A LIGHTWEIGHT NEURAL NETWORK FOR CROWD ANALYSIS OF IMAGES WITH CONGESTED SCENES

Xiangyu Ma[1], Shan Du[2], Yu Liu[1]

1.School of Miciroelectronics, Tianjin University, Tianjin 300072, China. 2.Department of Computer Science, Lakehead University, ON, Canada.

1. Introduction

Recently, various CNN-based approaches have been proposed for the task of crowd analysis, which usually have a large number of parameters and require huge computing resources. In order to reduce the computational time and save computing resources, we focus on low-complexity approaches and propose a lightweight end-to-end network for crowd analysis in this paper, which only contains **0.86M** parameters (Lightweight). According to our experiments, our proposal obtains a better result than other existing methods on several testing sequences.

2. Density Map Generation and Metric

Suppose there is a head at pixel x_i in the image with N labelled head, then the result of convolving a delta $\delta(x-x_i)$ function with a Gaussian function is used to represent a person/head.

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

The mean absolute error (MAE) and the mean squared error (MSE) are used to evaluate the performance on the test datasets C_i and C_i^{GT} represent the estimated and the ground truth crowd count respectively corresponding to the i_{th} image.

$$MAE=\frac{1}{N}\sum_{i=1}^{N}\left|C_i-C_i^{GT}\right|,\quad MSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|C_i-C_i^{GT}\right|^2}$$

