A Systematic Review on Crowd Density Estimation Using Deep Learning Techniques: State-of-the-Art Methods and Future Challenges

Norah Aloufi*, Liyakathunisa Syed
Department of Computer Science-College of Computer Science and Engineering, Taibah University, Medina, Saudi Arabia

***Abstract*—Estimating crowd density is a cornerstone of modern urban management and public safety, particularly in the aftermath of catastrophic incidents, such as the 2015 Mina stampede. With the rapid advancement of artificial intelligence (AI) technologies, deep learning (DL) has emerged as a powerful tool for addressing these challenges. This systematic review provides a comprehensive evaluation of current crowd density estimation methodologies, analyzing model architectures, datasets, and research trends. The review was conducted in accordance with PRISMA 2020 guidelines, and the search encompassed five major electronic databases (IEEE Xplore, Scopus, Google Scholar, Web of Science, and ScienceDirect) for the period 2020 to 2025. The selection process relied on rigorous eligibility criteria, including English-language publications that offer methodological contributions or empirical assessments in the field of computer vision and machine learning (ML). Twenty final studies were included, 70% of which were published in scientific journals. The analysis revealed that 55% of the studies relied entirely on DL models, while 30% leaned towards hybrid modelling. The ShanghaiTech dataset remained the most frequently used benchmark, accounting for 50% of the studies, followed by UCF_CC_50 and WorldExpo'10 datasets. Although some models achieved a high accuracy of 99.88%, they still faced challenges in highly congested scenes and visual obstructions. This review reveals a growing shift towards edge intelligence and lightweight models to reduce latency, with a pressing need for more diverse datasets to minimize bias. This study concludes that bridging the gap between simulation and reality requires integrating contextual information and behavioral analysis to enable more reliable, proactive, and real-time crowd management.**

***Keywords*—Crowd density estimation; computer vision; deep learning; PRISMA 2020; systematic literature review**

I. Introduction

A. Context and Historical Challenges

Crowd management is a critical field of study that focuses on the organization, management, and safety of crowd events. Crowd management plays a vital role in ensuring public safety, optimally utilizing space, and managing risks associated with events like concerts, sports activities, religious activities, and public rallies. Efficient crowd management is critical in preventing accidents, but also in maximizing the overall experience of attendees. With the increased rate of mass gatherings in the world, adopting appropriate crowd management policies has been a necessity more than ever before. Efficient crowd management ensures that crises are well-handled, resources are efficiently utilized, and any potential risks are prevented in their inception [1]. As highlighted by recent research, crowd management has become a significant part of urban planning for modern societies to enable the planning of crowd movement, counting, and the prevention of future issues [2].

Improper crowd management can lead to catastrophic consequences, including injuries, fatalities, and excessive property damage. An example of this is the 2015 Hajj stampede at Mina, Saudi Arabia, triggered by inadequate crowd management and planning, which resulted in over 2,400 fatalities, and around 400 were reported missing [3]. This tragedy highlighted the imperative for advanced crowd management systems to prevent such catastrophes. Similarly, the 2021 Astro World Festival tragedy in Houston, Texas, exposed how unscientific crowd handling can make events turn fatal, as overcrowding and inadequate emergency response led to ten deaths and many injuries [4]. Such accidents point towards the absolute need for using crowd management practices based on advanced technologies to ensure public safety. There has been recent emphasis that the failure to properly manage crowds may have devastating consequences, including loss of lives and property, along with a loss of public confidence in event organizers [2].

B. Technological Evolution in Crowd Monitoring

In recent years, technology has revolutionized the crowd management sector. Artificial intelligence (AI), internet of things (IoT), edge computing, and digital twins are a few such technologies that are being implemented to predict and manage crowd behavior more effectively. AI-powered surveillance systems have the ability to track crowd density in real time and alert the authorities of possible threats, such as overcrowding or unauthorized access [5]. Additionally, machine learning (ML) algorithms are being used to analyze historical data and predict crowd behavior, enabling proactive decision-making [2]. The IoT further enhances these capabilities by connecting a network of smart devices, such as cameras, sensors, and wearables, enabling the collection of real-time data on crowd movement and environmental conditions. This data is then processed and integrated with AI systems to provide a comprehensive view of crowd dynamics, allowing for more precise and timely interventions [6].

A critical limitation in current literature is the performance gap between controlled benchmark environments and unpredictable real-world scenarios. While many models achieve high accuracy on standard datasets, they often struggle with the dynamic complexities of field deployment, such as extreme lighting variations and dense occlusions in uncontrolled environments [7]. To address these challenges, edge computing

*Corresponding author.

plays a pivotal role by enabling local data processing on devices such as cameras and sensors, reducing latency and ensuring real-time responsiveness [8]. This is particularly important in high-density environments where immediate action is required to prevent accidents or congestion. In parallel, digital twins create virtual replicas of actual spaces, allowing organizers to simulate crowd flow and predict congestion before an event [9]. These technologies not only enhance safety but also improve the efficiency of crowd management operations.

The resolution of the problem of accurate crowd density estimation is of immediate importance as it has direct implications for public safety, resource utilization, and overall planning of large-scale events. Accurate crowd density estimation can prevent overcrowding, eliminate safety threats, and enable timely intervention in the event of an emergency. For instance, in religious gatherings like Hajj or sporting events with large crowds, real-time crowd monitoring can significantly reduce the possibility of stampedes or other disasters. Moreover, advances in technology here can enable urban planning, traffic management, and disaster response systems to be more intelligent and safer cities. By addressing the limitations of current technologies through a systematic analysis of existing studies, this review aims to support and inform the development of more efficient, reliable, and scalable crowd density estimation systems, ultimately contributing to improved safety and quality of life in dense environments.

Unlike earlier studies that mainly focused on traditional convolutional neural network (CNN) architectures [10], [11], this review provides a more contemporary analysis by categorizing methodologies into ML, deep learning (DL), and hybrid approaches. While earlier literature reviews [12], [13] offered broad overviews of crowd counting, they often lacked a systematic framework like PRISMA 2020 [14] and did not account for the rapid advancements in hybrid modeling (e.g., ConvLSTM and GANs) or the critical shift toward edge intelligence required for real-time deployment in high-density scenarios. As demonstrated in Table II, our study bridges this gap by analyzing state-of-the-art contributions from 2020 to 2025, providing a specialized taxonomy that addresses both computational efficiency and architectural complexity.

C. Objectives and Scope of the Review

Given the aforementioned challenges and limitations, this systematic review presents a structured analysis of existing approaches to crowd density estimation, focusing on datasets, computational methods, and hybrid approaches. It also aims to achieve the following specific objectives: 1) review and categorize state-of-the-art methods for crowd density estimation using ML, DL, and hybrid approaches; 2) analyze datasets and metrics used for evaluating crowd density estimation research; 3) outline strengths and weaknesses in accuracy, generalization capability, and application feasibility in real time at very high densities or occlusions; and 4) outline current research gaps and future directions for developing robust, scalable, and computational-efficiency crowd density estimation models suitable for real-world and edge computing–based deployments.

The following research questions (RQs) were formulated to address the objectives of this review:

- RQ1: How can current crowd density estimation methodologies be categorized into ML, DL, and hybrid approaches, and what are the most common architectural structures in recent literature?
- RQ2: What are the most commonly used benchmark datasets, and how do visual challenges such as overcrowding and occlusion affect the accuracy and generalization of these models?
- RQ3: What current research gaps hinder the transition from simulation-based models to actual deployment in edge intelligence environments, and what requirements are necessary to ensure real-time response?

