Vol. 5, No. 1 | January – June 2022

# Multi-Scale Pooling In Deep Neural Networks For Dense Crowd Estimation

Ali Raza Radhan[1], Fareed Ahmed Jokhio[2], Ghulam Hussain[1,3], Kamran Javed[4], Arsalan Ahmed[1]

**Abstract:**

*State-of-the-art-methods for counting persons in dense crowded places lack in estimating accurate crowd density due to following reasons. They typically apply the same filters over a complete image or over big image patches. Only then the perspective distortion can be compensated by estimating local scale. It is achieved by training an additional classifier with the optimal kernel size chosen from limited choices. These methods are restricted to the context they are applied on because they are not end-to-end trainable; cannot justify quick scale changes because they allocate a single scale to big image patches; and can only utilize a narrow range of receptive fields for the networks to be of a feasible size. In this study, we bring in an end-to-end trainable deep architecture that merges features achieved from multiple kernels of different sizes and learns various essential features such as quick scale changes and to utilize the right context at each image location. This technique flexibly encodes scale of related information to precisely predict crowd density. The training and validation loss of the proposed approach is 5% and 4% lower than the state-of-the-art context aware method, respectively.*

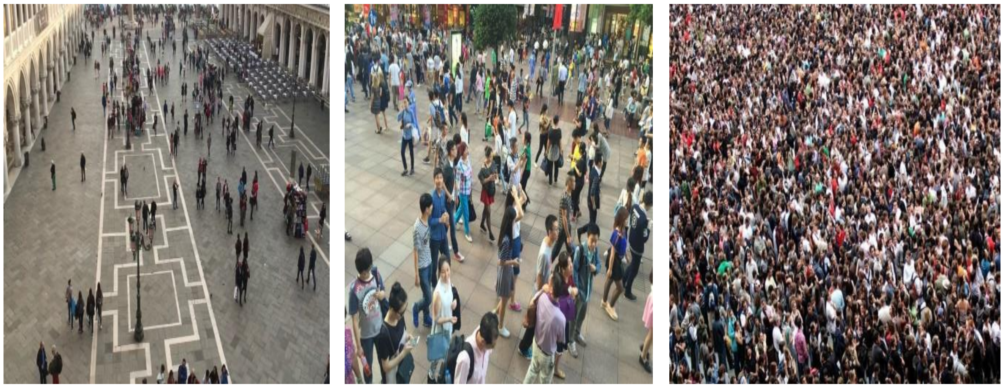
**Keywords:** *Perspective Distortion, local scale, image patches, crowd counting, deep learning*.

## 1 Introduction

Crowd counting has become an interesting topic in the recent years for researchers due to its broad applications, including crowd monitoring, traffic control, public safety, and event planning, video surveillance and city management. Over the last few years, the density map generation methods have been developed to count people in a scene by training regressors to estimate people density per unit area and get total number of counts by integration without detecting people individually. Currently deep learning methods [28-33] have become prevailing tool for crowd counting, due to powerful learning ability of convolutional neural networks (CNNs). Although crowd counting algorithms have been broadly examined by previous methods [20-22, 28- 33], but handling large density variance in crowd images which causes occlusion and perspective distortion that still remained a challenging problem. As illustrated in fig.1, the densities of a crowd vary significantly from low crowd (e.g. Venice dataset) to extremely dense crowd (Shanghai Part A dataset). Such large variation in density of people is a great challenge for CNN models and creates problems for predicting accurate density map. Most deep learning-based approaches [30-36] use same filters and

[1]Dept. Electronic Engineering, Quaid-e-Awam University, Larkana, Pakistan.
[2]Dept. Computer System Engineering, Quaid-e-Awam University, Nawabshah, Pakistan
[3]Dept. Electronic Engineering, Sungkyunkwan University, Suwon, South Korea
[4]National Centre of Artificial Intelligence (NCAI), Saudi Data and Artificial Intelligence
Corresponding Author: fajokhio@quest.edu.pk

pooling operations on a complete image. These depend upon fix sizes of receptive fields. However, one ought to alter receptive field size over an image to get better results. Stunning advancement has been attained by learning a density map by designing multi-scale structures [39] or accumulating multi-scale features [40,41], which shows capacity to manage with density variation is tuff for crowd counting approaches and remains a huge challenge. Figure 2 compares the Mean Absolute Error of several techniques on three standard benchmark crowd counting datasets having different crowd densities. Result indicates the strength of proposed method to high scale variance.



