

# MULTI-SCALE CONVOLUTIONAL NEURAL NETWORKS FOR CROWD COUNTING

Lingke Zeng, Xiangmin Xu\*, Bolun Cai, Suo Qiu, Tong Zhang

School of Electronic and Information Engineering  
South China University of Technology, Guangzhou, China

z.lineker@gmail.com, xmxu@scut.edu.cn, caibolun@gmail.com, q.suo@foxmail.com, tony@scut.edu.cn

## ABSTRACT

Crowd counting on static images is a challenging problem due to scale variations. Recently deep neural networks have been shown to be effective in this task. However, existing neural-networks-based methods often use the multi-column or multi-network model to extract the scale-relevant features, which is more complicated for optimization and computation wasting. To this end, we propose a novel multi-scale convolutional neural network (MSCNN) for single image crowd counting. Based on the multi-scale blobs, the network is able to generate scale-relevant features for higher crowd counting performances in a single-column architecture, which is both accuracy and cost effective for practical applications. Complementary results show that our method outperforms the state-of-the-art methods on both accuracy and robustness with far less number of parameters.

**Index Terms**— Multi-scale CNN, scale-relevant architectures, crowd counting.

## 1. INTRODUCTION

Crowd counting aims to estimate the number of people in the crowded images or videos feed from surveillance cameras. Overcrowding in scenarios such as tourist attractions and public rallies can cause crowd crushes, blockages and even stampedes. It has been much significant to public safety to produce an accurate and robust crowd count estimation using computer vision techniques.

Existing methods of crowd counting can be generally divided into two categories: *detection-based methods* and *regression-based methods*.

*Detection-based* methods generally assume that each person on the crowd images can be detected and located by using the given visual object detector [1, 2, 3], and obtain the counting result by accumulating each detected person. However, these methods [4, 5, 6] need huge computing resource and they are often limited by person occlusions and complex background in practical scenarios, resulting at a relatively low robustness and accuracy.

*Regression-based* methods regress the crowd count from the image directly. Chan et al. [7] used handcraft features to translate the crowd counting task into a regression problem. Following works [8, 9] proposed more kinds of crowd-relevant features including segment-based features, structural-based features and local texture features. Lempitsky et al. [10] proposed a density-based algorithm that obtain the count by integrating the estimated density map.

Recently, deep convolutional neural networks have been shown to be effective in crowd counting. Zhang et al. [11] proposed a convolutional neural network (CNN) to alternatively learn the crowd density and the crowd count. Wang et al. [12] directly used a CNN-based model to map the image patch to its people count value. However, these *single-CNN*-based algorithms are limited to extract scale-relevant features and hard to address the scale variations on crowd images. Zhang et al. [13] proposed a *multi-column CNN* to extract multi-scale features by columns with different kernel sizes. Boominathan et al. [14] proposed a *multi-network CNN* that used a deep and shallow network to improve the spatial resolution. These improved algorithms can relatively suppress the scale variations problem, but they still have two shortages:

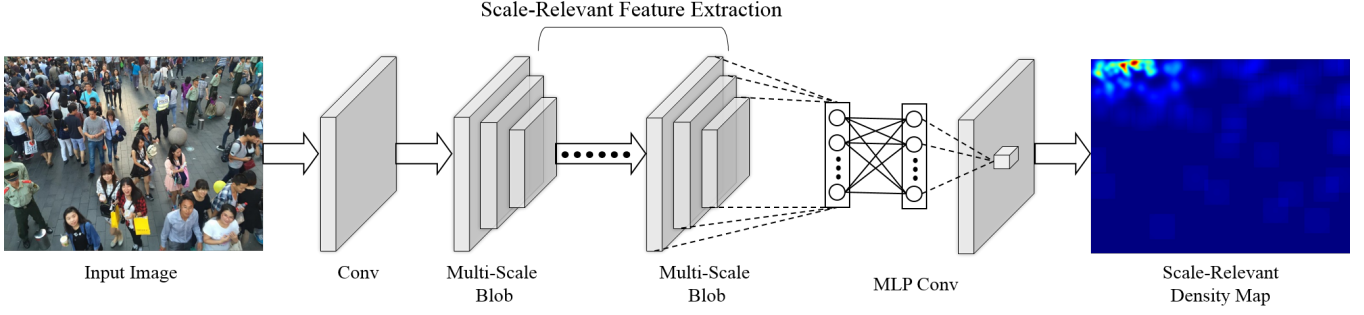
- Multi-column/network need pre-trained single-network for global optimization, which is more complicated than end-to-end training.
- Multi-column/network introduce more parameters to consume more computing resource, which make it hard for practical application.

