Dilated Convolution to Capture Scale-Invariant Context in Crowd Density Estimation

Thishen Packirisamy
School of Computer Science and Applied Maths
University of The Witwatersrand
Johannesburg, South Africa
1839434@students.wits.ac.za

Richard Klein
School of Computer Science and Applied Maths
University of The Witwatersrand
Johannesburg, South Africa
Richard.Klein@wits.ac.za

Abstract—**Crowd density estimation or crowd counting is a challenging problem due to problems such as occlusion and massive scale variations. This research looks to create, evaluate and compare different approaches to crowd counting focusing on the ability for dilated convolution to extract scale-invariant contextual information. We go about conducting this research by building and training three different model architectures: a Convolutional Neural Network (CNN) without dilation, a CNN with dilation to capture context and a CNN with an Atrous Spatial Pyramid Pooling (ASPP) layer to capture scale-invariant contextual features. We train each architecture multiple times to ensure statistical significance and evaluate them using the Mean Squared Error (MSE) and Mean Average Error (MAE) on the ShanghaiTech and UCF_CC_50 datasets. Comparing the results between approaches we find that applying dilated convolution to more sparse crowd images with little scale variations does not make a significant difference but, on highly congested crowd images, dilated convolutions are more resilient to occlusion and perform better. Furthermore, the ASPP model, in capturing features at various scales, improves scale invariance and overall performance on dense crowd images. The code for this research is available at https://github.com/ThishenP/crowd-density.**

Index Terms—**Crowd density estimation, convolutional neural networks, dilated convolution**

I. Introduction

Countless unnecessary deaths are caused by stampedes in dense crowds. This may be due to poor crowd management and a lack of information about the formation of the crowd. The risk of similar tragedies is ever increasing due to the rapid increase in population around the world. Understanding the spatial density distribution of crowds at all times can play a massive role in ensuring the safety of those within the crowds. It is, therefore, imperative that automated methods for accurate Crowd Density Estimation (CDE) are found. These density maps could be used to identify areas with densities above a safe level and allow for the issuing of warnings to these areas and the surrounding areas. It also allows safety officials or algorithms to determine which surrounding parts of the crowd to disperse in order to safely reduce the pressure towards the high-density areas.

Crowd counting and crowd density estimation are different formulations of a similar problem in which a given image is mapped to a crowd count. The step of creating a density map alongside an image count is what differentiates the problems. This mapping can be seen in Figure 1 with an image and its corresponding ground-truth density map. It is however common in the field to refer to both problems under the umbrella term of crowd counting.

This work considers only crowds of people but the field of research does also branch into many other object counting and density estimation fields such as traffic estimation and prediction, cell microscopy and animal crowd analysis [1]. Each sub-field suffers and benefits from similar techniques as found in the overall dense object counting field.

The crowd counting problem was often formulated as a detection based problem [2]–[4], these methods were superseded by traditional regression methods [5], [6]. The success of deep learning has had a large benefit in this field and now the vast majority of the state of the art CDE methods make use of Convolutional Neural Networks (CNN) [7]–[12]. Most of the CNN models take in an image of a crowd and output a density map representing the spatial distribution of objects within an image.

The primary goal of this research is to explore deep learning approaches for crowd density estimation in hopes of finding clarity on the impacts of dilated convolutional kernels in extracting valuable contextual information from crowd images. We do this by building and comparing three models: a baseline CNN that does not capture context, a CNN that captures context through dilated convolution and a CNN that captures scale-invariant context through Atrous Spatial Pyramid Pooling.

We find that applying Dilated convolutions to the problem retains more contextual information and allows the models to be robust to images with a high level of occlusion than traditional convolutions. In addition to this, we find that

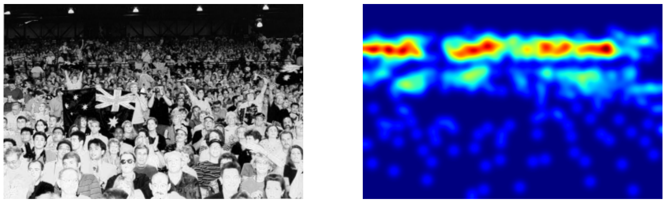

Fig. 1. Example crowd image and corresponding density map

including an Atrous Pyramid Pooling layer within the network captures contextual information at many scales and adds a level of scale invariance which is beneficial to the type of data found in the field of Crowd Density Estimation.

