedestrians and measuring crowd density play an essential role for crowd monitoring applications including physical security, public space management, and retail space design [1]. Traditional detection-based methods attempt to detect individuals in the crowd via either direct localization or trajectory clustering [2, 3]. Detection-based methods often suffer from severe occlusions, cluttered backgrounds, and drastic illumination changes. As an alternative, regression-based counting methods recently gain more attention and reported state-of-the-art performance [4, 5]. By learning a mapping function from feature representations of crowd segments to corresponding counts, regression-based methods emphasize the holistic depiction of the crowd, sidestepping the challenging task of localizing individuals in complex scenes.

This work was partly funded by NSFC (No.61571297, 61521062, 61420106008), State Key Research and Development Program (2016YFB1001003), the 111Program (B07022), and STCSM (14XD1402100).

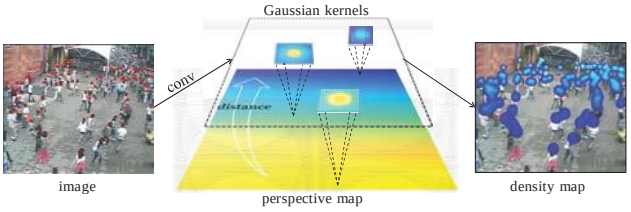

Fig. 1. Generation of crowd density map of an image. The right-side density map is obtained by the convolution of the left-side sample image and the location-aware Gaussian kernels, whose size are determined by the perspective map. Best viewed in color.

Perspective distortions need to be compensated in regression-based crowd counting methods. Due to perspective variations in surveillance scenes, features extracted from objects close to the camera will account for a larger portion of the scene than that extracted from objects far away [4]. To mitigate this issue, perspective normalization is applied before local features are fed into the regression function [1, 6, 7]. However, when the normalization is imposed as a weight for each pixel, the estimation performance becomes very sensitive to inherent normalization errors.

Recently, deep learning has shown strong performance in various visual understanding tasks [8, 9]. As a consequence, it has also been introduced for crowd counting. Still, how to incorporate scene perspective under deep learning frameworks is remaining an open question. To this end, Zhang *et al.* [5] make the first attempt by extracting candidate training/test patches based on perspective. Each patch is extracted at a size proportional to the perspective value at its corresponding location and then normalized into a fixed size. In this way, scale variations of people are compensated outside the network. However, this scheme cannot be naturally merged into the deep architecture, thus it is laborious not only in the training phase but also in the testing stage. It is also inevitable that spatial and contextual information will lose after warping patch proposals [10].

We have also observed that density map based regres-

sion, which has been a common framework of crowd counting methods [7, 11, 5], is closely related to the perspective. A crowd density map assigns each pixel a likelihood score of being a part of the crowd in the input image. As for the ground truth of regression functions, density maps are generated via the convolution of correspondingly-sized Gaussian kernels at each annotation point of pedestrians with the original image. Figure 1 illustrates this process. To accurately model people in various sizes, Gaussian kernel parameters are selected based on the perspective. Our intuition is that the regression objectives implicitly encode the scene perspective, thus incorporating perspective directly in the inference process of the neural network would provide imperative guidance, boosting the density map accuracy.

Motivated by the above observation that the ground truth regression objective (crowd density map) is generated based on the perspective, and also to fully exploit perspective within deep learning framework, we propose a perspective-embedded deconvolution network for crowd counting. Instead of imposing perspective correction and feature learning process separately, we employ perspective conjuncted with the deconvolution network jointly. This allows more accurate interpretation from high-level feature maps to the crowd density map. Our network is built on top of convolution layers. We construct a three-layer perspective pyramid and incorporate each of them into the deconvolution network by inserting multiple fusion layers. Unlike previous deep learning based methods that produce downsamples output and use hard-coded interpolations, our network generates output that has the same size as the input, in an end-to-end fashion. This in-network guided upsampling can also be viewed as a generalization of existing methods.

In the next section, we give an overview of the related work on density map based crowd counting, and recent approaches for object counting using convnets. Then, we introduce our architecture with perspective-embedded deconvolution network, and describe our experimental settings. Finally, we present experiment results on the WorldExpo2010 dataset.

