Congested Scene Analysis using Dilated Convolutional Generative Adversarial Networks

Prateek Malhotra[1][0000−0001−6747−9149]

Pune Institute of Computer Technology, Pune 411043
prateekdbst@gmail.com

Abstract. This paper proposes a deep-learning based novel crowd counting network to estimate the number and understand the distribution of people in densely crowded images by producing high-quality density maps. This application of semantic segmentation methods is very important in controlling the problems that arise from overcrowding including particularly dangerous ones like stampedes and collapsing of structures. The generator model in our GAN-based approach consists of kernels with varying dilation rates leading to aggregation of information from different-sized receptive fields. Our proposed approach not only produces high quality density maps but achieves a significant parameter reduction compared to other contemporary methods. We have tested our approach on the ShanghaiTech dataset and have achieved very good results for both: part A and part B while using our model of size 511 kB - achieving a size reduction of more than 500 percent when compared to the state-of-the-art model.

Keywords: Deep Learning · Semantic Segmentation · Crowd Counting · Generative Adversarial Networks · Computer Vision.

1 INTRODUCTION

Analysis of crowded scenes is very important in case of stampedes, cases of overcrowding and riots. Each year overcrowding causes several preventable accidents due to the presence of poor management strategies and an unforeseen, sudden increase in the number of people. A crucial part of developing better crowd management strategies and regulations is identifying the number of people present in a densely crowded scene at a given time. These strategies are particularly important in stadiums, festivals, political rallies and the like.

Identifying just the number of people, though, is only a part of the problem. This is because crowd density is almost never uniform across the entire scene: they cluster at important regions and are less dense in other regions. Thus, keeping this property in mind, this becomes a two task problem: identifying the number of people and creating a density map depicting their distribution across the scene.

The major challenges faced in the above task are the differences in lighting, irregular distribution of people , presence of other objects like buildings and trees, and different camera perspectives for each image. Earlier, head detection methods like Histogram of Oriented Gradients [4] was used but recently, owing to the success Neural Network based methods have had on semantic segmentation tasks, most approaches use CNNs which significantly outperform traditional methods.

We build on contemporary CNN based approaches to introduce a dilated convolution based Generative Adversarial Network to create a compact, small-sized model which is suitable for deploying on surveillance devices with limited hardware functionalities. As investigated by [8], most multi-column CNN architectures, introduced initially by [20], lead to a significant increase in training time and number of parameters and that their improved performance is mostly due to their increase in number of parameters. Hence, keeping this result in mind, this paper focuses on tackling the problem without creating multiple networks trained on different crowd densities.

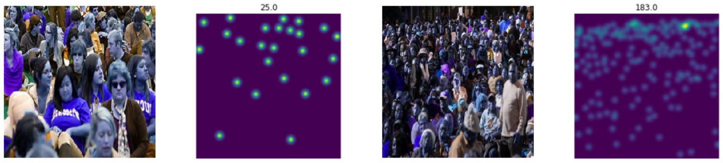


Fig. 1. Sample ShanghaiTech [20] training image patches along with the generated ground-truth

Generative Adversarial Networks [3] have performed extremely well on many image-to-image translation tasks [5] and our network builds on the work done by [7] where the authors used a pix2pix [5] based architecture for generating crowd density maps. Our end-to-end density estimator network uses a generator similar to [10] where different branches of different dilation rates are used to incorporate spatial awareness and integrate local, intermediate, and global features of an image to ensure that we achieve a high quality output with relatively less number of parameters.

By integrating the ideas of both Generative Adversarial Networks and dilated convolutional kernels, this paper achieves results not very far from state-of-the-art methods on major crowd counting datasets while using less than 1/10th the number of parameters. We test our approach on the entire ShanghaiTech dataset [20].

2 RELATED WORK

Before the advent of CNN-based solutions to tackle the task of crowd counting, researchers focused on methods which required trained networks to extract local features from human body parts [17, 2]. Recently, LSTM based approaches have also been used to generate bounding boxes for the same head-detection task [14]. These solutions, however, do not perform well as the number of people increase and the crowd becomes denser (leading to higher occlusion of the targeted human body features). Later, to combat the challenges presented by dense crowds, the focus was shifted to using CNN-based regression methods to generate the crowd count from extracted image features [16] giving, however, no information about the distribution of the crowd in the image.

Recently, Zhang et al. [19] proposed a system where a CNN alternatively trains on two objective functions: count and density estimation in order to reach a better local optima. To adapt the network to a new scene, it is fine-tuned using training examples similar to the target scene. To further improve results, Walach and Wolf [15] introduced layered boosting to iteratively add CNN layers in order to estimate residual error of the previous layer along with selective sampling to get rid of outliers and low quality samples.