The remainder of this paper goes on to explain the necessary concepts for understanding the paper as well as outline the related work in Section II. The way in which we went about conducting this research is explained in Section III. The specific experiments performed as well as evaluation of the experiments can be found in Section IV. Lastly, the document is concluded in Section V.

II. Background and Related Work

A. Background

1) Common problems in Crowd Density Estimation: One common problem that research in crowd density estimation struggles with is occlusion, which refers to when an object cannot be recognised by sensors used to feed the algorithm. In this case, the camera may be blocked by another object. This is often caused by dense crowd images as many people overlap and block others from the camera. This can therefore make it more difficult for algorithms to recognise all people within the image.

Scale variations also form a large challenge for Crowd Density Estimation. This problem refers to the fact that in many real-world crowd photographs people stand at a variety of distances and angles from the camera. This causes different parts of the images to be at vastly different scales. Researchers, therefore, have to implement scale-invariant algorithms which tends to be difficult and can lead to a higher computational cost.

The field also deals with all the problems that may come with traditional computer vision tasks such as lighting changes, computational cost and low-resolution images.

2) Dilated Convolution: Dilated Convolution refers to convolution in which a filter's receptive field is increased without increasing the filter's overall area. This is achieved by skipping pixels within the filter's grid. The following formula found in [13] illustrates how the pixel positions within the convolutional kernel are being skipped.

$$(G *_l k)(p) = \sum_{s+lt=p} G(s)k(t)$$

where G represents the image, k represents the kernel and l represents the dilation factor. The dilation factor causes the kernel to skip pixels and create a more spread out filter as can be seen in Figure 2. A dilated convolutional layer could take into account a larger receptive field without taking more pixels into account. This can therefore give more contextual information at the same computational efficiency. This also and somewhat more importantly captures contextual information without decreasing the resolution since the same number of pixels are being used to create the output. This could be used to create a deep network that takes into account a large amount of context and does not lose resolution in its mappings. This is, therefore, useful in the crowd density estimation field.

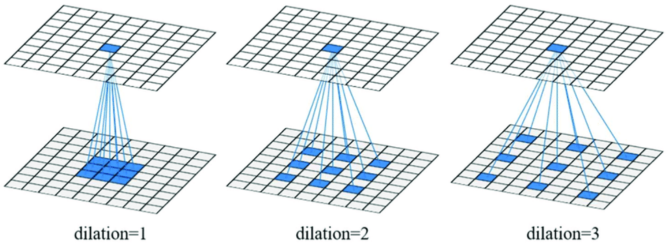

Fig. 2. An illustration of receptive fields for dilated kernels [14]

3) Atrous Spatial Pyramid Pooling: The process of Atrous Spatial Pyramid Pooling(ASPP) [15] involves filtering an input using a variety of dilated kernels each with a different dilation rate and subsequently concatenating the feature maps before being sent to the next layer. This combination of features accounts for context at various scales and allows for a variety of receptive fields to be taken into account further down the network. An illustration of the combination of dilated features can be found in Figure 3. Contextual information becomes very useful in something like crowd density estimation as models could pick up on information around objects that may be obscured or at a very low resolution. It is however difficult to extract useful contextual information when there is a large amount of scale variance. ASPP goes a long way in solving these problems as it can provide similar contextual information at many scales using a variety of receptive fields. This, therefore, helps the model generalise towards multiple scales and achieve a level of scale invariance.

B. Related Work

1) Traditional Approaches: Early crowd counting approaches fell into the category of detection based methods [2]–[4] in which individual objects would be identified and counted to create a final crowd count of the image. These methods were however superseded by regression-based methods [5], [6] that directly learn a crowd count. In recent years deep, fully convolutional methods have dominated the state of the art. These models make use of CNNs to create a density map given a crowd image. This density map is then summed to get the crowd count

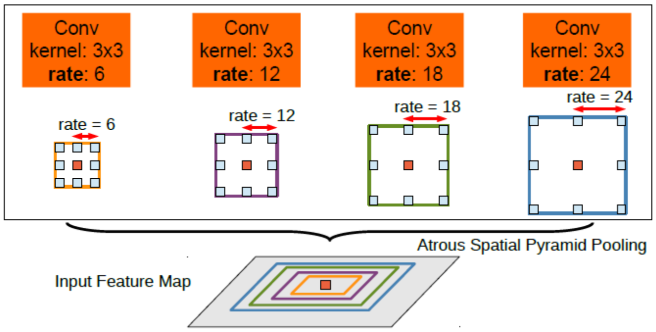

Fig. 3. An illustration of the concatenation of various dilated feature maps [15]

