2022 IEEE 2nd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA) Sabratha, Libya 23-25 May 2022

Transfer Learning For Crowed Counting

Othman Omran Khalifa
Libyan Center for Engineering
Research and Information Technology
Baniwalid
Libya
khalifa@iium.edu.my

Abdulgani Albagul
Libyan Center for Engineering
Rresearch and Information Technology
Baniwalid
Libya
albagoul@yahoo.com

Aisha Hassan Abdallah Hashim
Electrical and Computer Engineering
Departmet, Faculty of Engineering,
International Islamic University
Malaysia Kuala Lumpur, Malaysia
aisha@iium.edu.my

Noreaha Abdul Malik Hashim
Electrical and Computer Engineering
Departmet, Faculty of Engineering,
International Islamic University Malaysia
Kuala Lumpur, Malaysia

Kwan Nur Sakinahbt Wan Zainuddin
Electrical and Computer Engineering
Departmet, Faculty of Engineering,
International Islamic University Malaysia
Kuala lampur, Malaysia

***Abstract*— Transfer learning for crowd counting via CNN is explored in this research to minimize training time and computational cost. The Mall dataset is used to evaluate the effectiveness of the transfer learning approach and is compared with other recent techniques that use their own deep architecture. By using transfer learning, less computation is involved because the pretrained model has already learned the necessary weightage and architecture thus reducing time consumption for training. The ResNet50 model is fine-tuned to be applying to the Mall dataset. The result of this project shows that using ResNet50 for transfer learning achieved mediocre MAE and MSE compared to other recent techniques such as the DeepCount and DecideNet models. Therefore, application of transfer learning in crowd counting using CNN for ResNet50 on Mall dataset was proven inefficient and further improvements needs to be carried out for this application to be beneficial**

Keywords*— *Crowd counting, convolutional neural network, Deepcount, MAE, MSE, DecideNet, ResNet50

I. INTRODUCTION

A crowd is defined as a large number of persons collected together in one place. Crowds can form for a various number of reasons including but not limited to entertainment purposes, protests, rallies, and religious pilgrimages. In a tight area with a high concentration of people, erratic human behaviour can lead to human stampedes which commonly sacrifices innocent lives. One of the worst recorded incidents of human stampede occurred in 2015 which caused the loss of more than 767 lives near Mecca, Saudi Arabia [1]. In 2020 alone, there have been seven separate incidents of stampede where the incident that recorded as many as 56 lives lost occurred in Iran [2]. Especially now in the middle of a worldwide pandemic, crowd counting has become more vital than ever to disperse crowds by issuing early warning. Crowd counting is an important instrument especially in public safety, urban development planning, and traffic monitoring and video surveillance. Crowd counting also involves distributing crowd density over the area of the gathering to identify crucial regions that are above the safety limit so that early warnings can be issued to prevent potential human stampedes. One of the most effective methods for crowd control and crowd mitigation is by using crowd counting systems. Modern crowd counting system is credited to Herbert Jacobs, a journalism professor from University of California, Berkeley, who derived the basic crowd density rule, Jacob's Method, where an area is broken up into smaller sections, finding the density of people in each small section and calculating the average density thus inferring the average density of the entire area using the average density that has been calculated. Under Jacob's Method categorization, one person for every 0.25 meter squared is already considered 2 mosh-pit densities [3]. Nowadays there are many other methods to estimate crowd size, but the general principle has not changed, area times density. Crowd counting systems are mainly divided into two categories, supervised and unsupervised crowd counting. Under unsupervised crowd counting, counting is carried out using clustering. Whereas under supervised crowd counting, the category can be further divided into four different methods, namely, detection, regression, density estimation and Convolutional Neural Networks (CNN). The detection method, including methods such as Histogram Oriented Gradient (HOG), does well in low density crowds but performance declines when used in high density crowds. Counting using regression involves feature extraction of foreground area and edge features. Linear regression model is then formulated to map the actual number and the predicted number. Counting using density estimation, a prediction is obtained from collected information from an imperceptible probability-density function. Detection and localization of a single object is then made easier by calculating image density. Counting using CNN uses convolutional layers, pooling layers, Rectified Linear Unit (ReLU) layers and Fully Connected Layers (FCL) to extract properties that are later mapped to a density map. Usage of CNN is more efficient and accurate compared to using detection, regression, or density estimation. CNNs are better equipped to learn the deeper and more significant features involved in crowd counting. However, the usage of CNN comes at the cost of high computational complexity [4]. The transfer learning concept can be used to improve the performance of crowd counting systems that use CNN. Hence the aim is to decrease the high computational cost and reduce training time by applying transfer learning to pretrained models for crowd counting systems that use CNN.

