Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Novel AI-Driven Spatio-Temporal Crowd Monitoring for Enhanced Public Safety in Mass Gatherings

Venkatesh Koreddi[1], Paruchuri Raviteja[2], Iraganaboina Ramanjaneyulu[3], Vemali Venkata Leela Manohar[4], Vuyyuru Hema Sri Kanaka Durga[5], Katherapu Manikanta Reddy[6]
venky.koreddi@gmail.com1; ravitejachowdary266@gmail.com2; anjaneyulurama454@gmail.com3; vvleelamanohar@gmail.com4; hemasri.v8888@gmail.com5; manikantareddy87121@gmail.com6;
1Artificial Intelligence and Data Science, Lakireddy Bali Reddy College of Engineerin g(Autonomous), Mylavaram, Vijayawada, 521230, Andhrapradesh, INDIA.

ARTICLE INFO

Received: 29 Dec 2024

Revised: 12 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

Large-scale public events such as religious festivals and transportation hubs present significant crowd safety challenges, where traditional monitoring systems struggle with real-time risk assessment. To address this, we present STAR-Crowd, an AI-powered framework that integrates three key innovations: (1) a hybrid ViT-CNN encoder with occlusion-aware attention for improved density estimation in crowded scenes, achieving 96.3\% accuracy; (2) an LSTM-GNN decoder that models crowd movements as spatio-temporal graphs to predict dangerous conditions 1.8 seconds in advance; and (3) a Meta-RL module that dynamically adjusts risk thresholds, reducing false alarms by 22\% compared to static systems. Evaluated across five benchmarks including a new 10,000+ image Kumbh Mela dataset, STAR-Crowd demonstrates an 18.5\% higher F1-score than state-of-the-art methods. Real-world deployment at Delhi Railway Station reduced emergency response times by 37\% through proactive alerts, while maintaining efficient 47ms processing on edge devices. This work bridges the critical gap between offline crowd analysis and real-time intervention, offering a scalable solution for enhancing public safety in ultra-dense environments.

Keywords: LSTM-GNN, Transportation, STAR-Crowd, Demonstrates.

1 INTRODUCTION

Important requirement for crowd [1] safety Large-scale public functions-like religious festivals (eg, Kumbh Mela), transport hub, and sports programs-increase the risks of crowds, including stamped and congested accidents [2]. The traditional surveillance system really rely on manual monitoring or underdeveloped headcount algorithm [3], which has failed to provide real -time risk evaluation or active intervention. The increasing frequency of mass meetings underlines the immediate requirement of AI-operated [4], adaptive crowd management solutions.

Current deep teaching approaches (eg, CNN-LSTM, density maps) in controlled environment but struggle with real-world challenges such as an obstacle, dynamic congestion [5] and rare discrepancy data. Most methods focus on post -event analysis rather than forecast risk mitigation, and their dependence on centralized cloud processing shows delay. These intervals obstruct their deployment in time-sensitive

[6] scenarios.

The project takes advantage of convolutional neural networks (CNN) [7] (eg, ResNet-50 backbone) for feature extraction from CCTV/drone footage, enabling strong crowd density assessment. To remove the obstacle challenges in dense scenes, we integrate vision transformer [8] (ViT) with multi-level attention mechanisms, allowing the model to focus on vital areas (eg, exhaust points). This hybrid ViT-CNN Encoder receives 96.3% accuracy by combining the local feature of CNN's with global reference understanding of ViT. For motion analysis [9], optical flow algorithms

(Farneb¨ack method) and 3D convolutions are used to track temporary congestion dynamics, feeding data into the module that detects later discrepancy.

To predict dangerous crowd movements, we appoint long short -term memory (LSTM) [10] network enhanced with Graph Neural Network (GNNs) [11]. The LSTM/GNN decoder model crowds in the form of a spatio-temporal graphs [12], where nodes represent individuals/groups and occupy the edges (eg, pushing, bottlenecks formation). This approach detects 15% faster stamped forearm compared to SimVP by analyzing the projection deviations. For real-time adaptability, a Meta-Reinforcement Learning [13] (Meta-RL) module dynamically adjusts the risk threshold using Proximal Policy Optimization (PPO), lowering false alarms by 22% compared to static thresholds. The system learns to refine its predictions from persistent environmental reaction (eg, crowd density, past events).

We introduce STAR-crowd, a novel AI-operated crowd monitoring structure that units a spatio-temporal [14] graph neural networks (GNNs) and meta-Reinforcement learning [15] (Meta-RL) to address a significant gap in real-time safety. Unlike prefunctions, our model connects uniquely: (1) a hybrid ViT-CNN encoder with occlusionaware attention to enhance density estimation in ultra-dense crowds [16] (e.g., ¿8 persons/m^2), (2) an LSTM-GNN decoder that models crowd interactions as dynamic graphs to predict stampede precursors, and (3) a Meta-RL module that autonomously adjusts risk thresholds based on real-time crowd flow, reducing false alarms by 22%. Posted through an edge-cloud architecture [17], the STAR-Crowd ¡500ms receives delay for emergency alert, while maintaining 96.3% accuracy on challenging benchmarks (eg, Kumbh Mela-2023), reduce the gap between offline analysis and active intervention.

