16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

Perspective-Aware Loss Function for Crowd Density Estimation

Bedir Yilmaz
Fac. of Info. Sci. and Tech.
National University of Malaysia, Malaysia
p90306@siswa.ukm.edu.my

Ven Jyn Kok
Fac. of Info. Sci. and Tech.
National University of Malaysia, Malaysia
vj.kok@ukm.edu.my

Mei Kuan Lim
School of Information Technology
Monash University, Malaysia
Lim.meikuan@monash.edu

Siti Norul Huda Sheikh Abdullah
Fac. of Info. Sci. and Tech.
National University of Malaysia, Malaysia
snhsabdullah@ukm.edu.my

Abstract

Estimation errors caused by perspective distortions are a long-standing problem in the domain of crowd counting. In this paper, we propose a novel loss function to allow filters in convolutional neural networks to learn features that are adaptive to the scale and perspective variation of individuals in crowd images. By exploring the crowd count error from regions close to the vanishing point of a perspective distorted image, we are able to penalize under-estimations. This is useful to train a network that is robust against perspective distortion for accurate density estimation. The proposed method is scene-independent and can be applied effectively to crowd scene with a variety of physical layout. Extensive comparative evaluations demonstrate that our proposed method achieves significant improvement over the state-of-the-art approaches on the challenging ShanghaiTech and UCF-QNRF datasets.

1 Introduction

One of the most intriguing researches of human behaviour focuses on the crowd phenomenon. What makes crowd phenomenon interesting is that each individual is self-organized but they tend to act and gather as a united mass without prior awareness [1]. This notion of crowd collectiveness is commonly being applied in visual crowd analysis for crowd segmentation [2, 3], crowd behaviour analysis [4, 5] and crowd density estimation [6, 7]. Among these analyses, substantial effort have been made in recent years toward crowd density estimation, largely in response to rising anxieties of recurrent fatal crowd tragedy at mass gatherings, e.g. pilgrimages [8] and parades [9]. The endeavour is further intensified to meet the need for a proactive crowd management to anticipate disasters.

The complexity of estimating crowd density increases disproportionately in relation to the number of individuals in a crowd [11]. This is not surprising since individuals in crowd are often severely occluded due to excessive number of individuals in the scene. To further complicate the matter, individuals in crowd scenes often experience drastic variations in their visual appearance owing to different illumination condition and camera orientations. In addition, crowdedness and distribution of individuals are rarely uniform due to the viewpoint of the scene captured and/or the un-

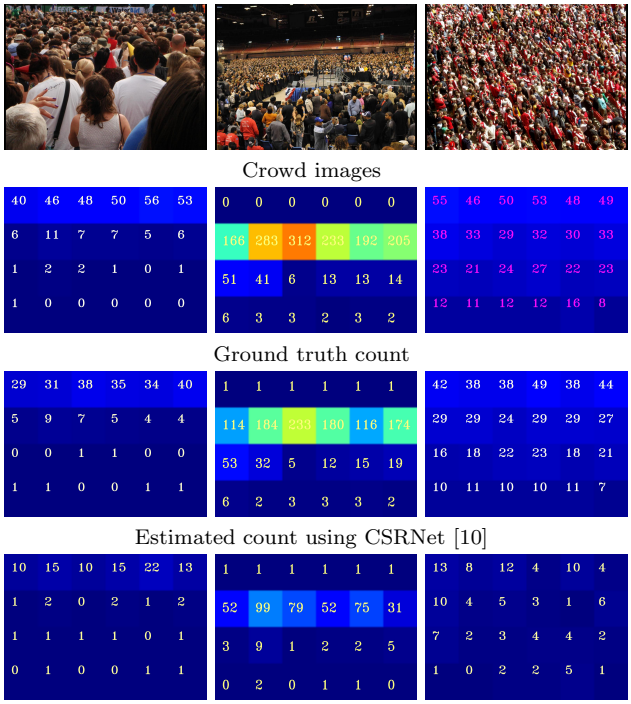

Figure 1: Crowd images with variations in terms of perspective, illumination, crowd density and physical layout of the environment with the respective ground truth count, estimated count using CSRNet [10] and absolute difference between ground truth count and estimated count. Crowd regions close to the vanishing point of perspective distorted images are regions with highly inaccurate density estimation, i.e. highly under-estimated. Best viewed in colour.