Low Medium High

**Fig. 1.**Crowd Types

Such large variation in density of people is a great challenge for CNN models and creates problems for predicting accurate density map. Most deep learning-based approaches [30-36] use same filters and pooling operations on a complete image. These depend upon fix sizes of receptive fields. However, one ought to alter receptive field size over an image to get better results. Stunning advancement has been attained by learning a density map by designing multi- scale structures [39] or accumulating multi- scale features [40,41], which shows capacity to manage with density variation is tuff for crowd counting approaches and remains a huge challenge. Figure 2 compares the Mean Absolute Error of several techniques on three standard benchmark crowd counting datasets having different crowd densities. Result indicates the strength of proposed method to high scale variance.

In this paper, we introduce a deep learning method that gets features from different receptive field sizes and learn the importance of each feature at different image locations and accounts for rapid scale changes. Our method can alleviate the problem of density variation, occlusion, and perspective distortion by using multi-scale pooling operation and give better results than other state-of-art methods. This is contrast to crowd counting methods that work on the density variation as in [19, 42], but different only in the loss function as we get the accurate multi-scale features with minimum error. Experiments are done on several standard crowds counting benchmark datasets such as ShanghaiTech Part A, Part B and Venice which show large density variation as shown in figure 1.

## 2 Letrature Review

Most of the initial research was centered on detection-based crowd counting, where rectangular bounding boxes were used to detect persons in the scene [1] and this information was used to count the number of persons [2]. People are counted either by whole body detection or by different body parts-based detection. Same class detection techniques [3,4] generally are conventional pedestrian detection ways which train a classifier using attributes (such as Haar wavelets [5], histogram-oriented gradients [4], edgelet [6] and shapelet [7]) taken out from a full body. Several learning methods such as Support Vector Machines, boosting [8] and random forest [9] have been used with different level of success. Though fruitful in small scale crowd scenes, these methods are badly affected in dense crowd places. Analysts have endeavored to address this issue by embracing part-based detection methods [10, 11], where one applies boosted classifiers for particular body parts such as the head and shoulder to approximate the people counts in that specific area [2].

Though part-based detectors were used to reduce the problems of overlapping, these mechanisms had not fruitful results in the presence of highly congested crowds and high background litter.

