Journal of Global Research in Electronics and Communication
Volume 2, No. 1, January, 2026
Available Online at: www.jgrec.info
RESEARCH PAPER

Deep-Learning-Based Crowd Density Estimation: A Comprehensive Survey

Shrihari Pathak
Scholar, Department of Computer Science and Engineering
Sri Aurobindo Institute of Technology
Indore, India
shrihari17@gmail.com

***Abstract*—The importance of crowd density estimate (CDE) in event management, urban planning, and public safety has made it a crucial area of study in computer vision. Conventional computer vision techniques are susceptible to challenging situations in the crowd including occlusions, scaling, and light conditions. Since the implementation of deep learning, convolutional neural networks (CNNs) and, more recently, transformer-based and hybrid designs, have demonstrated state-of-the-art performance in the estimate of crowd density and person count. The paper will provide an extensive review of the deep-learning-based methods of crowd density estimation published between 2015 and 2025 in terms of architectural development, the variability of the datasets, and the advancement of the methodology. The trends, benchmarks, and challenges are consolidated in the survey, which will serve as a roadmap in the future research of deep-learning-based crowd analysis.**

Keywords—crowd density estimation, deep learning, CNN, YOLO, computer vision, image processing, surveillance.

I. Introduction

Crowd density estimation (CDE) is one of the essential applications of intelligent surveillance, crowd management, and smart infrastructure of a city[1][2][3]. The main goal here is to determine how many people there may be in an image or video frame or how many there are in a density map[4][5]. The manual monitoring systems are expensive and ineffective in big public events. Therefore, the move to automated solutions based on AI is now a necessity.

In early methods of analysis of crowds were predominantly handcrafted, using background subtraction techniques and feature lists[6]. These methods were however poor at dealing with perspective distortion and occlusion[7][8]. CNNs are deep-learning-based methods that have transformed CDE by the means of learning the spatial and contextual representation directly through data.

The paper is intended to summarize the current studies of 20182024, examine the changes in methods of analysis, contrast databases usage, and define future research perspectives along with the focus on the use of new-generation models such as YOLO and high-end CNN versions.

II. Literature Review

A number of studies have explored deep-learning-based CDE under a variety of methodological perspectives, such as regression-based, detection-based and hybrid frameworks.

A. Regression-Based Approaches

Regression models are used to estimate density maps directly using input images[9]. The first step toward this direction incorporated multi-column CNNs (MCNN) that scale-invariantly represent crowds with many receptive fields [10]. Subsequently, CSRNet proposed dilated convolutions which can be used to expand the receptive field efficiently, without raising computational complexity[11] [12].

Recent studies such as the literature "AI Enhanced Surveillance to identify and recognize crowd behavior" used CNNs and RNN to provide improved density estimation of real time crowd[13].

B. Detection-Based Approaches

Detection-based methods use object detection networks to measure the number of people directly[14][15][16]. The initial models were based on Faster R-CNN and SSD, and the subsequent models were based on YOLO and Retina Net variants. Nevertheless, the most recent models, including YOLOv8 and future YOLOv11, will be better because they have an anchor-free detection model and transformer backbones[17].

A new article in IET Image Processing [https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/ipr2.13328] demonstrated a transformer-CNN hybrid with a high level of counting performance with varying densities, suggesting the shift to fusion architectures.

C. Hybrid and Transformer Models

Hybrid architectures either use a combination of regression and detection systems or apply attention-based transformers to contextual learning[18]. Transformer backbones like Swin-Transformer and Vision Transformer (ViT) have been adapted for CDE tasks, showing promising generalization under varying scales and scenes.

The study in "Video based crowd analysis using deep learning" emphasized the importance of combining CNN feature extraction with transformer-based reasoning for dynamic crowd scenes [19].