constrained physical layout of the environment. Nevertheless, one of the foremost challenges in crowd density estimation is the effects of perspective distortion owing to camera orientation. For instance, the images in Figure 1(row 1) depicts that the individuals who are closer to the camera view appear larger than those further away from the camera.

Traditional crowd density estimation approaches address the variation of visual appearance problem by relying on handcrafted features to count by detecting each individual in a crowd [12, 13, 14] or count by regression [15, 16, 17]. The latter approaches obviate the need to segregate individuals by estimating the crowd density based on collective description of crowd patterns (e.g. texture features). The crowd density estimation problem is formulated as estimating a continuous density function whose integral over any crowd image region gives the count of individuals within that region [18]. In order to address the problem of perspective distortion, image space is divided into different pixel-grid or multiscale pixel-grid where each grid is modelled by a regression function. This technique has been extended to convolutional neural network (CNN) based approaches [19, 20, 21, 22, 23] in the form of multi receptive fields provided by convolutional filters of different sizes. Commonly, these filters are adopted using multi-column CNN architecture.

Despite the promising results of these techniques, as noted by Li et al. [10], each receptive field in such network learns nearly identical features. In essence, the filters are unable to adapt to the scale and perspective variation in crowd scenes due to perspective distortion. Moreover, in the presence of low and high density crowd images, multi-column architectures tend to either under-estimate or over-estimate crowd count [23]. In order to overcome these problems, a deeper, single column network architecture [10] has been proposed and significant improvements have been achieved. Nonetheless, similar to multi-column architectures, as illustrated in Figure 1(row 3 and row 2), this network still requires perspective information to cope with perspective distortion in crowd scenes.

Interestingly, we observe in Figure 1(row 4) that crowd regions close to the vanishing point of perspective distorted images are regions with highly inaccurate density estimation, i.e. highly under-estimated. Individuals in these regions are represented with fewer pixels per target, thus, making it difficult to discern each individual. This observation can be embedded in a loss function that serves to improve network training. This is to allow each kernel filter to learn features that are adaptive to the large scale and perspective variation of individuals in crowd images. Accordingly, the overall network is robust to perspective distortion in crowd images for density estimation. Specifically, in this work, the error counts from regions close to the vanishing point are used to further penalize underestimation of crowd count prominent in the region. To our knowledge, the notion of using loss function to cope with perspective distortion is generally unprecedented in the existing crowd density estimation studies.

The main contribution of this paper is to propose a novel loss function that enables accurate density estimation on perspective distorted crowd images. The loss function allows each filter in a network to learn features that are adaptive to the scale and perspective variation of individuals in crowd images. Instead of designing multi-column architectures like most existing works [21, 22], we formulate an approach to harness the error count in perspective distorted crowd images to improve crowd density estimation performance. Extensive experimental evaluations on ShanghaiTech [21] and UCF-QNRF [7] datasets in Section 4 demonstrate the effectiveness of the proposed loss function to allow receptive fields to adapt to perspective distortion in crowd images. The proposed approach achieved state-of-the-art performance for crowd density estimation.

2 Related Work

Existing crowd density estimation can be divided into two main approaches. The first approach infers crowd count by tracking or detecting individuals in a scene. For instance, Rabaud and Belongie [12] perform clustering of coherent trajectories to determine the number of individuals in the scene. Using a similar concept, Ge and Collins [13] proposed a Bayesian marked point process to detect individuals in crowd for density estimation. Such approaches require that individuals in the scenes are clearly visible. However, in high-density crowd scenes, tracking and detection tend to fail due to severe occlusion and background clutter. Therefore, the first approach generally works well in low-density crowd scenes.