Boominatan et al. [1] proposed a two-column fully convolutional network for density map generation while using extensive data augmentation. Sindagi et al. [12] introduced a CNN that enhanced density estimation using high-level prior information. Zhang et al. further added to the dual-column idea by building a multi-column CNN architecture (MCNN) [20] for density estimation and crowd counting.

More recent approaches employ more complex architectures to perform the same task with more precision. The Switch-CNN [11] proposed by Sam et al. involves sending image patches to independent CNN regressors which have different receptive fields and a switch classifier relays a particular image patch to the most suitable regressor. Sindagi et al. [13] introduced a Contextual Pyramid CNN which produces high quality density maps and an overall lower count error by estimating the context at different levels. As pointed out by [8], these approaches using MCNN based architecture are difficult to train and involve very high (often redundant) number of parameters.

The approach taken by CSRNet [8] to achieve state-of-the-art is, however, simpler and more effective than the above approaches. It uses a CNN as a front-end 2D feature extractor and a dilated CNN at the back end to generate the final density map. While this approach reduces the number of parameters as compared to above methods and attains higher accuracy, the size of the network is still very big as the front-end is composed of a VGG network and the back-end consists of several stacks of dilated convolutional layers. Another interesting approach taken by [7] used Generative Adversarial Networks for image-to-density translation as a special case of an image-to-image translation task by using a pix2pix [5] like architecture and training strategy. While their count results were sub-par compared to the more recent approaches, they achieved faster training time and good quality density maps.

This paper's approach to tackle the counting and density map generation problem is a direct extension of the two latest outlined approaches. We use a Generative Adversarial Network with a generator containing multiple dilated convolutional layers but significantly less number of parameters. To our knowledge, this is the first time that such an adversarial network is being used for the task of crowd density estimation.

3 PROPOSED APPROACH

3.1 Dilated Convolutions

Dilated Convolutions [18] introduced by Fisher et al. has achieved state of the art on many semantic segmentation datasets and it was very recently adopted by CSRNet [8] for the purpose of crowd counting and density estimation. It is motivated by the fact that it supports exponentially expanding receptive fields without losing resolution or coverage. The following figure visually explains the concept of dilated convolutions. Each element in the output for 1-dilated kernel has a receptive field of 3x3 while for 2-dilated kernel and 4-dilated kernel the receptive field is 7x7 and 15x15 respectively (as described by Fisher et al). For our model, we apply 2-D dilated convolutions as follows:-

$$y[m,n] = \sum_{i=1}^{M} \sum_{j=1}^{N} x[m + r \cdot i, n + r \cdot j] w[i,j]$$

Where $y[m,n]$ is the output signal, $x[m,n]$ is the input signal and $w[i,j]$ is the weight associated with the convolution kernel of size $[M \times N]$. In the above formula, if r is set to 1, the dilated convolution becomes a regular convolution.

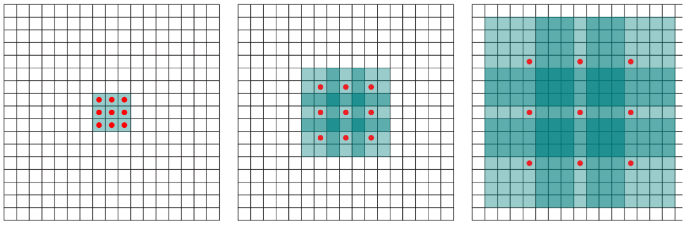


Fig. 2. From left to right: 1-dilated kernel, 2-dilated kernel and 4-dilated kernel. The number of parameters associated with each layer is identical. The receptive field grows exponentially while the number of parameters grows linearly[18]

3.2 Generator Architecture

The architecture for the Generator network is very similar to the one suggested by Perone et al. [ ] for the task of gray matter segmentation. This architecture was chosen specifically for its low number of parameters, small network size and distinct branches with different dilated kernel sizes. Their architecture also very recently achieved state-of-the-art in the gray matter segmentation challenge[10]. In this setup, the input image is first fed to a regular 3x3 convolution layer and then to a 3x3 convolution layer having a dilation rate of 2.

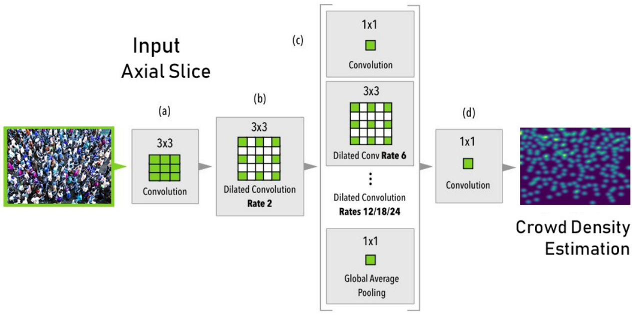


Fig. 3. In the above architecture, the input height and width dimensions are same as the output's height and width dimensions (H x W). Padding is used throughout to ensure that (h x w) of all intermediate feature maps is also (H, W) [10].