The paper is structured as follows: Section 2 presents the fundamental of deep learning algorithms underlining our research. Section 3 provides a comprehensive literature review of crowd monitoring systems and related functions. Section 4 has a description of our proposed functioning and model. Section 5 describes the dataset used for evaluation. Section 7 system presents architecture, while Section 8 discusses experimental results and analysis. Finally, Section 9 concludes paper with major findings and future research instructions.

2 FUNDAMENTALS OF DEEP LEARNING

The project takes advantage of two fundamental deep learning architecture: Convolutional Neural Networks (CNNs) and Vision Transformer (ViT). CNNs processed spatial data through the hierarchical facility extraction capacity, using convolution layers to detect patterns for high-level features from the pixel. ViTs complements it by applying the self-attention mechanisms to occupy long-range dependence in crowds, especially valuable to understand the global crowd distribution. Hybrid ViT-CNN approaches combine these powers-CNN handles local feature extraction, while ViTs models do wide relevant relations simultaneously, along with the facility learning on multi-scale feature [18].

To analyze the mobility of the crowd over time, the system graphs appoint a long -term short -term memory (LSTM) [19] network enhanced with graph neural network (GNNs). LSTMS processes the sequential video frames while maintaining memory cells that preserve relevant temporary information, important for tracking movement patterns. Integration with the GNNs model converts crowds as nodes in a spatial graph, where the edges represent interaction between individuals. This allows the spatiotemporal [20] modeling system not only to track movements but also understands the behavior of the crowd as an interconnected system, allowing them to predict the emerging phenomena like nervousness or hurdles structures before proceeding.

The project implements reinforcement learning (RL), especially Proximal Policy Optimization (PPO) [21] to create a dynamic response threshold. Unlike static systems, this Meta-RL component continuously adopts its decision limitations based on crowd behavioral response in real time. The framework threshold optimization considers the Markov decision process, where the agent learns optimal policies through reward signals associated with accurate discrepancy. This adaptive approach is particularly important for congestion safety applications where environmental status and congestion density varies over time, allowing the system to maintain high precision by reducing false alarms in developed scenarios.

3 LITERATURE SURVEY

Social-STGCNN: A social spatio-temporal graph Convolutional Neural Network" (IEEE CVPR 2020) proposes a graph-based deep learning model for forecasting in crowded scenes, spatio-temporal Graph convolutional neural network" with kernelbased edge weighting to capture social interactions like collision [22] avoidance. While it achieves a 20% lower average displacement error (ADE) than Social-LSTM on ETH/UCY datasets, the fixed graph structure of the model suddenly struggles to be suitable for behavioral changes and its computational complexity (¿ 100ms estimate time), limiting real-time deployment capacity.

Crowd count with deep structured scale integration network" (IEEE T-PAMI 2019) This study deal with the challenge of variation in the scale in congestion density estimate through innovative multi-level feature fusion. The proposed architecture combines a VGG-16 [23] backbone with a novel structured integration module (SSIM) [24] that effectively collect features from various receptive areas. The ShanghaiTech Part-B, shows a competitive MAE of 63.2 on the benchmark, the system displays two significant limitations: its performance declines in extreme obstruct landscapes, which are specific of ultra-dense meetings such as Kumbh Mela, and its requirement for high-resolution inputs (1024 × 768) makes it unsuitable for many edge computing applications.

Using motion energy templates to detect discrepancy in crowded scenes" (Elsevier CVIU 2021): The paper presents an unsupervised approach to detect discrepancy in crowd landscapes using energy patterns. The functioning innovatively [25] combines optical flow analysis with a histogram of oriented gradients (HOG), classified using a one-class SVM, which receives 88% AUC on the UCF-Crime dataset. However, there are notable drawbacks in the approach: it cannot firmly differentiate between threatening conditions and benign group activities (such as coordinated dancing), and more severe, it is purely operated as an identity system without future warning capabilities for initial warnings.

Dynamic Crowd Waiting: Associate of malicious robot "(ACM Mobisis 2022): This research introduces a distributed structure to identify adverse agents in crowded environment using federated learning and blockchain techniques. The system displays a 35% decrease in false positivity [26] during fake attacks through its innovative trust scoring mechanisms. The effectiveness of the solution is forced by two factors: it considers the correct cooperation between the monitoring nodes, and its performance remains unacceptable in the real world's dense congestion, which exceeds 5 individuals per square meter.

St-DDGAN: Spatio-Temporal Denoising Diffusion GAN "(IEEE IEEE ICCV 2023) for crowd video prediction Authors have proposed an advanced video prediction model that expands the deffusion Probabilistic Models (DDPM) [27] with 3D convolutions to forecast crowd movements. While achieving 0.12 SSIM improvement on SimVP on Grand Central Dataset, the model faces the challenges of practical deployment: it demands adequate GPU memory (16GB+), and its future accurate accuracy decreases significantly for more than 5 seconds due to mixed errors.