In order to eliminate the need to track or detect individuals in crowd scenes, most existing density estimation approaches [16, 24, 17] emphasise on extracting a set of low-level image features (i.e. texture features) and learn a direct mapping from the features to estimate crowd density. Chan et al. [24] propose to extract dynamic texture features and map the features to a number of people by using Bayesian regression. However, a common problem in regression approach is perspective distortion where features of individuals extracted at different depth in an image would have high discrepancy in crowd density value. One of the common approaches to deal with this problem is to divide the image space into different pixel-grids, where each pixel-grid is modelled by a regressor to mitigate the effects of perspective distortion. For instance, Chen et al. [16] and Idrees et al. [17] rely on modelling of local features to analyse pixel-grids for density estimation.

The idea of pixel-grids-based regression approach has been implemented in CNN-based approaches in the form of multi-column [21, 22, 23] or multi-scale architectures [20]. The multi receptive fields provided by the

convolutional filters of different sizes are dedicated to different types of scales in perspective distorted crowd scenes. Despite the flexibility of multi receptive fields and promising results, Li et al. [10] show that the filters, in essence, learn similar features. In an effort to cope with perspective distortion for density estimation, Li et al. [10] propose a deep network with dilated kernels to generate high-quality density maps for density estimation. In contrast to the aforementioned approaches, we propose a novel loss function to curb the effects of perspective distortion for a more accurate density estimation. Error counts from crowd regions close to the vanishing point in images are used to penalize under-estimation of crowd count.

3 Proposed Method

In this paper, we propose a novel loss function to deal with the problem of perspective distortion in crowd images for crowd density estimation. The fundamental idea of the novel loss function is to improve network training with the aim to have kernel filters that are adaptive to the scale and perspective variations of individuals in crowd images. This is achieved by penalizing under-estimation of crowd count which is prominent in crowd regions close to the vanishing point in crowd images.

3.1 Perspective-Aware Loss

In a perspective distorted crowd image, individuals that are nearer to the vanishing point in the image are usually represented with only few pixels per individual. Hence, it is extremely challenging to discern individuals within these regions. As illustrated in Figure 1, we observe from experiments that the regions are commonly under-estimated. Therefore, in this work, we incorporate the error count from these regions as a loss function to improve the network training.

Formally, given a crowd image, $I \in \mathbb{R}^P$, P is the number of pixels and c^{GT} is the ground truth count of an image. To this end, we propose to learn Θ by minimizing a combination of two losses:

$$L_T(\Theta) = L_{l1}(\Theta) + L_C(\Theta) \tag{1}$$

$$L_{l1}(\Theta) = \frac{1}{2N}\sum_{i=1}^{N}\left\|Y_{I_i}(\cdot;\Theta) - Y_{I_i}^{GT}(\cdot)\right\|_1 \tag{2}$$

$$L_C(\Theta) = \sum_{i=1}^{N}\left\|[Y_{I_i}(\cdot;\Theta) - Y_{I_i}^{GT}(\cdot)] \bullet M_{I_i}\right\|_1 \tag{3}$$

where $\bullet$ denotes element-wise multiplication. $L_{l1}(\Theta)$ is the $l1$ norm that measures the absolute count difference between an estimated count and the ground truth count. $L_C(\Theta)$ is the error count from crowd regions which are close to the vanishing point of an image. N is the total number of training images. $Y_{I_i}^{GT}(\cdot)$ and $Y_{I_i}(\cdot;\Theta)$ correspond to the ground truth density map and estimated density map of image I_i. The density map is generated using geometry-adaptive kernels [21] to adaptively determine the spread parameter of Gaussian kernel based on local crowd density. Specifically, the spread parameter for each individual is based on its average distance to its neighbour. Note that however, when the estimated count of an image, c^E, is greater than the ground truth count, c^{GT}, the loss function $L_T(\Theta) = L_{l1}(\Theta)$ will be used instead.