III. Methodology

This paper systematically surveys works from 2015 to 2025 by analysing architectures, datasets, and evaluation metrics. The methodology follows a three-phase approach:

Keep your graphic and text files apart until the text has been styled and formatted. Avoid using hard tabs and just use one hard return at the conclusion of a paragraph. The document should not contain any pagination of any type. Text heads should not be numbered; the template will do that for you.

A. *Data Collection*

Open-access repositories are among the primary data sources (e.g., Crowd Counting Datasets like ShanghaiTech, UCF_CC_50, and WorldExpo'10), peer-reviewed journals, and conference proceedings. Only works explicitly employing deep learning or CNN-based architectures were included.

B. *Comparative Framework*

The survey classifies studies based on methodological dimensions:

- Architecture Type: Regression, Detection, or Hybrid.
- Model Backbone: VGG, ResNet, Efficient Net, Transformer, or YOLO variants.
- Dataset Diversity: Range of environmental and density conditions.
- Evaluation Metrics: Mean Absolute Error (MAE) and Mean Squared Error (MSE).

C. *Analytical Synthesis*

TABLE I. SUMMARY OF KEY CNN-BASED CDE MODELS

Year	Model	Architecture	Dataset	MAE/ MSE
2018	MCNN [1]	Multi-Column CNN	ShanghaiTech	110.2/ 173.2
2019	CSRNet [4]	Dilated CNN	UCF_CC_50	68/115
2021	CANNet [9]	Context-Aware CNN	WorldExpo'1 0	7.8/12.3
2023	TransCro wd [7]	Transformer + CNN	UCF-QNRF	55/92

IV. DISCUSSIONS

A. *Architectural Trends*

Between 2018 and 2021, CNN-based regression models dominated research, primarily using VGG or ResNet backbones[20][21]. However, post-2022, hybrid and transformer-integrated models began outperforming pure CNNs in complex scenes. Transformer integration allows models to attend to global context, reducing over-counting in dense regions.

B. *Dataset Expansion*

Datasets such as ShanghaiTech, UCF_CC_50, and JHU-CROWD++ remain standard benchmarks. However, it is limited to generalizability by dataset imbalance where Asian cityscapes are favored. After 2020, a tendency towards synthetic dataset creation with the help of GANs and simulation tools developed [22]. The bigger and broader data sets are essential in the real-world implementation.

C. *Real-Time Application Challenges*

Although regression models are highly accurate, detection-based variants of YOLO are real-time, which is required by surveillance systems [23]. Yet, it does not cope with severe occlusions.

These limitations can be greatly alleviated with the introduction of the latest release YOLOv11, that has better contextual embedding and dynamic anchor-free detection[24][25].

D. *Generalization and Domain Adaptation*

Cross-scene generalization is another problem[26]. Adversarial learning and self-supervised fine-tuning are techniques of domain adaptation that have been investigated but not yet defined by benchmarks[27]. Future versions of YOLO would be able to be scaled to multi-environment operation by using attention-based scene adaptation[28].

V. PROPOSED IMPROVEMENTS

According to the gaps found in the literature, this paper introduces the following improvements:

A. *Dense Crowd Detection*

Recent advancements of YOLO's transformer-assisted spatial reasoning can outperform earlier model's performance in crowded and low-light conditions[29].

B. *Enhanced CNN Backbones*

Adoption of EfficientNet-V2 and ConvNeXt architectures for better efficiency-accuracy trade-offs.

C. *Augmented Datasets*

Expansion of training datasets using synthetic crowd simulation and domain randomization.

D. *Multi-modal Learning*

Combination of visual and contextual (e.g. environmental or temporal) information to more detailed representations.

E. *Benchmark Standardization*

Single review framework among datasets to make equal comparisons.

VI. LIMITATIONS

Although this survey covers a broad spectrum of research from 2018–2024, certain limitations persist:

A. *Dataset Accessibility*

Limited accessibility to proprietary datasets.

B. *Cross Comparison*

Variations in evaluation protocols make cross-comparison difficult.

VII. FUTURE WORK

Future research should focus on:

A. *Real-time CDE under Edge Computing Constraints*

Lightweight transformer-CNN hybrids for embedded systems.

