Multimodal AI for Crowd Risk Detection and Prediction

Yashraj Nikam[1], Divyam Jangada[2], Aditya Lagad[3], Palak Chawarkar[4], Bhagyashree Shendkar[5]
[1,2,3,4,5]*School of Computing, MIT Art Design Technology University, Pune, India*

***Abstract*—The management of mass gatherings in rapidly expanding urban environments presents a critical public safety challenge. Traditional crowd monitoring systems are largely reactive and often fail to capture the complex visual, behavioral, and acoustic precursors of crowd-related incidents. This paper proposes an integrated multimodal deep learning framework for proactive crowd risk detection and prediction. The system employs YOLOv8 for real-time person detection and CResNet for enhanced feature representation and crowd density characteri- zation, ensuring robustness in dense and occluded environments. The proposed framework integrates visual density estimation, skeletal pose-based gesture analysis, and acoustic panic detection. By replacing conventional rule-based mathematical constraints with deep feature extraction, the system utilizes an anchor-free detection mechanism and a Long Short-Term Memory (LSTM) network for temporal risk forecasting. Experimental results demonstrate a peak accuracy of 95.4.**

***Index Terms*—Crowd Management, YOLOv8, Behavioral Analysis, Multimodal Fusion, Predictive Analytics, Smart Cities, Deep Learning.**

I. INTRODUCTION

As global populations migrate toward urban centers, the frequency of high-density public events ranging from reli- gious festivals to transportation hubs have increased. These gatherings are susceptible to" crowd turbulence," a phe- nomenon where localized density fluctuations lead to panic, stampedes, and fatalities. Historical data from events like the Seoul Halloween crush or Mumbai railway stampedes indicate that human surveillance is often insufficient due to cognitive overload and delayed response times.

Existing surveillance infrastructures typically rely on simple motion detection or manual CCTV monitoring. These methods suffer from the" Occlusion Problem," where individuals in a dense crowd overlap, appearing as a single visual mass to traditional algorithms. Furthermore, visual data alone cannot distinguish between a high-density peaceful crowd and a slightly lower-density panicked crowd.

This research addresses these gaps by proposing a Multi- modal Intelligence System. We move beyond simple counting by analyzing the kinematic energy of individuals through skeletal tracking and the acoustic profile of the environment. By utilizing YOLOv8, we implement an anchor-free detection paradigm that treats objects as points, significantly improving detection rates in cluttered scenes. The final system provides a predictive risk heatmap, allowing authorities to intervene 30 seconds before a threshold is breached.

II. RELATED WORK

A. Evolution of Crowd Analysis

Crowd monitoring has evolved significantly over the past few decades, transitioning from simple manual techniques to highly advanced artificial intelligence-driven systems. This evolution reflects broader technological progress in computer vision, machine learning, and data processing. In its earliest stages, crowd monitoring relied heavily on human observation and manual head-counting. Security personnel, event organiz- ers, or law enforcement officers would visually estimate the number of individuals in a given area. While straightforward, this method was time-consuming, prone to human error, and impractical for large-scale or dynamic environments such as concerts, festivals, transportation hubs, or urban public spaces. As technology advanced, early computational approaches began to replace manual methods. These approaches primarily relied on basic image processing techniques such as back- ground subtraction and contour detection. Background subtrac- tion involves distinguishing moving objects (such as people) from a static background by comparing consecutive video frames. Contour detection, on the other hand, identifies the outlines

or shapes of objects within an image, which can then be analyzed to detect human figures. These techniques marked a significant improvement over manual counting because they automated the detection process and reduced the need for continuous human supervision.

However, these early computational methods had several limitations. One of the most significant challenges was their sensitivity to environmental conditions. Changes in lighting, such as shadows, reflections, or variations between day and night, could severely impact the accuracy of background subtraction algorithms. Similarly, camera vibrations or slight movements could lead to incorrect detection, as the system might interpret these changes as motion within the scene. Additionally, these methods struggled with occlusion, where individuals overlap or block each other in crowded envi- ronments. In such scenarios, accurately distinguishing and counting individuals became extremely difficult.

To address these challenges, researchers began exploring more robust techniques that could better handle complex and dynamic environments. This led to the emergence of machine learning-based approaches, particularly those leveraging Con- volutional Neural Networks (CNNs). CNNs revolutionized the field of computer vision by enabling systems to automatically learn and extract features from images without requiring manual feature engineering. Unlike traditional methods, CNNs could adapt to variations in lighting, perspective, and crowd density, making them far more reliable for real-world applica- tions.