II. RELATED WORK

Boominathan et al [8] proposed combining deep and shallow networks to better capture individuals at a variety of different

978-1-6654-7918-9/22/$31.00 ©2022 IEEE

2022 IEEE 2nd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA) | 978-1-6654-7918-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/MI-STA54861.2022.9837673

scales that results in a predicted density map and augmenting training samples of highly dense crowds. This combination enables the system to capture individuals at differing scales more effectively. The emphasis on scales in images taken is also researched on by Idrees et al [9]. Idrees et al employed context using locally- consistent scale and related confidences to improve human detection in crowds. Tripathi et al [10] and Sindagi et al [11] both carried out surveys on crowd analysis methods. Both focused their areas of research on crowd counting using convolutional neural networks later categorizing the methods into four main categories and evaluating crowd counting methods that used tailor made representations as well as newly created datasets respectively. Tripathi et al reviewed and compared the performances between the different CNN networks used and studied the datasets commonly used in crowd counting while Sindagi et al. Liu et al [12] came up with a model that uses Deep Recurrent Spatial Aware Network that creates variations of crowd density and pose variations. Liu et al plans to use this model for crowd map refinement that can be applied in other crowd flow prediction research. Zhang et al [13] focuses on crowd counting for situations where the camera might be obstructed or unseen scenes. Zhang et al uses a CNN that has been trained with two different objectives that are related, crowd counting and crowd density estimation. Using a CNN that has been trained with these two objectives in mind enables the CNN to acquire a better local optimum. Liu et al [14] and Shao et al [15 focused on attributes and contextual information for crowds. Results obtained by Liu et al [14] showed better density estimates and increased crowd counting performance while Shao et al hopes that what the predictors have learned from the dataset they created can be applied elsewhere. Zhang et al [16] and Pu et al [17] concentrated on crowd density estimation maps. Zhang et al [16] used Multi column CNN architecture to receive input image of various sizes to be mapped to respective crowd density map whereas Pu et al used classic deep convolutional neural network (ConvNets) and built a new dataset to evaluate cross-scene crowd density accuracy. Pardamean at al used transfer learning for a smart building management system where transfer learning is applied to a deep learning model that was trained beforehand using the ImageNet dataset. The pretrained deep learning model is then trained on a handcrafted dataset. The results of the smart building system that used transfer learning for the crowd counting function are compared to when five other popular CNN models are use as the pretrained models. Zhang et al [18] applies transfer learning to reduce time taken for model convergence, a direct way to train the ResNet-DC, which is used in the end-to-end structure. Zhang et al [18] combines ResNet-DC with PCM to estimate the amount of people and the location. The backend is calibrated from ResNet-18 and used to extract features to up sample extracted features into maps using PCM. PCM maintains crowd distribution and location data. This method has been proven to obtain good crowd counting performance and accurate location data. Liu et al [26] uses DecideNet which estimates the crowd density through generation and detection of regression based density maps. DecideNet combines an attention module that evaluates the reliability of the both the crowd density and regression estimations. The final crowd counts are acquired using the attention module to choose suitable estimations from the either density maps. Ma et al [28] proposes a new patch-wise regression loss (PRL) to improve the initial pixel-wise loss. The following table is a summary of the methodology and advantage or disadvantage of the journals and articles that are used as reference for this literature review