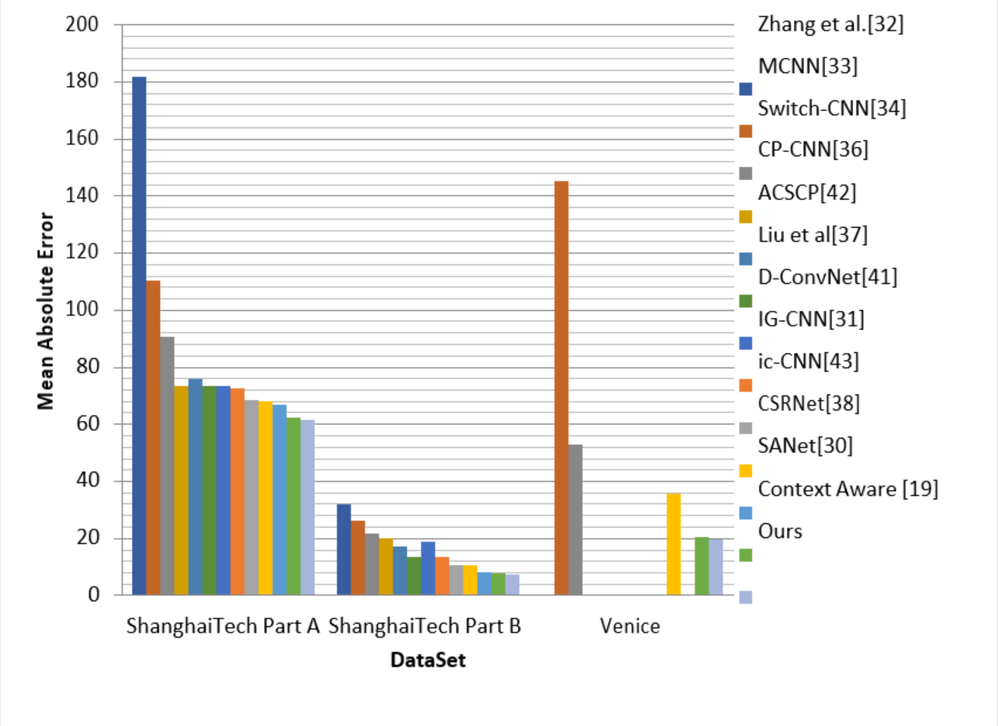


**Fig. 2.**The performance comparison of the state-of-the-art methods

To solve these problems researchers tried to count by regression where they apply a mapping between features brought out from local image patches to their researcher's counts [12, 13]. By counting using regression, these strategies dodge reliance on teaching detectors which could be a generally complex task. In recent research, Idrees et al. [23] recognized that in congested crowds, no any detection method is effective enough to give precise data for counting people due to problems of overlapping, low quality images and perspective distortion. Furthermore, they noticed that there exists a spatial relationship that can be utilized to oblige the count approximation in adjacent local regions. With these ideas in mind, they suggest finding features using various methods that gather different information. By treating thickly populated crowds of people as irregular and non-homogeneous surface, they applied

Fourier analysis along with head detections and Scale-Invariant Feature Transform (SIFT) interest point based counting in local adjacent regions. The three sources, i.e., Fourier, interest points and head detection are then merged with their respective confidences and counts at localized patches are determined independently.

As the previous methods were successful in addressing the problems of occlusion and clutter, they mostly neglected important spatial information as they were regressing on the global count. In variance, Lempitsky et al. [24] proposed a method to learn a linear mapping between local patch features and object density maps, thereby absorbing spatial information in the learning procedure. Perceiving that it is hard to learn a linear mapping, Pham et al. [25] proposed a method to learn a non-linear mapping between local patch features and density maps. They take multiple image patches using random forest regression to vote for densities of various target objects to learn a non-linear mapping.

Similar to the above technique, Wang and Zou [26] proposed a fast method for density estimation based on subspace learning viewing the computational complication point of view of the existing methods. In a recent approach, Xu and Qiu [27] perceiving that the existing crowd density estimation strategies utilized a smaller set of features results in limiting their capability to perform better. They put forward a method to increase the accuracy of crowd density estimation by utilizing a large set of features. As the regression methods used by previous methods (based on Gaussian process regression or Ridge regression) are computationally complex and are not able to operate very high-dimensional features, they utilized random forest as the regression model whose tree structure is speedy and flexible. Unlike conventional methods to random forest construction, they inserted random projection in the tree nodes to tackle the problem of dimensionality and to introduce randomness in the tree structure.

Now density-based counting methods are mostly superseded by convolutional neural networks-based methods where instead of looking at the patches of an image, researchers form an end-to-end regression method using CNNs. Wang et al. [28] applied CNNs firstly for the task of crowd density estimation. Wang et al. gave an end-to-end deep CNN regression model for counting persons from images in highly dense crowds. In addition, to minimize false responses background like trees and buildings in the images, training data is increased with extra negative samples having ground truth count is fixed as zero. In another approach, Fu et al. introduced to divide the image into one of the five categories: very high density, high density, medium density, low density and very low density rather than estimating density maps, they utilized a combination of two classifiers to acquire boosting in which the first classifier samples misclassified images and the other one reclassifies rejected samples.

The approach of [29, 30] utilizes image patches taken out at multiple scales as source to a multi stream network. They then either combine the features for final density prediction [29] without continuous scale changes or introduce an adhoc term in the training loss function [30] predict consistency across scales forcibly. This, however, does not take contextual information into the features achieved by the algorithm, so has limited effect. Whereas approach [31] learns multi-scale features, by utilizing various receptive fields, they merge all of these features to estimate the density.

