IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CNN BASED METHODS FOR CROWD COUNTING – A COMPREHENSIVE STUDY

[1]Ruchika Lalit, [2]Priyanka Bhutani, [3]Neha Verma, [4]Yatharth Sharma
[1]Assistant Professor, [2]Assistant Professor, [3]Associate Professor, [4]Student

[1]The NorthCap University, Gurugram, India,
[2] University School of Information, Communication &
Technology (USIC&T), Guru Gobind Singh Indraprastha University (GGSIPU), New Delhi, India,
[3] VIPS, Guru Gobind Singh Indraprastha University(GGSIPU), New Delhi, India,
[4] The NorthCap University, Gurugram, India

Abstract: Crowd counting is a used to count the number of individuals are visible in an image or video. It has recently gained a lot of interest because to its usefulness in several domains, such as security, transportation, event management, and urban planning. The variability of crowd density, size, and appearance, as well as occlusion and perspective distortion, make crowd counting a difficult task. Recently, computer vision and deep learning-based techniques have led to the development of various methods for counting crowd, such as detection-based methods, density-based methods, and regression-based methods. These methods employ a variety of techniques to estimate the crowd size, including direct prediction, density estimation, individual detection and counting. To assess the effectiveness of crowd-counting approaches, several metrics are used, such as the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and F1 score. Despite the tremendous advances made in the field recently, crowd counting remains a challenging field, and advanced research is necessary to provide more precise and effective methods.

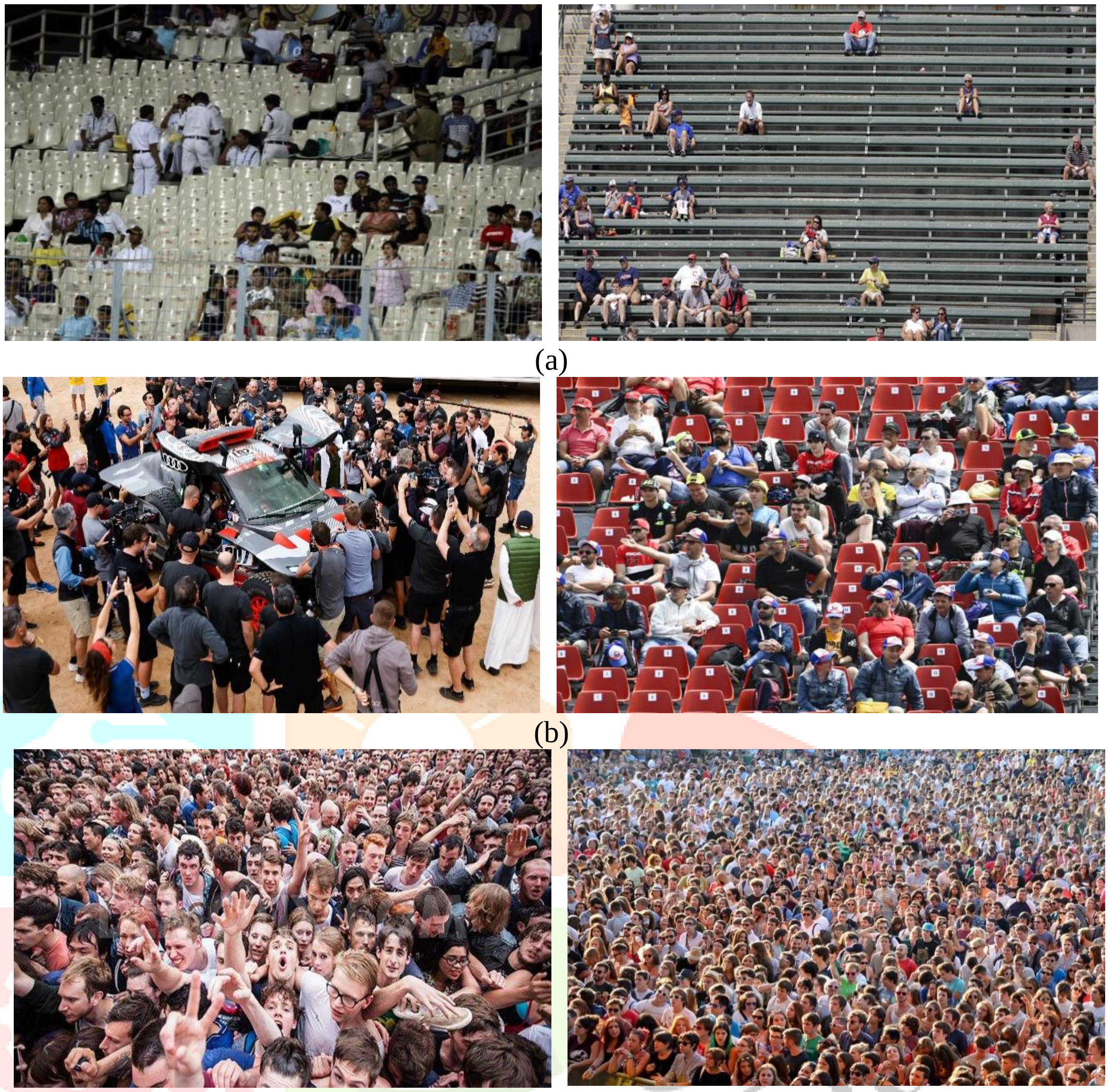
Index Terms **- Crowd Counting, Computer Vision, CNN, Crowd density estimation, Deep learning.**

I. Introduction

Crowd is the gathering of humans sharing same geographical area. The crowd can be dense, medium, or sparse as shown in the figure 1. Crowd counting can be manual as well as automatic. Manual counting is possible when the count of humans is low, but for medium to high density crowded places on hand it is crucial to count the people to avoid any mishappening and on the other hand it is difficult to do it manually, so various basic to machine learning and deep learning-based techniques have been introduced. In computer vision it has wide range of applications, is a challenging task. Since the necessity for monitoring, safety, and security in public spaces has increased, crowd counting has taken on more significance in recent years. Event management, urban planning, transportation, and security are just a few of the industries where crowd counting is used.