We adopt a binary mask, M, to determine the under-estimated crowd regions which are close to the vanishing point of an image, where:

$$M_{I_i} = 1_{[Y_{I_i}^{GT} > \alpha_{I_i}]} \tag{4}$$

The indicator function $1_{[\cdot]}$ returns 1 when the value of ground truth density map, Y^{GT}, is greater than the average density per pixel in an image, α. The average density per pixel in an image, α, in this work, is defined as the number of individuals per foreground (i.e crowd) pixels, where λ is a parameter.

$$\alpha_{I_i} = \lambda \frac{c_{I_i}^{GT}}{P_{I_i} - \left\|1_{[Y_{I_i}^{GT}=0]}\right\|_1} \tag{5}$$

The loss function, L_T is optimized by backpropagating the network via stochastic gradient descent (SGD). Minimizing the loss function serves to improve network training with the aim that each kernel filter is adaptive to the large scale and perspective variation of individuals in crowd images.

3.2 Model Settings and Architecture

In all the experiments, the CNN architecture proposed by Li et al. [10] serves as the base architecture of our work. Note that however, our novel loss function is not limited to any particular base CNN. We empirically set the parameter $\lambda = 5$. We do not use any data augmentation in data preparation phase. During training, the batch size is set to 1. Implementation of the proposed framework and its training are based on the PyTorch framework.

4 Experimental Results

Evaluation on the propose novel loss function for crowd density estimation are conducted on two challenging benchmark datasets: ShanghaiTech [21] and UCF-QNRF [7]. The ground truth count for images in each dataset has been provided by the respective authors. Each individual in the images is manually annotated where each head position is marked. The crowd images vary in terms of perspective, illumination, resolution and physical layout of the environment. Most

importantly, there is a large range of crowd density between images making these dataset challenging to achieve accurate density estimation.

Similar with existing approaches [17, 21, 22, 10], the performance of the novel loss function on crowd density estimation are evaluated by accessing the similarity between the actual count and the estimated count of individuals in a scene. Specifically, we use Mean Absolute Error (MAE) and Mean Squared Error (MSE):

$$MAE = \frac{1}{M}\sum_{j=1}^{M}|c_{I_j}^{E} - c_{I_j}^{GT}| \tag{6}$$

$$MSE = \sqrt{\frac{1}{M}\sum_{j=1}^{M}(c_{I_j}^{E} - c_{I_j}^{GT})^2} \tag{7}$$

where M is the number of test images. MAE is a measure of the accuracy of the estimated crowd count across the test images, whereas MSE is used to indicate the robustness of the estimated count.

4.1 ShanghaiTech Dataset

ShanghaiTech dataset [21] is a crowd counting dataset containing $1,198$ annotated images with a total of $330,165$ individuals. There are two parts in this dataset: Part A & B which consist of 482 and 716 images, respectively. Only the images in Part A contains high-density crowd scenes with the number of individuals ranges between 33 and $3,319$. The number of individuals in Part B ranges between 9 and 578. Images in Part A were crawled from the Internet, while images in Part B are surveillance footages taken from a busy street of metropolitan areas in Shanghai.

Consistent with [21], Part A is partitioned into chunk of 300 images for training and 182 images for testing. Similarly, Part B is partitioned into chunk of 400 images for training and 316 images for testing. The comparisons on the ShanghaiTech dataset are presented in Table 1. The proposed method significantly outperforms existing state-of-the-art approaches both in MAE and MSE. The improvement of the MAE and MSE alludes that penalizing the error count from crowd regions close to the vanishing point is significant for density estimation.

Evaluation on crowd images with perspective distortion (as shown in Figure 2 and Figure 3) shows that our propose approach is able to cope with varying scale of individuals in the crowd for accurate density estimation. When perspective distortion is less prominent in the crowd images (as shown in row 3 and 4 of Figure 2), the propose approach also accurately estimate the number of individuals in the crowd. This demonstrates the effectiveness of the novel loss function in enhancing the network training process for better density estimation.