B. *Model Advancements*

Exploring vision-language models for joint crowd reasoning.

C. *Ethical AI and Privacy*

Incorporating anonymization layers in surveillance pipelines.

D. *Cross-modal Adaptation*

Improving 3D-aware density estimation with LiDAR, drone, and CCTV imagery.

VIII. Conclusion

Crowd density estimation with deep learning has developed from handcrafted models of features to highly flexible CNN-transformer hybrids. The 20182024 period has witnessed an intensive innovation of architectural design and a major increase in data. Nevertheless, there are still issues of real-time scalability, dataset bias and domain generalization. The future modelling of YOLOv11 with other new detection models can potentially redefine the accuracy and computational performance of dense scene understanding.

This survey contributes to consolidating the progress and identifying forward-looking pathways for research and application in intelligent surveillance and crowd analytics.

References

[1] L. Sujihelen, S. Subhadra, S. P. Kota, and S. Vignesh, "Video based crowd analysis using deep learning," *AIP Conf. Proc.*, vol. 3257, no. 1, p. 20064, 2025, doi: 10.1063/5.0276180.

[2] R. Patel and P. Patel, "A Survey on AI-Driven Autonomous Robots for Smart Manufacturing and Industrial Automation," *Tech. Int. J. Eng. Res.*, vol. 9, no. 2, pp. 46–55, 2022, doi: 10.56975/tijer.v9i2.158819.

[3] K. Seetharaman, "Incorporating the Internet of Things (IoT) for Smart Cities: Applications, Challenges, and Emerging Trends," *Asian J. Comput. Sci. Eng.*, vol. 08, no. 01, pp. 8–14, 2023, doi: 10.22377/ajcse.v8i01.199.

[4] B. Jeganathan, "Exploring the Power of Generative Adversarial Networks (GANs) for Image Generation: A Case Study on the MNIST Dataset," *Int. J. Adv. Eng. Manag.*, vol. 7, no. 1, pp. 21–46, Jan. 2025, doi: 10.35629/5252-07012146.

[5] S. Kumara, "Post-Quantum Identity Mesh For Autonomous 5g, Iot, And National Connectivity Systems: Implications For Future-Resilient Digital Infrastructure," *Int. J. Adv. Res. Comput. Sci.*, vol. 16, no. 6, pp. 105–113, Dec. 2025, doi: 10.26483/ijarcs.v16i6.7390.

[6] 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)*, IEEE, Jun. 2016, pp. 589–597. doi: 10.1109/CVPR.2016.70.

[7] S. B. R. Karri, V. K. Devalla, R. K. Bojja, and M. S. Pandey, "An Architecture for Model Monitoring System with Automated Data Validation and Failure Handling," in *2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI)*, IEEE, Apr. 2025, pp. 1960–1966. doi: 10.1109/ICCSAI64074.2025.11064092.

[8] S. Dodda, N. Kamuni, P. Nutalapati, and J. R. Vummadi, "Intelligent Data Processing for IoT Real-Time Analytics and Predictive Modeling," in *2025 International Conference on Data Science and Its Applications (ICoDSA)*, IEEE, Jul. 2025, pp. 649–654. doi: 10.1109/ICoDSA67155.2025.11157424.

[9] S. Garg, "Next-Gen Smart City Operations with AIOps & IoT : A Comprehensive look at Optimizing Urban Infrastructure," *J. Adv. Dev. Res.*, vol. 12, no. 1, 2021, doi: 10.5281/zenodo.15364012.

[10] M. Wang, "A comprehensive survey of crowd density estimation and counting," no. September 2024, pp. 1–32, 2025, doi: 10.1049/ipr2.13328.

[11] Z. Huo, Y. Wang, Y. Qiao, J. Wang, and F. Luo, "Domain adaptive crowd counting via dynamic scale aggregation network," *IET Comput. Vis.*, vol. 17, no. 7, pp. 814–828, Oct. 2023, doi: 10.1049/cvi2.12198.