While the earlier approaches account for scale, they neglect the reality that the suitable scale changes smoothly over the image should be controlled dynamically. This was proposed in [32] by weighting various density maps generated from input images at different scales. As the density map at each scale just relies upon features extracted from a specific scale, and consequently be ruined by the absence of versatile-scale reasoning. Here, we content that one should rather extract features at various scales and figure out how to adaptively merge them. While this, basically was also the inspiration of [33] , which train an additional classifier to dole out the best receptive field for each image patch, these techniques stay restricted in many significant manners. Firstly, they depend on classifiers, which need pre-training the network before training the classifier, and in this way it is not end-to-end trainable. Secondly, they commonly assign a single scale to a whole image patch that can still be large and in this manner don't represent quick scale changes. Lastly, the range of receptive field sizes they reliance remain limited in part because using much bigger ones would need using more complex architectures, which may be difficult to train given sort of networks being used.

In this study, we bring in an end-to-end trainable deep architecture that merges features achieved from multiple kernels of different sizes and learns various essential features such as quick scale changes and to utilize the right context at each image location. This technique flexibly encodes scale of related information to precisely predict crowd density. The training and validation loss is relatively low than above methods.

## 3 Method

### 3.1 Overview

As discussed above, we target the perspective distortion problem in crowd counting methods where density of people varies from low to extreme level. For decreasing the training and validation loss and to improve the generalization of the model with density variation, we trained it on optimal number of epoch and batch size to get the final density map.

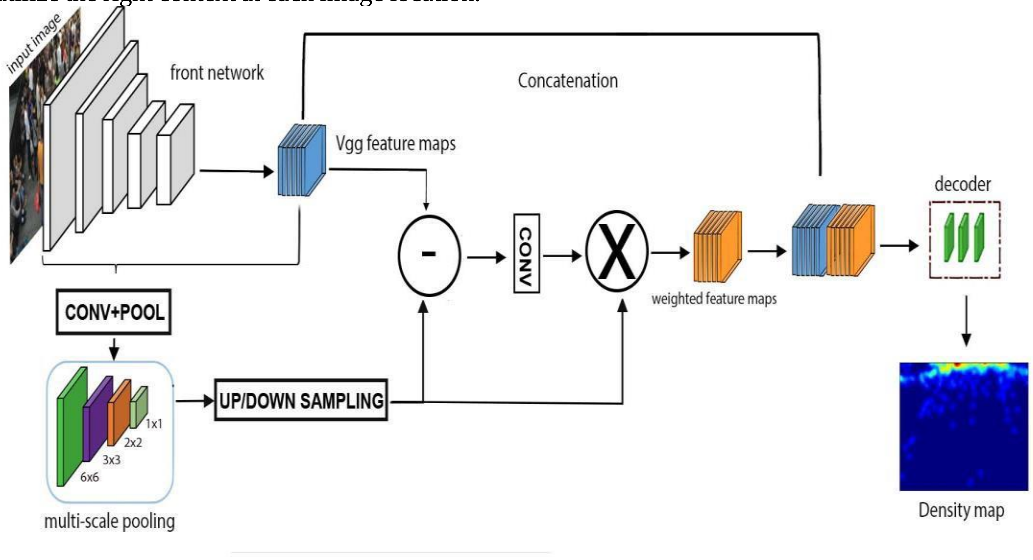


**Fig. 3.**The proposed methodology

### 3.2 Scale Learning Module

We aim to generate an estimated density map that is very similar to ground-truth density map of the given set of M training images with corresponding ground-truth density maps. Images are given to initial point of our network having first ten layers of pre- trained *VGG-16* network to get *Vgg* feature maps as shown in fig.2. These *Vgg* features are the base to our learning-scale model. As discussed earlier in section 2, *Vgg-16* network use same filters and pooling operations on a complete image and depends upon fix sizes of receptive fields. So, it is less efficient in rapid scale change scenarios. To solve this, we alter receptive field size over an image to get learning-scale features by applying multi- scale pooling 1X1, 2X2, 3X3 and 6X6 on Vgg features. This forms a pyramid structure. Each pooling feature shows a density map, hence four different densities extract information from the given image.

These multi-scale densities are then up sampled to the size of convolution layers by linear interpolation and finally combined to give the weighted features. In the end weighted features are concatenated with *Vgg* features to predict the final density map of underlying image as shown in Figure 3.

