Bayesian Multi Scale Neural Network for Crowd Counting

Abhinav Sagar[*]
Vellore Institute of Technology
Vellore, Tamil Nadu, India
abhinavsagar4@gmail.com

Abstract

Crowd Counting is a difficult but important problem in computer vision. Convolutional Neural Networks based on estimating the density map over the image has been highly successful in this domain. However dense crowd counting remains an open problem because of severe occlusion and perspective view in which people can be present at various sizes. In this work, we propose a new network which uses a ResNet based feature extractor, downsampling block which uses dilated convolutions and upsampling block using transposed convolutions. We present a novel aggregation module which makes our network robust to the perspective view problem. We present the optimization details, loss functions and the algorithm used in our work. On evaluating on ShanghaiTech, UCF-CC-50 and UCF-QNRF datasets using MSE and MAE as evaluation metrics, our network outperforms previous state of the art approaches while giving uncertainty estimates in a principled bayesian manner.

1 Introduction

Crowd Counting has attracted a lot of attention of late in the computer vision community due to a range of applications like counting the number of participants in political rallies, social and sport events, etc. Also the same methodology can be used for other problems like counting cells in microscopic images, cars in satellite imagery etc. Crowd Counting is a difficult problem especially in dense crowds due to two main reasons 1) there is often clutter, overlap and occlusions present 2) in perspective view it is difficult to take into account the shape and size of object present with respect to the background.

A lot of algorithms have been proposed in the literature for tackling this problem. Most of them use some form of convolutional neural network along with a density map estimation which predicts a density map over the input image and then summing to get the count of objects. The datasets which are used for training crowd counting only provide point annotations for each training image, i.e., only one pixel of each person is labeled (typically the center of the head).

The early works used the concept of detecting the individual objects by using some kind of object detection architecture or by segmenting the objects. This was an inefficient approach due to the huge computations required. Also this approach was abandoned due to low accuracy as it fails to give correct results in dense crowds. To tackle this problem, regression based methods were used by removing the detection of individual objects and instead using a direct scalar mapping from the input image to the count. This is made possible by learning the low level features and by regressing over it to give a measure of count of objects present. This approach was used to tackle the occlusion

[*] Website of author - https://abhinavsagar.github.io/

problem which was faced with detection based methods however still it lacked the information that in perspective objects can be present in different sizes.

The next approach used a density estimation concept in which a density map was learned which preserves the information present when an object is present in different scales. This method also tackles the occlusion problem by learning a direct mapping from the input image to a density map over it. This approach has become state of the art in crowd counting and most recent algorithms use some kind of density map estimation using convolutional neural networks.

2 Related Work

(Zhang et al., 2015) one of the first works on crowd counting uses switchable learning process with two learning objectives, crowd density maps and crowd counts. (Sam et al., 2017) proposed an end-to-end network to predict crowd density for a crowd. A multi column CNN is used in (Zhang et al., 2016) by replacing the fully connected layer with a convolution layer whose filter size is 1×1. (Boominathan et al., 2016) used a combination of deep network as well as a shallow network which works well for detecting people under large scale variations and severe occlusion. (Ranjan et al., 2018) proposed a two-stage CNN framework for crowd density estimation and counting.

(Sindagi and Patel, 2017a) proposed a multi-task cascaded CNN network for jointly learning crowd count classification and density map estimation. A similar end-to-end training method was used by (Zeng et al., 2017) with no requirement for multicolumn network and shows pre-training works. (Liu et al., 2019b) showed that encoding multi-scale context, along with providing an explicit model of perspective distortion results in substantially increased crowd counting performance. (Zhang et al., 2018) proposed a method which concatenates multiple feature maps at different scales to produce a strong scale-adaptive crowd counting method. On the other hand (Shang et al., 2016) uses a contextual information to predict both local and global count.

A top-down feedback was used by (Sam and Babu, 2018) which carries high-level scene context to correct wrong detections. (Shi et al., 2019) uses perspective maps which are encoded as perspective-aware weighting layers to adaptively combine the multi-scale density outputs. (Jiang et al., 2019) proposed a network using a multi-scale encoder and a multi-path decoder to generate high-quality density estimation maps. Detection and regression based count estimations was done by (Liu et al., 2018) under the guidance of attention mechanism. (Liu et al., 2019a) used a new loss function which learns the local correlation within regions of various sizes thus producing locally consistent estimation. (Hossain et al., 2019) used the attention mechanism to softly select the appropriate scales at both global and local levels.

