Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 218 (2023) 2448–2458

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

Crowd counting analysis using deep learning: a critical review

Akshita Patwal[a], Manoj Diwakar[a], Vikas Tripathi[a], Prabhishek Singh[b]

[a] Department of CSE, Graphic Era Deemed to be University, Dehradun, Uttarakhand, India

[b] School of Computer Science Engineering and Technology, Bennett University, India

Abstract

The term "crowd counting" refers to the practise of counting the number of people present in a certain area. Urban planning, medical services, emergency preparedness, public security, strategic planning, and defence all seem to be domains where this method may be used. Occlusion, size and perspective distortion, and non-uniform distribution are all problems that crowd counting approaches face. As the population density grows, so does the complexity of the calculations. Great advances in deep convolution neural networks (CNNs) and datasets are largely responsible for the tremendous development in crowd count approaches seen in the past few years. In this paper we assess recent efforts and provide a complete evaluation of modern deep learning-based crowd counting systems. This paper discusses some classic and deep learning-based crowd counting approaches. We examine detection-based, regression-based, and classic density estimation approaches briefly. For the purpose of estimating the crowd density and count for the provided crowd scene image, we have evaluated the recent 10 publications on crowd counting using deep learning. We also go through the most widely used datasets. In conclusion, these investigations demonstrated a high degree of precision and offered good illustrations of the potential of AI in crowd counting. From paper to paper, there were significant differences in the approaches and algorithms used to address the crowd counting and density mapping challenge. We also examine the possible uses of crowd counting as well as the difficulties associated with it. As a result, fresh study is being conducted in this area.

Keywords: Detection; Density Estimation; CNN; Regression; Crowd counting; Deep learning

1. Introduction

The world's population is increasing at an exponential rate. The current annual population increase is predicted to be 72 million people, and it continues to rise. Crowd stampede incidents have increased significantly in recent years. These potentially fatal circumstance occurs when there is a large crowd, there is inadequate crowd control, there is a rush for help, or it appears to occur for no apparent reason. The movement of people, whether pedestrians or crowds, seems to be improperly controlled in certain circumstances.[1] Sporting events, political rallies, and music concerts

10.1016/j.procs.2023.01.220

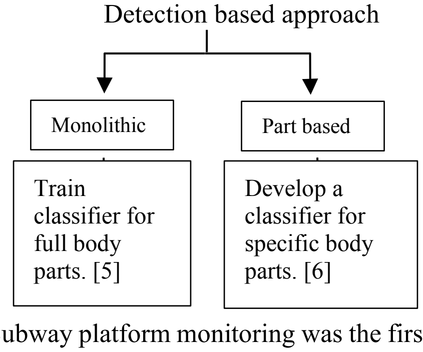
are all occasions where large crowds congregate. For the safety and security of big groups of people, greater management is required which includes installation of CCTVs for several purposes, including traffic management, monitoring public spaces, for anomaly detection, crowd counting [2]. Crowd analysis can be used to improve management. In the Houston Texas crowd crush led to the death of 8 people. Public panic, mob crushes, and a breakdown of control can all contribute to crowd turbulence tragedy. This type of accident can be avoided by detecting an uncontrolled crowd flow early on. Humans can also identify and respond to strange behavior in a monitoring area. However, there are significant constraints to simultaneously observing multiple signals in a highdensity crowd. In order to overcome this constraint, tools in the field of crowd analysis were established. The academic community has put a lot of effort into establishing several frameworks for automatic crowd counting in video surveillance, which is a burgeoning field in AI today but to properly train a deep network for people counting in extremely crowded photos, a large dataset is essential [2]. The various approaches that have historically been used to solve the crowd counting problem are: a crowd-oriented method, regression-based method and a density maporiented approach. An object detector that uses the sliding window technique is used in the crowd-oriented approach to count the number of persons in the picture [5,6]. And a regression function is used in regression-based methods.[7] When there are a large number of people present, these methods have trouble providing reliable findings. A density map-oriented technique is utilized to address this flaw. A density map contains the geographical data that may be efficiently used to show the overall number of people in the picture.[3]. Deep learning, in particular have been widely employed to solve crowd counting and has achieved great progress as a result of its abilities to accurately simulate the changes and the variance in crowd densities between locations. This review examines papers that have recently been published and is structured as follows: Reviewing the conventional crowd counting and density estimate techniques is the focus of Section 2. We discuss the most recent deep learning techniques in Section 3. In Section 4, we go over the most significant publicly available datasets. The Section 5 brings our survey to a close describing the results and discussion. Finally, Section 6 concludes the review.