In recent years, several approaches have been proposed for crowd counting [1]–[3], ranging from traditional methods to deep learning-based technique. Conventional approaches use hand-crafted features and conventional machine learning algorithms, while deep learning-based approaches [2] leverage the power of convolutional neural networks (CNNs) to learn features automatically from the input data.

In this survey paper, we provide a comprehensive overview of crowd counting methods, including recent advances in deep learning-based technique[1], [2], [4]–[6]. We first discuss traditional methods for crowd counting, followed by an overview of deep learning-based methods. We then review various datasets and evaluation metrics commonly used for crowd counting. Finally, we discuss some of the current challenges in crowd counting and potential directions for future research.



(a)

(b)

(c)

Figure 1: (a) Sparse Density Crowd, (b) Medium Density Crowd, (c) High Density Crowd

II. RELATED WORKS

The field of crowd counting has been a dynamic and rapidly evolving research area within computer vision. Over the years, Numerous creative solutions have been suggested in academic literature to address this difficult issue. Density-based, regression-based, and detection-based methods have been used traditionally to count crowds. Density-based methods extract a density map from the input image to infer the population density. A regression-based algorithm predicts the crowd count directly from the input image, whereas a detection based algorithm detects individuals in an image and aggregates their counts to estimate the total number of people present in the crowd. Many CNN-based approaches for crowd counting have been developed because of deep learning advances, including ResNet-based methods, dilated CNN-based methods, and VGG-based methods. Traditional methods have been shown to be less accurate and less robust to changing crowd density and perspective in comparison to these methods. CNN-based density estimation is a popular technique for crowd counting. To create a density map, a CNN is trained on an input image. Typically, the CNN is made up of multiple convolutional layers that extract characteristics from the image. One or more layers are fully interconnected in order to create the density map. The multi-column CNN (MCNN) [7] method is one of the most common CNN-based density estimation methods for crowd counting. This method involves training multiple CNNs with different architectures on the same dataset, and then combining their density maps to generate a final estimate. Another popular approach is the scale-adaptive counting (SA-CNN) method, which uses a single CNN to generate density maps at multiple scales. The networks and inference algorithms used in this study can be divided into groups based on their architecture, as well as the process depicted in Figure 2.

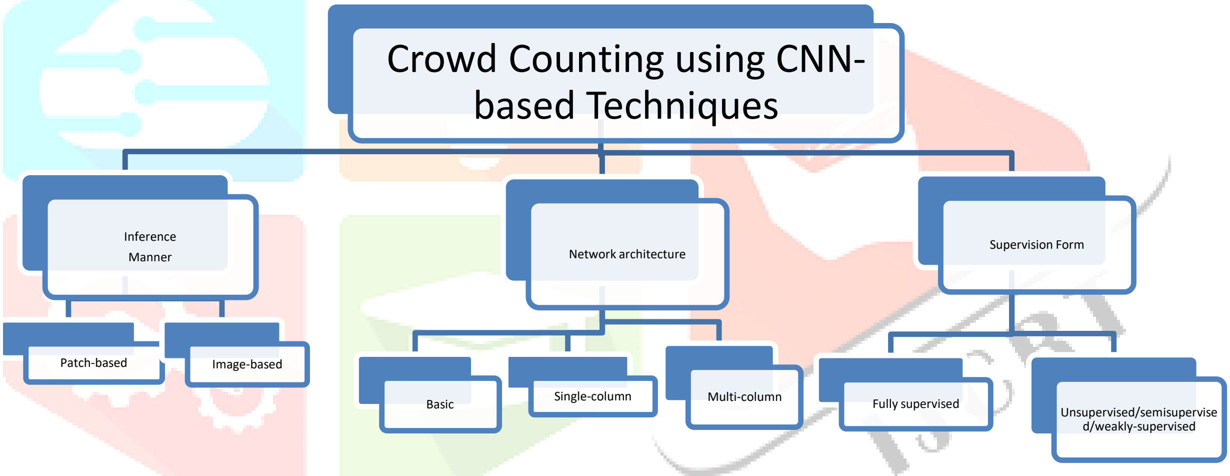


Figure 2: Crowd Counting using CNN-based Techniques

Earlier crowd counting methods, cited in references [8]–[11] employed sliding windows to detect people or heads on images using detection-based methods. Modern object detectors, like R-CNN [11] and YOLO [12], [13] often perform poorly, where the crowds are dense or there is background clutter. The development of regression-based methods has been for identifying the mapping between an image patch and a count. Either global or local features as well as regression techniques like linear regression or Gaussian mixture regression are used by these methods. However, these methods do not take spatial information into account. In order to resolve this problem, Lemptisky et. al. [1] used both a non-linear mapping method utilizing random forest regression and a crowdedness prior, as well as a linear mapping method based on the local characteristics and the density map in order to solve the problem. It is important to note that although these approaches incorporate geographical information, they still rely on conventional, handcrafted characteristics that may not be sufficient to ensure precise counting.