(Sam et al., 2019) differs from the previous methods by using autoencoder to learn several layers of useful filters from unlabeled crowd images. (Cheng et al., 2019) method is able to capture the spatial variations by finding the pixel-level subregion with high discrepancy to the ground truth. A novel network by (Sindagi and Patel, 2019) involves two sets of attention modules: spatial attention and global attention module at various scales. (Oh et al., 2020) uses uncertainty quantification at the same time while estimating the count using a density map based on extracting features at various scales. A new loss function proposed by (Ma et al., 2019) gives an uncertainty estimate at the same time while estimating the count. (Idrees et al., 2018) approach estimates counts, density maps and localization in dense crowd images.

We summarize our main contributions as follows:

• We propose a new model architecture which is based on a ResNet based feature extractor, down-sampling part using dilated convolutional layers and upsampling part using transposed convolutional layers.

• We present layer wise details, a new aggregation module, optimization details, loss functions, evaluation metrics and algorithms used in this work.

• On evaluating our network on ShanghaiTech, UCF-CC-50 and UCF-QNRF datasets using MSE and MAE as evaluation metrics our model outperforms previous state of the art model architectures with much less number of parameters.

• Our network along with giving the count of the people present in the image also gives epistemic uncertainty and aleatoric uncertainty quantification.

3 Proposed Method

3.1 Dataset

Experimental evaluations are conducted using three widely used crowd counting datasets: ShanghaiTech part A and part B, UCF-CC 50 and UCF-QNRF. These datasets are described as follows:

• ShanghaiTech is made up of two datasets labelled as part A and part B. In part A, there are 300 images for training and 182 images for testing while Part B has 400 training images and 316 testing images. Most of the images are of very crowded scenes such as rallies and large sporting events. Part A has a significantly higher density than part B.

• UCF-CC-50 contains 50 gray images with different resolutions. The average count for each image is 1,280, and the minimum and maximum counts are 94 and 4,532, respectively.

• UCF-QNRF is the third dataset used in this work which has 1535 images with 1.25 million point annotations. It is a challenging dataset because it has a wide range of counts, image resolutions, light conditions and viewpoints. The training set has 1,201 images and 334 images are used for testing.

3.2 Model Architecture

The network architecture is made up of a ResNet based feature extractor with dilated convolutions which is defined as a downsampling block. This helps in extracting the details of objects at various scales hence solving the perspective view problem faced by earlier approaches. Next the upsampling block uses transposed convolutions with skip connections in between the two creating an additional pathway. The last part has three heads: output of density map which when integrated gives the absolute count, epistemic uncertainty and aleatoric uncertainty heads. The network architecture along with layerwise details used in this work is shown in Figure 1:

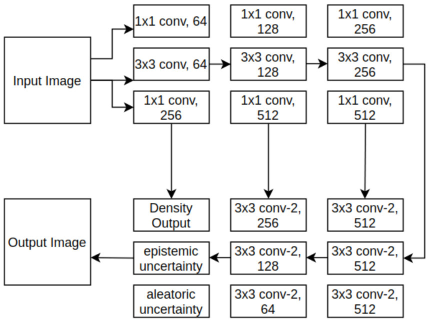

Figure 1: Our Neural network architecture

Where 1×1, 3×3 denotes Filters, 64, 128, 256 denotes Recpetive Field, conv denotes Dilated Convolutional layer and conv-2 denotes Transposed convolutional layer.

3.3 Optimization

While training the network, vanishing gradient problem showed up ie weights of the connections were turning out to be zero. To alleviate this, instance normalization was used after both convolutional and transposed convolutional layers as defined in Equation 1.

$$y = ReLU\left(\sum_{i=0}^{d} w_i \cdot ReLU\left(\gamma_i \cdot \frac{x_i - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} + \beta_i\right) + b\right) \tag{1}$$

where w and b are weight and bias term of the convolution layer, γ and β are weight and bias term of the Instance Normalization layer, μ and σ are mean and variance of the input.