[12] R. P. Mahajan and N. Jain, "Enhancing the Deep Learning-Based Pet Imaging Super-Resolution for Facial Expression Images," in *2025 International Conference on Intelligent and Cloud Computing (ICoICC)*, 2025, pp. 1–6. doi: 10.1109/ICoICC64033.2025.11052125.

[13] Y. Li, X. Zhang, and D. Chen, *CSRNet : Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes*. 2018.

[14] A. Khan *et al.*, "A Deep Learning Approach for Crowd Counting in Highly Congested Scene," *Comput. Mater. Contin.*, vol. 73, no. 3, pp. 5825–5844, 2022, doi: 10.32604/cmc.2022.027077.

[15] R. Patel and P. B. Patel, "Mission-critical Facilities: Engineering Approaches for High Availability and Disaster Resilience," *Asian J. Comput. Sci. Eng.*, vol. 8, no. 3, pp. 1–9, 2023, doi: 10.22377/ajcse.v10i2.212 Authors:

[16] P. B. Patel, "Energy Consumption Forecasting and Optimization in Smart HVAC Systems Using Deep Learning," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 4, no. 3, pp. 780–788, 2024, doi: 10.48175/IJARSCT-18991.

[17] A. Bansal and K. S. Venkatesh, "People Counting in High Density Crowds from Still Images," pp. 1–7, 2015.

[18] B. D. Sunil, R. Venkatesh, and S. Todmal, "Density Estimation and Crowd Counting," 2025.

[19] W. Liu, M. Salzmann, and P. Fua, "Context-Aware Crowd Counting," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5094–5103. doi: 10.1109/CVPR.2019.00524.

[20] M. H. Rahima Khanam, "Yolov11: An Overview Of The Key Architectural Enhancements," vol. 2024, pp. 1–9, 2024.

[21] B. S. Prakash, B. Tamilsudar, and T. V. Kani, "AI Enhanced Surveillance for Identifying and Recognizing Crowd Behavior," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 13, no. 5, pp. 4008–4017, May 2025, doi: 10.22214/ijraset.2025.71063.

[22] D. Liang, X. Chen, W. Xu, Y. Zhou, and X. Bai, "TransCrowd : weakly-supervised crowd counting with transformers," 2022.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, " You Only Look Once: Unified, Real-Time Object Detection ," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[24] S. Thangavel, "Method for Real-Time Image Processing in IoT-Based Smart Home Security Systems," 2024

[25] S. Amrale, "Anomaly Identification in Real-Time for Predictive Analytics in IoT Sensor Networks using Deep," *Int. J. Curr. Eng. Technol.*, vol. 14, no. 6, pp. 526–532, 2024, doi: 10.14741/ijcet/v.14.6.15.

[26] S. Phalke, Y. D. Athave, and B. N. Ilag, "A Multi-Layered Approach to IT Infrastructure Governance and Compliance-Security, Hardening, and Audit Readiness V3," *Int. J. Comput. Appl.*, vol. 187, no. 12, pp. 29–33, Jun. 2025, doi: 10.5120/ijca2025925133.

[27] K. P. Smithashree, M. G. Meghana, R. Shamitha, B. S. Suhasini, and M. U. Varsha, "Secure Crowd AI- Crowd Estimation and Surveillance System," vol. 12, no. 5, pp. 517–524, 2025, doi: 10.17148/IARJSET.2025.12583.

[28] Q. Wang, J. Gao, and W. Lin, "Learning from Synthetic Data for Crowd Counting in the Wild," 2019.

[29] S. K. Chintagunta and S. Amrale, "A Deep Learning Framework for Adaptive E- Learning : Integrating Learning Style Detection in Web-Based Platforms," *Int. J. Adv. Res. Sci. Commun. Technol.*, pp. 716–727, 2024, doi: 10.48175/IJARSCT-19397.