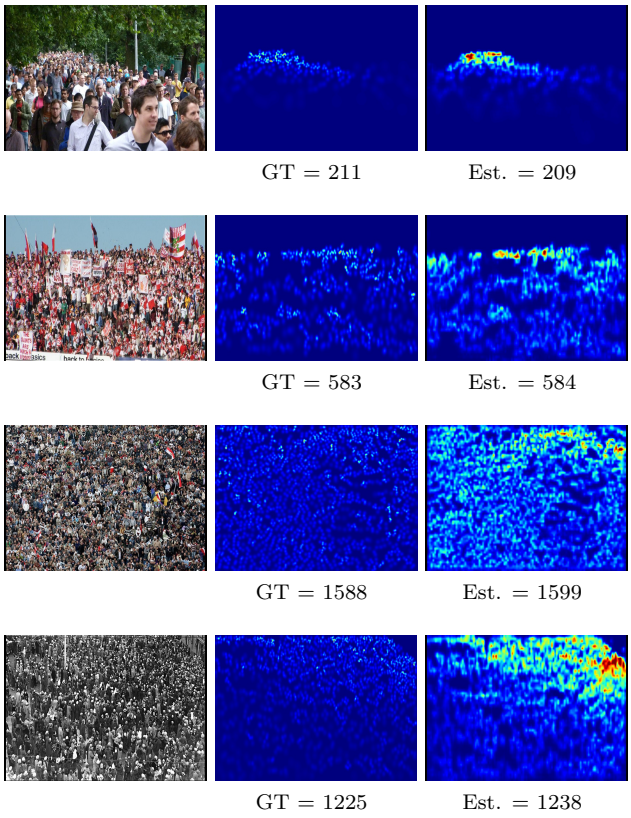

Figure 2: Example outputs on ShanghaiTech Part A dataset. (Left) Crowd images. (Centre) Ground truth density maps with the respective ground truth count. (Right) Estimated density map using proposed approach with the respective estimated count. Best viewed in colour.

4.2 UCF-QNRF Dataset

UCF-QNRF is a new large-scale crowd counting dataset consisting of $1,251,642$ individuals in $1,535$ images. The average number of individuals in the images is lower compared to the existing benchmark datasets, signifying that the images are real crowd scenes captured in the wild consisting of background clutters such as sky, buildings, roads and vegetation (see Figure 4). The average resolution in this dataset (i.e. 2013×2902) is also the largest compare to existing dataset.

Table 2 summarizes the crowd density estimation results which demonstrate that the proposed approach outperforms existing multi-column CNNs approaches. This shows that the proposed novel loss function can improve network training. This allows kernel filters to learn features that are adaptive to the large scale and perspective variation of individuals in crowd images. When compare with counting-by-detection approach [7], our proposed method achieves comparable results. Given that the average resolution in this dataset is