Table I Summary of Related Work

Authors/Year	Methodology	Advantages	Limitation
L. Boominathan, S. S. S. Kruthiventi, and R. V. Babu (2016)	CNN, data augmentation, combination of shallow and deep convolutional architectures	Effectively capture people at various scales and overcome undersupply of training samples of dense crowds using augmentation	Underestimates count when there are more than 2500 people
H. Idrees, K. Soomro, and M. Shah (2015) Crowd analysis, Markov	Random Field, combination-of-parts detection, Global Occlusion Reasoning	Context used on locally consistent scale and the associated confidence priors, improves human detection in dense crowds.	- Inability to detect in low resolution - High confidence detection in first iteration results in hypersensitivity and scale degradation
G. Tripathi, K. Singh, and D.K. Vishwakarma (2019)	CNN, crowd behaviour, deep learning, anomaly detection	- Explored major public crowd datasets - Categorization of crowd analysis methods	Non-applicable
W. Ouyang and L. Lin (2018)	Deep Recurrent Spatial Aware Network,	Proposed method achieves superior performance compared to other methods	Using recurrent refinement performance drops slightly after 30 iterations
C. Zhang, H. Li, X. Wang and X. Yang (2015)	CNN, crowd density map normalization, new dataset	More adept at describing crowd scenes	Ridge regression produces unsatisfactory result, distribution of density in the first 60 training frames significantly different than other test frames.
W. Liu, M. Salzmann and P. Fua (2019)	CNN, Scale Aware Contextual Features	Improves crowd counting performance and obtains improved density estimates	Less dense crowds provide less context and the method looses it's advantage
V. A. Sindagi and V. M. Patel (2018)	Survey, CNN, Density estimation	- Improved performance acquired using scale-cognizant and context-cognizant models - Reduction in count error motivated by increasingly	Addressed lack of uniform density datasets that cater to large density crowds

		complex 15 CNN models	
J. Shao, K. Kang, C. Chang Loy and X. Wang (2015)	New dataset, multitask deep learning model	- Deep models show improved performance for cross-scene feature identification - Deeply learned features carry out superior execution in multitask learning.	Deep model performs poorly in identifying features that are complicated and have various appearance and motion
Y. Zhang, D. Zhou, S. Chen, S. Gao and Y. Ma (2016)	MCNN	Proposed model is readily transferable to be used for other datasets	Network is affected by data in target domain, inadequate training data causes degraded performance
S. Pu, T.Song, Y.Zhan and D. Xie (2017)	ConvNets, density estimation	New crowd density estimation using deep ConvNets can perform well in practical applications	Misclassified samples occur in neighbouring levels
C. Wang, H. Zhang, L. Yang, S. Liu and X. Cao (2015) [19]	People counting, CNN, Crowd analysis, Deep regression	Improve robustness and decrease amount of false alarms	Less people in images causes unstable absolute difference and normalized absolute differences
B. Pardamean, H. H. Muljo, T. W. Cenggoro, B. J. Chandra, & R. Rahutomo (2019)	CNN, Transfer learning	Achieves lowest MSE using AlexNet	Dataset is too small and more complex CNNs suffer from overfitting
J. Zhang, , S. Chen, , S.Tian, W. Gong, G. Cai.,& Y. Wang. (2021) [20]	ResNet-DC, PCM	- Achieve higher crowd counting performance in highly dense areas - Accurately predict crowd location	Soft average precision (AP) causes degradation in MAE performance
N. Ilyas, B. Lee, and K. Kim (2021) [21]	CNN, HADF-Crowd, DFEM, CAM	Combination of local and global features improves crowd counting accuracy	Usage of CAM for high density areas which is meant for low density area causes high error rates
F. Xiong, X. Shi, and D.-Y. Yeung (2017) [25]	RN N, spatiotemporal modeling	Good generalization properties and is applicable to many other datasets	Spatiotemporal model does not annotate when head is not detected
J. Liu, C. Gao, D. Meng, and A. G. Hauptmann	CNN, quality aware density estimation	Able to achieve state of-the-art level result using three	Direct late fusion is not enough to obtain improved results across all datasets
(2018) [26]		different datasets	
Z. Chen, J. Cheng, Y. Yuan, D. Liao, Y. Li, & J. Lv (2020) [27]	Multilayer gradient fusion, CNN	Different level pixelation of density map improves SNR of training data and reduces estimation errors	When trained on sparse and non-uniform dataset, model doesn't always converge
Y.-J. Ma, H.-H. Shuai, and W.-H. Cheng (2021) [28]	CNN, density map regression, spatiotemporal modeling	New patch-wise regression loss (PRL) used to improve the	Difficult to precisely estimate density map at pixel level original pixel-wise loss.