2) CNN based methods: [10] Makes use of a multicolumn CNN with varied receptive fields between columns to account for the sizeable scale variations that are inherent to the CDE problem space. [16] also makes use of multiple columns but feeds in a pyramid of scale varied patches to the network. [11], dealing with the same problems of scale variations, proposed the use of a switching architecture in which an image would be split up into different patches with each patch being fed into a model trained on similar data. [7] accounts for low and high-level features by training two separate CNN columns. One shallow, for low-level features, and one deep, for high-level features. The results of both are merged to create a density map.

[8] applies Spatial Pyramid Pooling [17] to extract scale aware contextual features. [9] includes dilated convolutions network to make use of valuable contextual information. [12] makes use of ASPP layers to extract scale invariant contextual information in crowd images.

C. Problems with Related Work

The multicolumn or multi-network approaches allow for a level of scale invariance, however, they have a higher training and inference cost due to their complexity and model size. [9] takes contextual information into account but does not alleviate the problem of massive scale variations therefore in some cases the context from the dilated layers is not useful.

Previous approaches have various ways of accounting for the extraction of contextual information at various scales but there is still much room for improvement. Exploring the impacts and characteristics of dilated convolutional approaches could provide necessary information for progression towards fully scale-invariant context extraction in CDE.

III. Methodology

We focus this research on dilated convolutions due to their ability to increase the receptive field while retaining the same level of computational efficiency. We specifically aim to explore the contextual and scale-invariant phenomena that can be created through various configurations of dilated convolutions. We subsequently train and compare three different CNN architectures. We analyse how each model impacts the performance of crowd counting and the estimation of crowd density maps. We propose a baseline convolutional model containing no dilation, a basic dilated convolutional model and a dilated convolutional model with an ASPP layer.

A. Ground Truth Generation

The standard ground truth found in crowd counting datasets contains a list of coordinates that correspond to the positions of heads within the crowd image. The number of coordinates in each list gives the overall person count of its corresponding image. A network could output a single count for an image but this ignores the networks ability to predict the spatial distribution and density of crowds within an image. We, therefore, convert the ground truth into a density map.

In order to create the density map we, similarly to **??**, convert the lists of coordinates to a matrix of the same size as their corresponding input image we denote this as $H(x)$. We then convolve the matrix with a Gaussian kernel G_σ which gives us the ground truth density map $F(x)$ as found below.

$$F(x) = H(x) * G_\sigma$$

A σ value of 15 was found to work well in the problem space. Due to the nature of the Gaussian blur, the sum of the density map is roughly equal to the number of coordinates in the original ground truth list. This, therefore, means that summing the density map will return the image crowd count. The model, therefore, optimises for a density map $F(x)$ which captures spatial information. This map can simply be summed to produce the crowd count. In training, the ground truth values are scaled to the size of the model outputs and multiplied by the scaling factor to retain the property that the sum is equivalent to the crowd count.

Although the method of summing the ground truth density map to retrieve the true count is incredibly accurate there is sometimes a very small perturbation. This is fine for training but for testing, in order to have complete accuracy, the points are summed before Gaussian blurring to find the crowd count.

Another benefit of framing the problem in this way is that a fully connected layer is not required meaning the network can be fully convolutional and as a result of this can handle an input image of any shape or resolution.

B. Transfer Learning

The first seven layers of a pre-trained VGG-16 [18] make up the early layers of each proposed model. This was done to alleviate some of the required training time and cost needed to produce accurate deep vision models. We decided to only make use of the lower layers of VGG-16 to decrease the number of max-pooling layers present in the network. This allows for much of the resolution of the original image to be preserved. Preserving resolution is important in a problem space such as CDE due to the fact that important objects often appear incredibly small in crowd images.

C. Approaches

1) Baseline Approach: The baseline approach makes use of a Convolutional Neural Network (CNN) that does not include dilation. This corresponds to a dilation factor of 1. The architecture for this model can be seen in Figure 4 where the images are fed into the VGG layers. The output of the VGG layers is fed into the rest of the layers which we call the dilatable layers. In the case of the baseline, the dilation rate for all layers of the dilatable layers is set to 1. This model is intended to be used as a baseline that does not consider as much contextual information as subsequent dilated convolutional models.