This systematic review presents a novel analytical perspective by proposing a comprehensive taxonomy that classifies crowd density estimation techniques into ML, DL, and hybrid architectures. Distinct from a conventional survey, this research provides targeted benchmarking insights via a comparative examination of 20 leading studies (2020-2025), pinpointing the crucial accuracy–efficiency trade-offs essential for real-time edge computing applications. Through the synthesis of these methodological advancements, it formulates a strategic framework for the transition from theoretical modeling to proactive, intelligent crowd management in high-density environments, exemplified by events such as the Hajj.

II. Methods

This systematic review was performed following the guidelines of PRISMA 2020 [14]. This not only helps to make it transparent and rigorous but also facilitates their reproduction. The primary focus is placed on crowd density estimation, with selective inclusion of closely related crowd management studies. The following subsections list the processes of searching, selecting studies, the criteria of eligibility, and extraction of data that have been used during this review. Fig. 1 shows the PRISMA flow diagram considered in this review.

A. Search and Selection Strategy

The literature relevant to the topic of crowd density estimation and managing crowd has been systematically searched. The literature search has been done on the following five major electronic literature databases: IEEE Xplore, Google Scholar, Scopus, Web of Science, ScienceDirect, and Springer. The final search for literature was done in October 2025. Articles from journals as well as papers from conferences have been considered. This will help in obtaining all the important literature available on the topic. The research strategy was designed by incorporating a set of keywords associated with crowd analysis and density estimation. The major set of keywords are “crowd density estimation”, “crowd counting”, “crowd management”, “density estimation”, and “crowd”. These keywords are combined through the help of operators (AND and OR), according to the applicable syntax in each database. The research will only contain those studies formally published in English in the years from 2020 to 2025 that directly contribute to answering the pre-defined research questions (RQ1–RQ3). The process of selecting studies was carried out by two reviewers using an independent manual study selection process. The process involved screening all the retrieved records for duplication, which was removed before proceeding with

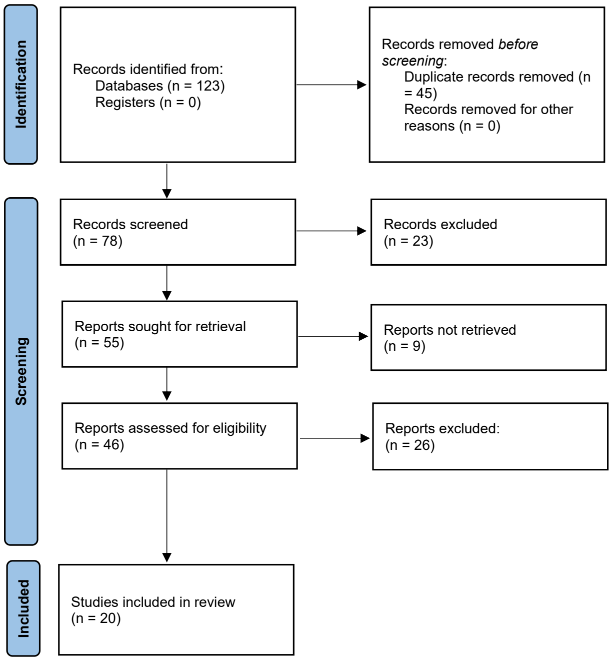


Fig. 1. PRISMA flow diagram of the survey methodology.

the study screening. There were two stages involved in the study screening process. The first stage involved title/abstract screening, whereby all inappropriate studies in relation to the topic under consideration were removed. The final stage involved full-text screening, whereby the identified potential studies were screened against the predetermined inclusion and exclusion criteria.

B. Eligibility Criteria

Eligibility criteria are established in a way that helps cover the necessary studies in accordance with the requirements of the systematic review. The criteria consider studies that address the topic of crowd density estimation, crowd counting, or crowd management. The studies particularly target research works in computer vision or ML fields using DL or hybrid models. Only journal publications and conference proceedings that are published in 2020-2025 in English and have undergone the peer-review process are eligible. Studies were required to present a clear methodological contribution, experimental evaluation, or performance analysis related to crowd density estimation. Research that reported quantitative results, comparative analysis, or evaluation metrics was considered eligible for inclusion. Both indoor and outdoor crowd environments were included, without restriction on the application domain. The studies that were not related to crowd density estimation or crowd analysis, or that lacked sufficient methodological detail, were excluded. Studies without experimental results were excluded unless they presented conceptual or framework-based contributions with clear and meaningful methodological insights. The non-peer-reviewed papers, review papers, editorials, abstracts, theses, or papers that were not in English were also excluded. Duplicate papers or papers containing overlaps were removed, keeping the most appropriate one.

C. Data Extraction

Extraction of data was done to ensure that relevant data from studies included in this review was gathered systematically. A data extraction form was created to ensure that a high level of accuracy was achieved during extraction of the data. Data extraction was done manually by two reviewers, once eligible studies had been finalized. For each study included, the essential information extracted was the details such as (authors and year of publication, type of publication: journal and/or conference proceedings, application field, datasets utilized, methodological paradigm: conventional, DL, or combination of both, model design, and principal experimental settings). Information on performance factors, namely evaluation criteria, level of accuracy, and computation time pertaining to real-time analysis, was also considered as part of the data extraction process. In order to ensure accuracy, the extracted information was checked for consistency. If different copies or overlapping reports for the same research study were found, only the most relevant copy was used for the analysis.

III. Public Datasets for Crowd Density Estimation

Crowd density estimation techniques are highly dependent on the availability of huge crowds with different levels of density for the purpose of benchmarking and comparing the methods used. Hence the availability of public crowd density estimation datasets plays a crucial role, enabling various comparative analyses. This section will introduce the popular crowd density estimation dataset, and then its detailed description will be provided.

A. Widely Used Datasets

Most crowd density estimation approaches have commonly relied on a set of well-established benchmark datasets, which have now almost become standard evaluation references in the literature. They are widely used due to their diversity in terms of crowd densities, challenging visual conditions, and standardized evaluation protocols. The UCF_CC_50 dataset is a cornerstone in the field of crowd density estimation [15]. The WorldExpo'10 dataset, is the largest so far for the comparison of cross-scene crowd counting algorithms [16]. The JHU-CROWD++ dataset has a massive number of unconstrained images and annotations in crowd counting and is one of the largest benchmarked to date [17]. It has been designed to address the challenges in its predecessors, and it has a huge number of 4372 images annotated with 1.51 million head annotations. The ShanghaiTech dataset, in [18], marks a leap forward in crowd counting. It was constructed in order to avoid the shortcomings of the previous datasets that had a lack of diversity in viewing angles, densities, and scenes. The SmartCity dataset is a significant scientific contribution by the researchers [19], it addresses a notable deficiency in existing databases that typically focus solely on dense crowds and outdoor scenes. The UCF-QNRF dataset marks a major step forward in dense crowd analysis literature [20]. The Beijing BRT dataset has made a great scientific contribution to intelligent transportation systems science [21]. The dataset is derived from video surveillance systems installed at Beijing's rapid transit bus stations. DroneRGBT is the first standard that focuses on crowd counting by drone with dual-mode (RGB +

Thermal) imaging, it allows for effective crowd surveillance even in complex lighting conditions like night or fog [22]. The NWPU-Crowd dataset represents the largest realistic crowd counting and positioning benchmark available to date [23].