2. Literature Survey

Crowd counting may be done in a number of ways, both classic and modern. The input data for supervised learning is known and labelled, whereas unsupervised classification uses unknown data and labels. Traditional and CNN-based algorithms are used in supervised classification. Traditional methods are divided into three categories.
They are: -

2.1. Traditional approaches

2.1.1. Methods that rely on detection: Here in detection-based approach frames are used for detection. Works well for the data having less count of people. The approach can be categorized as:


Detection based approach
Monolithic
Part based
Train classifier for full body parts. [5]
Develop a classifier for specific body parts. [6]

2.1.2. Regression based approaches: Subway platform monitoring was the first use of regression-based crowd counts. These approaches generally operate by eliminating the backdrop, measuring multiple foreground pixels, such as the total area or texture, and use of regression function. The regression function used are linear, piece-wise linear, or neural networks. Instead of using intermediate vision operations like object identification or feature tracking, Chan et al. [7] proposed that they use Bayesian regression to estimate the size of heterogeneous crowds made up of people walking in opposing directions Fig.1. depicts the framework, here on two big datasets, Peds1 and Peds2, regression-based counting was verified.

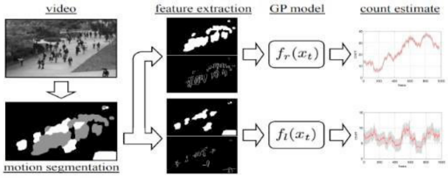


Fig. 1. Bayesian regression framework

2.1.3. Density estimation approaches: While previous methods were effective in addressing occlusion and clutter issues, most of them ignored critical spatial data in favour of a global count. The majority of crowd counting approaches entail density estimate. Fig 2. depicts the density map for an image with GT and prediction. Along with the count, density estimate provides a notion of the spatial distribution of individuals. In [8] a new loss function is introduced by Lempitsky et al. Here they developed a framework to count the bacterial cells and the pedestrians. The datasets used were Pedestrian dataset and cell dataset. Wang et al. [9] proposed that for the VOC problem, a rapid example-based solution is used. It counts with near-perfect precision while running quickly. The approach is based on the assumption that identifiable item distribution patterns are limited, therefore all counterpart neighbourhood embedding may be computed in the training phase rather of the testing phase. Datasets used were UCSD and cell dataset. The conventional crowd counting methods are summarized in Table 1.

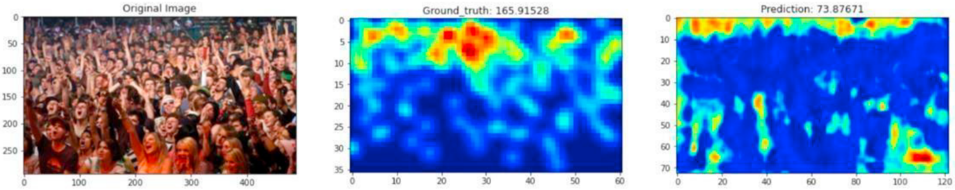


Fig. 2. Density map estimate for a particular picture is illustrated with GT and prediction.

Table 1: Comparison of Traditional Counting Methods

Traditional approach	Mechanism	Pros & Cons
Detection based	Sliding window is used for detecting body parts.	High accuracy for sparse crowd, but not applicable for the dense crowd.
Regression based	Two major components used are feature extractor and using a regression function.	Better accuracy, but neglected important spatial information in favour of a global count.
Density estimation	Use density map	Spatial information is used.

2.2. Deep learning methods: Deep learning has earned a huge interest from researchers around the globe. In image processing, CNNs have demonstrated significant learning capabilities, sparking a slew of CNN-based crowd counting projects. The first time CNN was used to count crowds was by Min et al. [4]. However, it simply calculated the density of people present and did not consider the count of people. In the recent year's CNN-based models for crowd counting has made considerable progress. These approaches outperform other crowd counting methods in circumstances including a plethora of different head scale, non-uniform density distributions, and significant alterations in perspective and scene, making them the most popular in current crowd counting research. Researchers have tried a variety of approaches to address the current crowd counting challenges. Section 3 examines the recent top 10 deep learning-based methods in depth.