Scale Learning Network is illustrated in figure 3. RGB images are at input to a front network that contains the first 10 layers of the *VGG-16* network. The resulting *Vgg* features are organized in blocks of various sizes by average pooling succeeded by a 1×1 convolutional layer. They are then up-sampled back to the original feature size to form the weighted features. Contrast or weighted features are further utilized to memorize the weights for the learning-scale features that are then fed to a back-end network to deliver the ultimate density map.

### 3.3 Training Details and Loss Function

Our network is end-to-end trainable which involves L2 loss defined as

$$L(\sigma) = \frac{1}{2B} \sum_{i=1}^{B} \{ \, D_i^{gt} - \, D_i^{pre} \, \}_2^2 \qquad (1)$$

Where B is the batch size, $\boldsymbol{\sigma}$ is the non-linear mapping parameter that maps an input image $\boldsymbol{I}$i to a predicted density map $\boldsymbol{D_i^{pre}}$. $\boldsymbol{D_i^{gt}}$ is the ground-truth density map that we get as in [19] by convolving an image having ones at people head's locations and zeros elsewhere with a Gaussian kernel $\boldsymbol{N^{gt}}$ $(\boldsymbol{p \backslash \mu, \delta^2})$ we have

$$\forall p \ \in I_i, D_i^{gt} \ (\mathrm{p} \backslash I_i) =$$

$$\sum_{j=1}^{ci} N^{gt} \ \left(p \backslash \mu = \ P_i^j, \delta^2\right) \qquad (2)$$

Where $\boldsymbol{\mu}$ and $\boldsymbol{\delta}$ are the mean and standard deviation of the normal distribution. To reduce the loss, we apply stochastic Gradient Descent with batch size 1 for ShanghaiTech Part_A dataset because it has various size images and got impressive results after training the model for 150 epochs. For other two Venice and ShanghaiTech_Part_B fixed image size datasets, we use Adam with batch size 16 and trained the model for 100 epochs

## 4 Experiments

### 4.1 Evaluation Matrices

We apply two standard error terms, i.e. Mean Absolute Error (MAE) and Mean Squared Error (MSE) for evaluation purpose and compare our results with other methods. These are defined as

$$MAE = \frac{1}{K} \sum_{i=1}^{K} |x_i - \hat{x}_i| \ , MSE = \sqrt{\frac{1}{K} \sum_{i=1}^{K} (x_i - \hat{x}_i)^2}, \qquad (3)$$

Where K is the total number of images of test images, $\boldsymbol{x_i}$ represents the ground-truth and $\boldsymbol{\hat{x}_i}$ is the predicted number of people in the $\boldsymbol{i^{th}}$ image.

### 4.2 Benchmark Datasets and Ground-Truth Data

We took three standard benchmark datasets including ShanghaiTech Part_A, ShanghaiTech Part_B and Venice to compare our method with other approaches. ShanghaiTech [27]. The shanghaiTech crowd counting dataset comprises of 1198 images with approximately 330,165 people in them. It is divided in two parts ShanghaiTech Part_A with 482 images which are randomly taken from internet and ShanghaiTech Part_B having 716 images taken from busy streets of metropolitan area in Shanghai city. In part_A 300 images and in Part_B 400 images are reserved for training set. The remaining images of both parts are used for testing purposes. The ground-truth density maps for part_A were generated by adaptive Gaussian kernels and for part_B by fixed kernels. Venice [19]. Venice dataset contains 167 fixed size 1280X720 resolution images from Piazza San Marco in Venice. In this dataset 80 images forms a training data set, and 87 images are used for testing purpose. The images of Venice dataset are more calibrated than

ShanghaiTech. The ground-truth density maps are generated by fixed Gaussian kernels as in ShanghaiTech part_B data set.

### 4.3 Ablation Study

Ablation study is mainly performed on ShanghaiTech and Venice datasets, as there could be the more problem of rapid scale change due to density variation. Here we confirm the benefits of specific features.

Table 1 shows comparison of Mean Absolute Error and Root Mean Square Error of different approaches on three standard benchmark crowd counting datasets having people density variation. Results in Figure 4 indicate the strength of proposed method on low, medium, and high types of crowds.