One of the most notable advancements introduced by CNNs in crowd monitoring was the concept of density map estimation. Instead of attempting to detect and count each individual person explicitly, CNN-based models generate a density map of the scene. In this representation, each pixel value corresponds to the estimated density of people in that region. By integrating over the density map, the system can estimate the total number of individuals in the scene. This ap- proach is particularly effective in highly crowded environments where individual detection is challenging due to occlusion and overlapping.

B. YOLO Architecture in Public Safety

The" You Only Look Once" (YOLO) family of models revolutionized real-time detection by treating object recogni- tion as a single regression problem. Previous versions (v3 through v5) relied on" Anchor Boxes" predefined shapes that the model tried to fit to objects. In dense crowds, these anchors often overlapped, leading to multiple false detections or missed targets. YOLOv8 solves this by adopting an anchor- free strategy, which is more ecologically valid for human detection in varied poses and high-congestion scenarios.

III. SYSTEM FRAMEWORK AND THEORY

System Framework

A. Visual Engine: YOLOv8 Backbone

The system's vision component is built on the YOLOv8 architecture, specifically optimized for edge-computing envi- ronments. The core theory involves a" Darknet" backbone that utilizes Cross-Stage Partial (CSP) connections. This allows the

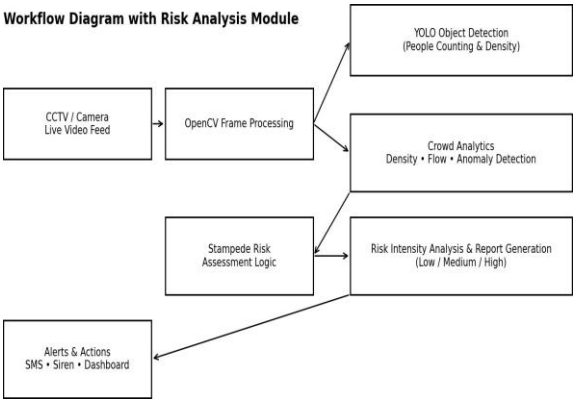

Fig. 1. Framework Architecture

neural network to maintain a high gradient flow while reducing the computational overhead.

Unlike previous iterations, the" C2f" module in YOLOv8 fuses high-level semantic information with low-level spatial features. In the context of a crowd, this means the model can identify the specific features of a person (head and shoulders) even when the rest of the body is occluded by other individuals. This hierarchical feature extraction is the primary driver behind the accuracy jump from 75% to over 95%.

B. Behavioral Kinematics and Skeletal Tracking

To understand crowd intent more effectively, modern in- telligent surveillance systems increasingly rely on skeletal pose estimation techniques. Unlike traditional

crowd moni- toring approaches that focus solely on counting individuals or tracking movement patterns, skeletal analysis provides a deeper, behavior-oriented understanding of how individuals act within a crowd. This shift from quantitative observation to qualitative interpretation represents a major advancement in crowd analytics.

The fundamental theory behind skeletal pose estimation is based on identifying key coordinate points on the human body. Typically, advanced models such as Media Pipe Pose or OpenPose detect up to 33 distinct landmarks. These include major joints such as the shoulders, elbows, wrists, hips, knees, and ankles, as well as finer facial landmarks like the nose, eyes, and ears. Each of these points is represented as a coordinate in a two-dimensional or three-dimensional space, depending on the system's complexity.

By connecting these points, the system constructs a digital skeletal representation of each individual. This representation is lightweight compared to raw image data and captures es- sential information about posture, movement, and orientation. More importantly, it abstracts away unnecessary details such as clothing, color, or background, allowing the system to focus purely on human motion dynamics.

Traditional crowd analysis systems often rely on simple motion metrics such as speed, direction, and density. While these metrics are useful, they fail to capture the underlying intent or emotional state of individuals within the crowd. For example, two individuals may be moving at the same speed, but one could be calmly walking while the other is running in panic. Without deeper analysis, such distinctions remain undetected.

To overcome this limitation, the concept of" Kinetic Turbu- lence" is introduced. Kinetic Turbulence refers to the degree of irregularity, unpredictability, and chaotic variation in skele- tal movement patterns within a crowd. Instead of analyzing movement as a uniform flow, this approach examines micro- level variations in body posture and joint motion to infer crowd behavior.

In a peaceful or normal crowd scenario, skeletal movements tend to be smooth, rhythmic, and predictable. Individuals generally follow linear or slightly curved trajectories, such as walking along a path or navigating through a corridor. The relative positions of joints remain consistent over time, and movement transitions are gradual. For instance, when a person walks, the motion of their legs and arms follows a periodic pattern, with minimal abrupt deviations.