III. METHODOLOGY

In training a neural network, the accurate weightage for the network is determined through several back and forth iterations. By applying transfer learning to pretrained models, the weightage and architecture used from the pretrained models can be used for the new model, saving time and resources, instead of training the new models from scratch. The methodology involved for this transfer learning project includes data augmentation, feature extraction and model fine tuning. Crowd counting systems using CNN are able to produce better results compared to crowd counting systems that uses traditional methods. and by applying transfer learning to pre-trained models, the crowd counting systems using CNN will be able to produce better results with smaller error margins. For this project, the Mall dataset [22] was chosen for this project that aims to test the effectiveness of transfer learning methods in crowd counting systems that uses CNN through quantitative analysis of results obtained from transfer learning of pretrained model, ResNet50. The proposed method to improve results of crowd counting using CNN is by using transfer learning. Using Liu et al's work of transfer learning using Bayesian Models as a frame of reference for the transfer learning process, the most appropriate pretrained model will be 19 chosen for the simulation. In this study, the Mall dataset will be used to analyse the results of using transfer learning on a pretrained model and compared to other recent techniques which were also evaluated using the Mall dataset. By using transfer learning onto pre-trained models, we hope to achieve lower MAE and MSE for the Mall dataset to verify the effectiveness of the transfer learning technique. The simulations involved will be implemented on ResNet50 using Python. The following figure shows the process flow of the methodology used in this project.

A. Dataset

This project will use the Mall dataset [22], a publicly available dataset. The Mall dataset is made up of 2000 images and all images have a resolution of 320 x 240. The smallest number of heads in the dataset is 11 people while the largest number is 53 people. Images used in the Mall dataset are collected from surveillance cameras located throughout the mall. The Mall dataset however does not have any variance in scene perspective and has a slightly higher density compared to the UCSD dataset. The table below is the properties of the dataset compared with the UCSD dataset.

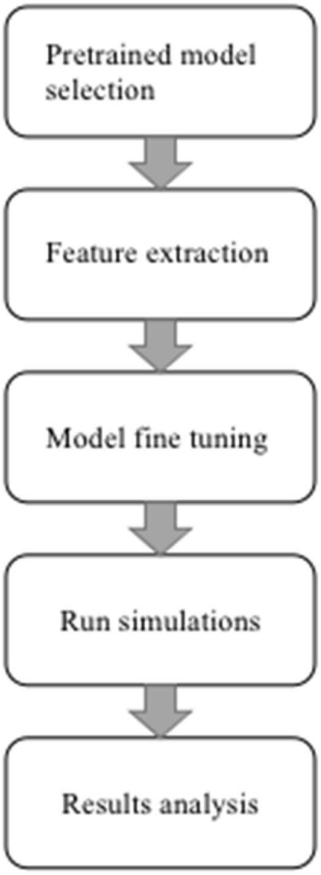

Fig.1 Project Methodology

Table II Comparison table of UCSD and Mall dataset

Dataset	Number of frames	Resolution	Frames per second	Density	Total number of people
UCSD	2000	238 x 158	10	11 - 46	49885
Mall	2000	320 x 240	<2	13 - 53	62325

B. Data Augmentation

To prevent problems caused by overfitting (explain overfitting), the Mall dataset that is used is augmented artificially by transforming the dataset using various minor modifications such as horizontal and vertical flips, rotations, gray scale and different colour saturations. Applying these minor modifications provide a new perspective to the object as well allows the model, ResNet50, to extrapolate unknown data based on these augmentations. Thus, the model is more adept at classifying the augmented images. For data augmentation, the Keras Image Data Generator class is used. The Image Data Generator class augments input images in real time while the model is being trained which ensures that less overhead memory is used. Image Data Generator generates the 21 augmented images in small batches instead of all together at once which saves on memory usage. For this project, the images are rescaled, normalized, randomly rotated, zoomed, and shifted horizontally and vertically. The input images are normalized by dividing the inputs by the dataset standard value. 20% of the augmented data is then assigned as the validation images.

Fig.2 Post-data augmentation view

C. Simulation

The proposed solution begins with feature extraction of ImageNet which is then applied to ResNet50. Model fine tuning is then applied to ResNet50 to be able to be used on the Mall dataset. Model optimization is also applied to ResNet50 to enable better results to be obtained. Finally the proposed solution is tested and evaluated where the results are analysed in the next fourth chapter.