B. Dataset Comparison and Analysis

Although the characteristics of individual crowd datasets are useful and offer insights into their corresponding crowd situations, comparison among these datasets is necessary for an analysis of their relative merits and demerits, and appropriateness for different models. This is especially true for crowd datasets, which may be sensitive to different factors, including the size of the dataset, density range, diversity of the crowd scenario, and quality of the annotation process, during training and testing stages. As a measure towards filling the current biases in crowd benchmarks, a comparison of popular crowd datasets is discussed in Table I, as well as Mall [24] and Hajjv2 [25] datasets. A sample of the crowd dataset is shown in Fig. 2.

IV. Methodologies for Crowd Density Estimation

This section presents the various ML and DL techniques used for crowd density estimation and the review of existing approaches for crowd density estimation.

A. Crowd Density Estimation ML and DL Techniques

The models used for crowd density estimation span a range of traditional ML techniques and modern DL architecture. Among the most popular:

- YOLOv3 [26], is a DL model commonly used for real-time object detection in surveillance footage.
- ResNet [27], especially with modified architectures like asymmetric convolutions, is favored for its strong feature extraction capabilities.
- CSRNet [28] is used in dense crowd counting, for its high accuracy though it is computationally heavy.
- Multi-Column Convolutional Neural Network (MCNN) [18] is another frequently used baseline that handles scale variations in crowds.
- SE-DenseNet [29] is applied for enhanced feature learning, often paired with attention mechanisms.
- DeepLabv3+ [30] is widely adopted for semantic segmentation, especially in video-based crowd mapping.
- Convolutional Long Short-Term Memory (ConvLSTM) [31] models are employed to handle temporal data by capturing spatial and time-based patterns.
- Self-Organizing Maps (SOM) [32] and Back Propagation (BP) neural networks are used in older or hybrid approaches, sometimes optimized with algorithms like genetic search.
- Lightweight models like LCDnet [33] and Single-Convolutional Neural Network with Three Layers (S-CNN3) [34] are designed for EC, balancing speed and accuracy with fewer parameters.
- Generative Adversarial Networks (GANs) [35] are used in self-learning setups to generate and refine crowd data, particularly for varying crowd densities.

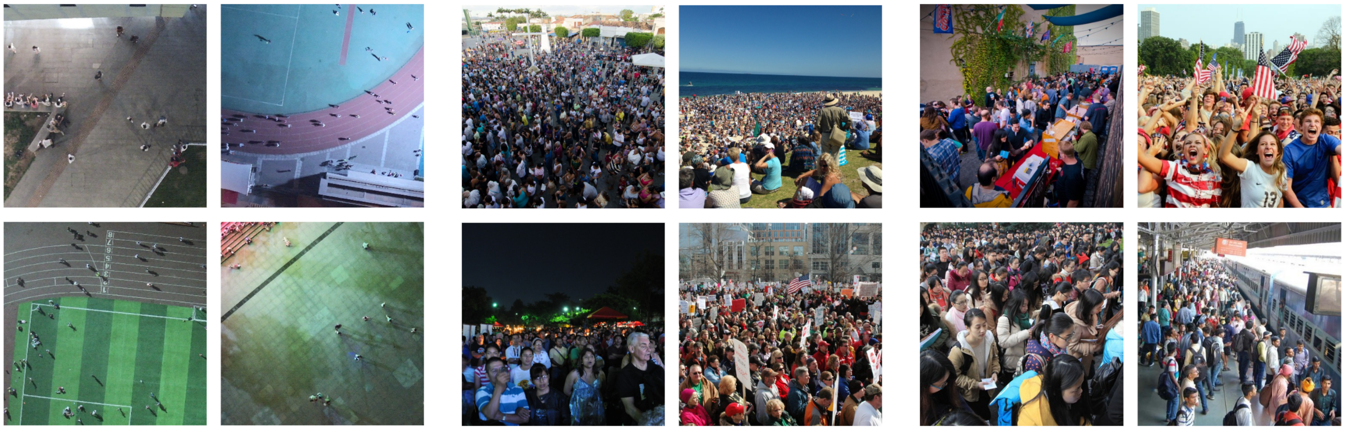
These models reflect a growing shift toward combining efficiency, accuracy, and adaptability in complex real-world scenarios.

B. Traditional Machine Learning Approaches

Almutairi et al in [36] aims to address the critical challenge of effectively organizing and managing crowded events during the COVID-19 and similar crises, especially given the failure of many previous gatherings to enforce necessary restrictions (such as social distancing and mask-wearing), making them hotbeds of infection. The methodology adopted is to propose a comprehensive and sustainable framework based on ML classification models. This framework is built on lessons learned from reviewing the organization of major events such as the Hajj and Kumbh Mela during the pandemic. The framework revolves around a detailed algorithm that begins by receiving an event request and calls a smart function to evaluate it. This function includes fetching historical and temporal location data (such as wireless sensor network readings and infection numbers) and applying preprocessing operations. Common ML models such as logistic regression and support vector machines are applied, and the accuracy of these models is compared with Threshold, which represents the minimum accuracy required for a model to be accepted. The researchers' findings present the framework and algorithm as a proposed solution that could be extremely useful in mitigating the spread of the virus. The main limitations are that true validation of this framework was not conducted in this paper due to the lack of actual data. Therefore, the detailed validation metrics presented in the paper are (schemes) for future validation when a suitable dataset becomes available.

Saxena et. al in [37] sought to investigate a ML-based assistance methodology, they used a framework that relies on managing and organizing the event from the proposal stage to implementation, taking into account various factors such as the event date, time, and location including the epidemic status in the specific area. Their methodology relies on data collection from various sensors, processing, and the application of ML models such as linear regression and decision trees. The results and conclusions demonstrate that crowd intelligence, through effective crowd management, can help protect individuals and save lives in public spaces, and that the proposed COVID-19 compliant framework can be adapted to manage crowds in various types of crises. Regarding limitations, one study indicated that there was no actual validation due to the unavailability of real data, and that the presented findings were merely recommendations for future validation.

Shah in [38] sought to develop a ML model to classify crowd density in Hajj video frames into three critical levels: moderate crowd, overcrowded and very dense crowd. The aim is to address the significant challenges facing managing large annual gatherings, ensuring public safety in critical areas such as Tawaf, and mitigating the risks of disasters such as stampedes and epidemics. The authors used the Hajjv2 dataset, which contains 18 videos from key Hajj sites such as Mina, Jamarat, Arafat, and Tawaf. They applied a Gradient Boosting Classifier (GBC)-based approach, focusing on extracting structured features including local binary pattern (LBP) texture analysis, edge density, and crowd coverage to



(a) DroneRGBT dataset samples

(b) UCF-QNRF dataset samples

(c) JHU-CROWD++ dataset samples

Fig. 2. Dataset samples.