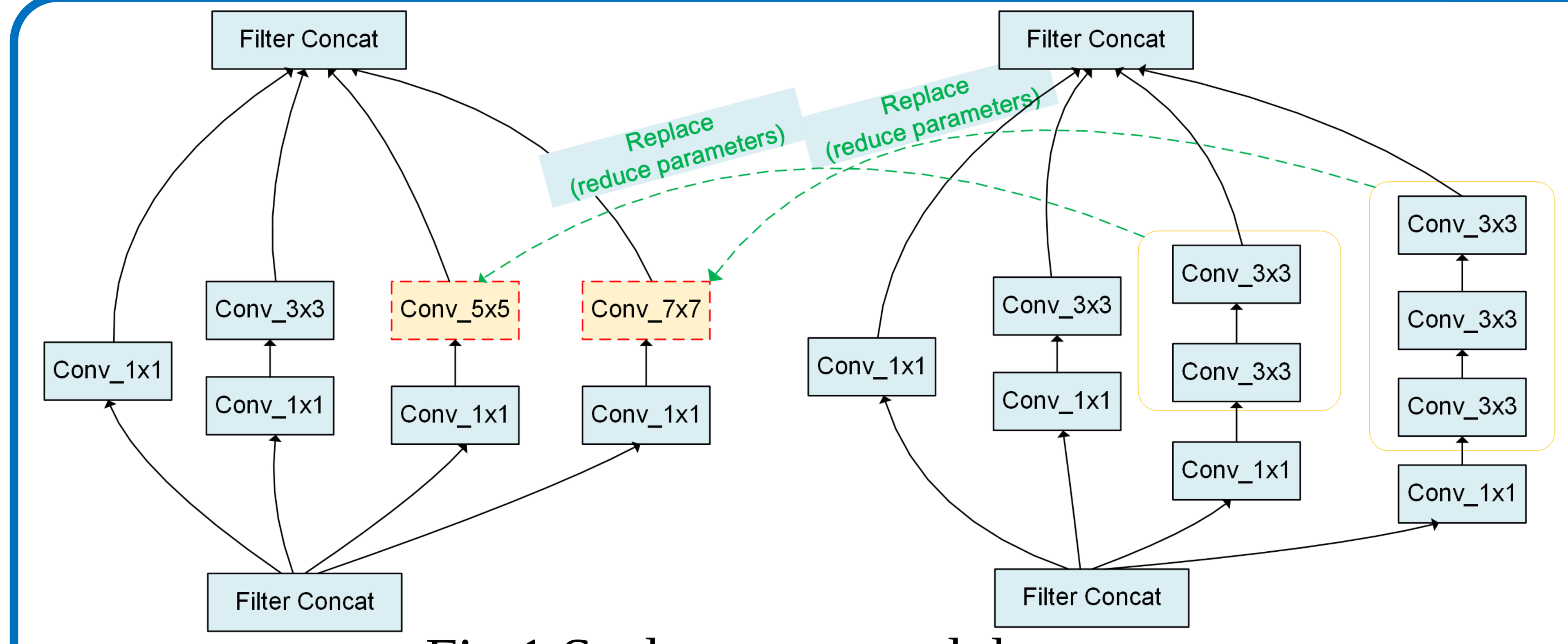


Fig.1 Scale-aware module

3. Lightweight Network

We use cascaded convolutions with smaller spatial filters instead of convolutions with larger spatial filters and design a scale-aware module as shown in Figure 1 to extract multi-scale features from the input image, and finally, an autoencoder network is used to correct the density map for better results. The complete network structure is shown in Figure 2.

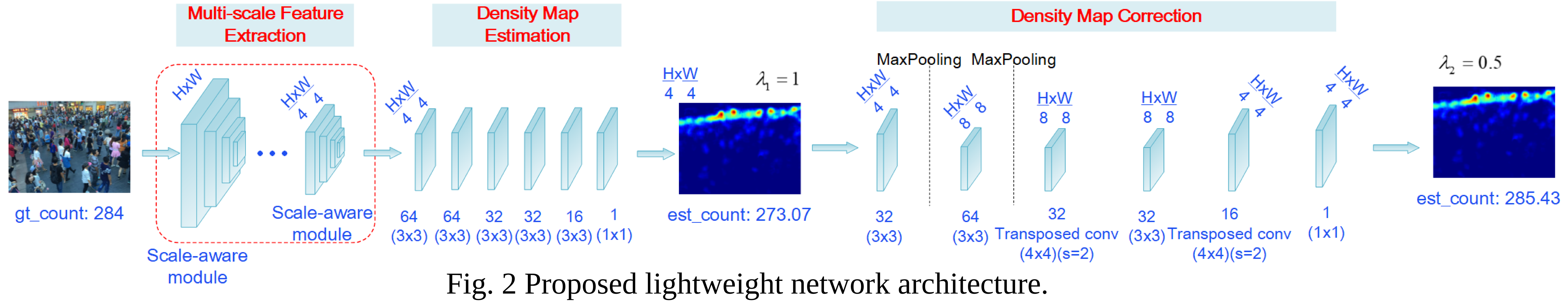
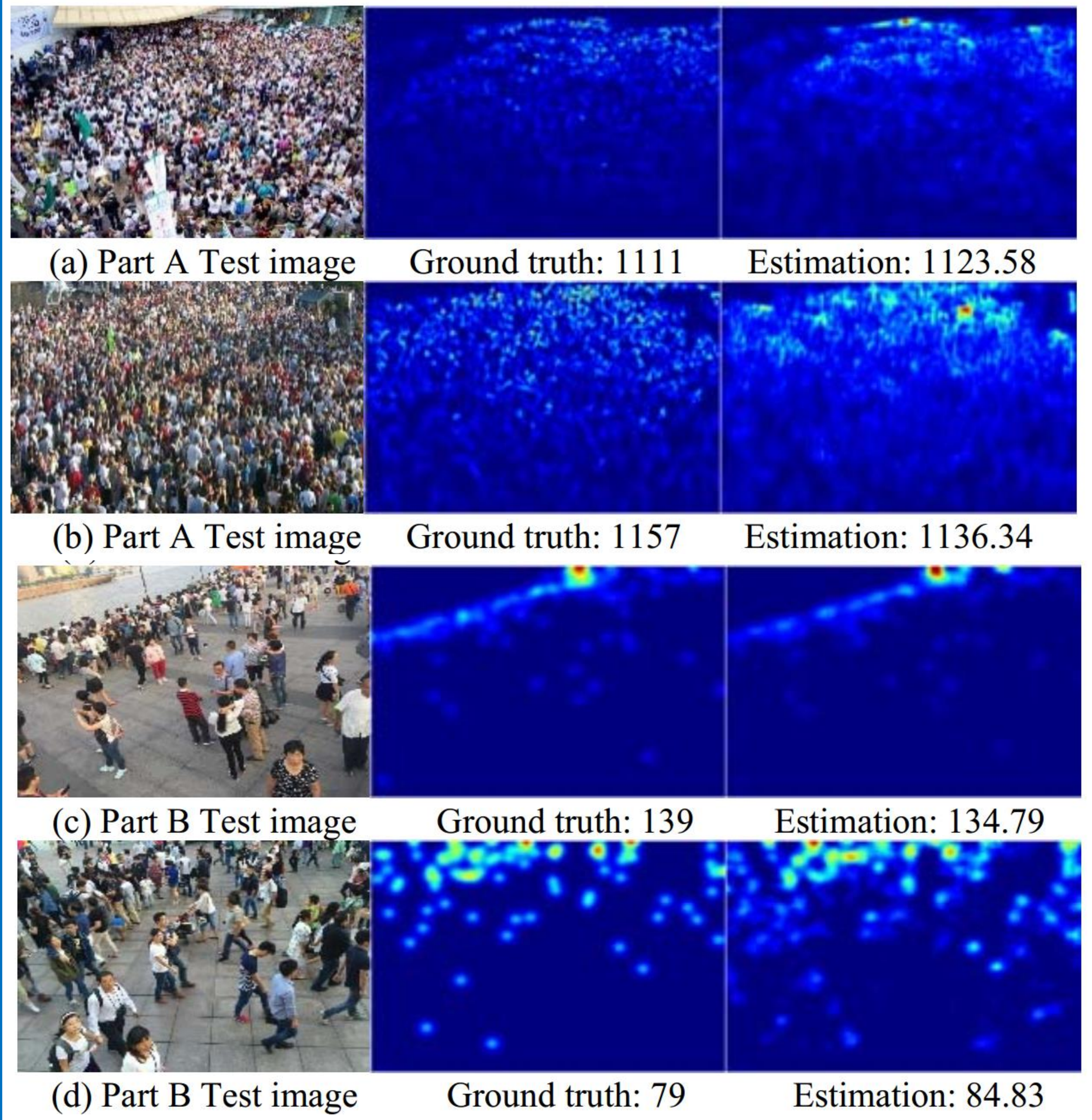