Related Previous Surveys

The earliest studies on crowd analysis were conducted by Zhan et al. [14] and Junior et al. [15], as mentioned. While Zitouni et al. [16] examined numerous approaches based on various criteria, Li et al. [17] highlighted the task of studying congested scenes using a variety of strategies. Using the same approach, Loy et al. [18] provided a thorough evaluation of contemporary crowd counting techniques based on video images, whereas Ryan et al. [18] assessed several image characteristics and regression models across multiple datasets. Grant et al. [19] looked at two different crowd analysis techniques, and Saleh et al. [20] evaluated the two basic methodologies, direct and indirect. These studies only addressed traditional methods even if they gave thorough analysis of traditional approaches with handcrafted elements. Using CNN based techniques, Sindagi et al [5] explored cutting-edge methods for crowd estimation and counting. Unfortunately, the most recent developments in CNN-based techniques, which were only up to 2017, were only briefly covered in this survey. Not just for crowd counts, Tripathi et al. [21] reviewed crowd analysis using CNN, but it was insufficiently thorough. Since new methods are being developed every month, it is important to track how crowd counting has changed over the previous five years.

III. UNDERSTANDING THE TAXONOMY OF CROWD COUNTING

A summary of many CNN-based methods for crowd counting and density estimation is given in this section. The methods are categorized based on their network architecture, as well as their training and inference approaches. The following summarizes the classification of these techniques.

3.1 Representative network architectures for crowd counting

Based on the various categories of network architectures used in crowd counting models, they can be bifurcated in three types. The three types are divided based on whether they use the basic architecture of CNN, or single column or multiple columns to extract the features of the image. These are detailed below:

3.1.1 Basic Architecture

One of the original deep learning approaches involved layers of convolution, pooling, and fully connected layers, to estimate the density and count the crowd. Density estimation and crowd counting are particularly well suited for these architectures, since they can effectively process visual information using convolutional filters.

- **Fu et al.** [23] proposed the first crowd-counting model based on CNN. Crowd counting is made more efficient and accurate by removing unused network connections from the feature maps and using two ConvNet classifiers in a cascade. This architecture aids in speeding up computations and enhancing model correctness. The significance and impact of Fu et al work is in this field are demonstrated by the fact that it has since served as the basis for other crowd counting models.

- **Wang et al.** [24]**,** based on the AlexNet architecture, proposed a deep network for counting crowds with high density. The model addresses the challenge of extremely dense crowds by adopting expanded negative samples, which are instances where the ground truth count is zero, to reduce interference in the network. Not only does this improve the crowd counting accuracy, the model can also be trained to distinguish between crowded and noncrowded scenes more effectively using these negative samples. Wang et al.'s work demonstrates the importance of incorporating negative samples in training deep networks for crowd counting tasks, especially when dealing with extremely dense crowds.

- **Yao et al.** [25]developed a cell counting framework by fine-tuning several architectures of deep residual networks (ResNet [26]). ResNet is a popular deep learning architecture that can efficiently learn features from large datasets, making it well-suited for image analysis tasks like cell counting. By fine-tuning ResNet, Yao et al. were able to create a highly accurate cell counting framework. With this approach, we are able to take advantage of ResNet's strengths and demonstrate the efficacy of transfer learning for developing deep learning models for new image analysis tasks. As well as contributing to cell counting, the work by Yao et al. has implications for other image analysis applications.

3.1.2 Multi-column Architecture

Multi-column architecture refers to a type of neural network architecture that is commonly used for crowd counting tasks. Unlike single-column architectures that use a single column to extract features from multiple scales, multi-column architectures use multiple independent columns to extract features from different scales. These columns are trained independently and then combined at a later stage to produce the final crowd count.

- **Hydra-CNN [22]:** To extract features at different scales, three parallel columns of different sizes are used in this architecture. In order to arrive at a final prediction, the outputs of each column are concatenated and passed through a fully connected layer.
- **CP-CNN[27] :** In order to draw out features at several scales this architecture uses a pyramid of convolutional neural networks. The features are mapped from each level of the pyramid are then combined using a channel-wise attention mechanism and passed through a fully connected layer for final prediction.
- **Multi-Column CNN** [7]**:** This architecture employs multiple columns with different filter sizes, for extracting features at multiple scales. Each column's output is then combined using a weighted sum and passed through a fully connected layer for final prediction.
- **Switch-CNN**[28] **:** As a result of this architecture, the appropriate column to process each part of the input image is dynamically selected using a switch mechanism. In order to determine the optimal switching points, the network learns the local density of the image.

3.1.3 Single-column Architecture

Crowd counting tasks are commonly carried out using single-column architectures of neural networks. Contrary to traditional multi-column architectures, single-column architectures extract features from multiple scales using a single column. The architectural design reduces the computational complexity of the model while maintaining high levels of accuracy in crowd counting.