3. Deep learning-based crowd counting methods

3.1. Jeong et al. [10]: They took on the task of accurately anticipating the crowd density in congested circumstances so that the individuals could be counted. Based on the scene geometry, offer an extended Bayesian loss method. Ablation tests were used to confirm the impact of suggested components. According to the findings of the ablation investigation, person-scale estimate improved crowd density localization accuracy while degrading counting performance. The suggested model was verified using datasets from UCF CC 50, UCF-QNRF and ShanghaiTech Part A and B.

3.2. Zhiheng et al. [11]: The authors used a loss function known as a Bayesian loss that contributed in making the density contribution probability model. Here only the locations of the head point annotations are used and each pixel is assigned a possibility of belonging to a person or a background. Adam optimizer is used with networks VGG-19 and Alex Net. The research makes use of the datasets.: UCF-QNRF, UCF-CC-50 , ShanghaiTech part A and part B.

3.3. Hafeezallah et al. [12]: A hybrid model, U-ASD is proposed which is based on U-Net and ASD. As shown in Fig.3. the backbone is made up of VGG16 and the model is divided into three parts, B1, B2, B3. The B1 part is for sparce crowd having the deconvolutional layers followed by five convolutional layers and max-pooling. The B2 layer is used for dense crowd having only the convolutional layers. The specifics of this method are presented in the B3 section. It starts with a global average pooling (GAP) layer and two fully connected layers (FCs), then ReLU and Sigmoid Normalization. The GAP may be used to compute each channel's global average value. In addition, Haramain, a novel dataset with hand annotations that includes three distinct scenes and densities, is introduced and utilized to evaluate the U-ASD Net.

3.4. Huang et al. [13]: The difficulty of large-scale variance in pictures was discussed, as well as the disadvantages of using multi-column frameworks to overcome the problem. Thus, to solve the problem SRNet was proposed which is having VGG16 as feature extractor and optimizer used is Adam. Fig.4. depicts the model where the encoder having VGG16 is used for feature extraction, which is then passed to the decoder part having the scaleaware feature learning module (SAM) is used to simulates the crowd's multi-scale characteristics at each level, with varying receptive field sizes and the pixel-aware up sampling module (PAM) enhances pixel wise semantic information.

3.5. Elharrouss et al. [14]: Primary contributions of this research are a novel dataset for crowd counting i.e., “FSCSet” dataset and a CNN-based approach named FSCNet for counting individuals and creating crowd density maps. It is using VGG-16 and the layers are passed into CP1 and CP2(Fig.5.). SHT Part A, SHT Part B, UCF QNRF, UCFCC-50, and UCSD datasets, as well as the suggested dataset FSC-Set, are used for training and testing.

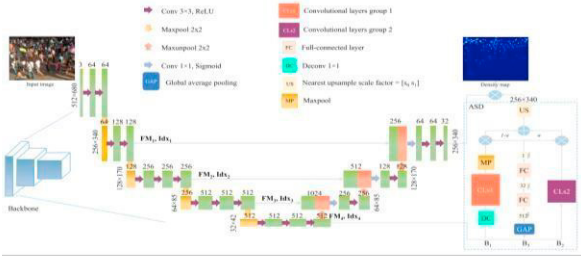


Fig.3.U-ASD Net Model [12]

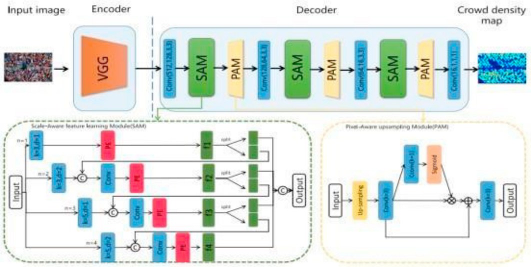


Fig.4.SRNet Model [13]

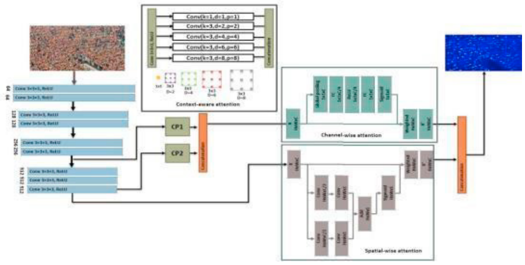


Fig.5.FSCNet architecture [14]