Fig. 2 Proposed lightweight network architecture.

4. Results and Analysis



Method	SHTA		SHTB		UCF_CC_50		PARAMS
	MAE	MSE	MAE	MSE	MAE	MSE	
MCNN	110.2	173.2	26.4	41.3	377.6	509.1	**0.13M**
Cascaded-MTL	101.3	152.4	20.0	31.3	322.8	**397.9**	2.3M
Switch-CNN	90.4	135.0	21.6	33.4	**318.1**	439.2	15.1M
MSCNN	**83.8**	**127.4**	17.7	30.2	363.7	468.4	2.9M
Ours	93.3	149.0	**15.3**	**25.2**	326.7	430.6	0.86M

Although the parameters of our method are 0.86 M that is more than MCNN, the MAE of our method is reduced by 16.9 on SHTA. The performance of MSCNN is about 9.5 better than ours, but the parameters are 2.04 M more. For SHTB, we use the least parameters, except MCNN, to acquire a better result than other methods. And our method also performs well even in extremely crowded scenarios (UCF_CC_50).

5. Conclusion

We propose a novel network structure with less complexity for crowd analysis in this paper. Compared with other networks that have large number of parameters, we get competitive results on *ShanghaiTech Part_A* and *UCF_CC_50* datasets and get the best results on *ShanghaiTech Part_B* using less than 1M parameters (only 0.86 M), the results show that our method is very useful in real life applications which lacks sufficient computing resources.