- **CSRNet [29]:** Using this architecture, features are extracted at different scales without downsampling the input image by using a single column of convolutional layers and dilated filters. Following the feature maps, a series of convolutional and pooling layers are applied to the feature maps;And then comes a fully connected layer to predict what will happen.
- **SANet [25]:** With this architecture, important regions of the input image are selectively focused on by using a single column of convolutional layers. Following the feature maps, a series of convolutional and pooling layers are applied to the feature maps; then a fully connected layer follows to predict what will happen.
- **ACSCP[30] :** A single column of convolutional layers is used by this architecture with a cascaded pyramid of dilated convolutions to extract features at multiple scales. For final prediction the feature maps are then moved through a sequence of convolutional and pooling layers; then a fully connected layer follows.
- **PACNN [31]:** For explicit modelling of the spatial layout of objects in the input image, this architecture uses a position-aware convolutional layer. Convolutional and pooling layers are then applied to the feature maps; and then a fully connected layer follows.

Table 1. Classification of CNN-Based Techniques for Computer Vision Tasks

Techniques	**Network architecture**	**Inference manner**	**Supervised form**
Fu et al. [23]	Basic	Patch based	Fully Supervised
Hydra-CNN [22]	Multi columned	Patch based	Fully Supervised
CP-CNN [27]	Multi columned	Image based	Fully Supervised
Multi-Column CNN [7]	Multi columned	Image based	Fully Supervised
Wang et al. [24]	Basic	Patch based	Fully Supervised
Switch-CNN [28]	Multi columned	Patch based	Fully Supervised
CSRNet [29]	Single columned	Image based	Fully Supervised
SANet [25]	Single columned	Image based	Fully Supervised
Yao et al. [25]	Basic	Patch based	Fully Supervised
ACSCP [30]	Single columned	Patch based	Fully Supervised
PACNN [31]	Single columned	Image based	Fully Supervised

3.2 Inference manner

Based on their training techniques, the CNN-based crowd counting techniques can be divided into two categories: whole image-based inference and patch-based inference.

- **Patch-based inference:** The original image's random selection process is used to train the patch-based model. A sliding window is used to predict each portion of the test image, during the test phase. The total of all the sectional counts is used to determine the final count. According to reference, the model is trained by selecting patches from the input photos that represent real-life 3 x 3 meter squares. The CNN model is then used to create a comparable population density map after resizing these patches to 72 by 72 pixels. Crowd density maps can be used in estimating the number of objects in each area depending on the data collected. Feature Fusion Networks (FFN), Feature Enhancement Layers (FEL), and Density-Aware Networks (DAN) are the three modules that make up PaDNet [32]. Nine patches are taken from each input image and used to train this network. For the purpose of collecting crowd density information, the DAN uses CNN sub-networks that have already been trained on images with different crowd densities. By combining the weighted local and global contextual characteristics from the FEL, the FFN is derived. To deal with the issue of crowd variety here, Sajid et al. presented a "plug-and-play-based patch rescaling module" (PRM) in reference [33]. Prior to utilising PRM, the classifier determines the crowd density level of the patch picture, which is used to decide the right scaler (Down-scaler or Up-scaler) to rescale the patch image. The PRM is bypassed by the medium-crowd without rescaling, the high-crowd and low-crowd regions travel directly through it, and the no-crowd regions are instantly deleted. In reference [34], Sam et al. developed a hierarchical CNN tree in which the CNN's offspring regressors are more accurate than any of their parents.

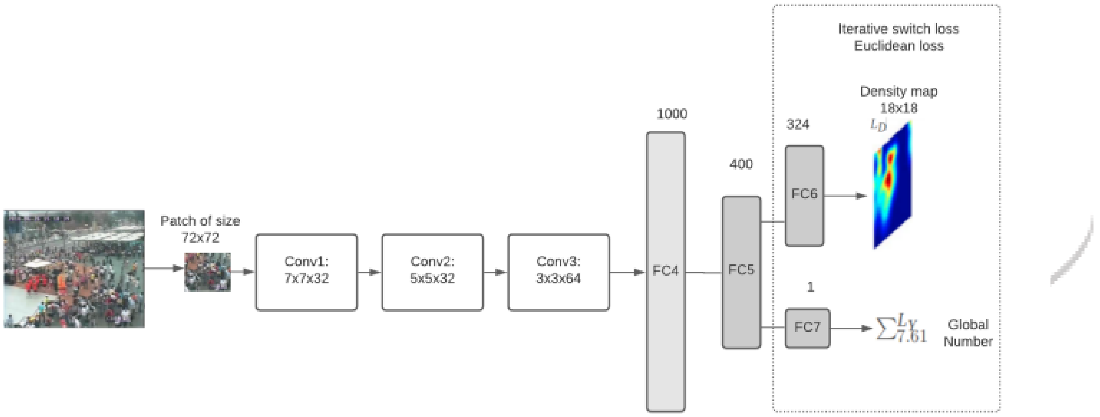
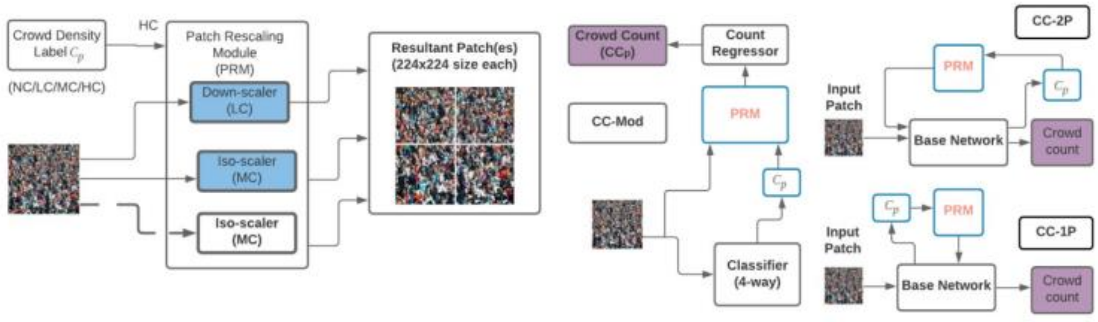