Previous works have used multi column architecture (Zhang et al., 2016) to deal with the various scales at which object might be present in the image. The problem with these methods is that the number of columns give a direct measure of the scale at which it can recognize individual objects. To tackle this, we propose a new technique to aggregate the filters with sizes 1×1, 3×3, 5×5. ReLU is applied after every convolutional and transposed convolutional layer. The filter branches make our network robust and can be extended by using more filters to tackle crowd counting in dense scenes. Our aggregation modules stacked on top of each other behave as ensembles thus minimizing overfitting which is a challenge with deep networks. The novel aggregation module used in our work is shown in Figure 2:

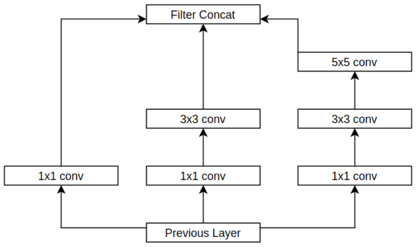

Figure 2: The architecture of our aggregation module

Table 1 shows the estimation error comprised of MSE and MAE for our network using image wise and patch wise test sample compared with L_E and L_S loss function.

Table 1: Estimation error of our network trained with different loss functions and tested with different samples

Loss function	Test sample	MAE	MSE
L_E	image	116.4	181.2
L_E	patch	71.3	107.7
L_E, L_S	image	87.1	134.1
L_E, L_S	patch	67.0	104.2

Where L_E refers to Euclidean loss and L_S refers to SSIM loss.

3.4 Loss Function

Most existing work uses pixelwise Euclidean loss for training the network. This gives a measure of estimation error at pixel level which is defined in Equation 2.

$$L_E = \frac{1}{N} \|F(X, \theta) - Y\|^2 \tag{2}$$

where θ denotes a set of the network parameters, N is the number of pixels in density maps, X is the input image and Y is the corresponding ground truth density map, $F(X, \theta)$ denotes the estimated density map. We also incorporate SSIM index in our loss to measure the deviation of the prediction from the ground truth. SSIM index is used in image quality assessment. It computes similarity between two images from three local statistics, i.e. mean, variance and covariance. The range of SSIM values is from -1 to 1 and it is equal to 1 when the two images are identical. SSIM index is defined in Equation 3.

$$SSIM = \frac{(2\mu_F \mu_Y + C_1)(2\sigma_{FY} + C_2)}{(\mu_F^2 + \mu_Y^2 + C_1)(\sigma_F^2 + \sigma_Y^2 + C_2)} \tag{3}$$

where C_1 and C_2 are small constants to avoid division by zero. Using this next term of the loss function can be written by averaging over the integral as shown in Equation 4.

$$L_S = \frac{1}{N} \sum_x SSIM(x) \tag{4}$$

where N is the number of pixels in density maps. L_S gives a measure of the difference between the network predictions and ground truth. The final loss function by adding the two terms can be written as shown in Equation 5.

$$L_{tot} = \alpha L_E + \beta L_S \tag{5}$$

where α_C and α_S are constants. In our experiments, we set both α_C and α_S as 0.5 to give equal weightage to both the terms.

3.5 Evaluation Metrics

For crowd counting, the count error is measured by two metrics, Mean Absolute Error (MAE) and Mean Squared Error (MSE), which are commonly used for quantitative comparison. These metrics are defined in Equation 6 and Equation 7 respectively.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left| C_i - C_i^{GT} \right| \tag{6}$$

$$MSE = \sqrt{\frac{1}{N} \sum_{i}^{N} \left| C_i - C_i^{GT} \right|^2} \tag{7}$$

where N is the number of test samples, C_i and CGT_i are the estimated and ground truth count corresponding to the i^{th} sample which is given by the integration of the density map. MAE shows the accuracy of predicted result while MSE measures the robustness of prediction.

3.6 Uncertainty Estimation

There are two main sources of uncertainty in model predictions: epistemic uncertainty is uncertainty due to our lack of knowledge and aleatoric uncertainty is due to stochasticity present in the data. Epistemic uncertainty is often called model uncertainty and it can be explained away given enough data. Using bayesian neural networks in which the weights are parameterized by distributions instead of point estimates, epistemic uncertainty can be computed. However crowd counting requires understanding the inherent nuances of the data like occlusions, scale ambiguity etc, hence aleatoric uncertainty is also important. To capture epistemic uncertainty in a neural network, we put a prior distribution over its weights. Taking this into account, uncertainty can be estimated using the loss function as defined in Equation 8 where the goal is to minimize the negative log likelihood.

$$\mathcal{L}(\theta) = \frac{1}{D} \sum_i \frac{1}{2\sigma^2} \left\| y_i - \hat{y}_i \right\|^2 + \frac{1}{2} \log \sigma^2 \tag{8}$$

where y_i is the i^{th} pixel of the output density y corresponding to input x and D is the number of output pixels. Note that the observation noise σ^2 which captures how much noise we have in the outputs stays constant for all data points.

