EPiC Series in Computing

Volume 85, 2022, Pages 89–103

Proceedings of 43rd Conference of the South African Institute of Computer Scientists and Information Technologists

Dilated Convolution to Capture Scale Invariant Context in Crowd Density Estimation

Thishen Packirisamy[1] and Richard Klein[2]

[1] The University of The Witwatersrand, Johannesburg, South Africa
1839434@students.wits.ac.za

[2] The University of The Witwatersrand, Johannesburg, South Africa
Richard.Klein@wits.ac.za

Abstract

Crowd Density Estimation (CDE) can be used ensure safety of crowds by preventing stampedes or reducing spread of disease which was made urgent with the rise of Covid-19. CDE 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. In this work we build and train 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), Mean Average Error (MAE) and Grid Average Mean Absolute Error (GAME) 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, we find that adding an ASPP layer improves performance in the case when there are significant differences in the scale of objects within the crowds. The code for this research is available at https://github.com/ThishenP/crowd-density.

1 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. In addition to this the Covid-19 pandemic has highlighted the need for understanding densities of crowds in order to decrease the spread of disease. 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 would, therefore, be valuable to develop automated methods for accurate Crowd Density Estimation (CDE).

Crowd counting and Crowd Density Estimation are different formulations of a similar problem in which a given image of crowd is mapped to a number of people within the crowd. 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.

The crowd counting problem was originally formulated as a detection based problem [24, 8, 14] but these methods were superseded by traditional regression methods [4]. 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) [3, 16, 15, 31, 21, 23]. 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. An example of such a density map is shown in Figure 1.

Many of the current approaches to CDE struggle with problems like occlusion and large scale variations [10]. The primary goal of this research is to address the problem of scale variations and occlusion by incorporating Dilated Convolutional layers and Atrous Spatial Pyramid Pooling into the CNNs. These techniques are known to help capture scale-invariant context [5] and we show that they adequately address the problems associated with scale and occlusion in CDE.

We find that applying Dilated convolutions within the network helps it retain more contextual information and allows the models to be robust to images with a higher level of occlusion than traditional convolutions. In addition to this, we find that including an Atrous Pyramid Pooling layer to capture scale invariant context increases performance when the crowd data contains large differences in object scales.

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 2. The way in which we went about conducting this research is explained in Section 3. The specific experiments performed as well as evaluation of the experiments can be found in Section 4. Lastly, the document is concluded in Section 6.

2 Background and Related Work

2.1 Background

2.1.1 Common problems in Crowd Density Estimation

One common problem in crowd density estimation is occlusion, which often occurs when objects overlap or appear to merge from the perspective of the camera. This can make it more difficult for algorithms to recognise all people within the image.

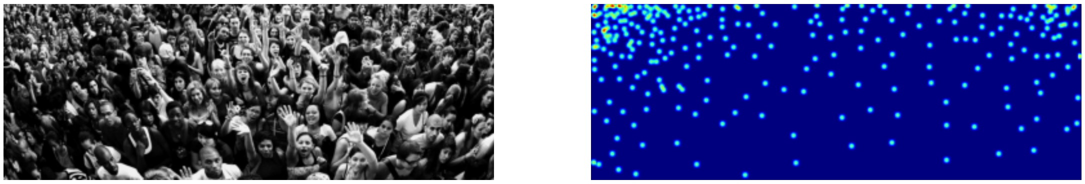


Figure 1: Example crowd image and corresponding density map

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.

2.1.2 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 [29] 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 without increasing the size of the network. 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.

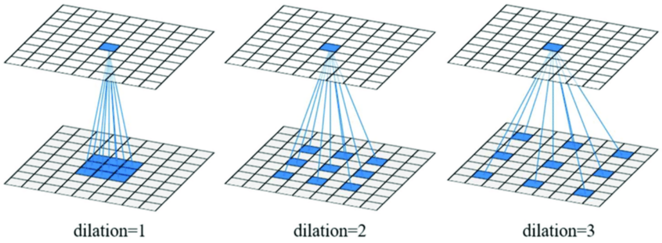


Figure 2: An illustration of receptive fields for dilated kernels [7]

2.1.3 Atrous Spatial Pyramid Pooling

The process of Atrous Spatial Pyramid Pooling (ASPP) [6] involves filtering an input using a variety of dilated kernels each with a different dilation rate. A 1x1 convolution and global average pooling are also applied to the input. The outputs of all previously mentioned operations are then concatenated. A 1x1 convolution is then applied to the concatenation of features to reduce the depth of feature maps. 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 can be useful in crowd density estimation field 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 variance in the scale of the objects being considered. ASPP may offer a way to address these problems as it can provide similar contextual information at many scales using a variety of receptive fields.