2) Dilated Convolution Approach: The dilated convolution approach adapts the model architecture of the baseline approach but makes use of a dilation factor of 2 for the dilatable layers. The dilation rate causes model to have a larger receptive

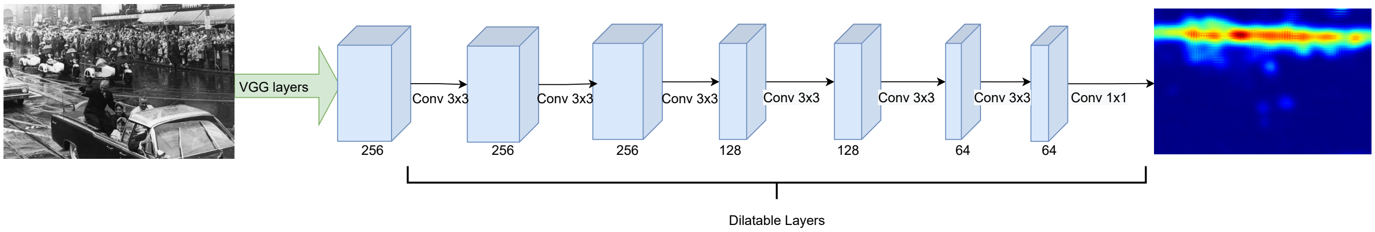

Fig. 4. Base architecture of models. If Dilatable layers have a dilation rate of 1 it is the Baseline model. If If Dilatable layers have a dilation rate of 2 it is the Dilated model. if an ASPP layer is inserted between VGG Layers and Dilatable layers it is the ASPP model

field and consider more contextual information within the crowd scenes.

3) ASPP Approach: The ASPP approach adapts the model architecture of the Dilated Convolution Approach but inserts an ASPP layer after the VGG-16 layers. The architecture of the ASPP layer can be seen in Figure 5. The ASPP layer applies 4 sets of 64 dilated convolutional filters and each set has a different dilation rate. The dilation rates for the 4 sets are 6, 12, 18 and 24. These sets of feature maps are concatenated and fed into the rest of the dilatable layers with a dilation rate of 2. We create this model in hopes of allowing the model to consider context multiple scales and therefore achieve a level of scale invariance.

IV. Experimentation

A. Dataset

We evaluate each model on two commonly used and publicly accessible crowd counting datasets. These being ShanghaiTech [10] and UCF_CC_50 [19].

1) ShanghaiTech: The ShanghaiTech [10] crowd counting dataset is a still image crowd counting dataset that contains 1198 images and 330,165 annotations [1]. The dataset is split into part A and part B. Part A contains high-density images scraped from the internet while part B contains more sparse crowd images collected by fixed street cameras. part A has more diversity as the count range is between 33 and 3139 as opposed to the range of 12 to 578 in part B. Part A and part B have predefined train and test splits with part A having 300 training images and 182 test images. Part B has 400 training images and 316 test images.

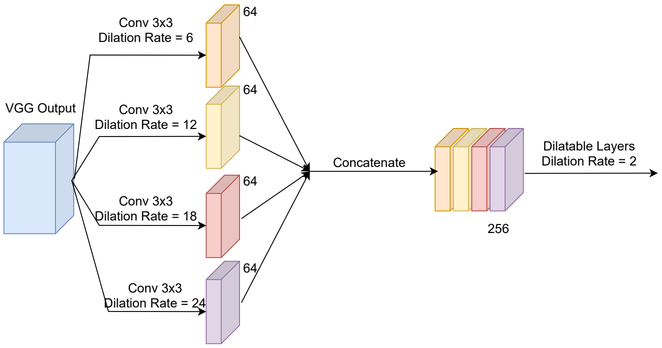

Fig. 5. Architecture of the ASPP Layer

2) UCF_CC_50: The UCF_CC_50 dataset [19] is an incredibly challenging crowd counting dataset which contains 50 images collected from the internet as well as their corresponding annotations. The scenes are often at a large scale including scenes such as stadiums, protests or concerts. The crowd counts of the images range from 96 to 4633.