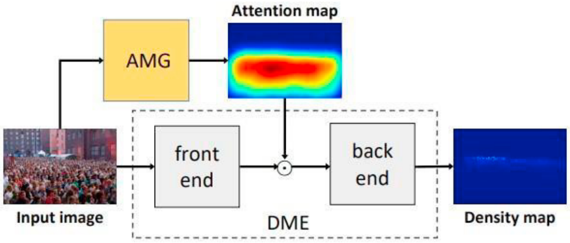


Fig.6. AdCrowdNet architecture [15]

3.6. Liu et al. [15]: For the crowd comprehension of congested noisy environments, they offer ADCrowdNet, a convolutional neural network-based architecture. ADCrowdNet used datasets from UCF CC 50, ShanghaiTech, and UCSD to evaluate crowd counting as well as the TRANCOS vehicle counting dataset. ADCrowdNet is made up of two linked networks. The first is the Attention Map Generator, an attention-aware network (AMG). It identifies congested areas and calculates the level of congestion. Second is the Density Map Estimator (DME) that creates a high-resolution density map. Fig.6. depicts the suggested method's architecture. ADCrowdNet is far more noise resistant and correctly captures crowd features. An image is divided into two groups by AMG: background and crowd. Create an attention map with higher values for busy locations as well. It depicts the amount of traffic congestion.

3.7. Wang et al. [16]: In the wild, it's a challenging task since the changing environment and the big number of individuals make present approaches ineffective. Thus created a large-scale, diversified synthetic dataset based on it. The dataset is named "GTA5 Crowd Counting" ("GCC"). Fig.7 depicts the spatial FCN architecture having VGG-16 as the backbone and on top of it a spatial encoder is added. Finally, the regression layer is added and the density map is formed. The optimizer used is Adam.

3.8. Yujun at el. [17]: Here in Fig.8. VGG16 is used to extract features. The DCB is the decoder block followed by the up-sampling operation and the RM module. The experiments are done on UCF-QNRF and on Shanghai tech.

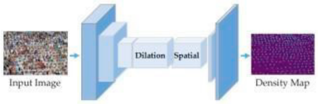


Fig.7.SFCN architecture [16]

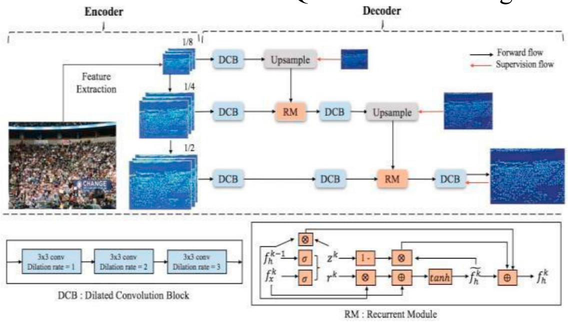


Fig.8.RSANet architecture [17]

3.9. Liu et al. [18]: This study proposes a Cross-stage Refinement Network (CRNet), which uses hierarchical multilevel density priors to improve projected density maps gradually. CRNet is made up of a lot of totally convolutional networks. Fully connected convolutional layers are used recursively to generate density map. The Fig.9 shows that with progressive refining, a high-quality density map may be created from a succession of recurrent fully convolutional networks.

3.10. Lingbo et al. [19]: The three subnetworks indicated in Fig.10 are used in the proposed model. It offers four SFEMs for refining multi scale characteristics from several subnetworks. The DMS-SSIM loss is also being used to improve the entire network.

3.11. Sindagi et al. [20]: The authors suggested a multi-level MBTTBF model that combines information from bottom to deeper layers and opposite (Fig.11.). They introduced a multi-level approach for dealing with scale variation, which can make crowd counting difficult in congested areas. The suggested technique takes a collection of scale complimentary characteristics from neighbouring layers before propagating them in a multi-level hierarchy. As a result, the merging of characteristics from several layers of the backbone network is more successful. Table 3 describes the pros and cons of the discussed deep learning framework.

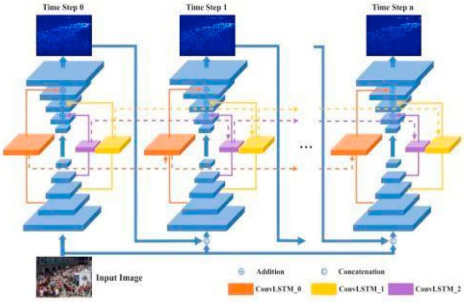


Fig.9.CRNet architecture [18]