In this paper, we propose a *multi-scale convolutional neural network (MSCNN)* to extract scale-relevant features. Rather than adding more columns or networks, we only introduce a multi-scale blob with different kernel sizes similar to the naive Inception module [15]. Our approach outperforms the state-of-the-art methods on the ShanghaiTech and UCF\_CC\_50 dataset with a small number of parameters.

## 2. MULTI-SCALE CNN FOR CROWD COUNTING

Crowd images are usually consisted of various sizes of persons pixels due to perspective distortion. Single-network is

\*Xiangmin Xu is the corresponding author.

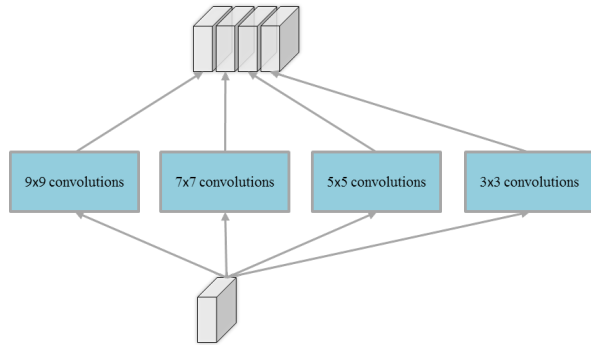


**Figure. 1.** Multi-scale convolutional neural network for crowd counting.

hard to counter scale variations with the same sized kernels combination. In [15], a Inception module is proposed to process visual information at various scales and aggregated to the next stage. Motivated by it, we designed a multi-scale convolutional neural network (MSCNN) to learn the scale-relevant density maps from original images.

### 2.1. Multi-scale Network Architecture

An overview of MSCNN is illustrated in Figure. 1, including *feature remapping*, *multi-scale feature extraction*, and *density map regression*. The first convolution layer is a traditional convolutional layer with single-sized kernels to remap the image feature. Multi-Scale Blob (MSB) is a Inception-like model (as Figure. 2) to extract the scale-relevant features, which consists of multiple filters with different kernel size (including  $9 \times 9$ ,  $7 \times 7$ ,  $5 \times 5$  and  $3 \times 3$ ). A multi-layer perceptron (MLP) [16] convolution layer works as a pixel-wise fully connection, which has multiple  $1 \times 1$  convolutional filters to regress the density map. Rectified linear unit (ReLU) [17] is applied after each convolution layer, which works as the activation function of previous convolutional layers except the last one. Since the value in density map is always positive, adding ReLU after last convolutional layer can enhance the density map restoration. Detailed parameter settings are listed in Table 1.