B. Training Details

1) Training and Validation sets: The training and validation sets are made of a combination of the predefined training sets from the ShanghaiTech dataset. There is no predefined set split for UCF_CC_50 meaning that training on the data would rule out the possibility of evaluation on the dataset. This dataset is an important evaluation benchmark in the field and we, therefore, do not include the dataset in training.

In exploring the ShanghaiTech dataset we find that there is a relative lack of diversity in crowd scale variations in part B when compared to part A. Therefore, to avoid overfitting to part B, we decide to make use of the whole training set from part A but only 100 of the 400 training images from part B.

The validation set is created by splitting the image counts into bins and splitting the validation data off from the training set in a similar way to a stratified split commonly used on classification data. The stratification is however done on intervals of crowd counts rather than classes.

2) Data Augmentation: A train image size is decided upon before training and each input image is randomly cropped to this size. This allows for the ground truth density maps to be scaled by a fixed amount to be used to calculate the loss. In addition to this, it causes more diversity in training, allowing the model to generalise better to the problem space. In some cases, it will flip the image to allow for an added level of diversity in the data.

3) Training Loss: In training the following loss function is optimised, where Θ represents the model parameters, N represents the batch size, $F(X_i;\Theta)$ represents the ith predicted density map in a batch, and F_i represents the ith ground truth density map.

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

The loss function finds the square of the euclidean distance between the two output maps. It, therefore, gives a mean squared error at the density map level. This takes into account every pixel and ensures the model does not optimise for overall crowd count but rather an accurate density map of the image. This train loss function is commonly used in the CDE field [7], [9], [10].

C. Experiments

We train each of the three models using the same training data. Each model is trained ten times on an NVIDIA GTX-3080 GPU for 800 epochs using the Adam optimizer [20]. Number of epochs, optimizer, batch size and learning rate were all optimized using the validation dataset.

- The baseline approach is used as a starting point to which we can compare the dilated approaches. This model is trained using a learning rate of 5×10^{-5} with a batch size of 8.
- The dilated convolution approach assesses the value of including more contextual information in the features maps without increasing computational cost. This model is trained using a learning rate of 2×10^{-5} with a batch size of 8.
- In testing the ASPP approach we hope to analyse the benefits of including multi-scale contextual information in the network. This model is trained using a learning rate of 2×10^{-5} with a batch size of 8.

D. Evaluation Metrics

The performance of each approach is evaluated on full-sized images as opposed to cropped images used for training. Mean Average Error(MAE) and Mean Squared Error(MSE) are used for the evaluation of the models. This is in line with common standards in the field. The formulas for the metrics are shown below in which a lower value indicates superior performance.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y_i}|,$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y_i})^2,$$

where n refers to the number of training examples, y_i represent a single image's labelled crowd count and $\hat{y_i}$ represents a single image's predicted crowd count.

We decide it is not necessary to incorporate patch-based or pixel-based methods to evaluate the spatial accuracy of maps due to the fact that, during training, the model optimises only for accuracy in spatial density distribution. This, therefore, means that any improvement in crowd count must come from improvement in the accuracy of the density map. MSE and MAE of crowd counts, therefore, contain the necessary information to analyse the models.

E. Results

The box plots in Figure 6 and Figure 7 depict the distribution of MSE and MAE obtained from the 10 runs and evaluations of each model. The orange line in the box plot signifies the median while the green triangle signifies the mean. The mean results of the models over 10 runs are presented in Table I. In addition to that the Density map outputs of all models can be found in Figure 8.

TABLE I
COMPARISON OF METHODS TESTED

	ShanghaiTechA		ShanghaiTechB		UCF_CC_50	
Approach	MAE	MSE	MAE	MSE	MAE	MSE
Baseline	91.2	149.6	29.1	40.7	511.0	751.5
Dilated	89.4	130.9	**24.02**	36.2	443.4	654.2
ASPP	**81.5**	**124.9**	25.4	**32.8**	**426.7**	**628.6**

F. Statistical Significance

To ensure insights learned from the research are statistically significant we train ten of each model architecture and combine the metrics into a distribution for each architecture. Each model has a distribution for both its MAE and MSE on each dataset. To test for significance the Mann-Whitney U Test [21] is leveraged. This can give us confidence that there is a significant difference between two distributions meaning it is not likely that the differences are based on chance.

