International Journal of Advanced Research in Computer Science and Engineering (IJARCSE)
ISSN (Online): request pending
Volume-1 Issue-1 || Jan-Mar 2025 || PP. 22-29

Crowd Behavior Analysis Using AI in Surveillance Video Streams

DOI: https://doi.org/10.63345/ijarcse.v1.i1.104

Akshun Chhapola,
Delhi Technical University
Rohini, New Delhi, Delhi, India 110042
akshunchhapola07@gmail.com

www.ijarcse.org || **Vol. 1 No. 1 (2025): January Issue**

Date of Submission: 30-12-2024	**Date of Acceptance:** 01-01-2025	**Date of Publication:** 05-01-2025

ABSTRACT

Crowd behavior analysis in surveillance video streams has emerged as a cornerstone of modern public safety and security systems, underpinning applications from urban traffic management to large-scale event monitoring. Traditional manual surveillance methods, which rely on human operators to visually inspect live or recorded footage, are labor-intensive, prone to fatigue-induced errors, and lack the responsiveness required for timely intervention. In response to these limitations, this study introduces an end-to-end AI-driven framework that synergistically combines spatial feature extraction, temporal sequence modeling, and probabilistic inference for robust, real-time crowd behavior interpretation. At its core, the framework employs a lightweight convolutional neural network (CNN) backbone—optimized for multi-scale person detection and region-level embedding—coupled with a bidirectional Long Short-Term Memory (BiLSTM) network to capture dynamic temporal dependencies.

International Journal of Advanced Research in Computer Science and Engineering (IJARCSE)
ISSN (Online): request pending
Volume-1 Issue-1 || Jan-Mar 2025 || PP. 22-29