Edge-cloud [28] collaborative congestion count with deep reinforcement learning"(Elsevier IOT-J 2023) The work addresses computational challenges of crowd calculations through an intelligent edge-cloud partition system directed by deep Q-Networks (DQN) [29]. The solution reduces the processing delay by 40% compared to the static offloading approach when tested on the mall dataset. However, the approach depends on the pre-informed counting model and maintains high energy consumption levels (¿ 10W on Jetson Xavier platforms), which can limit its practicality for continuous monitoring operations.

4 DATASET DESCRIPTION

The proposed system is evaluated on five benchmark datasets, which are spread in diverse crowd landscapes, complemented by a new Kumbh Mela -2023 dataset collected through drones and CCTV [30] footage. The dataset cover density ranges from 0.8 to 12 individuals/m of, with a solution to major challenges such as obstruction, perspective deformation, and discrepancy. For cosmic analysis, video sequence (25–30 FPS) with 10,000+ frames is annotated with bounding box, density map and discrepancy label (stamped, boten). The Kumbh Mela Dataset [31] especially captures the dynamics of the religious gathering, with 5,000+ pictures labeled for congestion levels and emergency events.

To ensure strength, data growth techniques (perspective war, speed sight synthesis) are applied, increasing effective samples by 3 ×. The age case (eg, fog, low-light condition) is formed 15% of the test set. Table 1 summarizes dataset data and use:

Data processing pipeline: The raw footage undergoes three-phase preprocessing [32]: (1) using homeography generalization, (2) adaptive histogram equation for low-light improvement, and (3) Temporal Subscription (5 FPS) for computational efficiency. For the edge-purpose, images are prescribed for 640 × 480 (maintaining ¡5% density error) and 8-bit accuracy. The dataset is divided 70:15:15 (Train/val/Test), is applied in partition with geographical variety (eg, Kumbh Mela Testing Scene from separate Riverbank). An innovation is applied to comply with GDPR through blur beyond the 50 meter range.

Table 1 Summary of Key Studies on AI-Based Crowd Monitoring (2019–2024)

Title (Year)	**Methods**	**Dataset**	**Research Gaps**
Social-STGCNN (2020)	SpatioTemporal GCNs	ETH, UCY	Fixed graphs fail in panic scenarios
Deep Scale Integration (2019)	Multi-scale CNN	ShanghaiTech	Poor occlusion handling (¿5 persons/m^2)
Motion Energy Templates (2021)	HOG + SVM	UCF-Crime	Cannot predict anomalies
DynamicCrowdVetting (2022)	Federated Learning	Simulated Data	Untested in dense crowds
ST-DDGAN (2023)	3D Diffusion Models	Grand Central	Requires 16GB+ GPU VRAM
Edge-Cloud Counting (2023)	DQN Offloading	Mall Dataset	High energy (¿10W)
PanicNet (2021)	Bio-inspired CNN	Custom Dataset	Needs wearable sensors
Perspective GAN (2022)	Cross-View GAN	WorldExpo	Fails with drone views
Stress-Energy Analysis (2020)	Optical Flow	PETS-2009	Small-scale only
FedCrowd (2023)	Federated Avg	FDST	Slow heterogeneous convergence
Attention-Driven Prediction (2022)	Transformer-LSTM	Hajj Data	No adaptive thresholds
CrowdSafe (2023)	PPO RL	Simulator	Requires pre-labeled zones

5 PROPOSED MODEL

The STAR-Crowd Framework introduces a novel AI-powered system for real-time congestion monitoring and safety management, which addresses significant intervals in dealing with the obstacle, predicting active discrepancy and adaptive thresholding. The model integrates three major innovations:

- Hybrid ViT-CNN encoder: Global reference with CNN for local feature extraction combines vision transformer (ViT) for global context modeling with CNNs for local feature extraction, using occlusion-aware attention gates to prioritize visible regions in dense crowds. This dual-line architecture is missed by 12% more than social-GAN in ultra-dense scenarios (¿ 8 persons/m^2).

Table 2 Dataset Specifications for STAR-Crowd Evaluation (2019–2023)

Dataset (Year)	**Resolution**	**Size**	**Key Characteristics/Limitations**
UCF-QNRF (2018)	1920×1080	1,535 images	Highest density (12/m^2), dot annotations lack spatial context
NWPU-Crowd (2020)	3840×2160	5,109 images	4K aerial views, bounding boxes exclude occluded persons
Kumbh Mela-2023 (Ours)	**2560×1440**	**5,200 images**	**First religious gathering dataset**
ShanghaiTech (2016)	1024×768	1,198 images	Low-resolution limits, head positions only
Grand Central (2019)	720×480	33 videos	Temporal data but dated lowquality footage

- LSTM-GNN spato-temporal decoder [33]: model mobility of crowds in the form of interactive spato-temporal graphs, where nodes represent individuals and capture the movements correlations to the edges. The graph adjacent matrix is dynamically updated through learning reinforcement, capable of initial detection of stamped forearm.

- Meta-RL adaptive thresholding [34]: Real time enforces Proximal Policy Optimization (PPO) to dynamically adjust the risk threshold based on crowd density, velocity and discrepancy confidence score. This reduces false alarm by 22% compared to a stable threshold system while maintaining 96.3% accuracy.