After these steps, the resultant intermediate feature map is fed to six parallel branches containing convolutional layers of different dilation rates. The result is then concatenated and passed through a final 1x1 convolutional block. Since the output map is a probability density, we apply a ReLU [9] function to ensure that all pixels hold a value greater than or equal to 0.

3.3 Discriminator Architecture

The discriminator is a fully convolutional neural network with input size (256 x 256 x 3). The output is a single valued scalar with values between 0 and 1. 0 denoting a fake image while 1 denoting a real image. Let C(k, s, f) denote a ReLU activation based convolutional layer with kernel size k, stride s and output filter size f. The architecture is as follows:
Input[256 x 256 x 3]-C(4, 2, 64)-C(4, 2, 128)-C(4, 2, 256)-C(4, 2, 512)-C(4, 2, 1024)-C(6, 1, 1) where the last C block is followed by a softmax layer.

3.4 Generation of Ground Truth

To generate ground truth density map, we apply a Gaussian blur with a fixed value of sigma = 4.0 for both dense and sparse crowd datasets. We extract image patches of size: 256x256 and most of these patches do not contain a very dense crowd distribution hence we have avoided the use of geometric kernel based methods to generate ground-truth density maps.

3.5 Data Preparation

Each image is first resized to 512 x 512 pixels. This resized image is then divided into four patches: each patch of size 256 x 256 pixels. Horizontal flipping based augmentation is applied to each patch before feeding it to the network. Multiplying each ground truth image by 510.0 before feeding it to the network and also dividing the generated image by 510.0 has also generated better results empirically. This is due to the increased L1 loss which prompts the network to learn faster.
Hence the final input size becomes: [5600, 256, 256, 3]

3.6 Training Method

Objective: Similar to [5], our GAN's objective function can be defined as

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[logD(x, y)] + \mathbb{E}_{x}[1 - logD(x, G(x)]$$

where G aims to fool the discriminator D. Moreover, in our approach, G is also tasked with producing outputs that are close to y in an L1 sense. It is noted that adding an L1 loss leads to less blurry outputs. Thus the second loss term is:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y}[|y - G(x)|]$$

and the final objective can be described as:

$$G^{*} = \arg\min_{G} \arg\max_{D} \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G)$$

3.7 Hyperparameters and Optimizers

Generator Objective hyperparameter: the final training objective described above, we fix the value of λ as 10.
Optimizers: We use an ADAM [6] optimizer for training both, the generator and the discriminator with an initial learning rate of 0.001 combined with a poly-learning rate strategy defined as:-

$$\eta = \eta_0 * (1 - n/N)^{p}$$

where N = number of epochs, n = current epoch, η_0 = initial learning rate and p is set to 0.9.
Batch size is set to 4.

4 RESULTS

This paper's approach is evaluated on the ShanghaiTech dataset: part A and B [20] and our results are compared with other contemporary CNN based crowd counting techniques.

4.1 Evaluation metric

To quantify the error in the number of people predicted, this paper follows recent approaches and uses Mean Absolute Error.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|E_i - C_i|$$

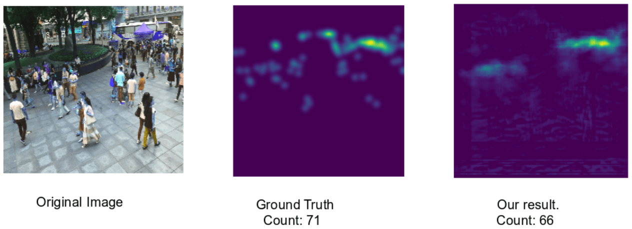
where E_i is the estimated count for the i^{th} image and C_i is the real count for the i^{th} image. To find out the estimated count E_i for the i^{th} image, we sum over the entire generated crowd heat-map pixels like so:-

$$E_i = \sum_{l=1}^{L}\sum_{w=1}^{W} p_{l,w}$$

where $p_{l,w}$ is the pixel value for the one-channeled output image coordinate (l, w)

4.2 ShanghaiTech dataset

The dataset [20] contains 1198 annotated images with a total of 330,165 people. The dataset is divided into two parts: Part A contains 300 training images of highly congested scenes and 182 testing images; Part B contains 400 training and 316 testing images of low density crowded scenes.