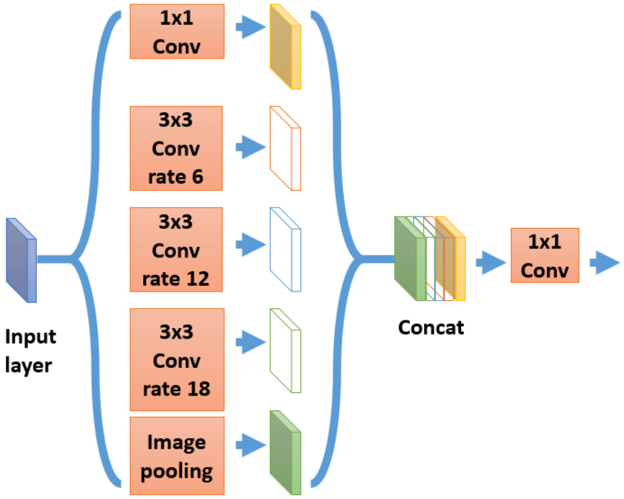


Figure 3: Architecture of an Atrous Spatial Pyramid Pooling Layer [2]

2.2 Related Work

2.2.1 Traditional Approaches

Early crowd counting approaches fell into the category of detection based methods [24, 8, 14] in which individual objects would be identified and counted to create a final crowd count of the image. Early detection methods [27] split images into cells and used techniques such as texture analysis to detect objects within the cells. The cells are subsequently summed to retrieve an overall crowd count. These methods were however superseded by regression-based methods [4] that directly learn a crowd count. Most modern formulations of the problem incorporate density estimation in which a crowd image is mapped to a density map [1]. 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

2.2.2 CNN based methods

Zhang et al [31] make 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. Oñoro-Rubio and López-Sastre [20] also make use of multiple columns but feed in a pyramid of scale varied patches to the network. Sam et al. [21], 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. Boominathan et al. [3] account 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.

Liu et al. [16] apply Spatial Pyramid Pooling [11] to extract scale aware contextual features. Li et al. [15] include dilated convolutions in their network to make use of valuable contextual information. Zhaoyi Yan et al. [28] leverage fractional dilated convolution in which dilation rates are guided by an estimated perspective of an image. This allows for the dilated kernels to be more robust to scale variations. Diptodip Deb and Jonathan Ventura [9] make use of a multi-column CNN similar to Oñoro-Rubio and López-Sastre [20] but leverage dilated convolutions to increase kernel size between layers without increasing the number of parameters. Yu-Jen Ma et al. [17], in estimating crowd density within videos, apply dilated convolutions to reduce the number of parameters and improve inference speed. Nguyen et al. [19] apply a multi-task learning approach in which a density map is created making use of dilated kernels of increasing dilation rate. In order to increase the scale invariance, by understanding perspective, a depth map is also learned. This depth map is taken into account when producing the final density map. Thanasutives et al. [23] make use of ASPP layers to extract scale-invariant contextual information in crowd images. Although much progress has been made in the field, accuracy in high-density images is still a problem due to issues such as high occlusion. This area needs further exploration and research into the contextual information provided by dilated convolution could be valuable.

3 Methodology

This research attempts to tackle scale invariant CDE by exploring the direct impact of dilated convolution and Atrous Spatial Pyramid Pooling. This is done by building and comparing similar models which only significantly differ by what is being tested. Ideas of dilated convolution and ASPP have been used in other CDE research but often form small parts of a larger model. This makes it difficult to understand the impact of the techniques in CDE. We, therefore, believe our more direct comparison is a valuable contribution to the field.

We 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.

3.1 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. In order to capture spatial information of the crowd we convert the ground truth into a density map.

In order to create the density maps we, as is common in the field [23, 25], follow Zhang et al. [31] using a fixed standard deviation $\sigma = 4$. Given N heads present in the image, each head annotation at pixel x_i can be represented as $\delta(x - x_i)$. This is then convolved with gaussian kernel $G_\sigma(x)$ as follows:

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_\sigma(x)$$

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 and 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 ground truth crowd count.

3.2 Approaches

To reduce training cost, the first seven layers of a pre-trained VGG-16 [22] make up the early layers of each proposed model.