Figure 4. Patched-based inference method [2].



(a) PRM Module

(b) Modular Scheme(CC-Mod)

(c) End-to-End Networks (CC-1P, CC-2P)

Figure 5. The PRM module [33].

- **Image-based inference:** Since patch-based methods allow for the processing of large images by dividing them into smaller patches, they are popular for the purpose of counting crowds. However, they may not be able to capture global information and can be computationally expensive. On the other hand, the end-to-end CNN model proposed by Chong et al. offers an alternative approach that can directly compute the final count using the complete image. In order to extract high-level characteristics from the input image, this model initially uses a pre-trained CNN. The recurrent neural network (RNN) with memory cells receives these features and converts them to local counting numbers. The RNN is trained to capture temporal dependencies in the features and generate accurate local counts. To consider contextual information, the model shares computation over overlapping regions. This means that the model processes each patch not only independently but also in relation to its surrounding patches. As a result, the model is better able to depict the spatial relationships between picture objects and produce more precise local and global counts. Overall, the end-to-end CNN model proposed by Chong et al. is a promising approach [34] for crowd counting that can leverage both local and global information while reducing the computational burden of patch-based methods.

3.3 Supervision form

Crowd counting methods are bifurcated into two types depending on whether they use human-labelled annotations for training, or not. The first category is fully-supervised methods, while the second category is that of methods that are unsupervised/self-supervised/weakly supervised/semi-supervised.

- **Fully-supervised methods:** Most crowd counting methods that use CNNs depend on huge amounts of accurately annotated and diverse data. Nevertheless, the collection and annotation of this type of data is an arduous and time-consuming task. Furthermore, due to the scarcity of labelled data, these methods may over fit and perform poorly when applied to new, untested environments or domains. Therefore, developing techniques for training CNNs with fewer or no labelled annotations is an important area for future research.
- **Self-supervised/Unsupervised /weakly supervised/semi-supervised methods:** Self-supervised methods include an auxiliary task that is unrelated with but nonetheless connected with the main supervised task. Unsupervised or semi-supervised methods learn with or without somewhat restricted truth labels. As compared to supervised methods, certain techniques that train using un-labelled data have shown comparable performance.

IV. DATASETS

In order to encourage the creation of algorithms that can handle a variety of difficulties such as changes in scale, background clutter in surveillance footage, and differences in illumination in outdoor settings, numerous datasets are being added to the area of crowd counting. The production of models that can overcome above-said difficulties and the development of crowd counting techniques have been supported by a number of datasets over the past ten years.

- **USCD[35]:** One of the earliest datasets to be gathered with the intention of counting individuals was this one. A stationary camera that was elevated and pointed down at pedestrian walkways was used to record it. The dataset consists of 2000 frames, each measuring 158 x 512. Every 1/5 of the frames in the dataset have pedestrian annotations, and linear interpolation is applied for the other frames. Each pedestrian's bounding box coordinates are also included in the dataset. The dataset contains 49,885 distinct instances of persons that are divided in different sections for training and test. The UCSD dataset includes a crowd with low-density, around 25 pedestrian instances on an average, and because all photographs were taken from the same place, the perspective is generally consistent across images.
- **UCF_CC_50[36]:** This dataset, which was produced using freely accessible web photos, is regarded as one of the first to be truly difficult. For diverse situations, including those at concerts, protests, stadiums, and marathons, it offers distinct distortions of perspective and a variety of densities. The evaluation method is a process of 5-fold cross-validation because there are a smaller number of pictures in the dataset i.e., only 50. Due to the tiny quantity of the dataset, even the most sophisticated CNN-based approaches struggle to provide good results
- **ShanghaiTech[7]:** It is a dataset having 1198 images subdivided into 2 parts- A and B. Part A comprises of 482 images, while Part B has 716 images. Since there is a huge variation in the background clutter, crowd density as well as scale, this dataset is considered quite a challenge.
- **Mall[37]:** It sounds like the Mall dataset is a challenging dataset to work with, as it contains a high density of people, significant changes in lighting, different activity patterns, and occlusions. Another layer of complication is added by the distortion of the scene's perspective, creating significant fluctuations in the objects' look and size. This dataset would be useful for researchers who are interested in developing computer vision algorithms that can handle challenging and diverse real-world scenarios. The dataset may also be valuable for evaluating the performance of existing algorithms and identifying areas for improvement.
- **WorldExpo'10[2]:** This dataset, also referred to as the "ShanghaiTech" dataset, was compiled from crowd counting data at the Shanghai 2010 World Expo. The dataset consists of 1,132 annotated video clips that 108

security cameras recorded. It has annotations for 199,923 individuals and is made up of 3,920 frames with a total resolution having 576 x 720 pixels.