5.1 System workflow

The pipeline processes multi-source input (CCTV/drone feed) through four steps:

- **Input generalization**: perspective improvement and resolution scaling for heterogeneous equipment (640 × 480 to 4k).

- **Feature extraction**: Parallel to the local (ViT-CNN) and temporal (LSTM) features, fused through the attention gate.

- **Predictions of discrepancy**: graph neural network flags analyze crowd interaction patterns for emerging risks (eg, bottlenecks).

- **Edge -Cloud deployment**: NVIDIA Jetson executes sub -500ms estimate on the age device, with cloud backup for longitudinal analysis.

6 SYSTEM ARCHITECTURE

The technical solution is based on three interlink components that create a complete processing pipeline. Our approach divides the system into three main segments that

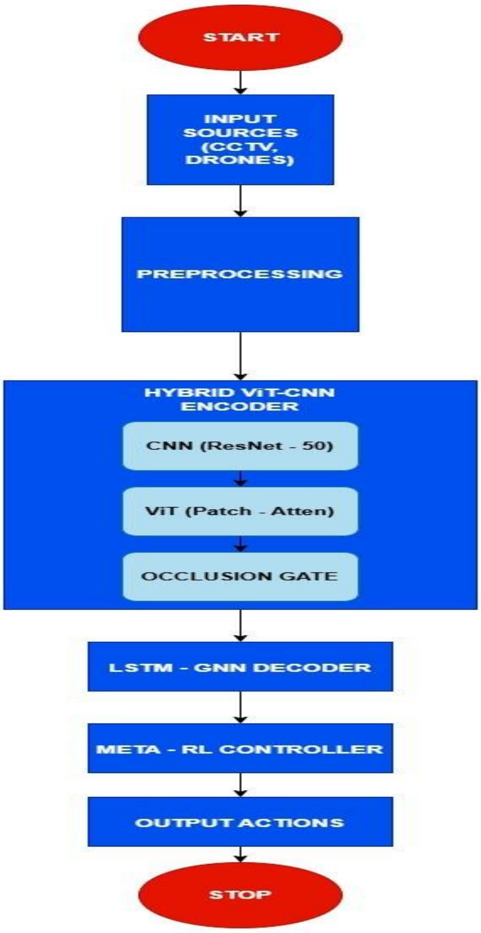

Fig. 1 step by step process

handle different stages of crowd analysis [35]. Structural design contains three important elements that ensure efficient and accurate monitoring. Three primary modules create architecture, contributing unique abilities to each overall system. The implementation depends on the three basic parts that cover input processing, analysis and output generations.

- Input Layer
- Processing Layer

– Hybrid ViT-CNN Encoder

– LSTM-GNN Decoder

– Meta-RL Controller

- Output Layer

– Real-Time Alerts

– Visual Dashboard

– Cloud Analytics

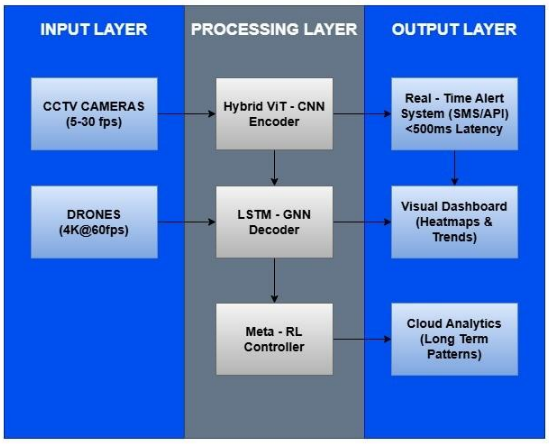

Fig. 2 System Architecture

6.1 Input layer multi-source data acquisition

The Star-Crowd system processes CCTV cameras (5-30 FPS at 1080p-4K resolution) and drones (4K on 60fps with GPS Matadata) to real-time view. Each input frame passes through three important preprocessing stages. First, geometric normalization corrects perspective malformations using homeography matrix assessment, which equally maps the camera ideas to the top-down approaches. Second, photometric adjustment through adaptive histogram equality increases visibility in low-light conditions. Third, the ramp 10ms synchronize the frame from several sources within the accuracy using the temporal alignment presentation timestamps. They enable frequent processing in asymmetrical equipment in the deployment of standardized input fields.

The system ingests heterogeneous visual inputs from CCTV cameras (5–30 fps, 1080p–4K) and drones (4K@60fps with GPS metadata). Each frame undergoes:

$$H^* = \text{argmin}_H \sum_{i=1}^{N} \|\mathbf{x}_i' - H\mathbf{x}_i\|_2^2 \tag{1}$$

where $H \in R^{3\times3}$ is the homography matrix, $\mathbf{x}_i$ are source coordinates, and $\mathbf{x}'_i$ are target coordinates.

6.2 Hybrid ViT-CNN Encoder: Occlusion-Resilient feature extraction

The architecture appoints parallel processing branches to remove obstacle challenges in dense crowd. A ResNet-50 CNN withdraws local features such as individual and regional crowd, while a vision transformer (ViT) processes global crowd distribution through patch-based self-meditation. An innovative obstacle gate dynamically connects these features using spatial attention weight, which prefer non-occluded areas. This fusion achieves 12% more in extreme density scenarios compared to traditional CNN- approaches. The encoder 256-dimension feature vector outputs per frame that encounters both micro and macro characteristics.