D. Results Analysis

To evaluate the performance of crowd counting models, mean absolute error (MAE) and mean squared error (MSE) are the most commonly used parameters. MAE and MSE are defined as

$$MAE = \frac{1}{N}\sum_{i=1}^{N}[zi - z^i]$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}[zi - z^i]^2$$

N is for the number of test images, zi is for the number of actual people in the image and z^i is the number of people estimated to be in the image [6]. The table below shows the MAE and MSE obtained at epochs 10, 20, 30, 40 and 50.

Table III MAE and MSE epoch

Epoch	MAE	MSE
10	3.6800	21.3869
20	3.2237	16.4517
30	3.2315	16.6704
40	3.2031	16.4972
50	3.2387	16.8727

Based on the table above this was obtained after every epoch, as there was a slight 0.0078 increase from the 20th epoch to the 30th epochs. The MSE also increases as much as 0.2187. From the 30th to the 40th epochs, there is a decrease for both MAE and MSE values however from the 40th to the 50th epochs, there is a 0.0356 increase and a 0.3755 increase for the MAE and MSE values respectively. Factors that can

affect MAE and MSE values are when there are outliers in the data obtained from the regression. The figure below is a mapping of the MAE of each epoch and the validation MAE.

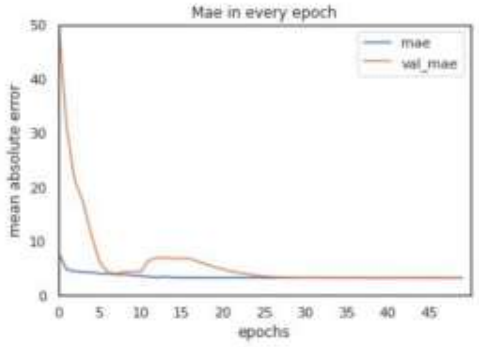

Fig.3 MAE in every epoch

By the fifth epoch, both MAE and validation MAE has smoothed out but at the tenth epoch the validation MAE increases while the MAE value constantly maintains a steady rate until 29 the 50th epoch. The figure below shows the training loss and validation loss of every epoch. BY the fifth epoch both training loss and validation loss has been reduced as much as possible.

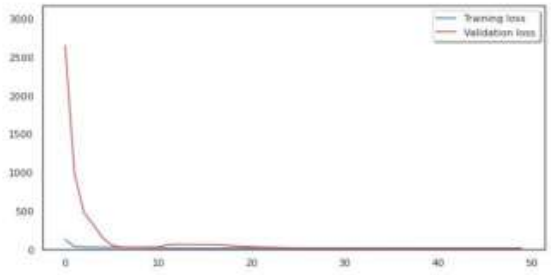

Fig.4 Training loss and validation loss of every epoch

In the figure below, the predicted values of each image is plotted using scatter plot against the actual values obtained to observe the relationship between the two outputs. The graph is linearly proportional between the predicted values and the true values.

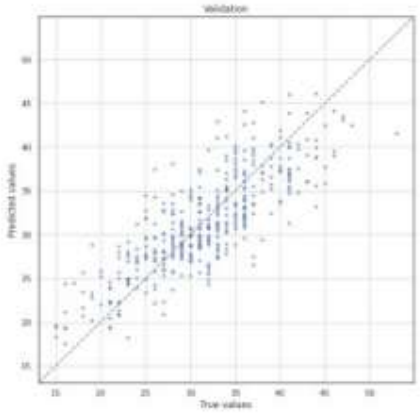

Fig.5 Training loss and validation loss of every epoch

The following sets of figures are the results obtained from the simulation of ResNet50 on the Mall dataset. The predicted number of people in each frame is stated on top of each frame.

Fig.6 Training loss and validation loss of every epoch

As can be observed from the images shown above, image 5 and image 6 both have an estimated crowd count of 41 people even though it can be seen that both images have slight differences. In the third image the crowd density in the left side of the image is a bit denser compared to the left side of image 5. From here it can already be seen that discrepancies occur even in just these four output images. There are many factors that can affect discrepancies such as mislabeling and overfitting.