- **UCF-QNRF[38]:** This dataset contains over 1.25 million annotated people instances in 1535 images, hence considered to be large-scale. The pictures comprising this dataset are collected from various sources. Not only this, the pictures also include a huge range of scenarios, ranging from low to high crowd densities, outdoor to indoor scenes, static to dynamic crowds. The dataset also includes annotations for head and shoulder locations, making it suitable for both the tasks of estimation of density and the counting of crowd.
- **ShanghaiTechRGBD[39]:** This dataset is also called RGB-D. The ShanghaiTechRGBD dataset is extensive in size, containing 144,512 labelled head counts over 2,193 pictures. The dataset's challenge lies in its varied crowd scenarios and lighting conditions. Hence, this dataset is considered the hardest in terms of head count numbers, as far as RGB-D crowd counting is concerned.
- **Smartcity[40]:** Tencent YouTu created a dataset consisting of 50 high-shot images captured in 10 different scenes, including sidewalk, office entrance, and shopping mall. The dataset's goal is to evaluate the model's capacity for generalization in extremely sparse scenarios, covering both indoor and outdoor settings.
- **NWPU-Crowd[41]:** The dataset comprises of annotated heads to the tune of 454,312 from a range of 5,109 images, captured from different crowded places. The images were collected from different sources, including the internet, surveillance cameras, and public events. The dataset includes various types of crowd scenes, such as shopping malls, stadiums, and streets. The number of individuals in each image and the positions of their heads are marked for every image of the dataset. Both manual and automatic methods were involved in the semi-automatic annotation process that generated the annotations.
- **JHU-CROWD++ [42]:** Over a set of 4,250 photos in the JHU-CROWD dataset, there are a total of 1.11 million head count annotations. The information presents a challenge for crowd counting because it was gathered in a variety of settings, including various weather and lighting situations. Moreover, the dataset contains an expansive number of distractor pictures too. The dataset's annotations are comprehensive and include counts for both the head and the image levels.

VI. COMPARATIVE ANALYSIS

- **Comparing CNN-based and Traditional Models for Crowd Counting:** Table 2 compares traditional crowd counting models with CNN-based models, and as expected, it shows that CNN-based methods significantly outperform traditional models, demonstrating the deep convolutional neural networks remarkable capacity for learning new features from annotated data of a huge-scale. The potential of techniques based on CNN for crowd counting problems is demonstrated by this gain in performance.

- **Comparing CNN-based Crowd Counting Models' performance:** Since the release of the first density map estimation model based on CNN for crowd counting in 2015, there has been significant progress in the field. The Cross Scene [47] model, which was among the initial researches to employ CNNs for crowd estimation, solved the problem of cross-scene by using a simple topology and by deploying trained CNNs in previously unknown scenes. This model performs the lowest among the deep models. Although this model could perform worse than domain-specific and single-scene models, it offers a good way to generalise learned models to new situations.

VII. CONCLUSION

Crowd counting is an essential computer vision problem having numerous practical uses, including crowd control, public safety, and event planning. Often requiring manual calibration for different scenes, the traditional crowd counting methods had various limitations. The deep learning methods have made significant progress in overcoming these limitations, hence providing more accurate and efficient results. Due to their capacity for extracting and learning pertinent information from input photos, CNNs have become a well-liked deep learning technique for crowd counting. The numerous CNN-based models that have been suggested and attained cutting-edge performance are reviewed in this study, along with a number of crowd counting datasets. The handling of occlusions, scale fluctuations, and environmental changes are a few of the difficulties that still need to be resolved in crowd counting. Additionally, there is a need for more diverse and larger datasets that encompass a variety of real-world situations to enable the development of more robust and generalizable models. Overall, the research into crowd counting is ongoing, the accessibility of massive datasets as well as the developments in deep learning techniques have created new opportunities for creating efficient and accurate crowd counting models.

Table 2 Comparative analysis of different datasets and models

Dataset	UCFCC50		UCF-QNRF		UCSD		ShanghaiTech Part A		ShanghaiTech Part B		WorldExpo10		Mall	
Methods	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Count-Forest[43]	-	-.	-	-	4.40	1.61	-	-	-	-	-	-	10.0	2.50
Multi-source Multi-scale Counting [36]	590.3	468.0	-	-	-	-	-	-	-	-	-	-	-	-
Hydra-CNN [22]	425.2	333.7	-	-	-	-	-	-	-	-	-	-	-	-
RANet[28]	319.4	239.8	190	111	-	-	102.0	59.4	12.9	7.9	-	-	-	-
MCNN [7]	509.1	377.6	-	-	1.35	1.07	173.2	110.2	41.3	26.4	-	11.6	-	-
Switching-CNN [28]	439.2	318.1	445	228	2.10	1.62	135	90.4	33.4	21.6	-	9.4	-	-
CP-CNN [40]	320.9	295.8	-	-	-	-	106.4	73.6	30.1	20.1	-	8.86	-	-
SANet [25]	334.9	258.4	-	-	-	-	104.5	67.0	12.6	8.4	-	-	-	-
CRSNet [29]	397.5	266.1	-	-	-	-	115	68.2	16	10.6	-	-	-	-
PACNN [31]	357.8	267.9	-	-	1.18	0.89	106.4	66.3	13.5	8.9	-	7.8	-	-
SaCNN [2]	424.8	314.9	-	-	-	-	139.2	86.8	25.8	16.2	-	8.5	-	-