The test finds a statistic ρ which offers information of how likely it is to sample values from the second distribution given the first. The lower the value the fewer similarities there are between the distributions. We consider the difference between distributions to be statistically significant if ρ is less than or equal to 0.05. The ρ values are calculated for pairs of models on each dataset to find out which differences are significant. Table II presents the ρ values given the MAE distributions and Table III presents the ρ values given the MSE distributions.

TABLE II
STATISTICAL SIGNIFICANCE - ρ VALUES USING MAE DISTRIBUTIONS

	ShanghaiTechA	ShanghaiTechB	UCF_CC_50
Approach Pair	ρ	ρ	ρ
Baseline and Dilated	0.43505	0.19234	0.00016
Baseline and ASPP	0.00050	0.26026	0.00009
Dilated and ASPP	0.03201	0.45486	0.00863

TABLE III
STATISTICAL SIGNIFICANCE - ρ VALUES USING MSE DISTRIBUTIONS

	ShanghaiTechA	ShanghaiTechB	UCF_CC_50
Approach Pair	ρ	ρ	ρ
Baseline and Dilated	0.00065	0.15374	0.00012
Baseline and ASPP	0.00009	0.09294	0.00009
Dilated and ASPP	0.00065	0.15374	0.01057

G. Analysis

We find that none of the differences between the models' metrics on ShangahaiTech Part B are statistically significant. The dataset is more sparse leaving less room for improvement from contextual information. In addition to that, the dataset does not contain the level of scale variations that can be found

