LDCrowdNet: A Lightweight Network for Dense Crowd Counting Supplementary Material

Shahbaz Ahmad[0009−0003−3434−8291], Yogesh Aggarwal[0000−0001−8263−9307], and Prithwijit Guha[0000−0003−2885−0026]

Department of Electronics & Electrical Engineering
Indian Institute of Technology Guwahati {shahbaz,yogesh,pguha}@iitg.ac.in

Abstract. Counting crowds is essential for applications such as urban planning, traffic monitoring, and video surveillance. However, accurately counting the number of individuals (people) is challenging within an image, especially for real-time applications. Recently developed Dense Crowd Counting systems incur high computational costs due to the utilize of computation intensive deep networks. Accordingly, this work proposes a novel lightweight model named Lightweight Network for Dense Crowd (LDCrowdNet), specifically designed for crowd counting and crowd density estimation. LDCrowdNet integrates components such as a ShuffleNetV2 backbone network, a proposed novel enhanced ECANet for efficient channel attention, an introduced feature pyramid network (FPN), and a context module (CM) that serves as a Density Map Generator (DMG). Extensive experiments have been conducted across multiple benchmark datasets, illustrate that LDCrowdNet surpasses current state-of-the-art methods in terms of accuracy and robustness. This superiority is evident across multiple crowd counting datasets: ShanghaiTech, UCF_CC_50, and UCF-QNRF.

Keywords: Crowd Counting · Density Estimation · Convolutional Neural Networks (CNN) · Lightweight Network.

1 Ground Truth Generation

This work adopts the approach proposed in [3] and utilizes the geometry-adaptive kernels to handle the extremely crowded scenes. To create the ground truth, we apply a Gaussian kernel (normalized at 1) for blurring each head annotation, taking into account the spatial dispersion across all images in each dataset. The parameters associated with the ground truth generation are presented in Table 1.

Assuming an annotated head is present at pixel x_i, this can be represented as a delta function:

$$\delta(x - x_i) = 1 \tag{1}$$

Table 1. The ground truth generation methods for various datasets.

Dataset	Generating Method
ShanghaiTech Part_A [3]	Geometry-adaptive kernels
UCF_CC_50 [2]	
ShanghaiTech Part_B [3]	Fixed Gaussian kernel: $\sigma = 15$
UCF-QNRF [1]	Fixed Gaussian kernel: $\sigma = 4$

An image containing N annotated heads can then be expressed by the following function:

$$H(x) = \sum_{i=1}^{N} \delta(x - x_i) \tag{2}$$

The geometry-adaptive kernel is described as:

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_{\sigma_i}(x), \quad \text{where} \quad \sigma_i = \beta \bar{d}_i \tag{3}$$

In the ground truth δ, we utilize $\bar{d}_i$ to represent the average distance between a head annotation and the k nearest head annotations for each target object x_i. We apply the convolution of $\delta(x - x_i)$ with a Gaussian kernel using parameter σ_i (standard deviation), where x denotes the pixel location in an image, for generating the density map. In our experiments, we adopt the configuration from [3] with $\beta = 0.3$ and $k = 3$. We adjust the Gaussian kernel to the typical head size to blur all annotations uniformly in the sparse crowd scenarios. The specific configurations for various datasets are presented in Table 1.

2 Evaluation Metrics

The Mean Absolute Error (MAE) and Mean Squared Error (MSE) are utilized for evaluation, defined as follows:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |c_i - \hat{c}_i| \tag{4}$$

$$\text{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} |c_i - \hat{c}_i|^2} \tag{5}$$

where N represents the number of testing images, c_i denotes the ground truth count, and $\hat{c}_i$ denotes the estimated (predicted) count which is described in the following manner:

$$\hat{c}_i = \sum_{h=1}^{H} \sum_{w=1}^{W} p_{h,w} \tag{6}$$

where H and W represent the height and width of the density map, respectively, while $p_{h,w}$ denotes the pixel at position (h, w) of the estimated (generated) density map, while $\hat{c}_i$ signifies the estimated count for image X_i. The total count of individuals is obtained by integrating the estimated (predicted) density map or summing the pixel values of the density image.

MAE calculates the mean absolute difference between the ground truth (actual) and estimated (predicted) counts, while MSE computes the mean of the squared differences. MAE denotes the counting accuracy and MSE denotes the robustness of our proposed model.

References

1. H. Idrees, M. Tayyab, K.A.D.Z.S.A.M.N.R.M.S.: Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
2. Idrees, H., Saleemi, I., Seibert, C., Shah, M.: Multi-source multi-scale counting in extremely dense crowd images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2547–2554 (2013)
3. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 589–597 (2016)