Table IV Comparison of MAE and MSE obtained

	MAE	**MSE**
Bidirectional ConvLSTM [25]	2.10	7.6
DecideNet[26]	1.52	1.90
DeepCount[27]	1.55	2.00
STDNet[28]	1.47	1.88
ResNet50 with transfer learning	3.31	15.5

The table above are the MAE and MSE obtained from the simulations using ResNet50 after transfer learning on the Mall dataset compared to other recent techniques. Based on the results obtained, it can be observed that using transfer learning for ResNet50 on the Mall dataset did not show any beneficial improvements compared to other CNNs that were studies recently. This can be the result of outliers that occurred during the regression. Furthermore, the Mall dataset did not show much variation between images compared to other datasets. The Mall dataset is also not very diverse and the training data is smaller compared to other datasets, which is why we initially put the data through data augmentation to artificially expand the data. However, data augmentation also has its limitations and can cause overfitting, which is one of the possible explanations as to why the MSE is quite high compared to the MSE of other recent techniques

IV. CONCLUSION

The purpose of this paper is to study, explore and examine the effects of transfer learning onto pretrained CNN models, to be used for crowd counting systems that uses CNN. As can be concluded from Liu et al, through transfer learning of a Bayesian Model for crowd counting using a newly created

dataset that is aimed to specifically test adaptation methods in counting crowds, where Liu et al found that transfer learning is useful for count transfer as well as the model that used transfer learning on average did not face negative transfers. Liu concludes that there is more room for improvement when using transfer learning via Bayesian Models. Pardamean et al used transfer learning for AlexNet in an intelligent human counting system for a handcrafted dataset and found that while the counting system was able to achieve the lowest MSE using AlexNet, the small size of the handcrafted dataset possibly contributed to the higher MSE that the other CNN models acquired. Conclusively, this Final Year Project was unable to reach its objectives which were to propose and simulate a better method for crowd counting using CNN. For this particular project, the main area of focus is the reduction of training time and computation as well as achieve a feasible MAE and MSE for a functioning crowd counting system. The MAE and MSE obtained through simulations were higher than MAE and MSE of other recent techniques, therefore rendering this method less favourable. The transfer learning process did however reduce training time to 2 hours and 26 minutes for the Mall dataset which contains 2000 images while running on a 4GB RAM. The high MAE and MSE makes this method unsuitable for public usage, as there is a possibility of false warnings. 34 This project was able to identify the weaknesses in crowd counting systems that used traditional methods and to evaluate the performance of the proposed solution against other recent techniques. However based on the results of from the testing and evaluation, it can be concluded that the hypothesis of using transfer learning for crowd counting via CNN using ResNet50 would improve the performance and decrease MAE and MSE has been disproven.

REFERENCES

[1] [1] "Hajj stampede: Saudis face growing criticism over deaths," BBC News, Sep. 25, 2015. [Online] Available : https://www.bbc.com/

[2] N. Karimi, A. Vahdat and J. Gambrell (2020, January 8). "At least 56 dead, over 200 hurt in stampede at funeral for Soleimani in Iran," Global News, Jan. 28, 2020. [Online] Available : https://globalnews.ca /

[3] H. Jacobs, "To count a crowd," Columbia Journalism Review, vol. 6, no. 1, p. 37, 1967.

[4] N. Ilyas, A. Shahzad, and K. Kim, "Convolutional-Neural Network-Based Image Crowd Counting: Review, Categorization, Analysis, and Performance Evaluation," Sensors, vol. 20, no. 1, p. 43, Dec. 2019, doi: 10.3390/s20010043.

[5] Rajkumar Buyya, R. N. Calheiros, and Amir Vahid Dastjerdi, Big data : principles and paradigms. Cambridge, Ma: Elsevier/Morgan Kaufmann, 2016.

[6] B. Liu and N. Vasconcelos, "Bayesian Model Adaptation for Crowd Counts," 2015 IEEE International Conference on Computer Vision (ICCV), Dec. 2015, doi: 10.1109/iccv.2015.475.

[7] B. Pardamean, H. H. Muljo, T. W. Cenggoro, B. J. Chandra, and R. Rahutomo, "Using transfer learning for smart building management system," Journal of

[8] Big Data, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0272-6. [8] L. Boominathan, S. S. S. Kruthiventi, and R. V. Babu, "CrowdNet," Proceedings of the 24th ACM international conference on Multimedia, Oct. 2016, doi: 10.1145/2964284.2967300.