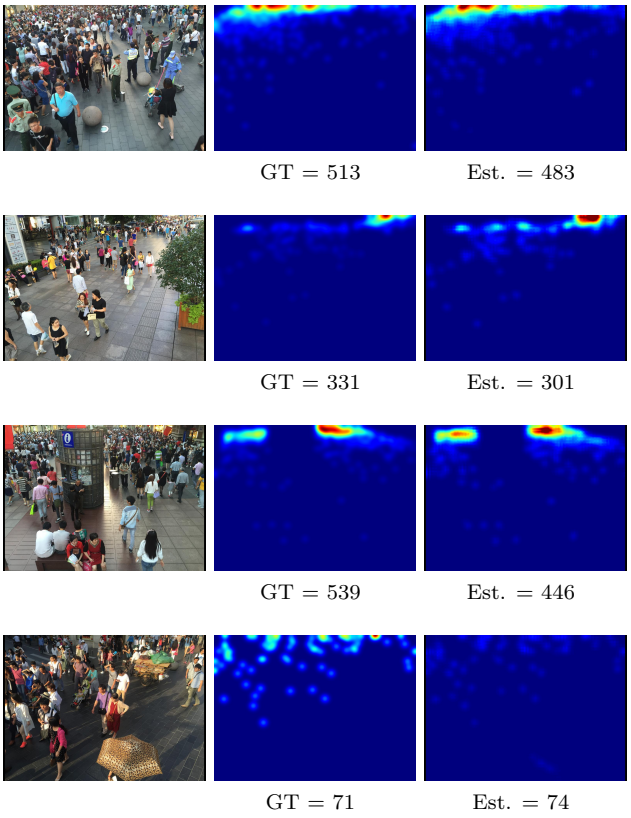

Figure 3: Example outputs on ShanghaiTech Part B dataset. (Left) Crowd images. (Centre) Ground truth density maps with the respective ground truth count. (Right) Estimated density map using proposed approach with the respective estimated count. Best viewed in colour.

2013×2902 where it is feasible for person detection based methods, the counting-by-detection approach [7] leverage on person detection count to achieve better density estimation.

5 Conclusion

In this paper, we proposed a novel loss function that incorporates the error crowd count from perspective distorted image to reduce estimation errors. We showed that crowd regions which are close to the vanishing point of perspective distorted images are regions with highly inaccurate density estimation, i.e. highly under-estimated. This observation can be embedded in a loss function which allows filters to learn features that are adaptive to the scale and perspective variation of individuals in crowd images. In contrast to the existing methods that focus on designing multi-column architecture, this work focuses on investigating the error count in perspective distorted crowd images to achieve lower crowd density estimation error. The proposed loss function is not limited by any particular base CNN. Extensive comparative evaluations demonstrate that our proposed method achieves significant improvement in reducing estimation errors as compared to the state-of-the-art approaches on the challenging ShanghaiTech and UCF-QNRF datasets.

Table 1: Comparative results with state-of-the-art approaches on ShanghaiTech dataset. The proposed method significantly outperforms other methods in reducing the MAE and MSE.

	Part_A		Part_B	
Method	MAE	MSE	MAE	MSE
Zhang et. al. [19]	181.8	227.7	32.0	49.8
Marsden et. al. [25]	126.5	173.5	23.8	33.1
Cascaded-MTL [26]	101.3	152.4	20.0	31.1
Switching-CNN [22]	90.4	135.0	21.6	33.4
CP-CNN [23]	73.6	106.4	20.1	30.1
CSRNet [10]	68.2	115.0	10.6	16.0
MCNN [21]	110.2	173.2	26.4	41.3
Proposed Method	**65.22**	**101.22**	**8.40**	**10.22**

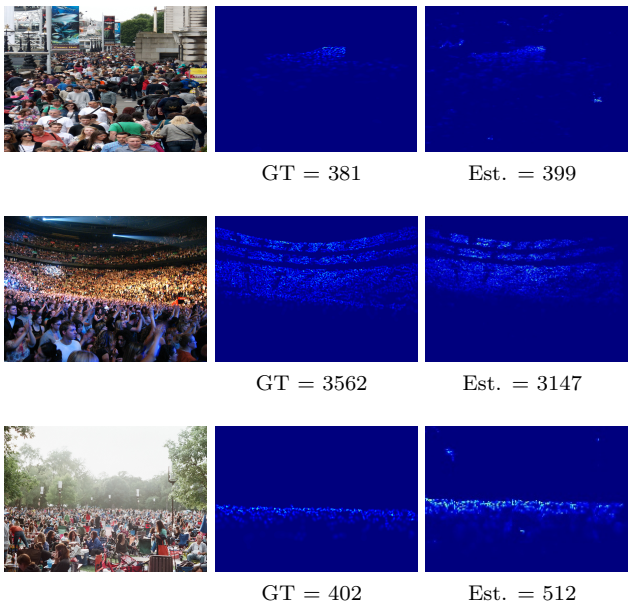

Figure 4: Example outputs on UCF-QNRF dataset. (Left) Crowd images. (Centre) Ground truth density maps with the respective ground truth count. (Right) Estimated density map using proposed approach with the respective estimated count. Best viewed in colour.