TABLE I. COMPARATIVE RESULTS ON THE VENICE DATASET & SHANGHAITECH DATASET VARIATION IN CROWD COUNTING. WE SUMMED UP

| | **Venice** | | **Shanghai Part-A** | | **Shanghai Part-B** | |
|---|---|---|---|---|---|---|
| **Method** | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| **Zhang et al.[32]** | - | - | 181.8 | 277.7 | 32.0 | 49.8 |
| **MCNN [33]** | 145.4 | 147.3 | 110.2 | 173.2 | 26.4 | 41.3 |
| **Switch-CNN[34]** | 52.8 | 59.5 | 90.4 | 135.0 | 21.6 | 33.4 |
| **CP-CNN[36]** | - | - | 73.6 | 106.4 | 20.1 | 30.1 |
| **ACSCP[42]** | - | - | 75.7 | 102.7 | 17.2 | 27.4 |
| **Liu et al.[37]** | - | - | 73.6 | 112.0 | 13.7 | 21.4 |
| **D-ConvNet[41]** | - | - | 73.5 | 112.3 | 18.7 | 26.0 |
| **IG-CNN[31]** | - | - | 72.5 | 118.2 | 13.6 | 21.1 |
| **Ic-CNN[43]** | - | - | 68.5 | 116.2 | 10.7 | 16.0 |
| **CSRNet[38]** | 35.8 | 50.0 | 68.2 | 115.0 | 10.6 | 16.0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **SANet[30]** | - | - | 67.0 | 104.5 | 8.4 | 13.6 |
| **Context aware[19]** | 20.5 | 29.9 | 62.3 | 100.0 | 7.8 | 12.2 |
| **Ours approach** | **19.8** | **29.3** | **61.5** | **100.0** | **7.2** | **12.0** |