References

[1] V. Lempitsky and A. Zisserman, “Learning To Count Objects in Images,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2010. Accessed: Mar. 21, 2023. [Online]. Available: https://papers.nips.cc/paper_files/paper/2010/hash/fe73f687e5bc5280214e0486b273a5f9-Abstract.html

[2] C. Zhang, H. Li, X. Wang, and X. Yang, “Cross-Scene Crowd Counting via Deep Convolutional Neural Networks,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 833–841. Accessed: Mar. 20, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2015/html/Zhang_Cross-Scene_Crowd_Counting_2015_CVPR_paper.html

[3] Ruchika and R. K. Purwar, “Crowd Density Estimation Using Hough Circle Transform for Video Surveillance,” in *2019 6th International Conference on Signal Processing and Integrated Networks (SPIN)*, Noida, India: IEEE, Mar. 2019, pp. 442–447. doi: 10.1109/SPIN.2019.8711692.

[4] V. A. Sindagi and V. M. Patel, “CNN-based Cascaded Multi-task Learning of High-level Prior and Density Estimation for Crowd Counting.” arXiv, Aug. 16, 2017. Accessed: Mar. 21, 2023. [Online]. Available: http://arxiv.org/abs/1707.09605

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

[6] D. B. Sam, N. N. Sajjan, R. V. Babu, and M. Srinivasan, “Divide and Grow: Capturing Huge Diversity in Crowd Images with Incrementally Growing CNN,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 3618–3626. doi: 10.1109/CVPR.2018.00381.

[7] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, “Single-Image Crowd Counting via Multi-Column Convolutional Neural Network,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 589–597. doi: 10.1109/CVPR.2016.70.

[8] I. S. Topkaya, H. Erdogan, and F. Porikli, “Counting people by clustering person detector outputs,” in *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Seoul, South Korea: IEEE, Aug. 2014, pp. 313–318. doi: 10.1109/AVSS.2014.6918687.

[9] M. Li, Z. Zhang, K. Huang, and T. Tan, “Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection,” in *2008 19th International Conference on Pattern Recognition*, Tampa, FL, USA: IEEE, Dec. 2008, pp. 1–4. doi: 10.1109/ICPR.2008.4761705.

[10] B. Leibe, E. Seemann, and B. Schiele, “Pedestrian Detection in Crowded Scenes,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, San Diego, CA, USA: IEEE, 2005, pp. 878–885. doi: 10.1109/CVPR.2005.272.

[11] M. Enzweiler and D. M. Gavrila, “Monocular Pedestrian Detection: Survey and Experiments,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009, doi: 10.1109/TPAMI.2008.260.

[12] Ruchika, R. K. Purwar, and S. Verma, “Analytical Study of YOLO and Its Various Versions in Crowd Counting,” in *Intelligent Data Communication Technologies and Internet of Things*, D. J. Hemanth, D. Pelusi, and C. Vuppalapati, Eds., Singapore: Springer Nature Singapore, 2022, pp. 975–989.

[13] J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 6517–6525. doi: 10.1109/CVPR.2017.690.

[14] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu, “Crowd analysis: a survey,” *Mach. Vis. Appl.*, vol. 19, no. 5–6, pp. 345–357, Oct. 2008, doi: 10.1007/s00138-008-0132-4.

[15] J. Silveira Jacques Junior, S. Musse, and C. Jung, “Crowd Analysis Using Computer Vision Techniques,” *IEEE Signal Process. Mag.*, p. 5562657, Sep. 2010, doi: 10.1109/MSP.2010.937394.

[16] M. S. Zitouni, H. Bhaskar, J. Dias, and M. E. Al-Mualla, “Advances and trends in visual crowd analysis: A systematic survey and evaluation of crowd modelling techniques,” *Neurocomputing*, vol. 186, pp. 139–159, Apr. 2016, doi: 10.1016/j.neucom.2015.12.070.

[17] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded Scene Analysis: A Survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 367–386, Mar. 2015, doi: 10.1109/TCSVT.2014.2358029.

[18] C. C. Loy, K. Chen, S. Gong, and T. Xiang, "Crowd Counting and Profiling: Methodology and Evaluation," in *Modeling, Simulation and Visual Analysis of Crowds*, S. Ali, K. Nishino, D. Manocha, and M. Shah, Eds., in The International Series in Video Computing, vol. 11. New York, NY: Springer New York, 2013, pp. 347–382. doi: 10.1007/978-1-4614-8483-7_14.

[19] J. M. Grant and P. J. Flynn, "Crowd Scene Understanding from Video: A Survey," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 13, no. 2, pp. 1–23, May 2017, doi: 10.1145/3052930.

[20] S. A. M. Saleh, S. A. Suandi, and H. Ibrahim, "Recent survey on crowd density estimation and counting for visual surveillance," *Eng. Appl. Artif. Intell.*, vol. 41, pp. 103–114, May 2015, doi: 10.1016/j.engappai.2015.01.007.