TABLE I. Summary of Datasets Used in Literature

Dataset name	Ref& Year	Type	Size	Annotations	Average count	Resolution
Mall	[24],2012	Images	2000	62,325	31	320 × 240
UCF_CC_50	[15], 2013	Images	50	63,974	1279	Different
WorldExpo'10	[16] ,2015	Videos	1132	225,216	56	576 × 720
JHU-CROWD++	[17],2015	Images	4372	1,515,005	346	910 × 1430
ShanghaiTech Part_A	[18],2016	Images	482	241,677	501	Different
ShanghaiTech Part_B	[13,]2016	Images	716	88,488	124	768 × 1024
SmartCity	[19],2018	Images	50	–	7	1920 × 1080
UCF-QNRF	[20],2018	Images	1535	1,251,642	815	2013 × 2902
Beijing BRT	[21],2018	Images	1280	16,795	13	640 × 360
DroneRGBT	[22],2020	Images	3600	175,698	49	512 × 640
NWPU-Crowd	[23],2021	Images	5109	2,133,238	418	2311 × 3383
Hajjv2	[25],2023	Videos	18	300,541	–	–

enhance classification accuracy. The proposed model demonstrated an accuracy rate of 87%, achieving a very low error rate of 2.14%, demonstrating its high reliability, especially in identifying "very dense" crowd conditions. The system triggers visual alerts (a red flashing indicator) in real time when these conditions are detected. Limitations noted include misclassifications between "moderate" and "dense" crowds, the need to improve real-time processing performance through hardware acceleration or model augmentation, and the potential for performance bias due to the underrepresentation of some density levels in the dataset.

C. Deep Learning-Based Methods

Crowd density estimation has seen significant advancements through the application of DL techniques, particularly CNNs. These methods have been instrumental in enhancing accuracy and efficiency across various scenarios. For instance, Z. Zhang and X. Sun in [39] propose a CNN-based model for crowd density estimation and counting, aiming to achieve a large receptive field without excessive parameters. They improve ResNet model by adding two sets of asymmetric long convolutions to replace large kernels, reducing computational costs while maintaining performance. The model is tested on ShanghaiTech, UCF-QNRF, and UCF_CC_50 datasets, showing significant improvements: on ShanghaiTech Part A, mean absolute error (MAE) and mean squared error (MSE) score was 73.5 and 127.9, while on Part B, they were 7.7 and 12.4. For UCF-QNRF, errors reduced by 179.1 MAE and 247.1 MSE compared to MCNN, and for UCF_CC_50, by 122.5 MAE and 95.7 MSE. The results demonstrate better accuracy and stability than other methods, validated by density map visualizations and error metrics. However, the model's dependency on labelled data and limited generalization to extreme occlusions or dynamic scenes remain challenges. The work highlights the effectiveness of asymmetric convolutions for crowd counting but acknowledges the need for further optimization in complex real-world scenarios.

Ranasinghe et al. in [40], introduced a novel approach to crowd density estimation using diffusion models, which generate high-quality density maps by reversing a noising process. Unlike traditional methods that rely on broad Gaussian kernels and suffer from noise or lost details in crowded scenes, their method employs narrow kernels to preserve accuracy. The model also leverages the randomness of diffusion to produce multiple density map variations, combining them intelligently to improve counting performance. Experiments across six datasets show significant improvements, with CrowdDiff outperforming existing methods, for example, reducing the MAE on ShanghaiTech Part A dataset to 47.4 compared to the previous best of 49.3. However, the iterative nature of diffusion models makes the method slower than real-time alternatives, and performance can still be affected by challenging lighting or heavy occlusions. Despite these trade-offs, the approach demonstrates stronger noise resistance and better handling of dense crowds than earlier techniques.

Similarly, Ding et al. in [41], discussed a symmetric encoder-decoder CNN for crowd density estimation, focusing on multi-layer feature fusion to improve accuracy and generate high-quality density maps. They address key challenges like occlusion, perspective distortion, and complex backgrounds by combining shallow (spatial) and deep (semantic) features during encoding and decoding. Their model uses fixed Gaussian kernels for density map generation, simplifying implementation compared to adaptive kernels while maintaining performance. They introduce Patch Absolute Error (PAE), a new metric to evaluate density map quality beyond traditional MAE/MSE, emphasizing local count accuracy. Experiments on datasets like ShanghaiTech, WorldExpo'10, and UCF_CC_50 show state-of-the-art results, with MAE of 69.8 on ShanghaiTech Part A and 10.2 on Part B. The model excels in cross-scene and transfer learning achieving an MAE of 5.9 on SmartCity dataset. Limitations include sensitivity to extreme occlusions and reliance on fixed kernels, which may not adapt well to all scenarios.

Addressing the need for lightweight models, Alashban et al. [34] proposed a S-CNN3 model. They created a 3-layer CNN model to count people in crowds and estimate density, which helps manage events like Hajj or concerts to avoid accidents. They used the ShanghaiTech dataset, the largest crowd-counting dataset, with over 1000 images labeled with head positions. They tested two versions: one with 20 density classes e.g., 0-5 people, 6-10, etc. and another with 33 classes for finer details. The model achieved 99.88% accuracy and a very low loss of 0.02, exceeding deeper models like 4-layer CNNs and Switch-CNN. It works well because of its simple design and focus on classification instead of complex regression. However, it might fail in extreme crowds or new environments since it was only tested on ShanghaiTech data. Also, splitting images into patches could miss global crowd patterns.

Kamra et al., in [42], tested three DL models such as Mobilenet SSD, YOLOv4, and Mask Region-based (RCNN) using real surveillance videos with different crowd sizes. To get the data ready, they resized and flipped the images to make the models stronger during training. Mobilenet SSD was the fastest and worked well on low-power devices, but it wasn't very accurate in crowded scenes. Its accuracy ranged from 43% to 84% across different video frames. YOLOv4 gave the best overall balance between speed and accuracy, reaching up to 95% accuracy and staying reliable even when people were close together. Mask RCNN offered the most detailed results by identifying each person clearly and reached up to 92% accuracy, but it was slower and needed more computing power. In short, YOLOv4 turned out to be most useful for real-time crowd analysis. However, all models struggled in some conditions, like bad lighting or when people were blocked from view.

Addressing edge computing constraints, Wang et al., in [43], introduces a smart and efficient way to count people in crowded places using edge computing and a lightweight DL model. Instead of sending video data to cloud servers, the authors designed a special CNN that runs directly on small edge devices like Raspberry Pi. This helps save internet bandwidth, protects people's privacy, and gives faster results. In real-world testing at a subway station in Beijing, their system worked well on Raspberry Pi 4 using real camera footage. It successfully tracked changes in crowd levels during morning rush hours. However, the model still has some limitations, like needing better handling for extremely dense crowds and reducing computational time even more.

Deokate et al., in [44], focused on improving how we count and monitor people in crowded areas using DL. They compared three different versions of the YOLO model for detecting and tracking people in video footage: 1) YOLOv3 combined with a Canny Edge detector, 2) YOLOv3 with Hungarian Algorithm and Kalman filters, and 3) YOLOv5 with OpenCV. In terms of performance, YOLOv3 with the Canny Edge detector gave the best results with 86.8% accuracy and 75% precision, meaning it was the most accurate at identifying people and had fewer false detections. YOLOv5 followed with 83.3% accuracy and 76.6% precision, while the third model, YOLOv3 with Hungarian and Kalman filters, had the lowest results: 79.2% accuracy and 70.1% precision. Although the system worked well overall, it had some weaknesses. It struggled with poor lighting, was sensitive to noise, and sometimes lost details when resizing video frames.