3.7 Algorithm

Next we present the algorithm used in this work:

Algorithm 1: Bayesian Multi Scale Neural Network for Crowd Counting

```
Require: Input images $\{x_n\}_{n=1}^{N}$ , GT density $\{y_n\}_{n=1}^{N}$
Initialize parameters $\theta$
for each epoch do
    for n = 1 to N do
        Sample $\theta, \phi \sim$ Uniform $\{1, \dots, K\}$
        Compute predictions $[y_n] = f_{\theta_k}(x_n)$
        Calculate loss: $L(\theta) = \frac{1}{D} \sum_i \frac{1}{2\sigma^2} \|y_i - \hat{y}_i\|^2 + \frac{1}{2} \log \sigma^2$
        Update $\theta_k$ using gradient $\frac{dL(\theta_k)}{d\theta_k}$
    end
end
```

4 Experimental Results

As shown in Table 2, our method obtains the lowest Mean Square Error (MSE) and Mean Absolute Error (MAE) on both subset of ShanghaiTech dataset.

Table 2: Comparison with state-of-the-art methods on ShanghaiTech dataset (lower is better)

Method	MAE	MSE	MAE	MSE
(Zhang et al., 2015)	181.8	277.7	32.0	49.8
MCNN (Zhang et al., 2016)	110.2	173.2	26.4	41.3
Cascaded-MTL (Sindagi and Patel, 2017a)	101.3	152.4	20.0	31.1
Switch-CNN (Sam et al., 2017)	90.4	135.0	21.6	33.4
CP-CNN (Sindagi and Patel, 2017b)	73.6	106.4	20.1	30.1
CSRNet (Li et al., 2018)	68.2	115.0	10.6	16.0
SANet (Cao et al., 2018)	67.0	104.5	8.4	13.6
SFCN (Wang et al., 2019)	64.8	107.5	7.6	13.0
CAN (Liu et al., 2019b)	62.3	100.0	7.8	12.2
DUBNet (Oh et al., 2020)	64.6	106.8	7.7	12.5
Ours	63.2	95.6	7.3	10.6

As shown in Table 3, our method obtains the lowest MSE and MAE on UCF CC 50 dataset.

Table 3: Comparison with state-of-the-art methods on UCF-CC 50 dataset (lower is better)

Method	MAE	MSE
MCNN (Zeng et al., 2017)	377.6	509.1
Cascaded-MTL (Sindagi and Patel, 2017a)	322.8	397.9
Switch-CNN (Sam et al., 2017)	318.1	439.2
D-ConvNet (Shi et al., 2018)	288.4	404.7
L2R (Wan et al., 2019)	279.6	388.9
CSRNet (Li et al., 2018)	266.1	397.5
ic-CNN (Ranjan et al., 2018)	260.9	365.5
SANet (Cao et al., 2018)	258.4	334.9
SFCN (Wang et al., 2019)	214.2	318.2
CAN (Liu et al., 2019b)	212.2	243.7
DUBNet (Oh et al., 2020)	243.8	329.3
Ours	216.7	225.1

As shown in Table 4, our method obtains the lowest MSE and MAE on UCF-QNRF dataset.

Table 4: Comparison with state-of-the-art methods on UCF-QNRF dataset (lower is better)

Method	MAE	MSE
MCNN (Zeng et al., 2017)	277	426
Cascaded-MTL (Sindagi and Patel, 2017a)	252	514
Switch-CNN (Sam et al., 2017)	228	445
CSRNet (Li et al., 2018)	135.5	207.4
SFCN (Wang et al., 2019)	102.0	171.4
CAN(Liu et al., 2019b)	107	183
DUBNet (Oh et al., 2020)	105.6	180.5
Ours	106.7	165.1

As shown in Table 5, the number of parameters of the proposed network is the least compared to previous works. Our method achieves superior results than other state-of-the-art methods with much less parameters.

Table 5: Number of parameters in millions (lower is better)

Method	(Sam et al., 2017)	(Sindagi and Patel, 2017b)	(Li et al., 2018)	(Cao et al., 2018)	Ours
Parameters	15.11	68.4	16.26	0.91	0.24

Figure 3 and Figure 4 respectively illustrate the qualitative results for sample images from the ShanghaiTech and UCF-QNFRF datasets respectively. The samples visualized along with estimated density maps and their epistemic and aleatoric uncertainty from test evaluations on the ShanghaiTech data and the UCF-QNRF dataset is shown.