[21] G. Tripathi, K. Singh, and D. K. Vishwakarma, "Convolutional neural networks for crowd behaviour analysis: a survey," *Vis. Comput.*, vol. 35, no. 5, pp. 753–776, May 2019, doi: 10.1007/s00371-018-1499-5.

[22] R. Gouiaa, M. A. Akhloufi, and M. Shahbazi, "Advances in Convolution Neural Networks Based Crowd Counting and Density Estimation," *Big Data Cogn. Comput.*, vol. 5, no. 4, p. 50, Sep. 2021, doi: 10.3390/bdcc5040050.

[23] M. Fu, P. Xu, X. Li, Q. Liu, M. Ye, and C. Zhu, "Fast crowd density estimation with convolutional neural networks," *Eng. Appl. Artif. Intell.*, vol. 43, pp. 81–88, Aug. 2015, doi: 10.1016/j.engappai.2015.04.006.

[24] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, "Deep People Counting in Extremely Dense Crowds," in *Proceedings of the 23rd ACM international conference on Multimedia*, Brisbane Australia: ACM, Oct. 2015, pp. 1299–1302. doi: 10.1145/2733373.2806337.

[25] *Computer Vision – ECCV 2018. Part V: Part V.* in Lecture notes in computer science Image processing, computer vision, pattern recognition, and graphics, no. 11209. Cham: Springer, 2018.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition." arXiv, Dec. 10, 2015. Accessed: Mar. 22, 2023. [Online]. Available: http://arxiv.org/abs/1512.03385

[27] 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.

[28] D. B. Sam, S. Surya, and R. V. Babu, "Switching Convolutional Neural Network for Crowd Counting," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 4031–4039. doi: 10.1109/CVPR.2017.429.

[29] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT: IEEE, Jun. 2018, pp. 1091–1100. doi: 10.1109/CVPR.2018.00120.

[30] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, and X. Yang, "Crowd Counting via Adversarial Cross-Scale Consistency Pursuit," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 5245–5254. doi: 10.1109/CVPR.2018.00550.

[31] M. Shi, Z. Yang, C. Xu, and Q. Chen, "Revisiting Perspective Information for Efficient Crowd Counting," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA: IEEE, Jun. 2019, pp. 7271–7280. doi: 10.1109/CVPR.2019.00745.

[32] Y. Tian, Y. Lei, J. Zhang, and J. Z. Wang, "PaDNet: Pan-Density Crowd Counting," *IEEE Trans. Image Process.*, vol. 29, pp. 2714–2727, 2020, doi: 10.1109/TIP.2019.2952083.

[33] U. Sajid and G. Wang, "Plug-and-Play Rescaling Based Crowd Counting in Static Images," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass Village, CO, USA: IEEE, Mar. 2020, pp. 2276–2285. doi: 10.1109/WACV45572.2020.9093561.

[34] C. Shang, H. Ai, and B. Bai, "End-to-end crowd counting via joint learning local and global count," in *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA: IEEE, Sep. 2016, pp. 1215–1219. doi: 10.1109/ICIP.2016.7532551.

[35] A. B. Chan, M. Morrow, and N. Vasconcelos, "Analysis of Crowded Scenes using Holistic Properties".

IJCRT

[36] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, “Multi-source Multi-scale Counting in Extremely Dense Crowd Images,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA: IEEE, Jun. 2013, pp. 2547–2554. doi: 10.1109/CVPR.2013.329.

[37] C. C. Loy, S. Gong, and T. Xiang, “From Semi-supervised to Transfer Counting of Crowds,” in *2013 IEEE International Conference on Computer Vision*, Sydney, Australia: IEEE, Dec. 2013, pp. 2256–2263. doi: 10.1109/ICCV.2013.270.

[38] H. Idrees *et al.*, “Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds.” arXiv, Aug. 02, 2018. Accessed: Mar. 22, 2023. [Online]. Available: http://arxiv.org/abs/1808.01050

[39] D. Lian, J. Li, J. Zheng, W. Luo, and S. Gao, “Density Map Regression Guided Detection Network for RGB-D Crowd Counting and Localization,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA: IEEE, Jun. 2019, pp. 1821–1830. doi: 10.1109/CVPR.2019.00192.

[40] L. Zhang, M. Shi, and Q. Chen, “Crowd counting via scale-adaptive convolutional neural network.” arXiv, Feb. 06, 2018. Accessed: Mar. 22, 2023. [Online]. Available: http://arxiv.org/abs/1711.04433

[41] Q. Wang, J. Gao, W. Lin, and X. Li, “NWPU-Crowd: A Large-Scale Benchmark for Crowd Counting and Localization,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2141–2149, Jun. 2021, doi: 10.1109/TPAMI.2020.3013269.

[42] V. A. Sindagi, R. Yasarla, and V. M. Patel, “Pushing the Frontiers of Unconstrained Crowd Counting: New Dataset and Benchmark Method.” arXiv, Oct. 27, 2019. Accessed: Mar. 22, 2023. [Online]. Available: http://arxiv.org/abs/1910.12384

[43] D. Oñoro-Rubio and R. J. López-Sastre, “Towards Perspective-Free Object Counting with Deep Learning,” in Computer Vision – ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., in Lecture Notes in Computer Science, vol. 9911. Cham: Springer International Publishing, 2016, pp. 615–629. doi: 10.1007/978-3-319-46478-7_38.

IJCRT