2. RELATED WORK

Counting by regression Many significant methods have been proposed for object counting. These methods can be mainly classified into three groups: counting by detection, counting by trajectory clustering and counting by regression. However, counting by detection or trajectory clustering [3, 2] methods are fragile in crowded scenes with severe occlusion and clutter background. As an alternative, counting by regression methods [7, 4] learn a regression function from image features to the corresponding count number and avoid the hard task of localization of individuals. One remarkable work is counting through density estimation [1]. The key idea is to learn a continuously-valued density for each pixel, denoting its probability of being as part of the object. This strategy

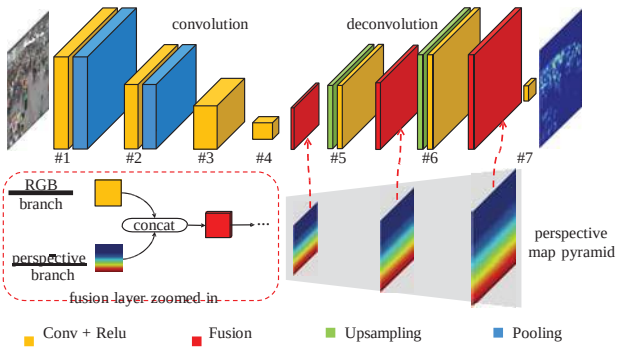

Fig. 2. Structure of the proposed network. Layer type is denoted by different colors. The fusion layer is implemented by concatenation, which is shown in the dashed box. Best viewed in color.

avoid the foreground segmentation task in traditional counting by regression methods [4] , which is very challenging with the clutter background. Based on this innovation, Fiaschi *et al.* [11] pose density estimation as a structure learning problem using random forest. In [12] an interactive system is proposed based on the simple ridge regression model instead of the costly regression model used in [1].

Counting by deep learning Deep learning has demonstrated powerful ability in many visual understanding tasks [8, 13] and has also been introduced into counting. Zhang *et al.* [5] first propose to learn the density map based on convolution neural network. However, their network incorporates the fully connected layer and is subject to fix-sized inputs, which is not very sufficient. Following this work, a multi-column CNN is proposed in [14] by stacking the feature maps generated by several convolution networks with different filter sizes. Similar network-stacking fashion is employed in [15], however with different input sizes additionally. These two methods both use the fully convolution network (FCN) to facilitate the end-to-end training process. However, due to the existence of pooling layers, the output density map has a down-sampled resolution and needs additional post-processing steps (*e.g.* interpolation).

Deconvolution network Recently, deconvolution network, which learns to upsample coarse feature maps for detailed information recovery, has gained much more attention on visual tasks with dense outputs. Long *et al.* [13] incorporate one deconvolution layer into the fully convolution network, to bilinearly upsample the coarse outputs to pixel-dense outputs. Noh *et al.* [16] dig deeper and stack more deconvolution layers on top of a vgg-16-layer net. Through layer-wise upsampling, detailed structures of objects can be identified more accurately. As far as we know, deconvolution network has not been explored for crowd counting yet, which similarly desires a dense output (*i.e.* crowd density map). It is expected the in-network upsampling with the deconvolution network

will not only benefit the regression accuracy, also will facilitate the directly full-resolution output despite the existence of max pooling layers in the network. We will show the benefits gained by the introducing of deconvlotuion network in our experiment in section 5.

3. CROWD COUNTING MODEL

Our objective is to solve the crowd counting problem given the RGB images and the perspective maps for different surveillance scenes. In this context, perspective value of each pixel denote the number of pixels in each location corresponding to $1m$ in practice. To this end, we aim to learn a regression function which maps the input RGB-P images to a dense crowd density map [1]. Denote $X_i (X_i \in \mathbb{R}^{H\times W\times D}$, where H, W and D denote the height, width and channels of the input image) as the i-th input image, the density estimation problem can be formulated as:

$$D_i^{pred} = F(X_i; \Theta) \tag{1}$$

where Θ is the learned parameter set by the proposed network. Given the positions of annotation dots for each object, the ground truth density map is defined as a summation of all the Gaussian kernels centering at each center of the objects. Due to the varying sizes of pedestrians caused by perspective distortion, it is necessary to incorporate specific scene geometric information to cover the size variations. Location-aware Gaussian functions with different kernel parameters are applied to each annotation dot respectively[1]. For each pixel p the ground truth density is defined as:

$$D_i^{gt}(p) = \sum_{A^t \in A_i} \mathcal{N}(p; A_i^t, \sigma_t), \ with\ \sigma_t = \alpha M(p) \tag{2}$$

where A_i^t denotes the annotation information of the t-th object in the annotation set A_i, for the i-th image. The Gaussian kernels are parameterized with σ_t, which is a scaling of the perspective value $M(p)$. In our experiment α is set to 0.15. Note the summation of the density value over all the pixels should be equal to the total number of the annotation dots C_i^{gt} in the image. Visualizations of location-aware Gaussian kernels and the density map are shown in Figure 1.