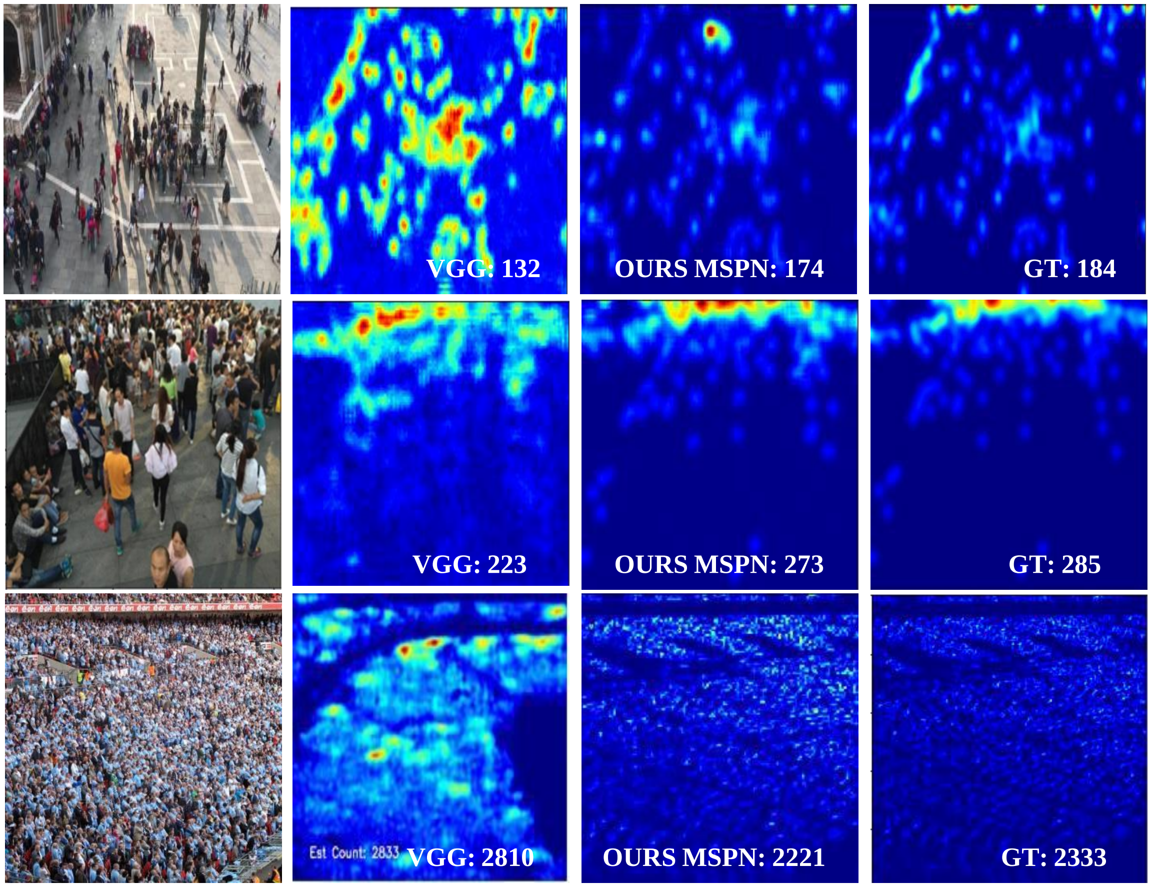


**Fig. 4.** The performance of the proposed approach on three different density level scenes for crowd density estimation

## 5 Conclusion and Future perspectives

In this paper, we propose a learning scale model by applying multi-scale pooling network to solve the problem of density four different scale density maps of an image to generate the final one. Experiments were performed on three different standard benchmark datasets having density variations and the generalization ability of the model was quite impressive than other state-of-art methods. These datasets were formed by fixed cameras, so in future we will work on the images taken from moving cameras e.g. drones. We will enhance our model to process consecutive images and their ground- truth data simultaneously

## Funding:

This research received no external funding.

## Conflicts of Interest:

The authors declare no conflict of interest.

## REFERENCES

[1] Dollar, P., Wojek, C., Schiele, B., Perona, P., 2012. Pedestrian detection: An evaluation of the state of the art. IEEE transactions on pattern analysis and machine intelligence 34, 743–761.

[2] Li, M., Zhang, Z., Huang, K., Tan, T., 2008. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection, in: Pattern Recognition, 2008. ICPR 2008.19th International Conference on, IEEE. pp. 1–4.

[3] Leibe, B., Seemann, E., Schiele, B., 2005. Pedestrian detection in crowded scenes, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, IEEE. pp. 878–885.

[4] Tuzel, O., Porikli, F., Meer, P., 2008. Pedestrian detection via classification on riemannian manifolds. IEEE transactions on pattern analysis and machine intelligence 30, 1713–1727.

[5] Viola, P., Jones, M.J., 2004. Robust real-time face detection. International journal of computer vision 57, 137–154.

[6] Wu, B., Nevatia, R., 2005. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors, in: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, IEEE. pp. 90–97.

[7] Sabzmeydani, P., Mori, G., 2007. Detecting pedestrians by learning shapelet features, in: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE. pp. 1–8.

[8] Viola, P., Jones, M.J., Snow, D., 2005. Detecting pedestrians using patterns of motion and appearance. International Journal of Computer Vision 63, 153–161.

[9] Gall, J., Yao, A., Razavi, N., Van Gool, L., Lempitsky, V., 2011. Hough forests for object detection, tracking, and action recognition. IEEE transactions on pattern analysis and machine intelligence 33, 2188–2202.

[10] Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part- based models. IEEE transactions on pattern analysis and machine intelligence 32, 1627–1645.

[11] Wu, B., Nevatia, R., 2007. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. International Journal of Computer Vision 75, 247–266.

[12] Chan, A.B., Vasconcelos, N., 2009. Bayesian poisson regression for crowd counting, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE. pp. 545–551.

[13] Ryan, D., Denman, S., Fookes, C., Sridharan, S., 2009. Crowd counting using multiple local features, in: Digital Image Computing: Techniques and Applications, 2009. DICTA'09., IEEE. pp. 81–88.

[14] Babu Sam, D., Surya, S. and Venkatesh Babu, R., 2017. Switching convolutional neural network for crowd counting. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5744-5752).

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

[16] Chen, J., Su, W. and Wang, Z., 2020. Crowd counting with crowd attention convolutional neural network. Neurocomputing, 382, pp.210-220.

[17] Zhang, Y., Zhou, D., Chen, S., Gao, S. and Ma, Y., 2016. Single-image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 589-597).