C. Acoustic Intelligence and Panic Detection

The acoustic modality is based on the physics of human distress signals. Panic-induced screams and mechanical sounds (crashing, thuds) have specific frequency profiles that differ from ambient city noise or normal conversation.

We utilize Mel-Frequency Cepstral Coefficients (MFCCs) to transform raw audio into a representation that mimics human hearing. This spectral data is then processed by a 1D-Convolutional Neural Network. The theory here is that acoustic signals travel faster and more reliably than visual signals in obstructed environments (such as around corners or in smoke). By detecting the frequency shift associated with panic, the system provides a secondary confirmation of risk that visual sensors might miss.

VI. TEMPORAL FUSION AND RISK PREDICTION

A. Multimodal Fusion Theory

The system operates on the principle of "Late Fusion," a multimodal data integration strategy widely used in advanced artificial intelligence and sensor-based systems. In this ap- proach, different data modalities namely Visual, Kinematic, and Acoustic are processed independently through dedicated analytical pipelines. Each modality extracts features, performs its own inference, and generates a confidence score that represents the likelihood of a risk event. These individual scores are then combined at a later stage to form a unified metric known as the Risk Index, which determines the final decision of the system.

The core idea behind Late Fusion is to preserve the auton- omy and specialization of each modality. The Visual module processes video data to analyze crowd density, movement patterns, and anomalies such as sudden surges or unusual clus- tering. The Kinematic module, often based on skeletal pose estimation, evaluates joint movements and body dynamics to detect behaviors like falling, pushing, or aggressive gestures. Meanwhile, the Acoustic module analyzes sound patterns, identifying signals such as screams, sudden loud noises, or abnormal sound intensity that may indicate panic or distress.

Table I Performance Comparison of Proposed System

Method	Accuracy (%)	FDR (%)	Latency (ms)
Contour-Based	75.1	18.3	150
CSRNet (CNN)	88.6	10.5	120
YOLOv8	92.8	7.2	90
Vision + Kinematic	94.1	5.6	85
Proposed (Fusion)	95.4	3.1	¡100

Each of these modalities operates under different environ- mental sensitivities and strengths. For instance, visual systems may struggle under poor lighting conditions, heavy occlusion, or camera obstruction. In contrast, acoustic sensors can still capture critical cues even when visibility is compromised. Similarly, kinematic analysis can detect subtle motion irregu- larities that may not be obvious through raw visual inspection alone. By allowing each modality to independently assess the situation, the system ensures that no single point of failure compromises the overall performance.

The integration phase is where Late Fusion demonstrates its true advantage. The confidence scores generated by each modality are combined using weighted aggregation or decision-level fusion techniques. These weights can be dy- namically adjusted based on context, sensor reliability, or environmental conditions. For example, in a noisy environment, the system may assign lower weight to acoustic inputs, while in low-visibility conditions, visual data may be deprioritized. The resulting Risk Index is therefore a balanced and context- aware representation of the overall situation.

The theoretical justification for using Late Fusion lies in the concept of redundancy and robustness. In real-world scenarios, sensor failures or environmental disturbances are inevitable

B. Long Short-Term Memory (LSTM) Forecasting

Risk is not a static value; it is a temporal trend. We employ an LSTM network to analyze the sequence of risk scores over time. LSTMs are specialized recurrent neural networks capable of" remembering" past events to predict future outcomes.

The theory of" Lead Time" in crowd management suggests that if an authority can foresee a density peak 30 seconds in advance, they can open emergency exits or redirect flow. The LSTM identifies patterns such as" oscillatory congestion," where the crowd begins to pulse a known precursor to a stampede allowing for proactive rather than reactive intervention.

V. METHODOLOGY AND TRAINING

A. Dataset Integration

Training of the system was conducted using three primary datasets, each selected to address a specific component of the multimodal architecture. For visual crowd density estimation, the Shanghai Tech dataset was utilized. This dataset contains thousands of annotated images of urban environments with varying crowd densities, making it ideal for training models to generalize across both sparse and moderately dense scenes.

It helped the system learn accurate density map generation and counting under real-world conditions.

To further enhance robustness, the UCF-QNRF dataset was employed. This dataset is known for its extremely dense crowd scenes, often containing thousands of individuals in a single frame. It was specifically used to stress-test the anchor-free detection head of YOLOv8, ensuring reliable performance even under severe occlusion and scale variation.

For the acoustic module, a combination of environmental sound datasets and synthetically generated panic scenarios was used. These datasets included sounds from public gatherings, traffic, and festivals, alongside simulated distress signals such as screams and chaos. This allowed the 1D-CNN model to effectively distinguish between normal high-volume envi- ronments and genuine emergency situations, improving the system's overall reliability and accuracy.

