MDCount: A Lightweight Encoder-Decoder Architecture for Resource-Saving Crowd Counting

Xiaolong Meng¹,²*, Zhengfei Ren³
¹Shanghai Institute of Tourism, Shanghai, China
²School of Tourism, Shanghai Normal University, Shanghai, China
³School of Mathematical Science, University College Cork, Cork, Ireland
*Corresponding author’s e-mail: mengxl@shnu.edu.cn

Abstract. This work put forwards a lightweight encoder-decoder architecture MDCount designed unambiguously for real-time high-precision crowd counting with constricted calculation resources. The lightweight backbone network MobileNetV2 is tailored to decrease parameter numbers in an acceptable accuracy. The encoder-decoder architecture with atrous spatial pyramid pooling modules is proposed to recover the spatial and contextual information at multiple scales. Experiments on realistic and challenging datasets and comparing contemporary approaches demonstrate that our method MDCount accomplishes equivalent performance with smaller computation costs.

1. Introduction
Population increasing and subsequent urbanization result in a chaotic crowd gettogether in countless scenarios such as pageants, shows, and arenas. In these scenarios, a task to estimate person numbers in congested scenes, called crowd counting, plays a requisite role in common safety and control organization. Numerous approaches are raised to tackle various crowd counting challenges, classified into three categories: detection-based methods, regression-based methods, plus density map estimation-based methods. Instead of the person detection via a sliding window on a picture or the overall regress count of the whole image, density map estimation-based techniques forecast the density at pixel-level and get hold of the crowd count by means of summing overall pixels. This method is trained with pixel-level supervision information, and it is more appropriate for high-density scenes. Our proposed architecture can as well be classified as a method of density map estimation-based.

In some tangible applications such as mobile or embedded devices, realizing rapid inference speed is almost as essential as attaining an acceptable accuracy under a restricted calculation budget. On the one hand, various lightweight neural networks have been offered for the following computer vision tasks, for instance, image classification, segmentation, and object detection. On the other hand, lightweight crowd counting networks have been infrequently studied regardless of their importance in surveillance. As a result, we design a resources-saving encoder-decoder architecture for crowd counting tasks to accomplish an optimized trade-off between accurateness and inference speediness.

2. Related work
We divide our discussion of related work into crowd counting approaches, lightweight networks, and encoder-decoder architectures.
2.1. Crowd Counting Approaches
The early works on crowd counting use detection-based methodologies [2] and usually put on a person or head detector with bounding boxes on an image. For instance, some extraordinary detectors, R-CNN, YOLO, as well as SSD, have been put forward to carry out dramatic detection accurateness in sparse areas. Unfortunately, these approaches hand out unsatisfactory products once coming upon obstruction and background confusion in tremendously dense scenes. Some works make known regression-based methods [3] to overcome the above problems, which directly pick up the mapping from extracted features to objects. They commonly first extract global or local features and then use some regression techniques such as linear regression as well as Gaussian mixture regression [4] to acquire final mapping functions. These works are up-and-coming in dealing with the above problems, but they constantly take no notice of spatial information.

Some studies propose density map estimation-based methods by building the mapping relationship between extracted features and corresponding final density maps. Nevertheless, the traditional hand-crafted feature extraction cannot drive the high-quality density map to estimate correct counting. More associates make use of it to improve the density map estimation in recent times, benefiting from the pungent feature representation of Convolutional Neural Networks (CNNs) [5]. MCNN carries out a three-branches multi-column structure [6] with different kernel sizes (large, medium, small) to capture matching head feature maps and employs a 1x1 convolution to blend the features maps into the density map estimation. Similarly, Switch-CNN introduces a switch classifier [7] that adaptively selects the best features map to estimate density. Such a method is robust to large-scale and perspective variations observed in typical crowd scenes. CMTL brings crowd counting classification and density map estimation [8] into a novel end-to-end and cascade-connected framework. This framework picks up globally relevant discriminative features in the direction of more sophisticated density map estimation and minor count miscalculation. CSRNet uses a single-column structure [9] with dilated convolution to collect the multi-scale contextual information in the jammed scenes. This structure is easy-taught because of its unadulterated convolutional structure.