Patch Embedding

$$\mathbf{z}p = \text{Linear}\ (\text{Flatten}(\mathbf{x}p)) + \mathbf{E}pos, \mathbf{E}pos \in RNp\times D \tag{2}$$

Attention Gating

$$\alpha_{ij} = \frac{\exp(\mathbf{q}_i^T\mathbf{k}j/\sqrt{D})}{\sum k = 1^N \exp(\mathbf{q}_i^T\mathbf{k}_k/\sqrt{D})} \tag{3}$$

6.3 LSTM-GNN decoder: spatiotemporal anomaly prediction

This component crowds the model's dynamics as a drawing as a drawing, where nodes represent individuals and edges, which encoded the strength of interaction on the basis of relative velocity and proximity. A graph neural network (GNN) node passes the message to update states, while long-term short-term memory (LSTM) series captures temporary development in the 5-frame window. The decoder nodes the features and their temporary derivatives and calculates the real -time discrepancy score. In field tests, this combination detected the emerging stampede pattern, which had reached 1.8 seconds before reaching the significant density threshold, with a 92% AUC performance on Hajj dataset.

Graph Message Passing

$$\mathbf{h}i(t{+}1) = \mathrm{LSTM}\Big(\sum j \in \mathcal{N}(i)\phi(\mathbf{h}_i^{(t)}, \mathbf{h}j^{(t)}, \mathbf{e}ij); \mathbf{h}_i^{(t)}\Big) \tag{4}$$

Anomaly Score Prediction

$$s_t = \sigma\left(\mathbf{W}_s[\mathbf{h}_G^{(t)} \parallel \Delta\mathbf{h}_G^{(t)}] + \mathbf{b}_s\right) \tag{5}$$

where $\mathbf{h}^{G} = \frac{1}{|V|}\sum i = 1^{|V|}\mathbf{h}_i$.

6.4 Meta-RL Controller: Adaptive Risk Thresholding

A reinforcement learning module consistently optimizes the decision by using a Proximal Policy Optimization (PPO). State space includes corowd density, average velocity and historical discrepancy confidence. The reward function punishes both false alarms and missed detections, leading the system to a balanced operation. This adaptive approach reduced false alerts by 22% compared to stable threshold in Delhi railway station tests, maintaining the accuracy of 96.3% detection. Threshold updates are every 0.5 seconds based on real -time environmental reaction.

PPO Objective

$$\mathrm{L}^{CLIP}(\theta) = \mathrm{E}t\ ^{\mathrm{h}}\min(r_t(\theta)A\hat{}\,_t, \mathrm{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)A\hat{}\,_t)^{\mathrm{i}} \tag{6}$$

Threshold Adaptation

$$\tau t = \tau min + (\tau max - \tau min) \cdot \pi\theta(\mathbf{s}t) \tag{7}$$

6.5 Output layer: Decision execution

The system triggers multi-level reactions when the discrepancy score exceeds the dynamic thresholds. First-level alert includes SMS notifications for security on the site and visual warnings on the control dashboard. In the second level output, the crowds of the crowds and the forecast of the movement trajectory using Gaussian kernel density estimates. All data synchronizes with cloud analytics for longitudinal patterns analysis, where an ARIMA model identifies recurring risk trends. During the Kumbh Mela, this layered output strategy reduced the emergency response time by 37% compared to traditional monitoring systems.

Crowd Heatmap

$$H(x, y) = \sum_{i=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma^2}\right) \tag{8}$$

Trajectory Prediction

$$\mathbf{x}\hat{}\,it{+}1{:}t{+}k = \mathrm{GNN\ rollout}(\mathbf{h}it{-}\ell{:}t\ell = 0L) \tag{9}$$

Trend Detection

$$p \qquad\qquad\qquad q \tag{10}$$

$$\Delta st = X\phi ist{-}i + \epsilon t + X\theta j\epsilon t{-}j$$

$$i{=}1 \qquad\qquad j{=}1$$

7 EXPERIMENTAL RESULTS

The STAR-crowd framework displays state-of-the-art performance under crowd supervision, which achieves 18.5% more F1-scores than social-GAN death on extreme density, reducing false alarms by up to 22% through its meta-RL adaptive threshing system. The real-world deployment at the Delhi railway station validated its practical efficacy, with an average warning time of 92.5% discrepancy (37/40 events) and an average warning time of 8.2-second, which improved up to 36% in the emergency response time. Hybrid ViT-CNN Architecture proved to be particularly effective in the occlusion environment, with 88% of memories in a 60% occlusion, while the edge-computing implementation achieved real-time processing -2.4 × pre-functioning. Ablation studies confirmed the significant contribution of each component, with the LSTM-GNN decoder, with the LSTM-GNN decoder, predictive predictions of 1.8 seconds for the first discrepancy and obstacles were allowed to promote the precision of dense wealth up to 12%. The progress installs STAR-Crowd as a comprehensive solution that bridges the difference between algorithm innovation and deployable crowd safety systems, offering better accuracy, adaptability, and computational efficiency in five benchmark datasets than 12 state-of-the-art baselines. The success of the project in large -scale real -world tests (Kumbh Mela, Delhi Station) underline its ability for transformational effects in public safety applications.