Wang et al., in [45], aims to enable real-time crowd density estimation and efficiently deploy this task on edge devices with limited computational resources within the framework of edge intelligence, addressing the challenge of the excessive computational complexity of traditional DL models. To achieve this, the researchers proposed a lightweight CNN model. Its architecture leverages a modified MobileNetV2 as a backbone for efficient extraction of shallow-layer features, utilizing depth-wise separable convolutions to reduce model size, and incorporates dilated convolution layers as a backend to extract deeper features and preserve spatial information. The results demonstrate that the proposed model achieved significantly faster inference speed with only a slight decrease in accuracy compared to similar methods. Regarding limitations, the researchers noted that performance was slightly worse at high density levels compared to baseline models, and that further improvements in inference speed are needed for true real-time performance on IoT devices. Zou in [46] aims to address fundamental challenges in crowd density estimation, particularly in crowded scenes characterized by significant and continuous variations in pedestrian scale within a single image, and the loss of spatial detail during feature extraction. To overcome these issues, the researchers propose a novel model called the Multi-Scale Perception Network for Dense Crowds (MSPN-DC), which utilizes the first 10 layers of VGG-16 as a backbone for initial feature extraction. Experimental results demonstrate that the proposed method outperforms other state-of-the-art approaches on challenging benchmark datasets (such as UCF-CC-50 and UCF-QNRF), achieving the best performance on the NWPU dataset with its wide range of pedestrian densities, thus demonstrating its robustness and good stability. However, the researchers acknowledge a slight performance decline in MAE on the dense ShanghaiTech Part-A dataset compared to the best-performing methods, attributing this to the significant background clutter in this dataset

D. Hybrid Models

Incorporating geographic information systems (GIS), Zhang et al., in [47], presented a Crowd Density Estimation and Mapping Method based on Surveillance Video and GIS (CDEM-M), a method for crowd density estimation and mapping by integrating surveillance video with GIS. They first developed a Crowd Semantic Segmentation Model (CSSM) using DeepLabv3+ and a Crowd Denoising Model (CDM) via CNN, achieving 96.7% segmentation accuracy and 86.29% noise removal accuracy. Using a homography matrix, they project crowd areas from video to geographic space, accounting for perspective distortions ("near large, far small"). Key innovations include GIS integration for unified visualization and the use of whole-body features (not just heads) for better accuracy in high-altitude views. Limitations include dependency on fixed camera angles and challenges with individual-group mixed scenes. The approach is practical for large venues like stadiums but requires further refinement for broader applications.

In the context of smart cities, Mansouri et al., in [48], propose a DL system called Deep Convolutional Neural Network-based Crowd Density Monitoring for Intelligent Urban Planning (DCNNCDM-IUP) to monitor crowd density in smart cities using Closed-Circuit Television (CCTV) and IoT sensors. Their goal is to improve urban planning and public safety by analyzing crowd movements in real time. To optimize performance, they use Red Fox Optimization (RFO) to fine-tune hyperparameters and ConvLSTM network to classify crowd density into four levels e.g., dense, sparse by analyzing both spatial and temporal data. They tested the system on a dataset of 1,000 images, the model achieves 98.4% accuracy, outperforming methods like GoogleNet and VGGNet. It's also faster, processing images in 9.8 seconds versus 15–21 seconds for other models. However, it struggles with extremely crowded scenes and relies heavily on the training dataset, which may limit real-world adaptability.

Exploring self-learning algorithms, Zhang et al., in [49], propose a self-learning soft computing system to predict crowd density using IoT sensors and AI. Their goal is to prevent overcrowding in places like airports or tourist spots. The system uses GANs to analyze video data from smart cameras. The model was tested on datasets like Shanghai-Tech and UCF_CC_50, achieving a low MAE of 290, better than previous methods discussed in the paper, and 93 on Shanghai-Tech Part A. It also processes data faster e.g., 9.8 seconds per image. However, training the GANs takes a long time, and the system struggles with extremely dense crowds. The authors suggest improving real-time performance and hardware compatibility in future work.

Focusing on social distancing, Fitwi et al., in [50], introduced a new system called E-SEC, which helps estimate how far apart people are from each other and how crowded a space is, using only one camera. They use a single CCTV camera and apply a DL model called YOLOv3 to detect people in video frames. Ther system gives real-time feedback, including alerts if people are too close to each other. They tested it using real videos and public datasets, and the results showed over 99% accuracy when the camera were properly setup. However, one limitation is that the system needs a good camera angle, proper height, and clear line of sight and may not perform accurately if people are directly under the camera or blocked by others. Still, E-SEC offers a low-cost and accurate solution for monitoring social distancing and crowd behavior using just one camera.

Zhu et al., in [51], aims to address the problem of crowd density estimation and counting in both sparse and dense crowd scenarios, overcoming significant challenges such as occlusions, non-uniform density, and scale and perspective variations caused by cameras capturing images from different distances and angles. To solve these problems, the researchers propose a two-stage approach called Patch Scale Discriminant Regression Network (PSDR) coupled with a Person Classification Activation Map (CAM). Their results demonstrate that the proposed method (PSDR + CAM) outperforms state-of-the-art methods, especially in sparse/moderate density scenes, achieving an improvement on the SmartCity dataset by reducing the MAE error from 8.6 to 7.5. Regarding limitations, the researchers acknowledge that their method performed slightly worse in highly dense scenes (such as ShanghaiTech PartA and UCF_CC_50) compared to the best methods in that category.

Alzahrani and Algethami, in [52], aimed to address the challenges facing optimal Hajj crowd management. High crowd density within specific temporal and spatial constraints renders traditional methods insufficient for providing immediate alerts or predictive preventive measures. The authors sought to develop a ML-DL based system to enhance crowd management by detecting anomalies and predicting future conditions. To address this problem, an integrated framework was proposed that utilized the Hajjv2 dataset, which contains annotated video frames of crowd behavior. The methodology involved extracting key features such as crowd density, movement speed, direction, and object space, applying an Isolation Forest algorithm to detect anomalies in real time, and using an Long Short-Term Memory (LSTM) neural network to predict future crowd behavior trends. The outputs of the two models were then combined to create a hybrid alerting system that triggers immediate (anomalies) or proactive (forecast) warnings when the prediction error exceeds an empirically defined threshold (95th percentile). The results showed that the Isolation Forest model achieved an average accuracy of 91% in identifying abnormal movement patterns. The LSTM model also achieved reliable predictions of average crowd speed with a low mean square error (MSE) of 0.000439, demonstrating the proposed system's potential to enable proactive and informed crowd management strategies. Regarding limitations, the authors note that the system was tested on a simulated dataset Hajjv2 rather than in a live deployment. The alert system currently relies on empirical thresholds that may need to be dynamically adjusted in a real-world environment

The DCNNCDM-IUP in Fig. 3 and E-SEC in Fig. 4 models are considered the best based on the literature for combining high efficiency and superior accuracy. The DCNNCDM-IUP model comprises sequential technologies, including Gaussian filtering for image cleaning, SE-DenseNet for feature extraction, RFO for parameter optimization, and a ConvLSTM network for spatial and temporal density classification. Its advantage lies in its accuracy of up to 98.4% and its high speed (9.8 seconds), which surpasses standard models such as GoogleNet. The disadvantage, however, lies in its difficulty in handling scenes that are extremely crowded and its reliance