3.2.1 Baseline Approach

The baseline approach makes use of a Convolutional Neural Network (CNN) that does not include dilation. 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.

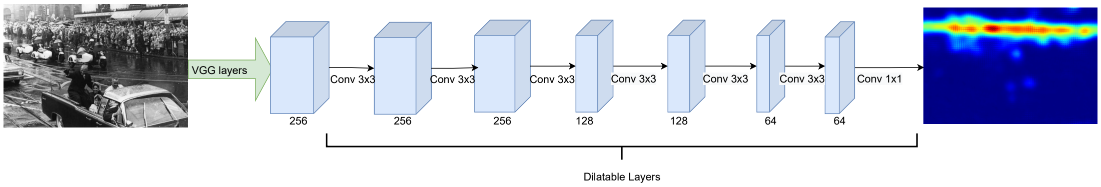


Figure 4: Base architecture of models. If Dilatable layers have a dilation rate of 1 it is the Baseline model. If If Dilatable layers have a dilation rate of 2 it is the Dilated model. if an ASPP layer is inserted between VGG Layers and Dilatable layers it is the ASPP model

3.2.2 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 field and consider more contextual information within the crowd scenes.

3.2.3 ASPP Approach

The ASPP approach adapts the model architecture of the Dilated Convolution Approach but inserts an ASPP layer immediately after the VGG-16 layers. The architecture of the ASPP layer follows the method used in DeeplabV3 [6] which was shown previously in Figure 3. The ASPP layer applies a 1x1 convolution, global average pooling and 3 sets of dilated convolutions with each set having a different dilation rate. The dilation rates for the 3 sets are 6, 12 and 18. These 3 sets of feature maps and the outputs of the pooling and 1x1 convolution are concatenated to create a set of feature maps with depth 1280. In order to decrease depth and combine information a 1x1 convolution layer is applied bringing the depth of the feature maps down to 256. This is then sent through to 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 at multiple scales and therefore achieve a level of scale invariance.

4 Experimentation

4.1 Dataset

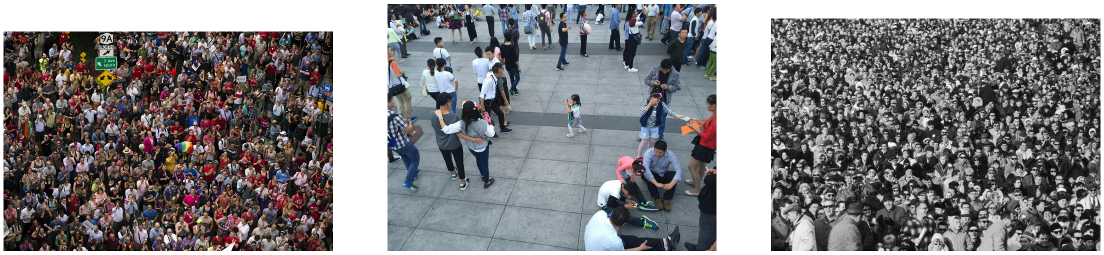
We evaluate each model on two commonly used and publicly accessible crowd counting datasets. These being ShanghaiTech [31] and UCF_CC_50 [12]. Each dataset contains many crowd images and each has corresponding annotations. These annotations are in the form of a list of coordinates of each head present within an image.

4.1.1 ShanghaiTech

The ShanghaiTech [31] crowd counting dataset is a still image crowd counting dataset that contains 1198 images and 330,165 head annotations [10]. 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. Examples from part A and part B can be found in Figure 5a and Figure 5b respectively.

4.1.2 UCF_CC_50

The UCF_CC_50 dataset [12] is a 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 with significant variations in the scale of people within the images. It includes scenes such as stadiums, protests or concerts. The crowd counts of the images range from 96 to 4633. An example from the dataset can be found in Figure 5c.



(a) ShanghaiTech Part A

(b) ShanghaiTech Part B

(c) UCF_CC_50

Figure 5: Examples crowd images from the datasets used in the research

4.2 Training Details

4.2.1 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.

4.2.2 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.

4.2.3 Training Loss

In training we make use of the L_2 loss function, 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 [3, 15, 31].

4.3 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 [13]. Number of epochs, optimizer, batch size and learning rate were all optimized using the validation dataset.

4.4 Evaluation Metrics

The performance of each approach is evaluated on full-sized images as opposed to cropped images used for training. Mean Absolute Error(MAE), Root Mean Squared Error(RMSE) and

Relative Average Error (RAE) are used for the evaluate the predicted crowd counts. 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.