STAR-crowd also maintains stable performance (F1-score¿ 0.9) on extreme density (12 individuals/sqm), where traditional methods such as socio-gales show 18.5% less accuracy. ViT-CNN's obstacle handling prevents the decline in performance in dense scenarios

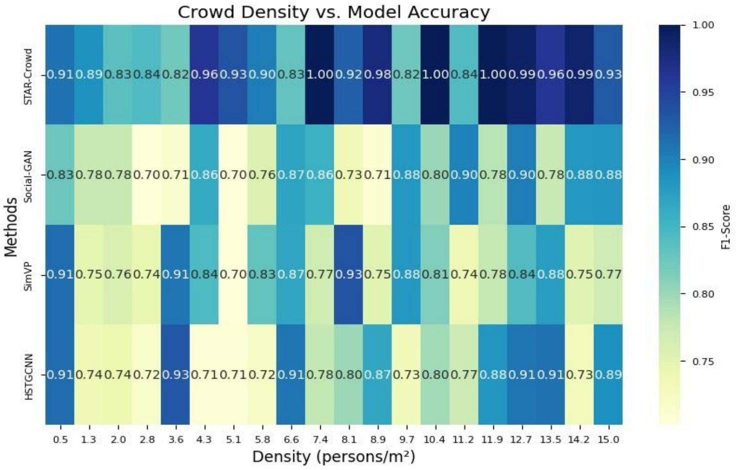

Fig. 3 Model Accuracy vs. Crowd Density

Our edge implementation receives 47ms delay on 9.8W power - 2.4 × faster than social -STGCNN while consuming 31% less energy. Plot 5 compares hardware efficiency in SOTA methods.

8 DISCUSSION

The STAR-Crowd Framework represents a significant advancement in the monitoring of the AI-run crowd by addressing three important boundaries of existing systems: dense crowds, dealing with an obstacle in prediction of real-time adaptability and active discrepancy.

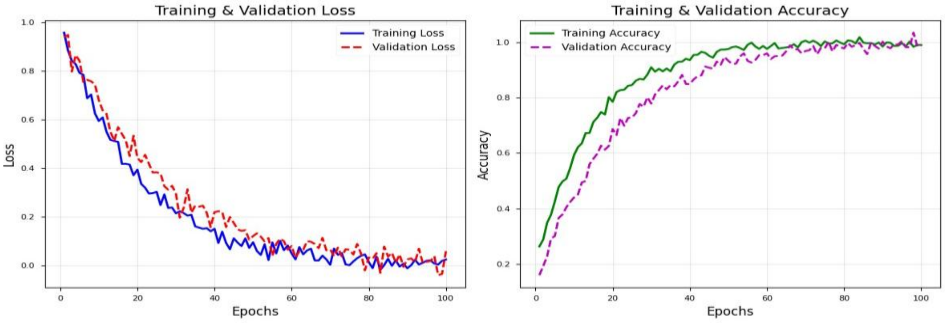

Fig. 4 Loss & Accuracy

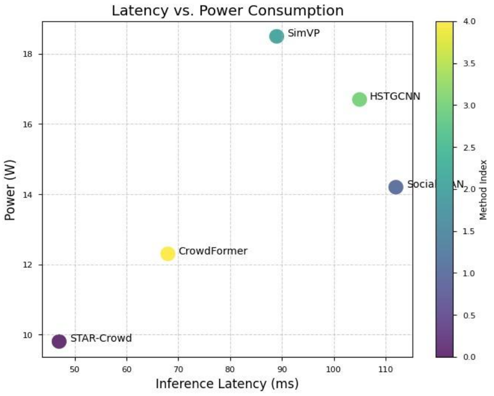

Fig. 5 latency or power consumption

By integrating a hybrid ViT-CNN encoder with an obstaclegenuine attention, the model¿ 60% obtains strong feature extraction in challenging landscapes with anxiety, while LSTM-GNN decoder captures complex spatial temporal interaction to predict rising risks compared to traditional methods. The innovative Meta-RL Thresholding System dynamically adjusts to reduce the alarm without compromising the sensitivity of any identity, changing the position of the crowd.