**Figure. 2.** Multi-scale blob with different kernel size.

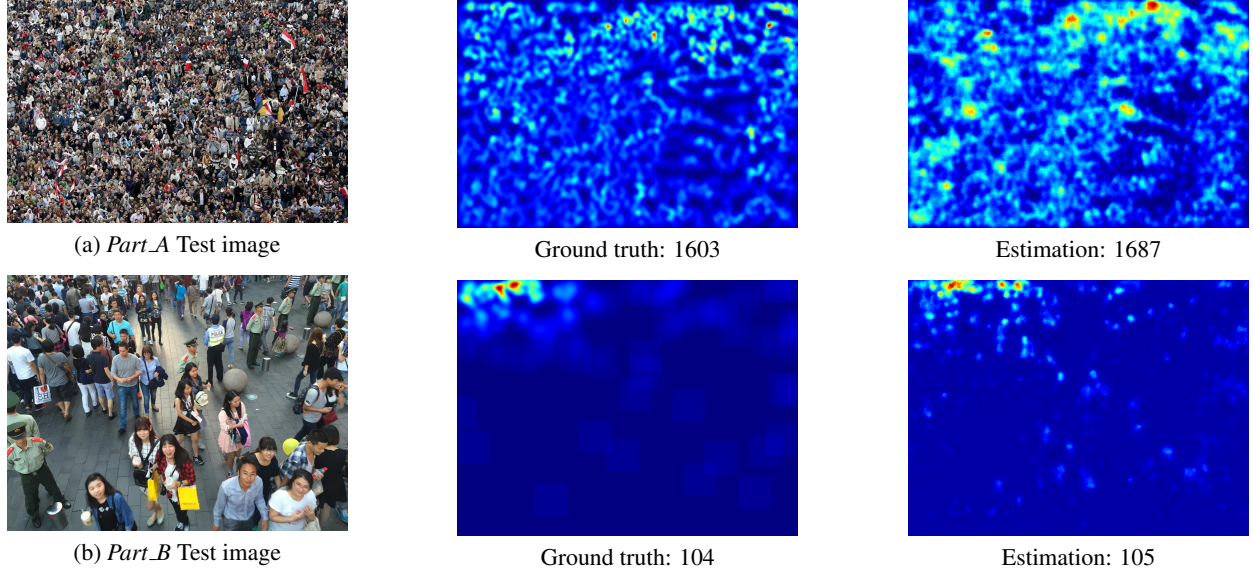
**Table. 1.** The multi-scale CNN architecture.

Formulation	Type	Num.	Filter Size	Pad
Feature Remap	Conv ReLU	64 -	$9 \times 9$ -	4 -
Multi-scale Feature	MSB Conv ReLU	$4 \times 16$ -	$(9/7/5/3) \times (9/7/5/3)$ -	$4/3/2/1$ -
Down-sample	MAX Pool	-	$2 \times 2$	0
Multi-scale Feature	MSB Conv ReLU MSB Conv ReLU	$4 \times 32$ - $4 \times 32$ -	$(9/7/5/3) \times (9/7/5/3)$ - $(9/7/5/3) \times (9/7/5/3)$ -	$4/3/2/1$ - $4/3/2/1$ -
Down-sample	MAX Pool	-	$2 \times 2$	0
Multi-scale Feature	MSB Conv ReLU MSB Conv ReLU	$3 \times 64$ - $3 \times 64$ -	$(7/5/3) \times (7/5/3)$ - $(7/5/3) \times (7/5/3)$ -	$3/2/1$ - $3/2/1$ -
Density Map Regression	MLP Conv ReLU Conv ReLU	1000 - 1 -	$1 \times 1$ - $1 \times 1$ -	0 - 0 -

### 2.2. Scale-relevant Density Map

Following Zhang et al. [13], we estimate the crowd density map directly from the input image. To generate a scale-relevant density map with high quality, the scale-adaptive kernel is currently the best choice. For each head annotation of the image, we represent it as a delta function  $\delta(x - x_i)$  and describe its distribution with a Gaussian kernel  $G_\sigma$  so that the density map can be represented as  $F(x) = H(x) * G_\sigma(x)$  and finally accumulated to the crowd count value. If we assume that the crowd is evenly distributed on the ground plane, the average distance  $\bar{d}_i$  between the head  $x_i$  and its nearest 10 annotations can generally characterize the geometric distortion caused by perspective effect using the Eq. (1), where  $M$  is the total number of head annotations in the image and we fix  $\beta = 0.3$  as [13] empirically.

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



**Figure. 3.** The ground truth and estimated density map of test images in ShanghaiTech dataset

### 2.3. Model Optimization

The output from our model is mapped to the density map, Euclidean distance is used to measure the difference between the output feature map and the corresponding ground truth. The loss function that needs to be optimized is defined as Eq. (2), where  $\Theta$  represents the parameters of the model while  $F(X_i; \Theta)$  represents the output of the model.  $X_i$  and  $F_i$  are respectively the  $i^{th}$  input image and density map ground truth.

$$L(\Theta) = \frac{1}{2N} \sum_{i=1}^N \|F(X_i; \Theta) - F_i\|_2^2 \quad (2)$$

## 3. EXPERIMENTS

We evaluate our multi-scale convolutional neural network (MSCNN) for crowd counting on two different datasets, which include the ShanghaiTech and UCF\_CC\_50 datasets. The experimental results show that our MSCNN outperforms the state-of-the-art methods on both accuracy and robustness with far less parameter. All of the convolutional neural networks are trained based on *Caffe* [18].