[9] H. Idrees, K. Soomro, and M. Shah, "Detecting Humans in Dense Crowds Using Locally-Consistent Scale Prior and Global Occlusion Reasoning," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 10, pp. 1986–1998, Oct. 2015, doi: 10.1109/tpami.2015.2396051.

[10] G. Tripathi, K. Singh, and D. K. Vishwakarma, "Convolutional neural networks for crowd behaviour analysis: a survey," The Visual Computer, vol. 35, no. 5, pp. 753–776, Mar. 2019, doi: 10.1007/s00371-018-1499-5.

[11] V. A. Sindagi and V. M. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," Pattern Recognition Letters, vol. 107, pp. 3– 16, May 2018, doi: 10.1016/j.patrec.2017.07.007.

[12] L. Liu, H. Wang, G. Li, W. Ouyang, & L. Lin, "Crowd counting using deep recurrent spatial-aware network," July 2018, arXiv preprint arXiv:1807.00601.

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

[14] W. Liu, M. Salzmann, & P. Fua,"Context-aware crowd counting," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5099-5108), Apr. 2019, arXiv preprint arXiv:1811.10452v2

[15] J. Shao, K. Kang, C. Change Loy, & X. Wang, "Deeply learned attributes for crowded scene understanding," In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4657-4666), Jun 2015.

[16] Y. Zhang, D. Zhou, S. Chen, S. Gao, & Y. Ma, "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), Jun 2016.

[17] S. Pu, T. Song, Y. Zhang, and D. Xie, "Estimation of crowd density in surveillance scenes based on deep convolutional neural network," Procedia Computer Science, vol. 111, pp. 154–159, Jan 2017, doi: 10.1016/j.procs.2017.06.022.

[18] B. Pardamean, H. H. Muljo, T. W. Cenggoro, B. J. Chandra, and R. Rahutomo, "Using transfer learning for smart building management system," Journal of Big Data, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0272-6.

[19] C. Wang, H. Zhang, L. Yang, S. Liu, & X. Cao, "Deep people counting in extremely dense crowds," In Proceedings of the 23rd ACM international conference on Multimedia (pp. 1299-1302), October 2015, doi: http://dx.doi.org/10.1145/2733373.28063370-12345-67-8/90/01.

[20] J. Zhang, S. Chen, S. Tian, W. Gong, G. Cai, and Y. Wang, "A Crowd Counting Framework Combining with Crowd Location," Journal of Advanced Transportation, vol. 2021, pp. 1–14, Feb. 2021, doi: 10.1155/2021/6664281.

[21] N. Ilyas, B. Lee, and K. Kim, "HADF-Crowd: A Hierarchical Attention-Based Dense Feature Extraction Network for Single-Image Crowd Counting," Sensors, vol. 21, no. 10, p. 3483, May 2021, doi: 10.3390/s21103483.

[22] K. Chen, C. Change Loy, S. Gong, and T. Xiang, "Feature Mining for Localised Crowd Counting," Bmvc, Vol. 1, No. 2, p. 3, September 2012, doi: http://dx.doi.org/10.5244/C.26.21. [23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016, doi: 10.1109/cvpr.2016.90.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks," 016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 630–645, 2016, Jul. 2016, doi: 10.1007/978-3-319-46493-0_38.

[24] F. Xiong, X. Shi, and D.-Y. Yeung, "Spatiotemporal Modeling for Crowd Counting in Videos," 2017 IEEE International Conference on Computer Vision (ICCV), Oct. 2017, doi: 10.1109/iccv.2017.551. (recent technique comparison)

[25] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, "DecideNet: Counting Varying Density Crowds Through Attention Guided Detection and Density Estimation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018, doi: 10.1109/cvpr.2018.00545.

[26] Z. Chen, J. Cheng, Y. Yuan, D. Liao, Y. Li, & J. Lv, "Deep Density-Aware Count Regressor," In ECAI 2020, pp. 2856-2863, Aug. 2020.

[27] Y.-J. Ma, H.-H. Shuai, and W.-H. Cheng, "Spatiotemporal Dilated Convolution with Uncertain Matching for Video-based Crowd Estimation," IEEE Transactions on Multimedia, pp. 1–1, 2021, doi: 10.1109/tmm.2021.3050059.

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