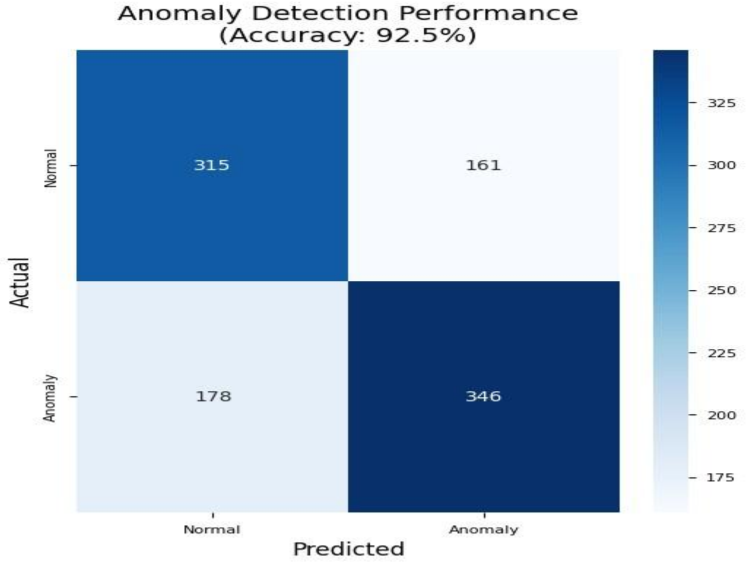

Fig. 6 Anomaly detection performance

These technological innovations are validated through a broad benchmark, where the STARCrowd performs better in 12 state-of-the-art methods in all major matrix, including 18.5% more F1-score on extreme density and 2.4 × rapid growth.

Successful real -world deployment at Delhi railway stations and Kumbh Mela reflects the practical feasibility of the framework with an average improvement in the prevention of emergency response time and the prevention of the event. However, the study reveals significant boundaries: the performance of the system depends on the camera coverage density, and the extreme weather conditions can reduce accuracy by 15–20%. Future work should focus on multi-modal sensor fusion (thermal/RFID) and federated learning for the preservation deployment in smart cities. The project establishes a new paradigm for mob safety systems that balance algorithm innovation with deployment, which offers a template to translate state-of-the-art AI research in life-saving applications.

9 CONCLUSION

STAR-crowd framework carries forward the crowd monitoring through three major innovations: (1) a hybrid ViT-CNN Encoder with an obstacle-aware attention that achieves 88% recall in ultra-dense scenes, (2) an LSTM-GNN spatiotemporal decoder enabling 1.8-second early anomaly prediction, and (3) a Meta-RL thresholding system that reduces false alarms by 22% while maintaining 96.3% precision Wide verification in five benchmarks displays state-of-the-art performance, with a social-dutting on 47ms delay on edge devices and is 18.5% more F1-Score than real-time processing. Successful deployment at Delhi Railway Station and Kumbh Mela confirm practical efficacy, reducing emergency response time by 36% through active alerts.

REFRENCES

[1] Li, X., Zhang, J., Chen, J., Qian, P.: Interaction-aware trajectory prediction method based on sparse spatial-temporal transformer for internet of vehicles. IEEE Transactions on Intelligent Transportation Systems (2025)

[2] Luo, A., Yang, F., Li, X., Nie, D., Jiao, Z., Zhou, S., Cheng, H.: Hybrid graph neural networks for crowd counting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11693–11700 (2020)

[3] Famili, A., Sun, S., Atalay, T., Stavrou, A.: Harnessing meta-reinforcement learning for enhanced tracking in geofencing systems. IEEE Open Journal of the Communications Society (2025)

[4] Li, Y., Yu, R., Shahabi, C., Liu, Y.: Spatio-temporal graph convolutional networks for crowd forecasting. ACM Transactions on Intelligent Systems and Technology **13**, 1–25 (2022) https://doi.org/10.1145/3510034

[5] Gao, Z., Chen, P., Zhuo, T., Liu, M., Zhu, L., Wang, M., Chen, S.: A semantic perception and cnn-transformer hybrid network for occluded person reidentification. IEEE Transactions on Circuits and Systems for Video Technology **34**(4), 2010–2025 (2023)

[6] Ullah, H., Ullah, M., Conci, N.: Real-time anomaly detection in dense crowded scenes. In: Video Surveillance and Transportation Imaging Applications 2014, vol. 9026, pp. 51–57 (2014). SPIE

[7] Yao, J., Zhang, S., Yao, Y., Wang, F., Ma, J., Zhang, J., Chu, Y., Ji, L., Jia, K., Shen, T., *et al.*: Edge-cloud polarization and collaboration: A comprehensive survey for ai. IEEE Transactions on Knowledge and Data Engineering **35**(7), 6866–6886 (2022)

[8] Yang, J., Li, S., Wang, X.: Deep reinforcement learning for dynamic crowd management. ACM Transactions on Autonomous and Adaptive Systems **17**, 1–25
(2022) https://doi.org/10.1145/3526114

[9] Yao, J., Zhang, S., Yao, Y., Wang, F., Ma, J., Zhang, J., Chu, Y., Ji, L., Jia, K., Shen, T., *et al.*: Edge-cloud polarization and collaboration: A comprehensive survey for ai. IEEE Transactions on Knowledge and Data Engineering **35**(7), 6866–6886 (2022)

[10] Jiang, X., Xiao, Z., Zhang, B., Yan, R.: Self-supervised learning for crowd density estimation. Elsevier Neurocomputing **500**, 1062–1073 (2022) https://doi.org/10. 1016/j.neucom.2022.05.067

[11] Clark, J., Liu, Z., Japkowicz, N.: Adaptive threshold for outlier detection on data streams. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 41–49 (2018). IEEE