TABLE II. COMPARATIVE ANALYSIS OF RELATED STUDIES

Technique	Ref & Year	Models	Datasets	Key findings	Limitations
ML	[36],2022	Common ML classification models	There is no real data to verify	Based on standard metrics for future verification	No real validation
ML	[37],2022	Traditional ML approaches	N/A	The system can track entry and exit numbers and determine occupancy levels	Poor data quality and insufficient training data
ML	[38],2024	Gradient Boosting Classifier (GBC)	Hajjv2	Accuracy: 87%, Low misclassification rate: 2.14%	Misclassification between moderate and heavy crowds. Real-time performance needs hardware acceleration.
DL	[39],2020	Improved ResNet-based model	ShanghaiTech, UCF-QNRF, UCF_CC_50	MAE: 97.9, MSE: 178.9	Performance is not good in very dense scenes or complex backgrounds
DL	[40],2023	Conditional diffusion models for generating density maps	JHU-CROWD++, ShanghaiTech, UCF_CC_50, UCF-QNRF, NWPU-Crowd	MAE: 5.7	Higher inference time due to the iterative nature of diffusion models
DL	[41],2021	CNN	ShanghaiTech, WorldExpo'10, Mall, UCF_CC_50, SmartCity, Beijing BRT	MAE: 1.36, MSE: 2.02	Data labeling errors, especially in high-density areas
DL	[34],2022	S-CNN3	ShanghaiTech	Accuracy: 99.88%	The difficulty of classification increases dramatically as the number of classes increases
DL	[42],2024	Mobilenet SSD, Yolov4, Mask RCNN	N/A	Accuracy: 95%	Affected by environmental factors and blockages
DL	[43],2021	ResNet50	ShanghaiTech	MAE: 11, RMSE: 17.6	Resource intensive compared to the capabilities of the edge
DL	[44],2024	YOLOv3	ShanghaiTech, UCF_CC_50, WorldExpo'10	Accuracy: 86.8%, Precision 75%	Inadequate handling of occlusions, leading to underestimation of density
DL	[45],2022	MobileNetV2	ShanghaiTech	MAE: 10.1, RMSE: 17.1	Limited computing resources of peripheral devices
DL	[46],2025	VGG-16	UCF_CC_50, UCF-QNRF, ShanghaiTech	MAE: 156.6, RMSE: 223.4	Background interferences
Hybrid	[47],2023	CSSM (DeepLabv3),+ CDM (CNN), and GA-BP NN	Images	CSSM accuracy 96.70%,CDM accuracy 86.29%, BP network average error 1.2	Suitable only for high altitude viewing
Hybrid	[48],2025	SE-DenseNet + ConvLSTM	N/A	Accuracy 98.4%	Rely on a fixed group size and a fixed learning rate, not optimal for dynamic problems
Hybrid	[49],2021	SFL-GAN + SFL	Custom dataset, UCF_CC_50, ShanghaiTech	MAE: 93	Density maps are blurry and of poor quality when using traditional regression loss. The network takes a long time to train
Hybrid	[50],2021	Modified YOLOv3 + E-SEC	Custom dataset, COVID-19, BriefCam, PETS2009	Personal distance calculation accuracy of over 99%	Cameras must be mounted at a height of at least 3 meters
Hybrid	[51],2020	PSDR + CAM	SmartCity, ShanghaiTech, UCF_CC_50	MAE: 7.5, MSE: 10.2	Worse performance in dense scenes
Hybrid	[52],2025	Isolation Forest + LSTM	Hajjv2	Accuracy: 91%, MSE: 0.000439	Tested on simulated data (Hajjv2) not live deployment

on training data, which may limit its practicality. The E-SEC model consists of the YOLOv3 system for detecting people and the "Pixel Per Metric" technology for estimating distances based on the average width of the human body. It is characterized by exceptional accuracy exceeding 99% and a low cost, as it utilizes only one camera without the need for physical measuring tools. Its disadvantages are highlighted by its need for a specific camera angle and height, as well as a clear line of sight, since its performance weakens if people are directly below the camera or if there are obstacles that obstruct the view.

V. CHALLENGES, LIMITATIONS, AND FUTURE RESEARCH

The major challenges encountered in crowd destiny estimation are presented in this section, based on the literature review of Section IV and comparative analysis presented in Table II. Collectively, this research review emphasizes the diverse approaches and ongoing challenges in crowd density estimation, highlighting the balance between accuracy, computational efficiency, and adaptability to varying real-world scenarios. Common limitations of research on crowd management and density estimation using ML and DL primarily center around three core challenges such as data inadequacies and validation, poor robustness of algorithms to visual and environmental complexity, and real-time performance and computational efficiency challenges.

- Lack of data quality and real-time validation is a key limitation. Some proposed frameworks rely on simulated datasets such as Hajjv2 or data with limited representation of specific scenarios or density levels, limiting the generalizability of models and leading to potential classification bias. Poor data quality itself is also a challenge, as are the potential for data labelling errors, especially in densely populated areas.

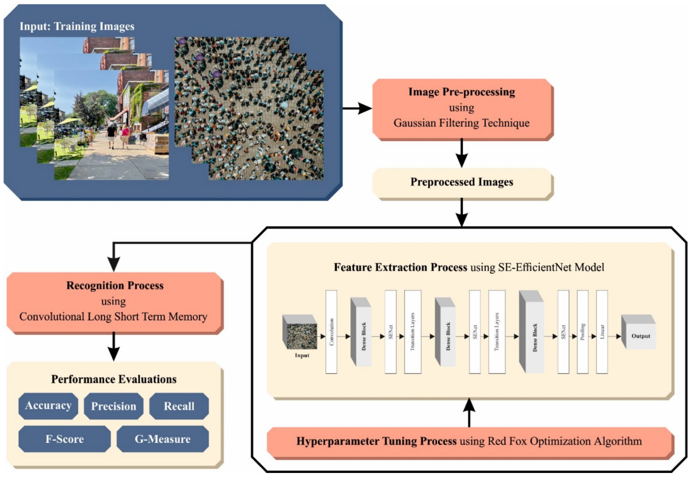


Fig. 3. The framework of DCNNCDM-IUP technique.

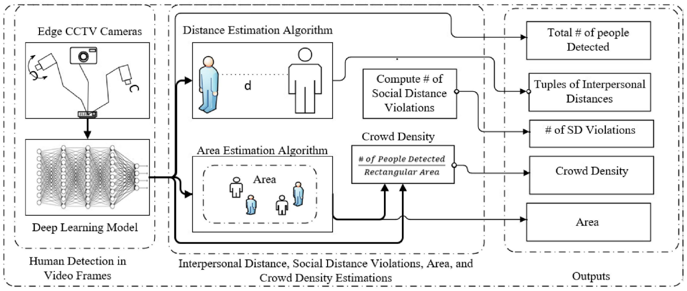


Fig. 4. The framework of E-SEC model.

- Visual and environmental challenges are the most common, with models failing to handle high-density scenes and facing severe occlusion, which often leads to underestimation of actual density. The constantly changing scale of the target and perspective effects also pose a significant challenge, requiring continuous processing to minimize the loss of spatial information, especially in high-altitude observation scenes where head features are less clear.
- Furthermore, the accuracy of algorithms is affected by environmental factors such as complex background interference, varying lighting changes, and image blur.
- Finally, performance and efficiency challenges arise, as many DL models are characterized by excessive computational complexity and a large number of parameters, which reduces real-time performance and causes unbearably high inference delay. This limitation is critical when attempting to deploy on-edge devices with limited storage and compute resources.

Furthermore, complex models such as GANs can be time-consuming during the training phase, and certain techniques such as dilatational convolution can lead to the gridding problem and loss of local information. Despite the significant advancements made by ML and DL techniques in crowd density estimation, the path to fully autonomous and intelligent monitoring systems remains fraught with fundamental challenges related to data quality and modelling methodologies.