4. PERSPECTIVE-EMBEDDED DECONVOLUTION NETWORK

4.1. Overview

An overview of the proposed perspective-embedded deconvolution network for counting tasks is shown in Figure 2. The network first comes with the convolution part extracting crowd features. On top of the convolution layers, we add the deconvolution network, in which feature maps are interpreted to the full-image resolution density map in a learning-to-upsampling fashion. Furthermore, We consider to fuse perspective maps at different resolutions into the deconvolution network, driving the network to adaptively learn to combine the underlying scene constraints for more consistent estimation.

$L2$ loss between the estimated and ground truth density maps is used to train our netowrk:

$$L_{density}(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \left\| F(X_i; \Theta) - D_i^{gt} \right\|^2 \tag{3}$$

At inference, the loss layer is removed and the output of last 1×1 convolution layer of the network is the estimated density map.

4.2. Baseline model: the counting FCN

We first deploy an effective fully convolution architecture [13] as a baseline model (CFCN). Shown in Figure 2, the CFCN network constitutes layers from conv1 to conv4, with filter sizes of $32\ 7\times 7\times 3, 32\ 7\times 7\times 32, 64\ 5\times 5\times 32$ for the first three layers. We replace the three fully connected layers in [5] with a 1×1 convolution layer (*i.e.* conv4) in our network for feature aggregation. All the previous conv layers are followed by rectified linear units (RELU). Max pooling layers with a 2×2 kernel are followed after Conv1 and Conv2. All the convolution layers are accordingly padded to keep the spatial resolution.

4.3. Deconvolution network

On top of CFCN, we add two deconvolution layers and build the deconvolution network (CFCN-DCN). conv5 with filter size 5×5 and conv6 with filter size 7×7 are learnable kernels for precisely dense output. The employment of the two deconvolution layers is mainly based on two considerations: 1) Instead of directly upsampling the feature maps by a factor of 4, this hierarchical fashion aggregates information at different levels and enables smooth estimation. 2) Introducing the deconvolution network will benefit learning of the underlying structural information between pixels, thus enabling more accurate density estimation. With CFCN-DCN, a full-resolution output map is directly accessible for arbitrary-sized inputs. Experimental results in Section 5 of CFCN-DCN demonstrate the effectiveness.

4.4. Perspective fusion

Unlike the traditional approaches that utilize the perspective as a separate normalizer, we consider to fuse the perspective into the network, driving the network to learn to compensate the distortions brought by the perspective. The perspective

fusion is a key ingredients of our proposed network. To this end, the most intuitive way is to directly stack the perspective map with the RGB image as an additional data channel. In this way, modification is only occurred to the first convolution layer by changing the filter depth from 3 to 4. Although simply to implement, it does not provide significant improvement in our experiments, possibly due to the reason that the perspective information inserted at the very initial place tend to disperse during the propagation through several layers.

We propose to incorporate the perspective information during the upsampling process of the deconvolution network, to final obtain the perspective-embedded deconvolution network (PE-CFCN-DCN). A perspective map pyramid is constructed at different resolutions according to the network. Then fusion layer is implemented by direct concatenation of the feature maps from the RGB input and the correspondingly-sized perspective map. Each fusion layer is inserted before each deconvolution block for guided interpolation. A 1×1 convolution layer with depth 2 (*i.e.* conv7 in Figure 2), is attached at the final end to return the single-depth density map D_i^{pred} for the input image. With the perspective-embedded deconvolution architecture, perspective errors are naturally compensated during in the feature propagation process. The proposed network can be viewed as a generalization of the traditional 'feature extraction-perspective normalization-regression' pipeline under the framework of deep learning. Experiment results demonstrate the effectiveness of perspective fusion and the competitive results obtained by the proposed architecture.

5. EXPERIMENT

5.1. Dataset

We evaluate the proposed network through extensive experiments on the publicly available WorldExpo'10 crowd dataset [5]. The dataset is composed of 1132 annotated video sequences captured by 108 surveillances, all from Shanghai 2010 WorldExpo. The severe occlusions of crowd and large layout variations between different scenes make crowd counting on this dataset a challenging task.

The training set is composed of 3380 576×720 RGB images sampled from 103 scenes, and the test set contains 600 images from another 5 scenes, with 120 images sampled from each scene respectively. Note the 5 test scenes are disjoint with the training scenes, which desires high robustness of the counting models. All the training and test images have been annotated with the total number and the exact position of each people in the image. For each scene, a region of interest (ROI) mask is provided and the evaluation will only involves this region. Also the labeled perspective map for each scene is also incorporated.