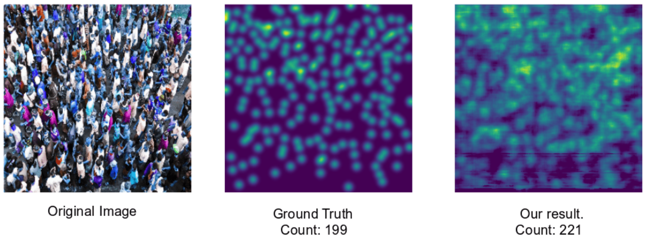


Original Image

Ground Truth
Count: 71

Our result.
Count: 66

Fig. 4. Sample results on a sparsely crowded test image



Original Image

Ground Truth
Count: 199

Our result.
Count: 221

Fig. 5. Sample results on a densely crowded test image

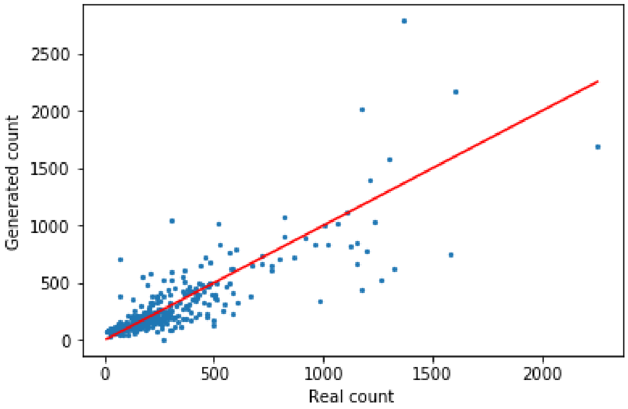


Fig. 6. The graph sheds light on change in the performance of our network as the number of people vary in the input image. The solid red line signifies ideal performance on the test dataset.

Approach	MAE: Part A	MAE: Part B
Zhang et al.	181.20	32.0
MCNN(S-M)	160.5	-
GAN: Li et al.	158.8	42.3
OUR METHOD	150.2	26.5
S-CNN	90.4	21.6
CSRNet	68.2	10.6

Table 1: Comparison with current CNN based approaches on the complete ShanghaiTech dataset

4.3 Parameter Reduction and storage size

This paper notes a very significant reduction in model size and the number of parameters used as shown in TABLE II which can be attributed to the use of multi-scale dilated convolution layers and generating high quality density maps with reduced parameters using a Conditional GAN. Moreover, this benefit makes the presented architecture especially useful for real-world CCTV camera applications with highly constrained memory requirements.

Approach	Parameters	Storage size
OUR	0.99 M	511 kB
CSRNet	276 M	139 MB

Table 2: Parameters and storage size comparison with state-of-the-art

5 CONCLUSION

In this project, a novel end-to-end crowd counting architecture using Dilated Convolutional Generative Adversarial Networks has been proposed. It is motivated by the removal of max-pooling operation and with having a branched architecture enabling encoding of immediate, local, intermediate and global features from an image. Parameters in our model can be learned in an end-to-end fashion and the presence of two loss functions acts as a regulariser to prevent overfitting on the training dataset. Our main contribution through this project has been achieving high quality density maps from crowded scenes and generating a reasonable count while working under high memory constraints. The model occupies only 511kB for storage which makes it apt for devices with low memory and processing power

References

1. 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*, pages 640–644. ACM, 2016.
2. Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2012.
3. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.
4. Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2547–2554. IEEE, 2013.

5. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.
6. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
7. J. Li, H. Yang, L. Chen, J. Li, and C. Zhi. An end-to-end generative adversarial network for crowd counting under complicated scenes. In *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–4, June 2017.
8. Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. *arXiv preprint arXiv:1802.10062*, 2018.
9. 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)*, pages 807–814, 2010.
10. Christian S Perone, Evan Calabrese, and Julien Cohen-Adad. Spinal cord gray matter segmentation using deep dilated convolutions. *Scientific reports*, 8(1):5966, 2018.
11. Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 6, 2017.
12. Vishwanath A Sindagi and Vishal M Patel. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*, pages 1–6. IEEE, 2017.
13. Vishwanath A Sindagi and Vishal M Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1879–1888. IEEE, 2017.
14. Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2325–2333, 2016.
15. Elad Walach and Lior Wolf. Learning to count with cnn boosting. In *European Conference on Computer Vision*, pages 660–676. Springer, 2016.
16. 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*, pages 1299–1302. ACM, 2015.
17. 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*, pages 3401–3408. IEEE, 2011.
18. Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.
19. Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 833–841. IEEE, 2015.
20. 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*, pages 589–597, 2016.