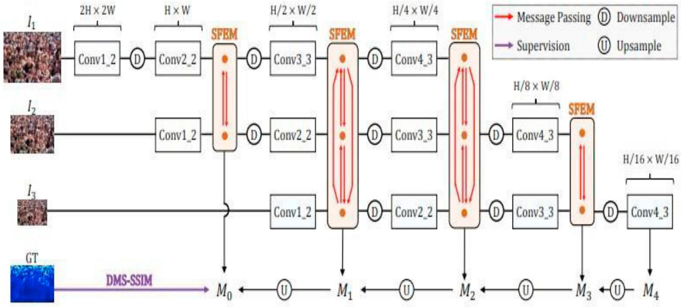


Fig.10.DSSINet framework [19]

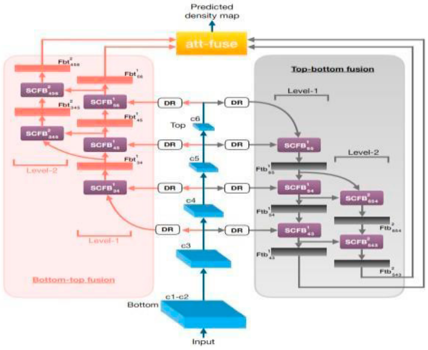


Fig.11.MBTTBF Model [20]

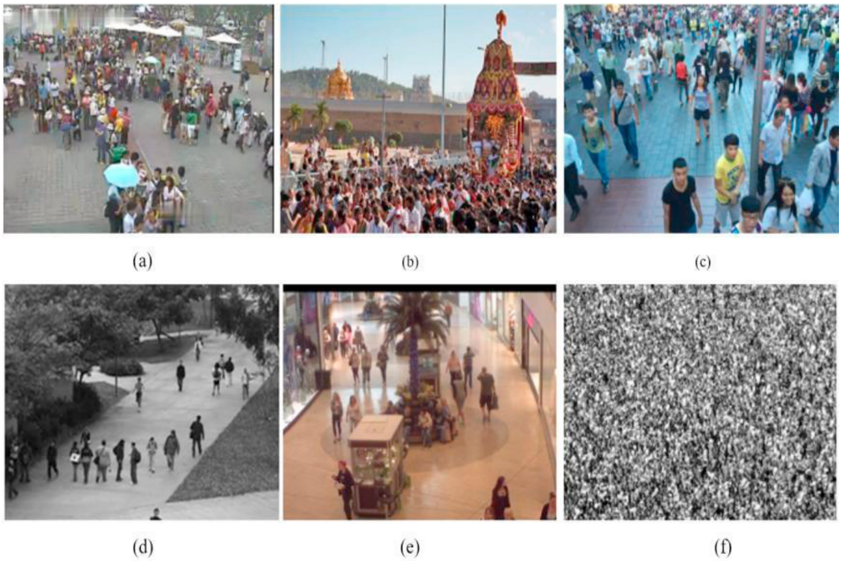


Fig.12.Various dataset's' sample images. (a)WorlExpo'10 (b)ShanghaiTech Part A (c)Shanghai Tech Part B (d)UCSD (e)MALL (f)UCF_CC_50

4. Commonly used datasets

4.1. *ShanghaiTech Dataset [21]:* It is the most popularly used dataset having 1198 images and 330,165 annotations. The dataset is separated into two portions based on various density distributions. Part a has a substantially higher density than part b. The high-density part comprises photos from shit on a busy street in the Shanghai metropolitan region, while the less dense part contains images randomly selected from the internet.

4.2. *UCSD [22]:* It is regarded as the first dataset to be developed for crowd counting. It was gathered via pedestrian cameras and it has 2000 frames. It has 800 training and 1200 testing data.

4.3. *UCF-CC-50[23]:* It’s the first truly difficult dataset made from freely available web photos. For various environments such as marathons, campaigns it has a range of densities and perspective distortions. As there are only 50 photos in this dataset, it is subjected to a 5-fold cross-validation procedure.

4.4. *UCF-QNRF [24]:* It includes 1535 different photos with around 1.25 million annotations. This dataset covers most of the views considering density and light variations. The data is collected from web search and Hajj images.

4.5. *MALL [25]:* It's a dataset gathered from a retail mall's CCTV camera. It has 2000 frames and a total of 62,325 pedestrians. 2012.Genuine units like glass item reflections, booths, interior plants, and so on, as well as human images, are included in the data gathering.

4.6. WorldExpo’10[26]: It has 3920 frames with a total size of 576720 pixels, with 199,923 people annotated. Fig.12. depicts the examples from the popular datasets and table 2 describes them.