A. *Challenges in Datasets*

Datasets are the cornerstone and essential engine that enables DL models to reason, but systematic examination of these datasets has revealed technical and organizational gaps that limit the models' ability to simulate complex realities. The most prominent of these challenges are as follows:

- Lack of diversity and comprehensiveness: The review reveals an over-reliance on standard datasets such as ShanghaiTech, which can lead to model biases toward specific environments. Current datasets lack diversity in scenarios, focusing either on multi-million-strong crowds or dispersed urban scenes, rendering models unable to adapt to sudden shifts in density within a single scene.
- The gap between simulation and reality: Some studies rely on simulation data (such as Hajjv2) to compensate for the lack of field data, but this creates a "realism gap". Virtually trained models fail to handle the complexities of real-world lighting or harsh weather conditions (such as fog and rain) available in assemblies like JHU-CROWD++.
- Challenges of annotation and human effort: Manual annotation remains a significant obstacle to building large assemblies. Assemblies like UCF-QNRF require thousands of hours of human labor. The likelihood of annotation errors also increases in densely populated areas, directly impacting the accuracy of density maps.
- Future direction: There is a pressing need to develop "cross-domain datasets" that combine optical and thermal imagery (such as DroneRGBT) to enable nighttime surveillance. Furthermore, "intelligent annotation" or semi-supervisory learning should be pursued to reduce reliance on arduous human effort.

B. *Challenges in ML and DL Models*

Despite rapid advancements in neural network models, crowd density estimation still faces fundamental technical and methodological challenges that prevent achieving absolute accuracy in dynamic and real-world environments. The most prominent of these challenges are as follows:

- Accuracy vs. Computational Efficiency Dilemma: The "critical balancing act" challenge arises between large-scale models (such as CSRNet) that achieve high accuracy but are computationally intensive, and lightweight models (such as LCDnet and MobileNetV2) designed for edge devices. Currently, the processing speed of devices like the Raspberry Pi remains a limitation for real-time emergency response.
- Optical Complexity and Severe Occlusion: The problem of "occlusion" and object overlap remains a major obstacle, as models tend to underestimate the actual numbers in very compact crowds. Furthermore, the "perspective effect" at high angles leads to the loss of fine vertices detail, which confounds traditional algorithms.

C. Future Research Directions

Research is moving towards "edge intelligence" to reduce reliance on cloud, focusing on attention mechanisms that enable models to distinguish humans from background "visual noise". Most current research remains confined to the "counting task" alone, lacking the ability to understand the context of movement. Issues such as detecting anomalous behaviors or predicting bottlenecks before disasters (as seen in the Mina) still require the integration of complex temporal and behavioral data. Integrating density estimation with spatial analysis (GIS) and trajectory prediction systems (LSTM) to provide proactive insights for crowd management, rather than simply monitoring cold, hard numbers. The future lies in creating an integrated ecosystem that links realistic and diverse datasets with hybrid algorithms that combine the accuracy of DL with the logic of behavioral analysis, to ensure the highest levels of public safety at major events

VI. Conclusion

This systematic review of the PRISMA 2020 guidelines aims to analyze and classify contemporary methodologies for estimating crowd density and developing intelligent management systems that prevent tragic disasters, such as the 2015 Mina stampede and the 2021 Texas incident. By examining 20 peer-reviewed studies published between 2020 and 2025, it was found that there is a significant reliance on standard datasets for model evaluation, with the ShanghaiTech dataset leading the way, used in 50% of the studies, followed by the UCF_CC_50, NWPU-Crowd, and UCF-QNRF datasets, which provide a variety of viewing angles and numerical densities. The results revealed the dominance of the DL models at 55% and a trend toward hybrid models at 30%. Software architectures, such as S-CNN3, achieved an exceptional accuracy of 99.88%, while lightweight models such as LCDnet and MobileNetV2, designed specifically for edge computing, emerged to minimize latency and provide real-time monitoring on resource-limited devices.

Beyond the theoretical synthesis, the findings of this review carry significant practical implications for real-world crowd management and urban safety. For practitioners and engineers, the transition toward hybrid models and edge intelligence, as identified in our analysis, provides a technical baseline for developing low-latency surveillance systems capable of immediate intervention in high-density zones. For policymakers and urban planners, the shift from reactive counting to proactive density estimation supported by digital twins and real-time AI offers a strategic framework for designing safer public spaces and optimizing resource allocation during mass gatherings. Furthermore, by addressing the gap between benchmark accuracy and field reliability, this research provides the academic community with a clear roadmap for developing more robust, context-aware architectures. Ultimately, these insights serve as a catalyst for advancing smart city initiatives where automated, scalable crowd monitoring becomes a cornerstone of public safety and operational efficiency.

Despite this technological advancement, significant challenges remain. Models often underperform under severe occlusion and complex environmental conditions, such as fluctuating lighting and fog, frequently resulting in underestimating numbers in very dense crowds. Current research also suffers from a "realism gap" due to an over-reliance on simulation data (such as Hajjv2) or static images from the internet, which limits the models' ability to generalize to dynamic field situations. Therefore, the review concludes that future trends should go beyond mere digital counting to include behavioral analysis and proactive detection of abnormal patterns, with the need to develop more comprehensive datasets and expand the use of attention mechanisms and self-learning to ensure the accuracy of density maps and the sustainability of public safety systems in smart cities.

References

[1] Asif, N. Crowd management–navigating challenges and implementing best practices in crowd management. *Available at SSRN 4655567*, (2023).

[2] Alasmari, A.M., Farooqi, N.S. & Alotaibi, Y.A. Recent trends in crowd management using deep learning techniques: a systematic literature review. *Journal of Umm Al-Qura University for Engineering and Architecture*, 1–29, (2024). Springer.

[3] Ganjeh, M. & Einollahi, B. Mass Fatalities in Hajj in 2015. *Trauma Monthly*, **21** (5) e43253, (2016).

[4] News, B. Travis Scott's Astroworld: Eight killed after crowd surge at Texas festival. *BBC News*, (2021).

[5] Khan, M.A., Menouar, H. & Hamila, R. LCDnet: a lightweight crowd density estimation model for real-time video surveillance. *Journal of Real-Time Image Processing*, **20** (2) 29, (2023). Springer.

[6] Al-Nabhan, N., Alenazi, S., Alquwaifili, S., Alzamzami, S., Altwayan, L., Alaloula, N., Alowaini, R. & Al Islam, A.A. An intelligent IoT approach for analyzing and managing crowds. *IEEE Access*, **9** 104874–104886, (2021). IEEE.

[7] Patwal, A., Diwakar, M., Tripathi, V. & Singh, P. Crowd counting analysis using deep learning: A critical review. *Procedia Computer Science*, **218** 2448–2458, (2023). Elsevier.

[8] Shi, W., Cao, J., Zhang, Q., Li, Y. & Xu, L. Edge computing: Vision and challenges. *IEEE internet of things journal*, **3** (5) 637–646, (2016). IEEE.

[9] Lorin, S. & others Digital Twins and Data Analysis for Crowd Management in High-Capacity Stations. *13th World Congress on Railway Research, Birmingham*, 06–10, (2022).

[10] Gao, G., Gao, J., Liu, Q., Wang, Q. & Wang, Y. Cnn-based density estimation and crowd counting: A survey. *arXiv preprint arXiv:2003.12783*, (2020).