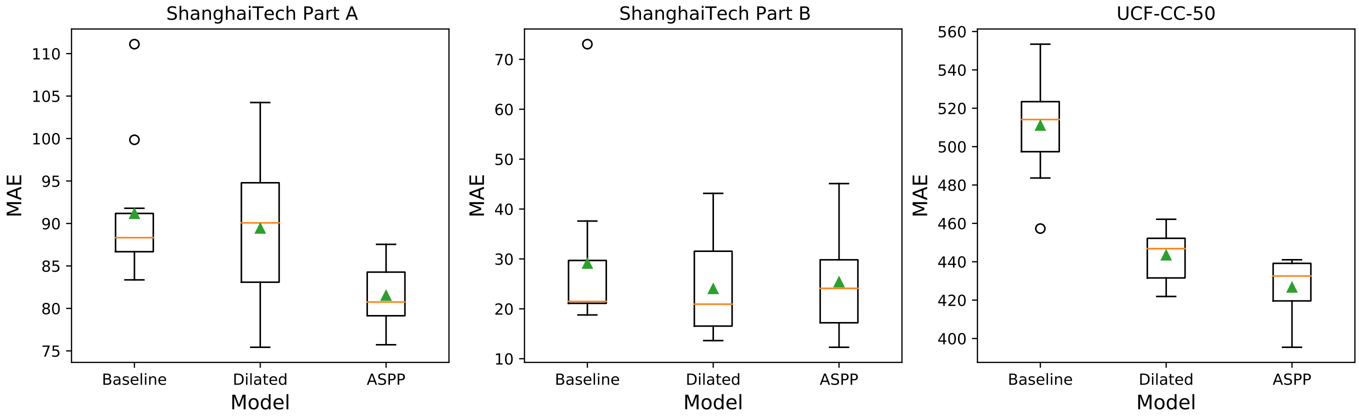

Fig. 6. Distribution of MAE values over 10 trains of each model

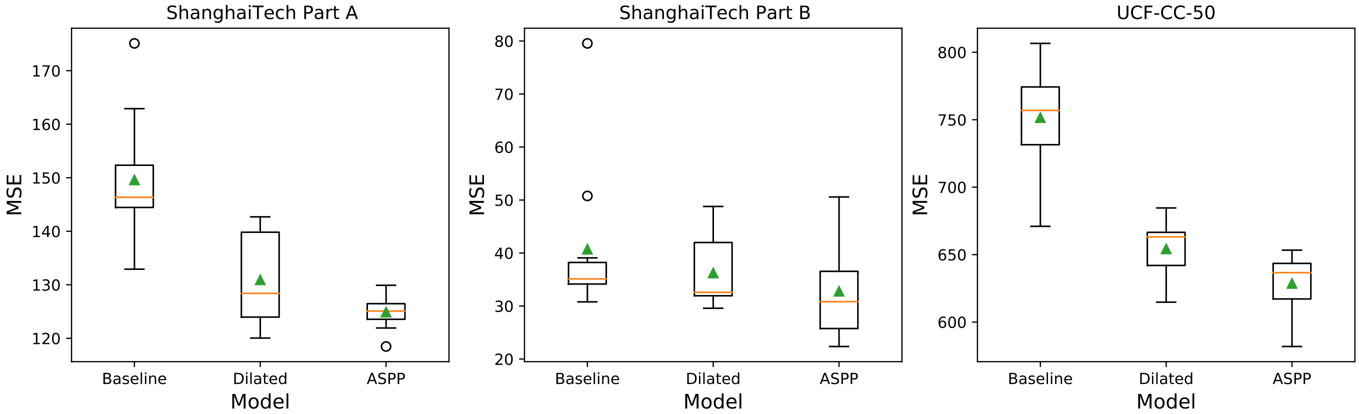

Fig. 7. Distribution of MSE values over 10 trains of each model

in the other two. These characteristics leave little room for the dilated and ASPP models to make an impact. We cannot draw meaningful insights from ShangahaiTech Part B so we consider the more dense datasets: ShanghaiTech Part A and UCF_CC_50.

We find that the difference between the MAE values of the dilated model and the baseline model on ShanghaiTech Part A is not statistically significant and we can therefore not draw insights from it. The dilated model however outperforms the baseline on all combinations deemed to be statistically significant. It outperforms the baseline on the more dense datasets including ShanghaiTech part A and UCF_CC_50. These datasets have a high level of occlusion which the baseline faces difficulty with. The dilated model captures more contextual information. This becomes very useful in the case of heavily occluded objects as the area around an object could provide information about an occluded object. In the case of people, a model could detect an obstructed face using contextual information such as the presence of shoulders. The results, therefore, illustrate the benefit of using dilated kernels to capture context in dense crowd counting.

The ShanghaiTech Part A and UCF_CC_50 datasets both have massive scale variations between images as well as within individual images. We observe that the ASPP approach outperforms the dilated model on these datasets for both MSE and MAE and the results are statistically significant. This illustrates that the ASPP approach successfully captures multi-scale contextual information to increase scale invariance which in turn produces a better crowd count. The Dilated model performs worse due to the fact that the context it learns only applies to a single scale and it, therefore, struggles to adapt to changes in scale.

Considering all the results, we find that when tackling more sparse crowds with fewer scale variations all models perform similarly. Therefore, simpler, less computationally expensive models would be preferred. These would be the baseline or dilated models. We, however, feel the technology could have a bigger impact In the area of more dense crowds with large scale variations. In this case, the ASPP model is preferable.

V. Conclusion

In this paper, we compare three different approaches to Crowd Density Estimation and Crowd Counting. These approaches include a CNN without Dilation, a CNN with dilation and a CNN with an Atrous Spatial Pyramid Pooling (ASPP)