2.2. Lightweight Networks
Designing a deep neural network for optimal accurateness and effectiveness has been an energetic research zone in recent years. SqueezeNet uses a bottleneck approach with the squeeze and expands modules [10] to trim down the parameters. The above approach achieves the equivalent accuracy as AlexNet on ImageNet with 50x fewer parameters. MobileNetV1 employs depthwise separable convolution [11] to advance calculation efficiency substantially. This method reaches 4% better correctness on ImageNet in the equal size as SqueezeNet. ShuffleNet utilizes point-wise group convolution and channel shuffle [12] to diminish operations numbers while maintaining accuracy. MobileNetV2 introduces a resource-efficient block with inverted residuals bottleneck as well as linear transformation [13]. Some state-of-the-art networks are put forward, but MobileNetV2 remains the most broadly used lightweight network.

2.3. Encoder-Decoder Architectures
The encoder-decoder architectures are broadly used in pixel-wise prediction tasks, particularly semantic segmentation, to learn feature representation of CNNs and engender high-definition output. Like a pixel-wise task, crowd counting also requests encoder-decoder architectures to come to the excellent density map of the equivalent size as the original input.

Dilated/Atrous convolutions have been extensively held in computer vision tasks, and voluminous recent publications report using this technique. DeepLab family [14-17] is a good representative of them. DeepLabV1 combines ideas from deep convolutional neural networks (DCNNs) and fully connected conditional random fields (CRFs) [14]. Such a novel method turns out accurate predictions semantically and is computationally efficient. DeepLabV2 proposes an atrous spatial pyramid pooling model (ASPP) [15] to robustly segment objects by taking on various parallel filters with various proportions, thus capturing expressive context at multiple scales. DeepLabV3 uses atrous convolution in cascade or
parallel to capture multi-scale context by adopting multiple atrous convolution rates [16]. Chen et al. in 2018 extend DeepLabV3 by engaging an encoder-decoder framework [17], named DeepLabV3+, which contains opulent semantic information from the encoder module and recovers the precise objective boundaries from the artless yet practical decoder module.

Benefiting from researches as mentioned earlier, we present a translucent FCNs architecture MDCount for resource-saving crowd counting tasks. This operation chooses MobileNetV2 as the backbone network and adopts the encoder-decoder architecture with atrous spatial pyramid pooling modules recommended in DeepLabV3+.

3. Our Approach
This section describes the suggested MDCount architecture for resource-saving crowd counting, including global architecture, objective loss functions, as well as ground truth generation.

3.1. MDCount Architecture
As shown in Figure 1, MDCount is an encoder-decoder architecture based on FCNs to estimate exact density maps, which contains the lightweight backbone network as the encoder for feature extracting and the decoder with atrous spatial pyramid pooling module for multi-scale information capturing.

**Lightweight Backbone Network as Encoder:** We practice MobileNetV2 as the backbone network then modify it precisely for resource-saving crowd counting. MobileNetV2 introduces a memory-efficient bottleneck residual block implementation with inverted residuals block and linear transformation, shown in Table 1.

| Input | Operator | Output |
|-------|----------|--------|
| $h \times w \times k$ | 1x1 conv2d, ReLU6 | $h \times w \times tk$ |
| $h \times w \times tk$ | 3x3 dwise s=s, ReLU6 | $h \times w \times tk$ |
| $h / s \times w / s \times tk$ | 1x1 conv2d, linear | $h / s \times w / s \times k'$ |

The bottleneck residual block consists of three steps: 1) 1×1 convolution with the ReLU6 transformation; 2) 3×3 depthwise separable convolutions; 3) 1×1 convolution with the linear transformation. The adoption of 3×3 depthwise separable convolutions makes the computational budget possible to be about 8 to 9 times smaller than that of normal full convolutions at merely a slight reduction inaccuracy. The linear transformation replaces the previous ReLU6 transformation for the reason that non-linear transformations destroy information in low-dimensional space. The shortcut connections are established between two bottleneck layers with low channels for the reason that the bottleneck layers perform better than the expanded layers.
These bottleneck residual blocks tolerate a predominantly memory-efficient execution which is essential for embedded devices applications. We remove the last three convolution layers of the standard MobileNetV2 structure for image classification tasks, as shown in Table 2, to save computation as well as improve accuracy. In addition, we find that using bottleneck3, bottleneck5, and bottleneck7’s feature maps can improve performance with minor computation costs through experimental study.