[11] Hassen, K.B.A., Machado, J.J. & Tavares, J.M.R. Convolutional neural networks and heuristic methods for crowd counting: A systematic review. *Sensors*, **22** (14) 5286, (2022). MDPI.

[12] Gouiaa, R., Akhloufi, M.A. & Shahbazi, M. Advances in convolution neural networks based crowd counting and density estimation. *Big Data and Cognitive Computing*, **5** (4) 50, (2021). MDPI.

[13] Sindagi, V.A. & Patel, V.M. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters*, **107** 3–16, (2018). Elsevier.

[14] Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Group, P. & others Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *International journal of surgery*, **8** (5) 336–341, (2010). Elsevier.

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

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

[17] Sindagi, V.A., Yasarla, R. & Patel, V.M. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE transactions on pattern analysis and machine intelligence*, **44** (5) 2594–2609, (2020). IEEE.

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

[19] Zhang, L., Shi, M. & Chen, Q. Crowd counting via scale-adaptive convolutional neural network. *2018 IEEE winter conference on applications of computer vision (WACV)*, 1113–1121, (2018). IEEE.

[20] Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N. & Shah, M. Composition loss for counting, density map estimation and localization in dense crowds. *Proceedings of the European conference on computer vision (ECCV)*, 532–546, (2018).

[21] Ding, X., Lin, Z., He, F., Wang, Y. & Huang, Y. A deeply-recursive convolutional network for crowd counting. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1942–1946, (2018). IEEE.

[22] Peng, T., Li, Q. & Zhu, P. Rgb-t crowd counting from drone: A benchmark and mmccn network. *Proceedings of the Asian conference on computer vision*, (2020).

[23] Wang, Q., Gao, J., Lin, W. & Li, X. NWPU-crowd: A large-scale benchmark for crowd counting and localization. *IEEE transactions on pattern analysis and machine intelligence*, **43** (6) 2141–2149, (2020). IEEE.

[24] Chen, K., Loy, C.C., Gong, S. & Xiang, T. Feature mining for localised crowd counting.. *Bmvc*, **1** (2) 3, (2012).

[25] Alafif, T., Hadi, A., Allahyani, M., Alzahrani, B., Alhothali, A., Alotaibi, R. & Barnawi, A. Hybrid classifiers for spatio-temporal abnormal behavior detection, tracking, and recognition in massive Hajj crowds. *Electronics*, **12** (5) 1165, (2023). MDPI.

[26] Redmon, J. & Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, (2018).

[27] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778, (2016).

[28] Li, Y., Zhang, X. & Chen, D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1091–1100, (2018).

[29] Jiang, M., Feng, C., Fang, X., Huang, Q., Zhang, C. & Shi, X. Rice disease identification method based on attention mechanism and deep dense network. *Electronics*, **12** (3) 508, (2023). MDPI.

[30] Chen, L., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European conference on computer vision (ECCV)*, 801–818, (2018).

[31] Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W. & Woo, W. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, **28**, (2015).

[32] Asan, U. & Ercan, S. An introduction to self-organizing maps. *Computational intelligence systems in industrial engineering: With recent theory and applications*, 295–315, (2012). Springer.

[33] Cattaneo, D., Vaghi, M. & Valada, A. Lcdnet: Deep loop closure detection and point cloud registration for lidar slam. *IEEE Transactions on Robotics*, **38** (4) 2074–2093, (2022). IEEE.

[34] Alashban, A., Alsadan, A., Alhussainan, N.F. & Ouni, R. Single convolutional neural network with three layers model for crowd density estimation. *IEEE Access*, **10** 63823–63833, (2022). IEEE.

[35] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, **27**, (2014).

[36] Almutairi, M.M., Yamin, M., Halikias, G. & Abi Sen, A.A. A framework for crowd management during COVID-19 with artificial intelligence. *Sustainability*, **14** (1) 303, (2021). MDPI.

[37] Saxena, D., Kumar, S., Tyagi, P.K., Singh, A., Pant, B. & Dornadula, V.H.R. Automatic Assisstance System Based on Machine Learning for Effective Crowd Management. *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 01–06, (2022). IEEE.

[38] Shah, A.A. A Machine Learning Model for Crowd Density Classification in Hajj Video Frames. *arXiv preprint arXiv:2501.04911*, (2025).

[39] Zhang, Z. & Sun, X. Crowd Density Estimation Based on Convolutional Neural Network. *2020 IEEE 20th International Conference on Communication Technology (ICCT)*, 1509–1513, (2020). IEEE.

[40] Ranasinghe, Y., Nair, N.G., Bandara, W.G.C. & Patel, V.M. CrowdDiff: Multi-hypothesis crowd density estimation using diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12809–12819, (2024).

[41] Ding, X., He, F., Lin, Z., Wang, Y., Guo, H. & Huang, Y. Crowd density estimation using fusion of multi-layer features. *IEEE Transactions on Intelligent Transportation Systems*, **22** (8) 4776–4787, (2020). IEEE.

[42] Kamra, V., Vaishnav, A., Verma, A., Khan, R. & Singh, S. A Novel Approach for Crowd Analysis and Density Estimation by Using Machine Learning Techniques. *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*, 1–6, (2024). IEEE.

[43] Wang, S., Pu, Z., Li, Q., Guo, Y. & Li, M. Edge computing-enabled crowd density estimation based on lightweight convolutional neural network. *2021 IEEE International Smart Cities Conference (ISC2)*, 1–7, (2021). IEEE.

[44] Deokate, S., Rajput, S., Nair, M., Parale, D., Mahamuni, A. & Nalla, P. Deep Learning-Based Crowd Surveillance and Density Estimation. *2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA)*, 1–8, (2024). IEEE.

[45] Wang, S., Pu, Z., Li, Q. & Wang, Y. Estimating crowd density with edge intelligence based on lightweight convolutional neural networks. *Expert Systems with Applications*, **206** 117823, (2022). Elsevier.

[46] Zou, L. Crowd Density Estimation Based on Multi-scale Feature Fusion and Information Enhancement. *IJLAI Transactions on Science and Engineering*, **3** (3) 1–11, (2025).

[47] Zhang, X., Sun, Y., Li, Q., Li, X. & Shi, X. Crowd density estimation and mapping method based on surveillance video and GIS. *ISPRS International Journal of Geo-Information*, **12** (2) 56, (2023). MDPI.

[48] Mansouri, W., Alohali, M.A., Alqahtani, H., Alruwais, N., Alshammeri, M. & Mahmud, A. Deep convolutional neural network-based enhanced crowd density monitoring for intelligent urban planning on smart cities. *Scientific Reports*, **15** (1) 5759, (2025). Nature Publishing Group UK London.

[49] Zhang, T., Yuan, J., Chen, Y. & Jia, W. Self-learning soft computing algorithms for prediction machines of estimating crowd density. *Applied Soft Computing*, **105** 107240, (2021). Elsevier.

[50] Fitwi, A., Chen, Y., Sun, H. & Harrod, R. Estimating interpersonal distance and crowd density with a single-edge camera. *Computers*, **10** (11) 143, (2021). MDPI.

[51] Zhu, L., Li, C., Yang, Z., Yuan, K. & Wang, S. Crowd density estimation based on classification activation map and patch density level. *Neural Computing and Applications*, **32** (9) 5105–5116, (2020). Springer.

[52] Alzahrani, R. & Algethami, N. Leveraging Machine Learning for Optimal Pilgrim Crowd Management. *Electronics*, **14** (13) 2507, (2025). MDPI.