### 3.1. Evaluation Metric

Following existing state-of-the-art methods [13], we use the mean absolute error (MAE), the mean squared error (MSE) and the number of neural networks parameters (PARAMS) to evaluate the performance on the testing datasets. The MAE and the MSE are defined in Eq. (3) and Eq. (4).

$$MAE = \frac{1}{N} \sum_{i=1}^N |z_i - \hat{z}_i| \quad (3)$$

$$MSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (z_i - \hat{z}_i)^2} \quad (4)$$

Here  $N$  represents the total number of images in the testing datasets,  $z_i$  and  $\hat{z}_i$  are the ground truth and the estimated value respectively for the  $i^{th}$  image. In general, MAE, MSE and PARAMS can respectively indicate the accuracy, robustness and computation complexity of a method.

### 3.2. The ShanghaiTech Dataset

The ShanghaiTech dataset is a large-scale crowd counting dataset introduced by [13]. It contains 1198 annotated images with a total of 330,165 persons. The dataset consists of 2 parts: *Part\_A* has 482 images crawled from the Internet and *Part\_B* has 716 images taken from the busy streets. Following [13], both of them are divided into a training set with 300 images and a testing set with the remainder.

#### 3.2.1. Model Training

To ensure a sufficient number of data for model training, we perform data augmentation by cropping 9 patches from each image and flipping them. We simply fix the 9 cropped points as top, center and bottom combining with left, center and right. Each patch is 90% of the original size.

In order to facilitate comparison with MCNN architecture [13], the network was designed similar to the largest column of MCNN but with MSB, of which detailed settings are described in Table 1. All convolutional kernels are initialized with Gaussian weight setting standard deviation to 0.01. As

described in Sec. 2.3, we use the SGD optimization with momentum of 0.9 and weight decay as 0.0005.

### 3.2.2. Results

We compare our method with 4 existing methods on the ShanghaiTech dataset. The LBP+RR method used LBP feature to regress the function between the counting value and the input image. Zhang et al. [11] designed a convolutional network to regress both the density map and the crowd count value from original pixels. A multi-column CNN [13] is proposed to estimate the crowd count value (MCNN-CCR) and crowd density map (MCNN).

In Table 2, the results illustrate that our approach achieves the state-of-the-art performance on the ShanghaiTech dataset. In addition, it should be emphasized that the number of our parameters is far less than other two CNN-based algorithms. MSCNN uses approximately  $7\times$  fewer parameters than the state-of-the-art method (MCNN) with higher accuracy and robustness.

**Table. 2.** Performances of methods on ShanghaiTech dataset.

Method	Part.A		Part.B		PARAMS
	MAE	MSE	MAE	MSE	
LBP+RR	303.2	371.0	59.1	81.7	-
MCNN-CCR [13]	245.0	336.1	70.9	95.9	-
Zhang et al. [11]	181.8	277.7	32.0	49.8	7.1M
MCNN [13]	110.2	173.2	26.4	41.3	19.2M
MSCNN	<b>83.8</b>	<b>127.4</b>	<b>17.7</b>	<b>30.2</b>	<b>2.9M</b>

### 3.3. The UCF\_CC\_50 Dataset

The UCF\_CC\_50 dataset [19] contains 50 gray scale images with a total 63,974 annotated persons. The number of people range from 94 to 4543 with an average 1280 individuals per image. Following [11, 13, 14], we divide the dataset into five splits evenly so that each split contains 10 images. Then we use 5-fold cross-validation to evaluate the performance of our proposed method.

#### 3.3.1. Model Training

The most challenging problem of the UCF\_CC\_50 dataset is the limited number of images for training while the people count in the images span too large. To ensure enough number of training data, we perform a data augmentation strategy following [14] by randomly cropping 36 patches with size  $225\times 225$  from each image and flipping them as similar in Sec. 3.2.1.