5.2. System settings

Training For each 576×720 training image, we randomly crop 10 256×256 patches for data augmentation. Each patch is normalized by subtracting the mean value. The density maps are also correspondingly cropped from the ground truth density map of each training image. We also multiply the density map labels by a factor of 100 since pixel values represented by Gaussian kernels are too small for effective regression.

Training the deconvolution network with perspective embedding from scratch is difficult due to the fusion layers existing in the network, hence the stage-wise training strategy is utilized [13]. First we pretrain the CFCN network with downsampled 64×64 ground truth density map. In the second stage, we fix the parameters of CFCN, add and train the deconvolution part of the CFCN-DCN architecture. At last, with inserted perspective fusion layer, we tune the whole network (PE-CFCN-DCN) end-to-end. During the training process, the batch size is set to 64, and the learning rate starts from 10^{-5} and is divided by 10 after every 20 epochs. Mini-batch gradient descent and back-propagation is used to minimize the loss function. After the whole network is converged, we input the full-resolution image instead of the extracted patches to smooth the parameters.

Test During inference an arbitrary-sized image and the correspondingly perspective map can be directly input the network and obtain a full-resolution density map. Given the predicted total number C_i^{pred} and the annotated ground truth count C_i^{gt} for the i-th image in the test data set, the Mean Absolute Error (MAE) is employed to evaluate the performance.

5.3. Quantitative Results

We experiment several components of the proposed architectures to demonstrate the effectiveness: the simple fully convolution network CFCN that contains the first 4 convolution layers, the deconvolution network CFCN-DCN with added deconvolution part, and the proposed perspective-embedded deconvolution network PE-CFCN-DCN with perspective fusion. Results of the extensive experiments are reported in Table 1.

We compare our results to three methods: one that based on the traditional regression method [17], and another two based on the deep learning frame work [5, 14]. It can be observed that the proposed PE-CFCN-DCN architecture obtains the lowest average MAE on the 5 test scenes. It is notable that during the testing process of [5], training images that share similarity on perspectives and crowd distributions with test images are specially selected out and are used to further fine tune the model, which is denoted as the *Fine-tuned Crowd CNN* model in the table. However, we don't use such machinery for the test scenes. The most close result to us is the method in [14]. They stack multiple columns of CNNs with different filter sizes together to cover the object sizes in the

Table 1. Mean absolute errors of the WorldExpo'10 crowd dataset

Models	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Average
Chen *et al.* [17]	2.1	55.9	9.6	11.3	3.4	16.5
Fine-tuned Crowd CNN [5]	9.8	14.1	14.3	22.2	3.7	12.9
MCNN [14]	3.4	20.6	12.9	13.0	8.1	11.6
CFCN	6.9	22.7	16.2	14.2	7.7	13.5
CFCN-DCN	4.6	18.8	17.2	13.5	4.8	11.8
PE-CFCN-DCN	4.0	16.6	17.3	13.9	4.1	**11.2**

images, which can also be viewed as an implicit compensation of the perspective distortions.

We also demonstrate the improvement brought by each component through ablation experiment. Several conclusions can be drawn by analyzing the quantitative results. Firstly, the introducing the deconvolution network is beneficial for density estimation. The average MAE of the baseline model CFCN is significantly improved when the deconvolution part is added on top. Secondly, with perspective map fusion of the model PE-CFCN-DCN, the average MAE further dropped down and the best performance is achieved across all the models. This indicates that the injection of perspective is able to drive the network to learn to incorporate the underlying scene constraints and tune the weights accordingly, thus enabling more consistent density estimation result.

5.4. Qualitative Results

The density estimation results are shown in Figure 3. The counting results for each single image in the 5 test scenes are plotted together with the ground truth counts for direct comparison. The derived density map for the sample image of each scene are also shown. Without foreground segmentation, the proposed network is still able to distinguish between the crowd and clutter background, and derive accurate counts for most of images. It seems that results of scene 3 is a little inferior compared with other scenes, and the errors occurred mainly due to the underestimation of crowd under the awnings. It can be observed that the illumination is this area is very dim and also the crowd is in severe occlusions and extremely small sizes, which increase the difficulty for the convolution network to accurately extract desired features. As a consequence, the density interpretation process of the deconvolution network is influenced. With increased depth of the convolution network and more robust feature abstraction ability, this problem could be alleviated.