Table 2: Datasets Specifications

Dataset	Year	Attribute	No of Images	Resolution	Min	Max	Average	Train/Test
ShanghaiTech Part A [21]	2016	Real World Congested	482	Varied	33	3139	501	300/182
ShanghaiTech Part B [21]	2016	Real World	716	768x1024	9	123	578	400/316
UCSD [22]	2008	Real World	2000	320x240	13	-	46	800/1200
UCF-CC-50[23]	2013	Real World	50	Varied	94	1279	4543	-
UCF-QNRF [24]	2018	Real World	1535	2013x2902	49	815	12865	1201/334
MALL [25]	2012	Real World	2000	320x240	13	53	31	800/1200
WorldExpo’10[26]	2015	Real World	3980	576x720	1	50.2	253	-

Table 3: Pros and cons of the deep learning methods.

Reference and Year	Methods	Pros	Cons
[10] 2022	CBL	Considers the person-scale and crowd-sparsity.	Pre-trained CNN model required.
[11] 2019	BL	The Bayesian loss function is introduced.	Problem caused by the various levels of occlusion that develop as a result of the diverse densities of people in a certain area.
[12] 2021	U-ASD Net	Resolved the perspective distortion and scale variation problem. Also introduced the manually annotated dataset called Haramain.	The strategy does not safeguard privacy, and the installed gadgets are easily destructible.
[13] 2021	SRNet	The methods involve extraction and fusion of features of different scales i.e., the fusion of scale aware feature learning module and the pixel aware up sampling module. Expanded the spatial resolution and modelled multiscale characteristics at different levels to increase the total counting accuracy.	Only suitable for discrete input and the sublevel resolutions they were trained on.
[14] 2022	FSCNet	Proposed a novel dataset FSCSet for football supporters crowd having manually annotated 6000 images	Problem due to scale variation and in unconstrained environment.
[15] 2019	ADCrowdNet	The model accurately detects crowd characteristics and is much more noise-resistant.	They ignore the global distributions of the density map, concentrating instead on localised areas of the crowd scene.
[16] 2019	SFCN	Created a large-scale synthetic crowd counting dataset (GCC dataset) with an autonomous data collector/labeller. The primary benefits of GCC are its ability to offer precise labels (point and mask) and a variety of situations.	There are several domain shifts/gaps between artificial and actual data, which limits its applicability.
[17] 2020	RSANet	The model generates a high-resolution density map with scale aware feature fusion approach.	Ignore the issue that making the network model architecture more complicated leads to an increase in the number of parameters, as well as feature correlation between many columns.
[18] 2020	CRNet	Explore a number of efficient crowd-specific data augmentation techniques	They do not prioritise addressing the domain shift issue.
[19] 2019	DSSINet	Solved the crowd counting scale variation problem and extracts features from multi-scale images.	Pre-trained VGG-16 model required and not give good results on large datasets.
[20] 2019	MBTTBF-SCFB	Resolved the issue of the dataset's and individual images' wide scale variations by fusing multi-level features together.	Do not scale well to large datasets

5. Results and discussion

On the most well-known datasets, we provide the outcomes of the deep learning-based techniques. Based on the following metrics, we contrast several approaches.

5.1. Assessment Metrics: MAE and MSE are two of the most commonly used crowd counting measures. MAE gives the accuracy of the count and MSE states the robustness of the estimation.

MAE: Mean absolute error; MSE: Mean square error. They are defined as follows:
Here N= No. of images in testing dataset.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|ci - \hat{c}\text{i}|, \ \ \text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(ci - \hat{c}\text{i})^2}$$

ci=Ground Truth (GT); ĉi=Estimated Count (EC)

5.2. Discussion

On the most popular datasets, we provide the findings of deep learning-based approaches. On the basis of the assessment measures, we compare several methodologies. Table 4 summarises the comparing results. Deep learning-based approaches performed well in a congested scenario with dense crowds and under a variety of scene circumstances (lighting, scaling, etc.). On almost all the datasets, CBL [10] and BL [11], as well as DSSI-Net [19] and MBTTBF [20], produced equivalent results. Fig.13 depicts the graphical representation, here we chose the best five models across three widely used datasets in terms of MAE and RMSE, and we discovered that using a loss function like in CBL [10], BL [10] it takes scale variation and crowd sparsity into account. Additionally, the single feature scale, such as FSCNet [14], outperforms the multiscale feature technique used in U-ASD [12]. Also, the huge size datasets must be collected (particularly for highly dense crowds), as deep learning requires large datasets for training. Even though there are other datasets available right now, only one (The UCF CC 50 [23]) is designed for crowds with a high density. The dataset isn't big enough to train deeper networks, though. The number of images per density level is not consistent, with many images being accessible for low density levels and relatively few samples for high density levels, despite Shanghai Tech's [12] attempts to record dense crowds. In addition to scale variation, occlusion, and background noise, crowd counting methods also struggle with the fact that some of the frameworks need a lot of parameters and that their results are inaccurate in unconstrained environments.