We train 5 models using 5 splits of training set. The MAE and the MSE are calculated after all the 5 models obtained the estimated results of the corresponding validation set. During training, the MSCNN model is initialized almost the same as the experiment on the ShanghaiTech dataset except that

the learning rate is fixed to be  $1e-7$  to guarantee the model convergence.

### 3.3.2. Results

We compared our method on the UCF\_CC\_50 dataset with 6 existing methods. In [20, 10, 19], handcraft features are used to regress the density map from the input image. Three CNN-based methods [11, 14, 13] proposed to used multi-column/network and perform evaluation on the UCF\_CC\_50 dataset.

Table 3 illustrates that our approach also achieves the state-of-the-art performance on the UCF\_CC\_50 dataset. Here our parameters number is approximately  $5\times$  fewer than the CrowdNet model, demonstrating that our proposed MSCNN can work more accurately and robustly.

**Table. 3.** Performances of methods on UCF\_CC\_50 dataset.

Method	MAE	MSE	PARAMS
Rodriguez et al. [20]	655.7	697.8	-
Lempitsky et al. [10]	493.4	487.1	-
Idrees et al. [19]	419.5	541.6	-
Zhang et al. [11]	467.0	498.5	7.1M
CrowdNet [14]	452.5	-	14.8M
MCNN [13]	377.6	509.1	19.2M
MSCNN	<b>363.7</b>	<b>468.4</b>	<b>2.9M</b>

## 4. CONCLUSION

In this paper, we proposed a multi-scale convolutional neural network (MSCNN) for crowd counting. Compared with the recent CNN-based methods, our algorithm can extract scale-relevant features from crowd images using a single column network based on the multi-scale blob (MSB). It is an end-to-end training method with no requirement for multi-column/network pre-training works. Our method can achieve more accurate and robust crowd counting performance with far less number of parameters, which make it more likely to extend to the practical application.

## 5. ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (61171142, 61401163), Science and Technology Planning Project of Guangdong Province of China (2014B010111003, 2014B010111006), and Guangzhou Key Lab of Body Data Science (201605030011).

## 6. REFERENCES

- [1] 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.

- [2] Navneet Dalal and Bill Triggs, “Histograms of oriented gradients for human detection,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 886–893.
- [3] Meng Wang and Xiaogang Wang, “Automatic adaptation of a generic pedestrian detector to a specific traffic scene,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3401–3408.
- [4] Weina Ge and Robert T Collins, “Marked point processes for crowd counting,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2913–2920.
- [5] 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*, vol. 37, no. 10, pp. 1986–1998, 2015.
- [6] 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.
- [7] 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.
- [8] Antoni B Chan and Nuno Vasconcelos, “Bayesian poisson regression for crowd counting,” in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 545–551.
- [9] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang, “Feature mining for localised crowd counting,” in *BMVC*, 2012, vol. 1, p. 3.
- [10] Victor Lempitsky and Andrew Zisserman, “Learning to count objects in images,” in *Advances in Neural Information Processing Systems*, 2010, pp. 1324–1332.
- [11] 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.
- [12] Chuan Wang, Hua Zhang, Liang Yang, Si Liu, and Xiaochun Cao, “Deep people counting in extremely dense crowds,” in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1299–1302.
- [13] 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.
- [14] Lokesh Boominathan, Srinivas SS Kruthiventi, and R Venkatesh Babu, “Crowdnet: A deep convolutional network for dense crowd counting,” in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 640–644.
- [15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.
- [16] Min Lin, Qiang Chen, and Shuicheng Yan, “Network in network,” *arXiv preprint arXiv:1312.4400*, 2013.
- [17] Vinod Nair and Geoffrey E Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.
- [18] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.
- [19] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah, “Multi-source multi-scale counting in extremely dense crowd images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2547–2554.
- [20] Mikel Rodriguez, Ivan Laptev, Josef Sivic, and Jean-Yves Audibert, “Density-aware person detection and tracking in crowds,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2423–2430.