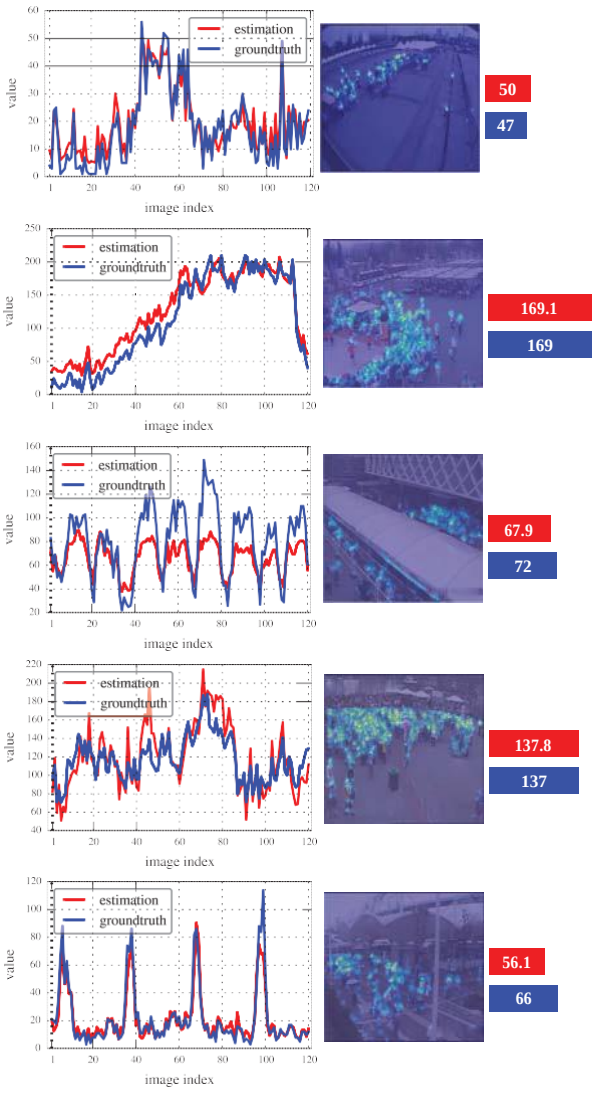

Fig. 3. Density estimation and counting results on the WorldExpo'10 dataset. The first column shows result curve for each of the 5 test scenes. The second column respectively shows one sample image masked by the estimated density map for each scene. The last column lists the estimated and ground truth count of the sample image.

6. CONLCUSION

In this paper we propose a perspective-embedded deconvolution network for crowd counting problem. The proposed network specially exploits and combines the powerful ability of feature learning of convolution network, and the guidance in-

formation provided by in-network perspective normalization with the deconvolution network. With extensive experiments, we show that the introducing of deconvolution network is beneficial to perform adaptive up-sampling. Furthermore, by fusing the scene constraints information underlying in the perspective map with the deconvolution network, more accurate location-aware densities and counts could be obtained.

7. REFERENCES

[1] Victor Lempitsky and Andrew Zisserman, "Learning to count objects in images," in *Advances in Neural Information Processing Systems*, 2010, pp. 1324–1332.

[2] Zhe Lin and Larry S Davis, "Shape-based human detection and segmentation via hierarchical part-template matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 604–618, 2010.

[3] Oliver Sidla, Yuriy Lypetskyy, Norbert Brandle, and Stefan Seer, "Pedestrian detection and tracking for counting applications in crowded situations," in *2006 IEEE International Conference on Video and Signal Based Surveillance*. IEEE, 2006, pp. 70–70.

[4] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–7.

[5] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 833–841.

[6] Sheng-Fuu Lin, Jaw-Yeh Chen, and Hung-Xin Chao, "Estimation of number of people in crowded scenes using perspective transformation," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 31, no. 6, pp. 645–654, 2001.

[7] Dan Kong, Douglas Gray, and Hai Tao, "A viewpoint invariant approach for crowd counting," in *18th International Conference on Pattern Recognition (ICPR'06)*. IEEE, 2006, vol. 3, pp. 1187–1190.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[9] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*. Springer, 2014, pp. 346–361.

[11] Luca Fiaschi, Ullrich Köthe, Rahul Nair, and Fred A Hamprecht, "Learning to count with regression forest and structured labels," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 2685–2688.

[12] Carlos Arteta, Victor Lempitsky, J Alison Noble, and Andrew Zisserman, "Interactive object counting," in *European Conference on Computer Vision*. Springer, 2014, pp. 504–518.

[13] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[14] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 589–597.

[15] Daniel Onoro-Rubio and Roberto J López-Sastre, "Towards perspective-free object counting with deep learning," in *European Conference on Computer Vision*. Springer, 2016, pp. 615–629.

[16] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.

[17] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang, "Feature mining for localized crowd counting," in *BMVC*, 2012, vol. 1, p. 3.