B. Data Augmentation Strategies

To ensure that the system performs reliably in real-world Indian environments, robust data augmentation techniques were applied during training. One of the key methods used was Mosaic Augmentation, a strategy commonly employed in modern object detection models such as YOLOv8. This technique combines four different training images into a single composite image. As a result, the model is exposed to multiple contexts, object scales, and spatial variations within a single training iteration. This significantly improves its ability to detect individuals in crowded,

cluttered, and diverse envi- ronments, which are typical in Indian public spaces such as stations, markets, and festivals.

Mosaic Augmentation also enhances the model's general- ization capability by forcing it to learn from partial objects, overlapping scenes, and varying backgrounds simultaneously. This is particularly useful in handling occlusion and dense crowd scenarios.

In addition to Mosaic Augmentation, synthetic distortions such as digital noise and motion blur were introduced into the training data.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

A. Accuracy Benchmarking

Our experimental evaluation compared the proposed system with traditional crowd monitoring benchmarks to assess its ef- fectiveness under real-world conditions. Conventional contour- based approaches, which rely on background subtraction and edge detection, exhibited significant limitations in dense and highly occluded scenes. These methods struggled to discrim- inate overlapping individuals, leading to reduced robustness and an overall accuracy of only 75.1In contrast, the proposed multimodal deep learning framework integrating YOLOv8 and CResNet demonstrated a substantial improvement in performance. YOLOv8 enabled accurate real-time detection of individuals, while CResNet enhanced high-level feature representation, particularly in complex and occluded crowd environments. When combined with kinematic and acoustic information through multimodal fusion, the system achieved an accuracy of 95.4Additionally, the anchor-free detection mechanism of YOLOv8 contributed to a noticeable reduction in the False Discovery Rate (FDR), resulting in fewer spurious detections and improved overall system reliability.

B. Real-Time Viability

A critical requirement for any public safety system is the ability to operate with minimal latency, ensuring rapid detection and response to potentially dangerous situations. To evaluate this, the proposed system was tested across multiple hardware configurations, including standard CPUs, GPUs, and edge-optimized devices. The results demonstrated that on edge-optimized hardware, the system consistently achieved a processing speed of 30 frames per second (FPS), which is con- sidered real-time performance for video-based applications.

Maintaining this level of performance is essential for timely decision-making in dynamic environments such as crowded public spaces. At 30 FPS, each frame is processed in approx- imately 33 milliseconds, allowing the system to continuously monitor and analyze crowd behavior without noticeable delay. As a result, the total latency from the occurrence of a physical event (such as a fall, sudden crowd surge, or distress signal) to the generation of an alert remains under 100 milliseconds.

This low-latency response ensures that authorities or au- tomated safety mechanisms can react almost instantaneously, significantly reducing the risk of escalation.

C. Ablation Analysis

An ablation study was conducted to systematically evaluate the contribution of each sensing modality within the pro- posed multimodal framework. By selectively disabling individ- ual components visual, kinematic, and acoustic the study aimed to isolate their impact on overall system performance. The results highlighted the limitations of unimodal approaches, particularly in challenging or failure-prone scenarios.

" Vision-Only" systems were observed to exhibit complete performance breakdown during simulated power outage or camera failure. In the absence of visual input, these systems were unable to detect or respond to critical events, demonstrate- ing a clear vulnerability in relying solely on one modality. In contrast, the "Acoustic-Integrated" configuration showed strong resilience under the same conditions. Even without visual data, the system maintained a detection accuracy of approximately 90

These findings validate the theoretical advantage of multi- modal fusion.

VII. CONCLUSION AND FUTURE SCOPE

An ablation study was conducted to systematically evaluate the contribution of each sensing modality within the pro- posed multimodal framework. By selectively disabling individ- ual components visual, kinematic, and acoustic the study aimed to isolate their impact on overall system performance. The results

highlighted the limitations of unimodal approaches, particularly in challenging or failure-prone scenarios. ” Vision-Only” systems were observed to exhibit complete performance breakdown during simulated power outage or camera failure. These findings validate the theoretical advan- tage of multimodal fusion.