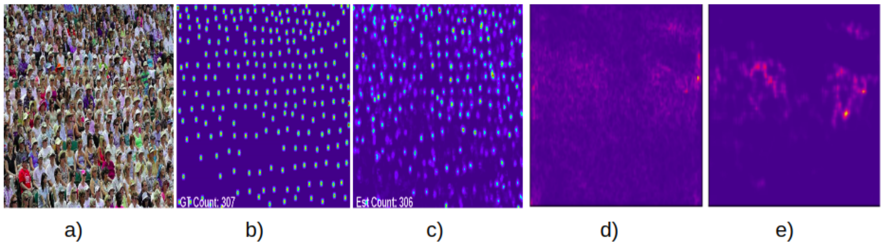

Figure 3: Sample results of the proposed method on ShanghaiTech dataset (a) Input. (b) Ground truth (c) Estimated density map (d) epistemic uncertainty and (e) aleatoric uncertainty quantification.

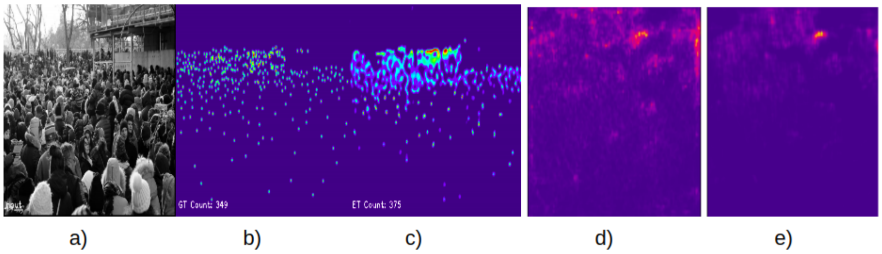

Figure 4: Sample results of the proposed method on UCF-QNRF dataset (a) Input. (b) Ground truth (c) Estimated density map (d) epistemic uncertainty and (e) aleatoric uncertainty quantification.

More red color means higher uncertainty. Epistemic uncertainty captures the model's lack of knowledge about the data while aleatoric uncertainty captures inherent noise in the data. From the above two figures, it is seen that both epistemic uncertainty and aleatoric uncertainty are co-related especially where the crowd density is high. This is natural as the problems of occlusion and perspective view of object size comes into picture. Also another thing to be noted is that the model is less certain in dense crowds hence uncertainty is high there.

5 Conclusions

In this work, we present a new network for crowd counting which is based on a ResNet based feature extractor and a new feature aggregation module. The downsampling blocks use dilated convolutional layers while upsampling blocks use transposed convolutional layers. Skip connections in between the blocks create an additional pathway thus preventing overfitting. We show the optimization details, loss functions and algorithms used in this work. Our method outperform previous state of the art on 3 publicly available datasets using MSE and MAE as the evaluation metrics. Our method also gives a measure of uncertainty thus solving the black box problem of neural networks.

Acknowledgments

We would like to thank Nvidia for providing the GPUs.

References

L. Boominathan, S. S. Kruthiventi, and R. V. Babu. Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 640–644, 2016.

X. Cao, Z. Wang, Y. Zhao, and F. Su. Scale aggregation network for accurate and efficient crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, and A. G. Hauptmann. Learning spatial awareness to improve crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6152–6161, 2019.

M. Hossain, M. Hosseinzadeh, O. Chanda, and Y. Wang. Crowd counting using scale-aware attention networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1280–1288. IEEE, 2019.

Y. Hu, H. Chang, F. Nian, Y. Wang, and T. Li. Dense crowd counting from still images with convolutional neural networks. *Journal of Visual Communication and Image Representation*, 38: 530–539, 2016.

S. Huang, X. Li, Z. Zhang, F. Wu, S. Gao, R. Ji, and J. Han. Body structure aware deep crowd counting. *IEEE Transactions on Image Processing*, 27(3):1049–1059, 2017.

H. Idrees, I. Saleemi, C. Seibert, and M. Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2547–2554, 2013.

H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–546, 2018.

X. Jiang, Z. Xiao, B. Zhang, X. Zhen, X. Cao, D. Doermann, and L. Shao. Crowd counting and density estimation by trellis encoder-decoder networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6133–6142, 2019.