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

[19] Liu, W., Salzmann, M. and Fua, P., 2019. Context-aware crowd counting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5099-5108).

[20] Zhang, C., Li, H., Wang, X. and Yang, X., 2015. Cross-scene crowd counting via deep convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 833-841).

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

[22] Boominathan, L., Kruthiventi, S.S. and Babu, R.V., 2016, October. Crowdnet: A deep convolutional network for dense crowd counting. In Proceedings of the 24th ACM international conference on Multimedia (pp. 640-644).

[23] Idrees, H., Saleemi, I., Seibert, C., Shah, M., 2013. Multi-source multiscale counting in extremely dense crowd images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2547–2554.

[24] Lempitsky, V., Zisserman, A., 2010. Learning to count objects in images, in: Advances in Neural Information Processing Systems, pp. 1324–1332.

[25] Pham, V.Q., Kozakaya, T., Yamaguchi, O., Okada, R., 2015. Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3253– 3261.

[26] Wang, Y., Zou, Y., 2016. Fast visual object counting via example-based density estimation, in: Image Processing (ICIP), 2016 IEEE International Conference on, IEEE. pp. 3653–3657.

[27] Xu, B., Qiu, G., 2016. Crowd density estimation based on rich features and random projection forest, in: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 1–8.

[28] Wang, C., Zhang, H., Yang, L., Liu, S., Cao, X., 2015. Deep people counting in extremely dense crowds, in: Proceedings of the 23rd ACM international conference on Multimedia, ACM. pp. 1299–1302.

[29] Daniel Onoro-Rubio and Roberto J. L´opez- Sastre. Towards Perspective-Free Object Counting with Deep Learning. In European Conference on Computer Vision, pages 615– 629, 2016.

[30] Xinkun Cao, ZhipengWang, Yanyun Zhao, and Fei Su. Scale Aggregation Network for Accurate and Efficient Crowd Counting. In European Conference on Computer Vision, 2018.

[31] Deepak Babu Sam, Neeraj N. Sajjan, R. Venkatesh Babu, and Mukundhan Srinivasan. Divide and Grow: Capturing Huge Diversity in Crowd Images with Incrementally Growing CNN. In Conference on Computer Vision and Pattern Recognition, 2018.

[32] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-Scene Crowd Counting via Deep Convolutional Neural Networks. In Conference on Computer Vision and Pattern Recognition, pages 833–841, 2015.

[33] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. In Conference on Computer Vision and Pattern Recognition, pages 589–597,2016.

[34] Deepak Babu Sam, Shiv Surya, and R. Venkatesh Babu. Switching Convolutional Neural Network for Crowd Counting.In Conference on Computer Vision and Pattern Recognition, page 6, 2017.

[35] Feng Xiong, Xinjian Shi, and Dit-Yan Yeung. Spatiotemporal Modeling for Crowd Counting in Videos. In International Conference on Computer Vision, pages 5161–5169, 2017.

[36] Vishwanath A. Sindagi and Vishal M. Patel. Generating High-Quality Crowd Density Maps Using Contextual Pyramid CNNs. In International Conference on Computer Vision,pages 1879–1888, 2017.

[37] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. Leveraging Unlabeled Data for Crowd Counting by Learning to Rank. In Conference on Computer Vision and Pattern Recognition, 2018.

[38] Yuhong Li, Xiaofan Zhang, and Deming Chen. CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. In Conference on Computer Vision and Pattern Recognition, 2018.

[39] Wei Liu, Dragomir Anguelov, Dumitru Erhan, ChristianSzegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C.Berg. SSD: Single Shot Multibox Detector. In European Conference on Computer Vision, 2016.

[40] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. arXiv Preprint, 2015.

[41] Zenglin Shi, Le Zhang, Yun Liu, and Xiaofeng Cao. Crowd Counting with Deep Negative Correlation Learning. In Conference on Computer Vision and Pattern Recognition, 2018.

[42] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. Crowd Counting via Adversarial Cross-Scale Consistency Pursuit. In Conference on Computer Vision and Pattern Recognition, 2018.

[43] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative Crowd Counting. In European Conference on Computer Vision,2018