**Decoder with atrous spatial pyramid pooling module:** The architecture gradually extracts the feature maps in the encoder route and recovers the spatial information in the decoder route. The atrous spatial pyramid pooling module captures the contextual information at multiple scales. Attempting to bring together the advantages from the above methods, we adopt an encoder-decoder with atrous spatial pyramid pooling module proposed in the paper [17]. Nonetheless, DeepLabv3+ architecture in the paper sets the significant atrous convolution rates for semantic segmentation, which causes the outsized gap among the different sizes of the receptive fields but is not desired for high-density crowd counting. For the reason that the scale variation of crowd scenes is nearly continuous, we set an increasing atrous convolution rate of 1, 2, 3 so that the model can cover most pixel-level information of the original feature map. This setting avoids irrelevant information through large spaces caused by significant atrous rates.

| Operate        | Size of Output | t | n | s | Channels of Output |
|----------------|----------------|---|---|---|--------------------|
| Image          | 1              | - | - | - | 3                  |
| Conv2d         | 1/2            | - | 1 | 2 | 32                 |
| Bottleneck1    | 1/2            | 1 | 1 | 1 | 16                 |
| Bottleneck2    | 1/4            | 6 | 2 | 2 | 24                 |
| Bottleneck3    | 1/8            | 6 | 3 | 2 | 32                 |
| Bottleneck4    | 1/16           | 6 | 4 | 2 | 64                 |
| Bottleneck5    | 1/16           | 6 | 3 | 1 | 96                 |
| Bottleneck6    | 1/32           | 6 | 3 | 2 | 160                |
| Bottleneck7    | 1/32           | 6 | 1 | 1 | 320                |

### 3.2. Objective Loss Functions

Our method MDCount uses mean squared error loss (MSE Loss) as the objective loss function, same with most previous works [6-8]. The MSE loss function takes care of pixel error, which is well-defined as follow:

$$L_e = \frac{1}{N} \sum_{i=1}^{N} \left\| G(X_i; \theta) - D_{i}^{GT} \right\|_2^2$$  \hspace{1cm} (1)

Where $N$ is the frame number, $G(X_i; \theta)$ is the output density map for training frame $X_i$ with parameter $\theta$, and $D_{i}^{GT}$ is the ground truth density map $X_i$.

### 3.3. Ground truth generation

Through blurring each person annotation with geometry-adaptive Gaussian kernel, the ground truth density maps is generated as follows:

$$D^{GT} = \sum_{i=1}^{N} \delta(x - x_i) \times G_{\sigma_{i}}(x), \text{ with } \sigma_{i} = 0.3\overline{d}_{i}$$  \hspace{1cm} (2)

Where $x$ is the location of the pixel, $N$ is the quantity of person annotations in the frame. For every targeted object $x_i$ in the ground truth $\delta$, $\overline{d}_{i}$ indicates the mean distance of $k$ nearest neighbors. The Gaussian kernel $\delta(x - x_i)$ with the factor $\sigma_{i}$ is convolved to produce the final density map.

### 4. Experiments

This section presents the execution details, introduces the evaluated five datasets and the adopted two evaluation metrics, and evaluates our architecture MDCount with three preceding approaches. Our
architecture achieves comparable calculating performance and resource-saving treating speed.

4.1. Execution details
Using the PyTorch framework, the experiments are performed on GeForce 2080Ti. We use Adam optimization with the adopted batch-size 6, the adopted initial learning rate $1 \times 10^{-4}$, in addition to the adopted weight decay rate $2 \times 10^{-2}$.

4.2. Datasets
ShanghaiTech [6] dataset includes 1198 images with a total of 330,165 persons. It is made up of two collections, one is Shanghai-Part_A with 482 images obtained from the Internet, and the other is Shanghai-Part_B with 716 images taken from the shopping street. WorldExpo10 [5] dataset comprises 3980 frames taken by 108 closed-circuit television (CCTV) during the Expo 2010 Shanghai China. UCF-CC_50 [18] dataset includes 50 images from different sights varying from smaller 94 to larger 4543. UCF-QNRF [19] dataset comprises 1535 with a broader diversity of scenes containing perspectives, compactness, and variants.