6 Acknowledgements

This research is supported by the Arus Perdana grant AP-2017-005/2 and GGPM grant GGPM-2017-024, from the National University of Malaysia (UKM).

Table 2: Comparative results with state-of-the-art approaches on UCF-QNRF dataset.

Method	MAE	MSE
Idrees *et al.* (2013) [17]	315	508
MCNN [21]	277	426
Encoder-Decoder [27]	270	478
Cascaded-MTL [26]	252	514
Switching-CNN [22]	228	445
Idrees *et al.* (2018) [7]	**132**	**191**
Proposed Method	206	348

References

[1] Ven Jyn Kok, Mei Kuan Lim, and Chee Seng Chan. Crowd behavior analysis: A review where physics meets biology. *Neurocomputing*, 177:342–362, 2016.

[2] Saad Ali and Mubarak Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR*, pages 1–6, 2007.

[3] Kai Kang and Xiaogang Wang. Fully convolutional neural networks for crowd segmentation. *arXiv preprint arXiv:1411.4464*, 2014.

[4] Mei Kuan Lim, Ven Jyn Kok, Chen Change Loy, and Chee Seng Chan. Crowd saliency detection via global similarity structure. In *ICPR*, pages 3957–3962, 2014.

[5] Jing Shao, Chen Change Loy, Kai Kang, and Xiaogang Wang. Crowded scene understanding by deeply learned volumetric slices. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3):613–623, 2017.

[6] Ven Jyn Kok and Chee Seng Chan. Granular-based dense crowd density estimation. *Multimedia Tools and Applications*, 77(15):20227–20246, 2018.

[7] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *ECCV*, pages 532–546, 2018.

[8] Knut Haase, Habib Zain Al Abideen, Salim Al-Bosta, Mathias Kasper, Matthes Koch, Sven Müller, and Dirk Helbing. Improving pilgrim safety during the hajj: an analytical and operational research approach. *Interfaces*, 46(1):74–90, 2016.

[9] Dirk Helbing and Pratik Mukerji. Crowd disasters as systemic failures: analysis of the love parade disaster. *EPJ Data Science*, 1(1):7, 2012.

[10] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*, pages 1091–1100, 2018.

[11] Haroon Idrees, Khurram Soomro, and Mubarak Shah. Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10):1986–1998, 2015.

[12] Vincent Rabaud and Serge Belongie. Counting crowded moving objects. In *CVPR*, pages 705–711, 2006.

[13] Weina Ge and Robert T Collins. Marked point processes for crowd counting. In *CVPR*, pages 2913–2920, 2009.

[14] Shih-Shinh Huang and Chun-Yuan Chen. Crowd pedestrian detection using expectation maximization with weighted local features. In *MVA*, pages 177–180, 2017.

[15] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, pages 1–7. IEEE, 2008.

[16] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *BMVC*, number 2, page 3, 2012.

[17] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *CVPR*, pages 2547–2554, 2013.

[18] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in neural information processing systems*, pages 1324–1332, 2010.

[19] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *CVPR*, pages 833–841, 2015.

[20] Daniel Onoro-Rubio and Roberto J López-Sastre. Towards perspective-free object counting with deep learning. In *ECCV*, pages 615–629. Springer, 2016.

[21] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, pages 589–597, 2016.

[22] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *CVPR*, volume 1, page 6, 2017.

[23] Vishwanath A Sindagi and Vishal M Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *ICCV*, pages 1879–1888, 2017.

[24] A. B. Chan and N. Vasconcelos. Counting people with low-level features and bayesian regression. *IEEE Transactions on Image Processing*, 21(4):2160–2177, 2012.

[25] Mark Marsden, Kevin McGuinness, Suzanne Little, and Noel E O'Connor. Fully convolutional crowd counting on highly congested scenes. *arXiv preprint arXiv: 1612.00220*, 2016.

[26] Vishwanath A Sindagi and Vishal M Patel. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *AVSS*, pages 1–6. IEEE, 2017.

[27] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.