D. Kang and A. Chan. Crowd counting by adaptively fusing predictions from an image pyramid. *arXiv preprint arXiv:1805.06115*, 2018.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

S. Kumagai, K. Hotta, and T. Kurita. Mixture of counting cnns: Adaptive integration of cnns specialized to specific appearance for crowd counting. *arXiv preprint arXiv:1703.09393*, 2017.

Y. Li, X. Zhang, and D. Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018.

J. Liu, C. Gao, D. Meng, and A. G. Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2018.

L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, and L. Lin. Crowd counting with deep structured scale integration network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1774–1783, 2019a.

W. Liu, M. Salzmann, and P. Fua. Context-aware crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5099–5108, 2019b.

Z. Ma, X. Wei, X. Hong, and Y. Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6142–6151, 2019.

M. M. Oghaz, A. R. Khadka, V. Argyriou, and P. Remagnino. Content-aware density map for crowd counting and density estimation. *arXiv preprint arXiv:1906.07258*, 2019.

M.-h. Oh, P. A. Olsen, and K. N. Ramamurthy. Crowd counting with decomposed uncertainty. In *AAAI*, pages 11799–11806, 2020.

D. Onoro-Rubio and R. J. López-Sastre. Towards perspective-free object counting with deep learning. In *European Conference on Computer Vision*, pages 615–629. Springer, 2016.

V. Ranjan, H. Le, and M. Hoai. Iterative crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 270–285, 2018.

D. Ryan, S. Denman, S. Sridharan, and C. Fookes. An evaluation of crowd counting methods, features and regression models. *Computer Vision and Image Understanding*, 130:1–17, 2015.

A. Sagar. Bayesian neural network via stochastic gradient descent. *arXiv preprint arXiv:2006.08453*, 2020a.

A. Sagar. Learning to detect 3d objects from point clouds in real time. *arXiv preprint arXiv:2006.01250*, 2020b.

D. B. Sam and R. V. Babu. Top-down feedback for crowd counting convolutional neural network. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

D. B. Sam, S. Surya, and R. V. Babu. Switching convolutional neural network for crowd counting. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4031–4039. IEEE, 2017.

D. B. Sam, N. N. Sajjan, H. Maurya, and R. V. Babu. Almost unsupervised learning for dense crowd counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8868–8875, 2019.

C. Shang, H. Ai, and B. Bai. End-to-end crowd counting via joint learning local and global count. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1215–1219. IEEE, 2016.

Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, and X. Yang. Crowd counting via adversarial cross-scale consistency pursuit. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5245–5254, 2018.

M. Shi, Z. Yang, C. Xu, and Q. Chen. Revisiting perspective information for efficient crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7279–7288, 2019.

Z. Shi, L. Zhang, Y. Liu, X. Cao, Y. Ye, M.-M. Cheng, and G. Zheng. Crowd counting with deep negative correlation learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5382–5390, 2018.

V. A. Sindagi and V. M. Patel. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017a.

V. A. Sindagi and V. M. Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1861–1870, 2017b.

V. A. Sindagi and V. M. Patel. Ha-ccn: Hierarchical attention-based crowd counting network. *IEEE Transactions on Image Processing*, 29:323–335, 2019.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958, 2014.

D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

J. Vandoni, E. Aldea, and S. Le Hégarat-Mascle. Evaluating crowd density estimators via their uncertainty bounds. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4579–4583. IEEE, 2019.

J. Wan, W. Luo, B. Wu, A. B. Chan, and W. Liu. Residual regression with semantic prior for crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4036–4045, 2019.

Q. Wang, J. Gao, W. Lin, and Y. Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8198–8207, 2019.

M. Xu, Z. Ge, X. Jiang, G. Cui, P. Lv, B. Zhou, and C. Xu. Depth information guided crowd counting for complex crowd scenes. *Pattern Recognition Letters*, 125:563–569, 2019.

L. Zeng, X. Xu, B. Cai, S. Qiu, and T. Zhang. Multi-scale convolutional neural networks for crowd counting. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 465–469. IEEE, 2017.

C. Zhang, H. Li, X. Wang, and X. Yang. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 833–841, 2015.

L. Zhang, M. Shi, and Q. Chen. Crowd counting via scale-adaptive convolutional neural network. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1113–1121. IEEE, 2018.

Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. 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.