4.3. Evaluation metrics
Our experiments take on the mean absolute error (MAE) as well as the root mean squared error (RMSE) [7-9] to evaluate the result. The definition of the above evaluation metrics are calculated as follows:

$$ MAE = \frac{1}{N} \sum_{i=1}^{N} |C_i - C_{GT}^i|, \quad RMSE = \frac{1}{N} \sqrt{\sum_{i=1}^{N} (C_i - C_{GT}^i)^2} \quad (3) $$

Where $N$ is the image quantity, $C_i$ denotes the count estimation, $C_{GT}^i$ signifies the ground truth count. What is more, the MAE reflects the architecture’s accuracy whereas the RMSE demonstrates the architecture’s robustness. For both of them, the lower value means better performance.

4.4. Results and analysis
Comparing with previous approaches, as shown in Table 3, our proposed lightweight encoder-decoder models MDCount of two different bottleneck combinations ([3, 7] or [3, 5, 7]) achieve upper accurateness in the case of fewer computation resources on the datasets called UCF-QNRF and UCF-CC_50. Our experiment results are worse than the best-acting technique but comparable on the other datasets named ShanghaiTech-Part_A, ShanghaiTech-Part_B, as well as WorldExpo10.

Table 3. Experiments Results on realistic and challenging datasets.

| Networks       | Parameter (MB) | ShanghaiTech Part_A | ShanghaiTech Part_B | UCF-QNRF | UCF-CC_50 | WE |
|----------------|----------------|---------------------|---------------------|----------|-----------|----|
| MCNN           | 0.13           | 110.2 173.2         | 26.4 41.4           | 277.0 426.0 | 377.6 509.1 | 11.6 |
| CMTL           | 2.46           | 101.3 152.4         | 20.0 31.1           | 252.0 514.0 | 322.8 397.9 | -   |
| CSRNet         | 16.26          | 68.2 115.0          | 10.6 16.0           | - -       | 266.1 397.5 | 8.6 |
| MDCount3,7     | 4.52           | 81.9 124.5          | 12.7 20.76          | 117.5 205.6 | 180.0 254.5 | 9.6 |
| MDCount3,5,7   | 5.33           | 84.2 130.7          | 11.8 19.15          | 111.3 203.0 | 103.1 158.1 | 9.4 |

5. Conclusion
This paper put forward a lightweight architecture named MDCount, which designed an optimal balance between accurateness and inference speed. Compared with previous approaches, the experiments performed on realistic and challenging datasets demonstrate that our method reaches equivalent performance using reduced calculation resources.

Acknowledgment
Our effort was supported by grants from the Scientific Research Project of Shanghai Tourism College (Project No. E3-0250-20-001-030).
References

[1] Gao G, Gao J, Liu Q, et al. Cnn-based density estimation and crowd counting: A survey[J]. arXiv preprint arXiv:2003.12783, 2020.

[2] Enzweiler M, Gavrila D M. Monocular pedestrian detection: Survey and experiments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 31(12): 2179-2195.

[3] Chan A B, Vasconcelos N. Bayesian poisson regression for crowd counting[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009: 545-551.

[4] Tian Y, Sigal L, Badino H, et al. Latent gaussian mixture regression for human pose estimation[C]// Proceedings of the Asian Conference on Computer Vision, 2010: 679-690.

[5] Zhang C, Li H, Wang X, et al. Cross-scene crowd counting via deep convolutional neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 833-841.

[6] Zhang Y, Zhou D, Chen S, et al. Single-image crowd counting via multi-column convolutional neural network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 589-597.

[7] Sam D B, Surya S, Babu R V. Switching convolutional neural network for crowd counting[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4031-4039.

[8] Sindagi V A, Patel V M. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting[C]// 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, 2017: 1-6.

[9] Li Y, Zhang X, Chen D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 1091-1100.

[10] Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size[J]. arXiv preprint arXiv:1602.07360, 2016.

[11] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[12] Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6848-6856.

[13] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4510-4520.

[14] Chen LC, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv preprint arXiv:1412.7062, 2014.

[15] Chen LC, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.

[16] Chen LC, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017.

[17] Chen LC, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]// Proceedings of the European Conference on Computer Vision. 2018: 801-818.

[18] Idrees H, Saleemi I, Seibert C, et al. Multi-source multi-scale counting in extremely dense crowd images[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013: 2547-2554.

[19] Idrees H, Tayyab M, Athrey K, et al. Composition loss for counting, density map estimation and localization in dense crowds[C]// Proceedings of the European Conference on Computer Vision, 2018: 532-546.