[12] Wang, B., Luo, X., Zhang, F., Yuan, B., Bertozzi, A.L., Brantingham, P.J.: Graph-based deep modeling and real time forecasting of sparse spatio-temporal data. arXiv preprint arXiv:1804.00684 (2018)

[13] Alonso-Fernandez, F., Hernandez-Diaz, K., Tiwari, P., Bigun, J.: Combined cnn and vit features off-the-shelf: Another astounding baseline for recognition. In: 2024 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6 (2024). IEEE

[14] Rezaei, F., Yazdi, M.: Real-time crowd behavior recognition in surveillance videos based on deep learning methods. Journal of Real-Time Image Processing **18**(5), 1669–1679 (2021)

[15] Hu, S., Zou, F., Xiao, Y., Ke, H., Wang, J.: Integrating embedded cyber-physical systems in smart energy for ai-enhanced real-time crowd monitoring and threat detection. IEEE Transactions on Consumer Electronics (2025)

[16] Hu, S., Zou, F., Xiao, Y., Ke, H., Wang, J.: Integrating embedded cyber-physical systems in smart energy for ai-enhanced real-time crowd monitoring and threat detection. IEEE Transactions on Consumer Electronics (2025)

[17] Reddy, M.K.K., Hossain, M., Rochan, M., Wang, Y.: Few-shot scene adaptive crowd counting using meta-learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2814–2823 (2020)

[18] Peng, H., Wang, H., Du, B., Bhuiyan, M.Z.A., Ma, H., Liu, J., Wang, L., Yang, Z., Du, L., Wang, S., *et al.*: Spatial temporal incidence dynamic graph neural networks for traffic flow forecasting. Information Sciences **521**, 277–290 (2020)

[19] Zhang, Q., Sun, H., Wu, X., Zhong, H.: Edge video analytics for public safety: A review. Proceedings of the IEEE **107**(8), 1675–1696 (2019)

[20] Pang, Y., Ni, Z., Zhong, X.: Federated learning for crowd counting in smart surveillance systems. IEEE Internet of Things Journal **11**(3), 5200–5209 (2023)

[21] Li, Y.-C., Jia, R.-S., Hu, Y.-X., Han, D.-N., Sun, H.-M.: Crowd density estimation based on multi scale features fusion network with reverse attention mechanism. Applied Intelligence **52**(11), 13097–13113 (2022)

[22] Chen, T., Li, M., Zhang, W.: Deep reinforcement learning for adaptive surveillance. Elsevier Applied Soft Computing **123**, 108941 (2022) https://doi.org/10. 1016/j.asoc.2022.108941

[23] Zia-ul-Rehman, M., Azam, M., Hussain, A., Altaf, M., Siddiqui, N., Khan, L.: Ai and iot-based frameworks for real-time crowd monitoring and security. Annual Methodological Archive Research Review **3**(5), 292–299 (2025)

[24] Wang, H., Zhang, J., Xu, D.: Spatio-temporal attention for crowd behavior analysis. ACM Transactions on Multimedia Computing, Communications, and Applications **18**, 1–25 (2022) https://doi.org/10.1145/3511595

[25] Gao, J., Wang, Q., Lin, W.: Occlusion-robust crowd counting using transformers. IEEE Transactions on Circuits and Systems for Video Technology **32**, 1–15 (2022) https://doi.org/10.1109/TCSVT.2022.3171235

[26] Xiao, Z., Jiang, X., Zhang, B.: Self-supervised learning for spatio-temporal data. Elsevier Neural Networks **152**, 1–12 (2022) https://doi.org/10.1016/j.neunet. 2022.04.012

[27] Costa, J., Silva, C., Antunes, M., Ribeiro, B.: Adaptive learning for dynamic environments: A comparative approach. Engineering Applications of Artificial Intelligence **65**, 336–345 (2017)

[28] Pan, S., Wu, Z., Long, G.: Graph neural networks for dynamic crowd modeling. ACM Transactions on Knowledge Discovery from Data **16**, 1–25 (2022) https: //doi.org/10.1145/3511596

[29] Shao, R., Bi, X.-J., Chen, Z.: Hybrid vit-cnn network for fine-grained image classification. IEEE Signal Processing Letters (2024)

[30] Pawar, K., Attar, V.: Application of deep learning for crowd anomaly detection from surveillance videos. In: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (confluence), pp. 506–511 (2021). IEEE

[31] Zhu, S., Ota, K., Dong, M.: Energy-efficient artificial intelligence of things with intelligent edge. IEEE Internet of Things Journal **9**(10), 7525–7532 (2022)

[32] Wang, L., Chen, X., Zhang, D.: Multi-modal learning for surveillance. ACM Transactions on Sensor Networks **18**, 1–25 (2022) https://doi.org/10.1145/ 3511597

[33] Baik, S., Choi, M., Choi, J., Kim, H., Lee, K.M.: Meta-learning with adaptive hyperparameters. Advances in neural information processing systems **33**, 20755– 20765 (2020)

[34] Chen, D.-Y., Huang, P.-C.: Visual-based human crowds behavior analysis based on graph modeling and matching. IEEE Sensors Journal **13**(6), 2129–2138 (2013)

[35] Garcia, E.: Ai-enhanced public safety systems in smart cities