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

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

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

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

The previously mentioned metrics evaluate the overall count accuracy of the predictions, this therefore does not accurately measure the the density distribution of the predicted density map. In order to better evaluate this we, similar to Thanasutives et al. [23], use the Grid Average Mean Absolute Error (GAME).

$$GAME(L) = \frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{4^L}|y_{l,i} - \hat{y}_{l,i}|,$$

In order to obtain the GAME metric the image is split into a grid 4^L non overlapping patches. The MAE value is then found for each patch and combined to get a measure that takes into account the predicted location of density in the map

4.5 Results

The box plot in Figure 6 depicts the distribution of RAE values obtained from the 12 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 tested in this research over 12 runs are presented in Table 1. How the results fit into the context of other literature can be found in Figure 2. In addition to that the Density map outputs of all models can be found in Figure 7.

Table 1: Comparison of methods tested

	ShanghaiTechA			ShanghaiTechB			UCF_CC_50		
Approach	MAE	RMSE	GAME	MAE	RMSE	GAME	MAE	RMSE	GAME
Baseline	99.1	163.7	120.9	31.3	43.4	42.4	534.4	770.6	566.5
Dilated	**81.7**	**130.4**	**99.8**	**27.2**	**38.2**	**36.6**	454.5	664.7	497.2
ASPP	85.7	134.3	104.3	38.5	48.5	47.5	**414.0**	**619.3**	**459.1**

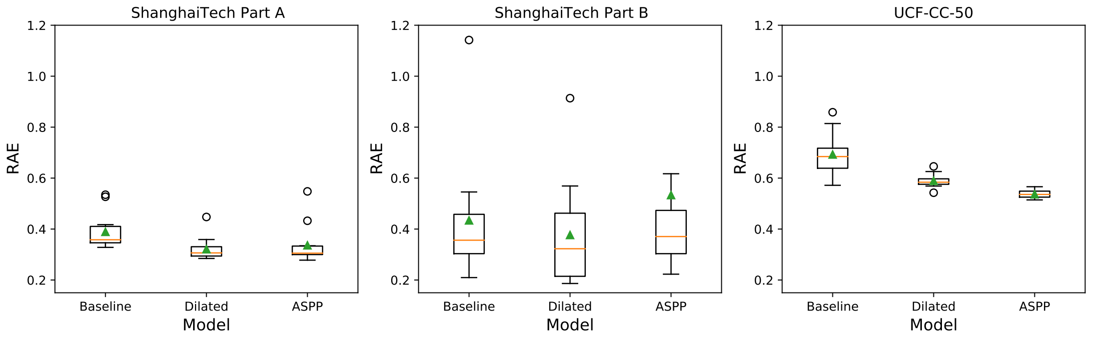


Figure 6: Distribution of RAE values over 12 trains of each model (lower is better)

Table 2: Comparison of methods tested to other methods in literature

	ShanghaiTechA		ShanghaiTechB		UCF_CC_50	
Approach	MAE	RMSE	MAE	RMSE	MAE	RMSE
Baseline (ours)	99.1	163.7	31.3	43.4	534.4	770.6
Dilated (ours)	81.7	130.4	27.2	38.2	454.5	664.7
ASPP (ours)	85.7	134.3	38.5	48.5	414.0	619.3
MCNN [31]	110.2	173.2	26.4	41.3	377.6	509.1
M-SFANet+M-SegNet [23]	**57.6**	**94.5**	6.32	**10.0**	**167.5**	**256.3**
CSRNet [15]	68.2	115.0	10.6	16.0	266.1	397.5
SGANet [26]	58.0	100.4	**6.3**	10.6	224.6	314.6
Zhang et al [30]	181.8	277.7	32.0	49.8	467.0	498.5

4.6 Statistical Significance

To ensure insights learned from the research are statistically significant we train 12 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 [18] is leveraged. We consider the difference between distributions to be statistically significant if ρ is less than or equal to 0.05. Tables presenting the ρ values given the MAE and MSE distribution can be found in Section A. In both tables the ρ value for statistically significant combinations is in bold. We find that all differences between models on UCF_CC_50 are statistically significant and all differences on ShanghaiTech B are not. On ShanghaiTech A the differences between the Baseline and Dilated models as well as Baseline and ASPP models are statistically significant.