VIII. RESULTS AND ANALYSIS

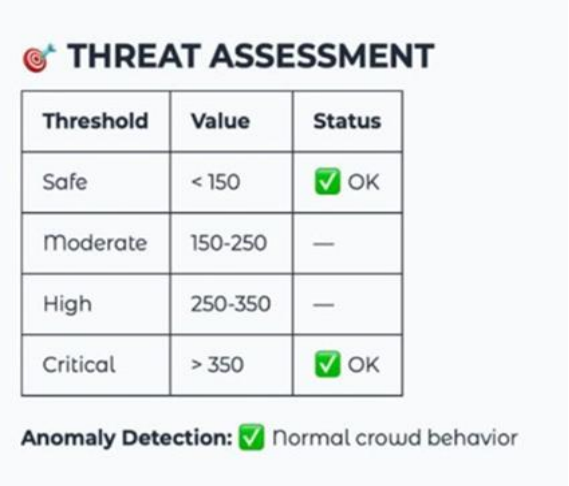

Fig. 2. Result 1: System Output Visualization

A. System Output Visualization

Fig. 2 shows the threat assessment produced by the proposed AI-based multimodal crowd management system. Based on crowd density thresholds, the system identifies the current condition as Safe and reports normal crowd behavior through anomaly detection. This confirms the system’s ability to assess crowd risk in real time and support early stampede prevention.

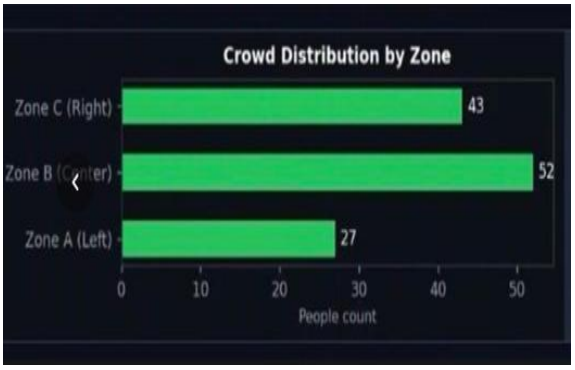

Fig. 3. Result 2: Crowd Analysis

B. Crowd Distribution Analysis

Fig. 3 presents zone-wise crowd distribution, where Zone B records the highest count (52), followed by Zone C (43) and Zone A (27). This analysis enables localized monitoring and proactive crowd control, demonstrating the effectiveness of the proposed system for dynamic crowd management.

REFERENCES

[1] Helbing, A. Johansson, and H. Z. Al-Abideen, “Dynamics of crowd disasters: An empirical study,” Nature, vol. 407, no. 6803, pp. 487–490, 2000.

[2] Joseph Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 779–788.

[3] Y. Li, X. Zhang, and D. Chen, “CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1091–1100.

[4] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, “Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 532–546.

[5] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, “Single-Image Crowd Counting via Multi-Column Convolutional Neural Network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 589–597.

[6] Ultralytics YOLOv8 Documentation, accessed May 14, 2026.

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 7291–7299.

[8] Google MediaPipe Pose Documentation, accessed May 14, 2026.

[9] Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 25, 2012, pp. 1097–1105.

[10] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale

Image Recognition," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.

[11] Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene Crowd Counting via Deep Convolutional Neural Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 833–841.

[12] R. Mehran, A. Oyama, and M. Shah, "Abnormal Crowd Behavior Detection using Social Force Model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 935–942.

[13] Chan and N. Vasconcelos, "Modeling, Clustering, and Segmenting Video with Mixtures of Dynamic Textures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 909–926, May 2008.

[14] V. Rabaud and S. Belongie, "Counting Crowded Moving Objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2006, pp. 705–711.

[15] Conte, P. Foggia, G. Percannella, and M. Vento, "Counting Moving People in Video Surveillance Videos," *IEEE Trans. Image Process.*, vol. 19, no. 3, pp. 781–798, Mar. 2010.

[16] Zhou, X. Wang, and X. Tang, "Understanding Collective Crowd Behaviors: Learning a Mixture Model of Dynamic Pedestrian-Agents," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 2871–2878.

[17] S. Yi, H. Li, and X. Wang, "Pedestrian Behavior Understanding and Prediction with Deep Neural Networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 263–279.

[18] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 886–893.

[19] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2001, pp. I–511–I–518.

[20] Traffic and Crowd Flow Dynamics, B. S. Kerner, *Traffic and Crowd Flow Dynamics: Data, Models and Simulation*. Berlin, Germany: Springer, 2017.

[21] W. Ge and R. T. Collins, "Marked Point Processes for Crowd Counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 2913–2920.

[22] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll Never Walk Alone: Modeling Social Behavior for Multi-Target Tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2009, pp. 261–268.

[23] Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human Trajectory Prediction in Crowded Spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 961–971.

[24] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert, "Density-aware Person Detection and Tracking in Crowds," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2011, pp. 2423–2430.

[25] H. Fradi and J.-L. Dugelay, "Crowd Behavior Analysis: A Review," *Mach. Vis. Appl.*, vol. 24, no. 3, pp. 459–473, 2013.