Table 4: Comparison of the deep learning-based approaches on various datasets.

Method	**Year**	**UCF_QNRF**		**SHT Part A**		**SHT Part B**		**UCF-CC-50**	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
CBL [10]	2022	87.0	155.8	61.8	101.7	7.7	13.1	191.7	283.0
BL [11]	2019	88.7	154.8	62.8	101.8	7.7	12.7	229.3	308.2
U-ASD [12]	2021	-	-	64.6	106.1	7.5	12.4	232.3	217.8
SRNet [13]	2021	108.2	177.5	66.0	96.7	-	-	184.1	232.7
FSCNet [14]	2022	90.8	131.5	55.6	92.75	8.25	17.79	194.3	226.6
ADCrowdNet [15]	2019	-	-	63.2	98.9	8.7	13.5	-	-
SFCN [16]	2019	102.0	171.4	64.8	107.5	7.6	13.0	214.2	318.2
RSANet [17]	2020	102.9	181.9	63.5	97.4	8.5	12.6	-	-
CRNet [18]	2020	101	162	56.4	90.4	7.4	11.9	203.3	263.4
DSSI-Net [19]	2019	99.1	159.2	60.6	96.0	6.8	10.3	216.9	302.4
MBTTBF [20]	2019	97.5	165.2	60.2	94.1	8.0	15.5	233.1	300.9

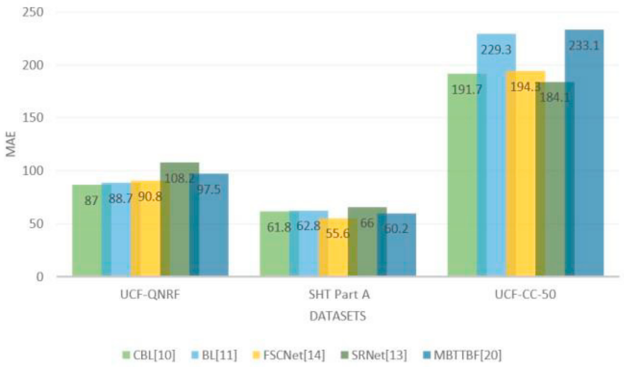


Fig 13. Graphical representation of the MAE of deep learning model on the most popular datasets. (Lower MAE corresponds to higher accuracy)

6. Conclusion

Research on crowd counting has considerably risen in recent years as a result of the need for crowd counting in a variety of fields. With the advancement of deep learning, crowd counting models' performance has substantially improved, opening up more opportunities for real-world applications. According to the assessment we did, the accuracy for crowd counting is improved using the deep learning methods. In this study, we looked at numerous crowd counting approaches, mostly based on CNN architecture. We have examined deep learning-based methods while taking into account the various feature methods, the network design as well as their benefits and drawbacks. We provide an overview of the most often used datasets for testing crowd counting methods. The approaches proposed by various authors were tested using datasets from ShanghaiTech, UCF CC 50, UCF_QNRF and others, and it was discovered that the methods under consideration performed differently depending on the dataset and the context in which they were utilised. In addition, we tested the most representative crowd counting algorithms' performance. Here we find that the model's performance is largely determined by how the dataset was pre-processed and how the processed data was fed into it. According to the analysis, each approach has benefits and limitations, and depending on the scenario, one may surpass the others. The creators of these approaches have done extensive study and have fine-tuned their model to provide consistent and efficient results.

It is clear from the review that each technique has pros and cons, and depending on the circumstances, one may perform better than the other models. Most datasets for the crowd counting challenge include a variety of items, including people, vehicles, and animals, which is not usually the case in real-world situations like airports or shows. In order to cover more realistic circumstances, future study must concentrate on brand-new datasets that have not previously been examined in crowd counting efforts. It would also be necessary to test this dataset on current models in order to compare and analyse the outcomes considering the real-world scenario.