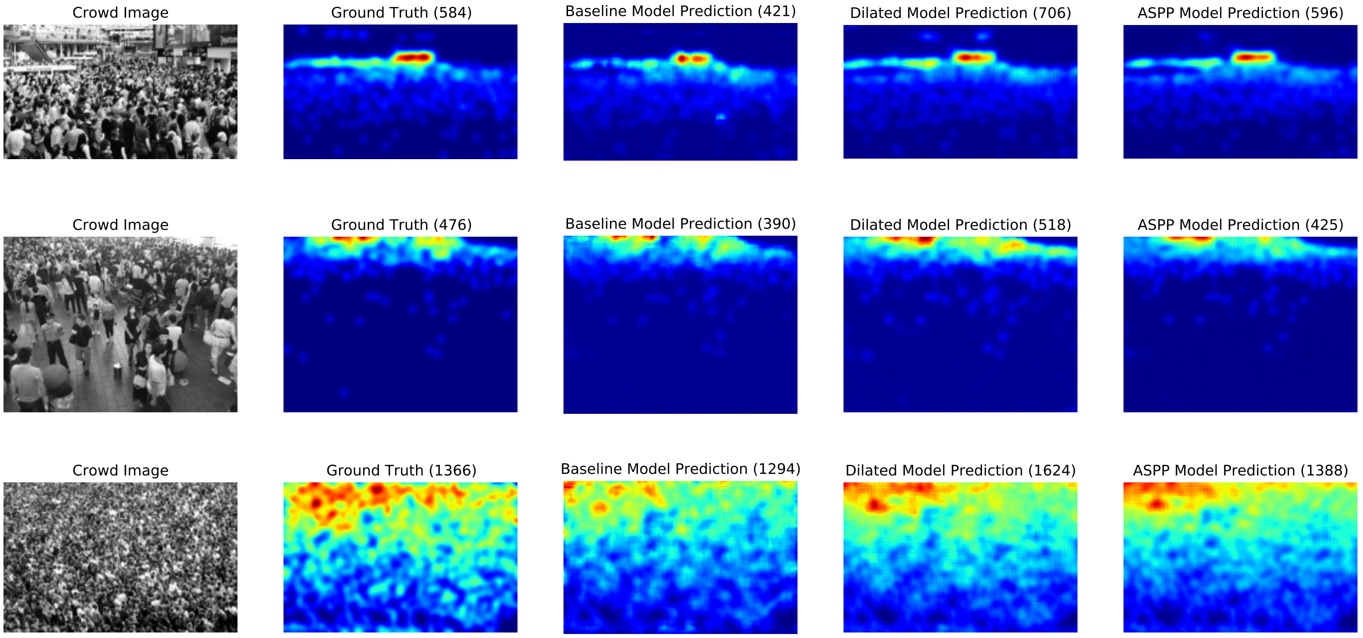

Fig. 8. Density Map Predictions by models on three test images. The numbers in brackets represents the overall crowd count

layer. The model architectures are trained multiple times and evaluated on the ShanghaiTech and UCF_CC_50 datasets. Given the results, we find that if dealing with relatively sparse crowds, dilated convolutional methods do not offer much improvement. However, in more dense crowds a larger receptive field capturing more contextual information is beneficial. Furthermore, in dense crowds, we find that the problem of scale variations can be combated by applying an ASPP layer.

References

[1] G. Gao, J. Gao, Q. Liu, Q. Wang, and Y. Wang, "Cnn-based density estimation and crowd counting: A survey," *CoRR*, vol. abs/2003.12783, 2020. [Online]. Available: https://arxiv.org/abs/2003.12783

[2] M. Wang and X. Wang, "Automatic adaptation of a generic pedestrian detector to a specific traffic scene," in *CVPR 2011*, 2011, pp. 3401–3408.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893 vol. 1.

[4] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection," in *2008 19th International Conference on Pattern Recognition*, 2008, pp. 1–4.

[5] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–7.

[6] ——, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–7.

[7] L. Boominathan, S. S. S. Kruthiventi, and R. V. Babu, "Crowdnet: A deep convolutional network for dense crowd counting," in *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016.* ACM, 2016, pp. 640–644. [Online]. Available: https://doi.org/10.1145/2964284.2967300

[8] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," *CoRR*, vol. abs/1811.10452, 2018. [Online]. Available: http://arxiv.org/abs/1811.10452

[9] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," *CoRR*, vol. abs/1802.10062, 2018. [Online]. Available: http://arxiv.org/abs/1802.10062

[10] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016.* IEEE Computer Society, 2016, pp. 589–597. [Online]. Available: https://doi.org/10.1109/CVPR.2016.70

[11] 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 2017, Honolulu, HI, USA, July 21-26, 2017.* IEEE Computer Society, 2017, pp. 4031–4039. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/CVPR.2017.429

[12] P. Thanasutives, K. Fukui, M. Numao, and B. Kijsirikul, "Encoder-decoder based convolutional neural networks with multi-scale-aware modules for crowd counting," *CoRR*, vol. abs/2003.05586, 2020. [Online]. Available: https://arxiv.org/abs/2003.05586

[13] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: http://arxiv.org/abs/1511.07122

[14] X. Cui, K. Zheng, L. Gao, D. Yang, and J. Ren, "Multiscale spatial-spectral convolutional network with image-based framework for hyperspectral imagery classification," *Remote Sensing*, vol. 11, p. 2220, 09 2019.

[15] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *CoRR*, vol. abs/1606.00915, 2016. [Online]. Available: http://arxiv.org/abs/1606.00915

[16] D. Oñoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 615–629.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *CoRR*, vol. abs/1406.4729, 2014. [Online]. Available: http://arxiv.org/abs/1406.4729

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9,*

2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[19] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2547–2554.

[20] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[21] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50 – 60, 1947. [Online]. Available: https://doi.org/10.1214/aoms/1177730491