4.7 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. We cannot draw meaningful insights from ShanghaiTech Part B so we consider the more dense datasets: ShanghaiTech Part A and UCF_CC_50.

The dilated model outperforms the baseline on the more dense datasets including ShanghaiTech part A and UCF_CC_50 and this difference is found to be statistically significant. The improvements in performance come in both the overall count and density distribution given the GAME score. 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.

In this case of ShanghaiTech Part A, we fail to reject the null hypothesis that the ASPP model differs significantly to the dilated model based on the tested metrics. This illustrates that the more scale invariant context from the ASPP layer does not alter the results significantly in the dataset. The dilated model's dilation rate of 2 is sufficient to capture context in this dataset and the ASPP model may be overkill for the level of object scale variation found in the dataset. We do, however, reject the null hypothesis on UCF_CC_50 where the ASPP model out performs the dilated model in both overall count and density distribution. The ASPP model seems to perform better on this dataset as it contains larger scale variations in a single dataset than the ShanghaiTech datasets. We believe that the increase in performance on UCF_CC_50 is due to the ability of the ASPP model to capture scale invariant context using multiple receptive fields.

Considering all the results, we find that when tackling more sparse crowds with fewer scale variations all models perform similarly. Therefore, the simpler CNN without dilation would be sufficient. We, however, feel the technology could have a bigger impact In the area of more dense crowds with large scale variations. The dilated model would be beneficial in higher density crowd images as it captures useful contextual information. However, if a high level of scale variation of objects within the crowds is expected the ASPP model would be preferable.

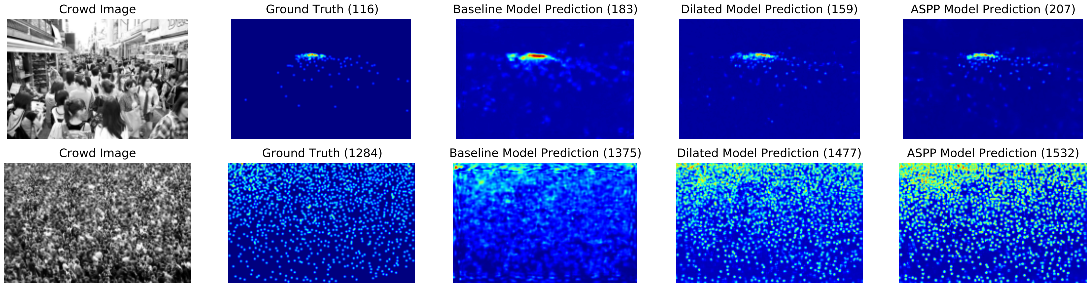


Figure 7: Density Map Predictions by models on two test images. The numbers in brackets represents the overall crowd count

5 Future Work

This research provides insights into the effects and benefits of configurations of dilated convolution in the context of crowd counting. There is however room to extend this research as problems such as scale variation and occlusion are still a challenge. We notice that the ASPP model, while providing a level of scale invariance in the smaller range of object sizes it, does not provide benefits on datasets in which the scale of objects are significantly larger. We therefore believe that a wider range of receptive fields could be added to the ASPP concatenation to address this. This therefore leads into further work in which data sets with larger variations in scale can be collected in order to optimise for more robust scale invariant crowd density estimation.

Dilated convolution allows for a network to capture a large receptive field while training fewer parameters than a dense kernel. We therefore suggest that research be done on the possible computational cost and speed benefits of the approach. This could be extended to tackle the problem of crowd counting in videos.

6 Conclusion

In this work, 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) 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 with significant scale variations, we find that an ASPP layer improves on the dilated model by capturing scale invariant context. We hope that this research will help to inform decisions about when and how to use configurations of dilated convolution in future work.

References

[1] Carlos Arteta, Victor S. Lempitsky, and Andrew Zisserman. Counting in the wild. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, volume 9911 of *Lecture Notes in Computer Science*, pages 483–498. Springer, 2016.

[2] Bilel Benjdira, Kais Ouni, Mohamad Alrahhal, Abdulrahman Albakr, Amro Al-Habib, and Emad Awwad. Spinal cord segmentation in ultrasound medical imagery. *Applied Sciences*, 10, 02 2020.

[3] Lokesh Boominathan, Srinivas S. S. Kruthiventi, and R. Venkatesh 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*, pages 640–644. ACM, 2016.

[4] Antoni B. Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2008.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016.