References

[1] Kefan, X., Song, Y., Liu, S., & Liu, J. (2018). Analysis of crowd stampede risk mechanism: A systems thinking perspective. *Kybernetes*.

[2] Tomar, A., Kumar, S., & Pant, B. (2022, March). Crowd Analysis in Video Surveillance: A Review. In *2022 International Conference on Decision Aid Sciences and Applications (DASA)* (pp. 162-168). IEEE.

[3] Shi, X., Li, X., Wu, C., Kong, S., Yang, J., & He, L. (2020, May). A real-time deep network for crowd counting. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2328-2332). IEEE.

[4] Fu, M., Xu, P., Li, X., Liu, Q., Ye, M., & Zhu, C. (2015). Fast crowd density estimation with convolutional neural networks. *Engineering Applications of Artificial Intelligence*, *43*, 81-88.

[5] Tuzel, O., Porikli, F., & Meer, P. (2008). Pedestrian detection via classification on riemannian manifolds. *IEEE transactions on pattern analysis and machine intelligence*, *30*(10), 1713-1727.

[6] 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*(9), 1627-1645.

[7] Chan, A. B., & Vasconcelos, N. (2011). Counting people with low-level features and Bayesian regression. *IEEE Transactions on image processing*, *21*(4), 2160-2177.

[8] Lempitsky, V., & Zisserman, A. (2010). Learning to count objects in images. *Advances in neural information processing systems*, *23*.

[9] Wang, Y., & Zou, Y. (2016, September). Fast visual object counting via example-based density estimation. In *2016 IEEE international conference on image processing (ICIP)* (pp. 3653-3657). IEEE.

[10] Jeong, J., Choi, J., Jo, D. U., & Choi, J. Y. (2022). Congestion-Aware Bayesian Loss for Crowd Counting. *IEEE Access*, *10*, 8462-8473.

[11] Ma, Z., Wei, X., Hong, X., & Gong, Y. (2019). Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6142-6151).

[12] Hafeezallah, A., Al-Dhamari, A., & Abu-Bakar, S. A. R. (2021). U-ASD Net: Supervised Crowd Counting Based on Semantic Segmentation and Adaptive Scenario Discovery. *IEEE Access*, *9*, 127444-127459.

[13] Huang, L., Zhu, L., Shen, S., Zhang, Q., & Zhang, J. (2021). SRNet: Scale-Aware Representation Learning Network for Dense Crowd Counting. *IEEE Access*, *9*, 136032-136044.

[14] Elharrouss, O., Almaadeed, N., Abualsaud, K., Al-Maadeed, S., Al-Ali, A., & Mohamed, A. (2022). FSC-Set: Counting, Localization of Football Supporters Crowd in the Stadiums. *IEEE Access*, *10*, 10445-10459.

[15] Liu, N., Long, Y., Zou, C., Niu, Q., Pan, L., & Wu, H. (2019). Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3225-3234).

[16] Wang, Q., Gao, J., Lin, W., & Yuan, Y. (2019). Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8198-8207).

[17] Xie, Y., Lu, Y., & Wang, S. (2020, October). Rsanet: Deep recurrent scale-aware network for crowd counting. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 1531-1535). IEEE.

[18] Liu, Y., Wen, Q., Chen, H., Liu, W., Qin, J., Han, G., & He, S. (2020). Crowd counting via cross-stage refinement networks. *IEEE Transactions on Image Processing*, *29*, 6800-6812.

[19] Liu, L., Qiu, Z., Li, G., Liu, S., Ouyang, W., & Lin, L. (2019). Crowd counting with deep structured scale integration network. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 17741783).

[20] Sindagi, V. A., & Patel, V. M. (2019). Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1002-1012).

[21] Zhang, Y., Zhou, D., Chen, S., Gao, S., & 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).

[22] Chan, A. B., Liang, Z. S. J., & Vasconcelos, N. (2008, June). Privacy preserving crowd monitoring: Counting people without people models or tracking. In *2008 IEEE conference on computer vision and pattern recognition* (pp. 1-7). IEEE.

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

[24] Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., & Shah, M. (2018). 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).

[25] Chen, K., Loy, C. C., Gong, S., & Xiang, T. (2012, September). Feature mining for localised crowd counting. In *Bmvc* (Vol. 1, No. 2, p. 3).

[26] Zhang, C., Li, H., Wang, X., & 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).