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[7] Ximin Cui, Ke Zheng, Lianru Gao, Dong Yang, and Jinchang Ren. Multiscale spatial-spectral convolutional network with image-based framework for hyperspectral imagery classification. *Remote Sensing*, 11:2220, 09 2019.

[8] 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)*, volume 1, pages 886–893 vol. 1, 2005.

[9] Diptodip Deb and Jonathan Ventura. An aggregated multicolumn dilated convolution network for perspective-free counting. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 195–204. Computer Vision Foundation / IEEE Computer Society, 2018.

[10] Guangshuai Gao, Junyu Gao, Qingjie Liu, Qi Wang, and Yunhong Wang. Cnn-based density estimation and crowd counting: A survey. *CoRR*, abs/2003.12783, 2020.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *CoRR*, abs/1406.4729, 2014.

[12] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2554, 2013.

[13] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[14] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu 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*, pages 1–4, 2008.

[15] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. *CoRR*, abs/1802.10062, 2018.

[16] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. *CoRR*, abs/1811.10452, 2018.

[17] Yu-Jen Ma, Hong-Han Shuai, and Wen-Huang Cheng. Spatiotemporal dilated convolution with uncertain matching for video-based crowd estimation. *IEEE Transactions on Multimedia*, 24:261–273, 2022.

[18] 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*, 18(1):50 – 60, 1947.

[19] Minh-Nghia Nguyen, Vu-Hoang Tran, and Ton-Nghia Huynh. Depth embedded and dense dilated convolutional network for crowd density estimation. In *2021 International Conference on System Science and Engineering (ICSSE)*, pages 221–225, 2021.

[20] Daniel Oñoro-Rubio and Roberto J. López-Sastre. Towards perspective-free object counting with deep learning. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 615–629, Cham, 2016. Springer International Publishing.

[21] Deepak Babu Sam, Shiv Surya, and R. Venkatesh 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*, pages 4031–4039. IEEE Computer Society, 2017.

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[23] Pongpisit Thanasutives, Ken-ichi Fukui, Masayuki Numao, and Boonserm Kijsirikul. Encoder-decoder based convolutional neural networks with multi-scale-aware modules for crowd counting. *CoRR*, abs/2003.05586, 2020.

[24] Meng Wang and Xiaogang Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *CVPR 2011*, pages 3401–3408, 2011.

[25] Qian Wang and Toby P. Breckon. Segmentation guided attention network for crowd counting via curriculum learning. *CoRR*, abs/1911.07990, 2019.

[26] Qian Wang and Toby P. Breckon. Crowd counting via segmentation guided attention networks and curriculum loss. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–11, 2022.

[27] Xinyu Wu, Guoyuan Liang, Ka Keung Lee, and Yangsheng Xu. Crowd density estimation using

texture analysis and learning. In *2006 IEEE International Conference on Robotics and Biomimetics*, pages 214–219, 2006.

[28] Zhaoyi Yan, Ruimao Zhang, Hongzhi Zhang, Qingfu Zhang, and Wangmeng Zuo. Crowd counting via perspective-guided fractional-dilation convolution. *IEEE Transactions on Multimedia*, 24:2633–2647, 2022.

[29] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[30] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 833–841, 2015.

[31] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi 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*, pages 589–597. IEEE Computer Society, 2016.

A Statistical significance ρ values

Table 3: Statistical Significance - ρ values using MAE distributions

	ShanghaiTechA	ShanghaiTechB	UCF_CC_50
Approach Pair	ρ	ρ	ρ
Baseline and Dilated	**0.00122**	0.25336	**0.00037**
Baseline and ASPP	**0.00363**	0.46549	**0.00002**
Dilated and ASPP	0.46549	0.14274	**0.00005**

Table 4: Statistical Significance - ρ values using MSE distributions

	ShanghaiTechA	ShanghaiTechB	UCF_CC_50
Approach Pair	ρ	ρ	ρ
Baseline and Dilated	**0.00012**	0.14274	**0.00100**
Baseline and ASPP	**0.00045**	0.37542	**0.00006**
Dilated and ASPP	0.46549	0.25336	**0.00255**

Table 5: Statistical Significance - ρ values using GAME distributions

	ShanghaiTechA	ShanghaiTechB	UCF_CC_50
Approach Pair	ρ	ρ	ρ
Baseline and Dilated	**0.00068**	0.08743	**0.00030**
Baseline and ASPP	**0.00213**	0.25336	**0.00001**
Dilated and ASPP	0.37542	0.23524	**0.00012**