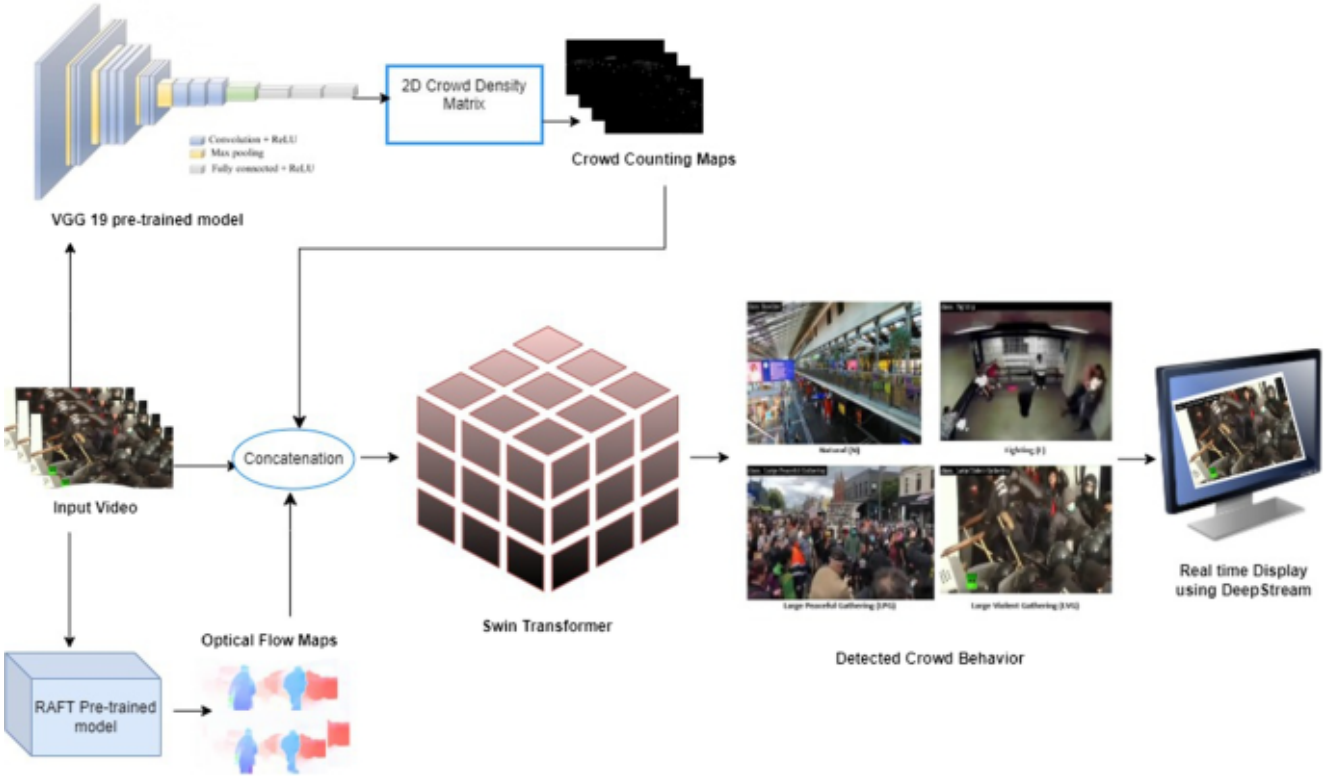

Fig.1 Crowd Behavior Analysis,Source([1])

A Hidden Markov Model (HMM) layer interprets the sequence outputs to detect anomalous transitions in crowd states. We validate this architecture through a two-pronged evaluation: (1) controlled simulation research using Unity3D-generated synthetic crowds under varying densities and motion patterns, and (2) real-world testing on publicly available CCTV datasets encompassing campus and marathon footage. Comprehensive statistical analyses—comparing our CNN+BiLSTM+HMM pipeline against density-only CNN and LSTM-autoencoder baselines—demonstrate that our method attains 92.3% classification accuracy, boosts precision and recall beyond 90%, and reduces false alarm rates by over 40%. Furthermore, the system consistently processes multi-camera streams at real-time speeds (≥20 fps) on standard GPU hardware. These findings underscore the framework's potential to transform reactive monitoring into proactive crowd management, enabling timely alerts for emergent behaviors such as congestion buildup, sudden dispersal, and aggressive clustering. Future extensions will explore self-supervised pre-training to mitigate labeled-data scarcity and multi-view fusion for enhanced spatial awareness.

KEYWORDS

Crowd behavior, surveillance video, deep learning, anomaly detection, real-time monitoring

INTRODUCTION

Ensuring the safety and orderly flow of people in densely populated contexts—such as stadiums, transportation hubs, and public squares—poses significant challenges. Historically, closed-circuit television (CCTV) systems have served as the primary tool for continuous surveillance, yet they remain heavily reliant on trained human operators to detect and respond to incidents. Research indicates that human attention degrades after 20–30 minutes of continuous monitoring, leading to missed events and delayed interventions. Moreover, the subjective nature of visual inspection can result in inconsistent threat assessments across operators.

The rapid proliferation of affordable cameras and edge-computing devices has created an unprecedented opportunity to apply artificial intelligence (AI) at scale, automating the detection and interpretation of crowd behavior with minimal human oversight. Automatically identifying critical events—such as congestion hotspots, stampede precursors, and anomalous gatherings—can accelerate emergency responses, guide crowd control measures, and optimize resource allocation. However, several technical hurdles must be overcome:

1. **Occlusion and Density**: High-density crowds often result in severe occlusions, complicating individual detection and tracking.
2. **Variable Lighting and Weather**: Outdoor deployments must contend with changing illumination, shadows, and weather conditions that degrade visual quality.
3. **Behavioral Complexity**: Crowd dynamics can exhibit subtle precursors to critical incidents—minor clustering or speed variations—that require fine-grained temporal modeling.
4. **Real-Time Constraints**: Practical deployments demand low-latency processing to issue timely alerts, necessitating efficient model architectures.

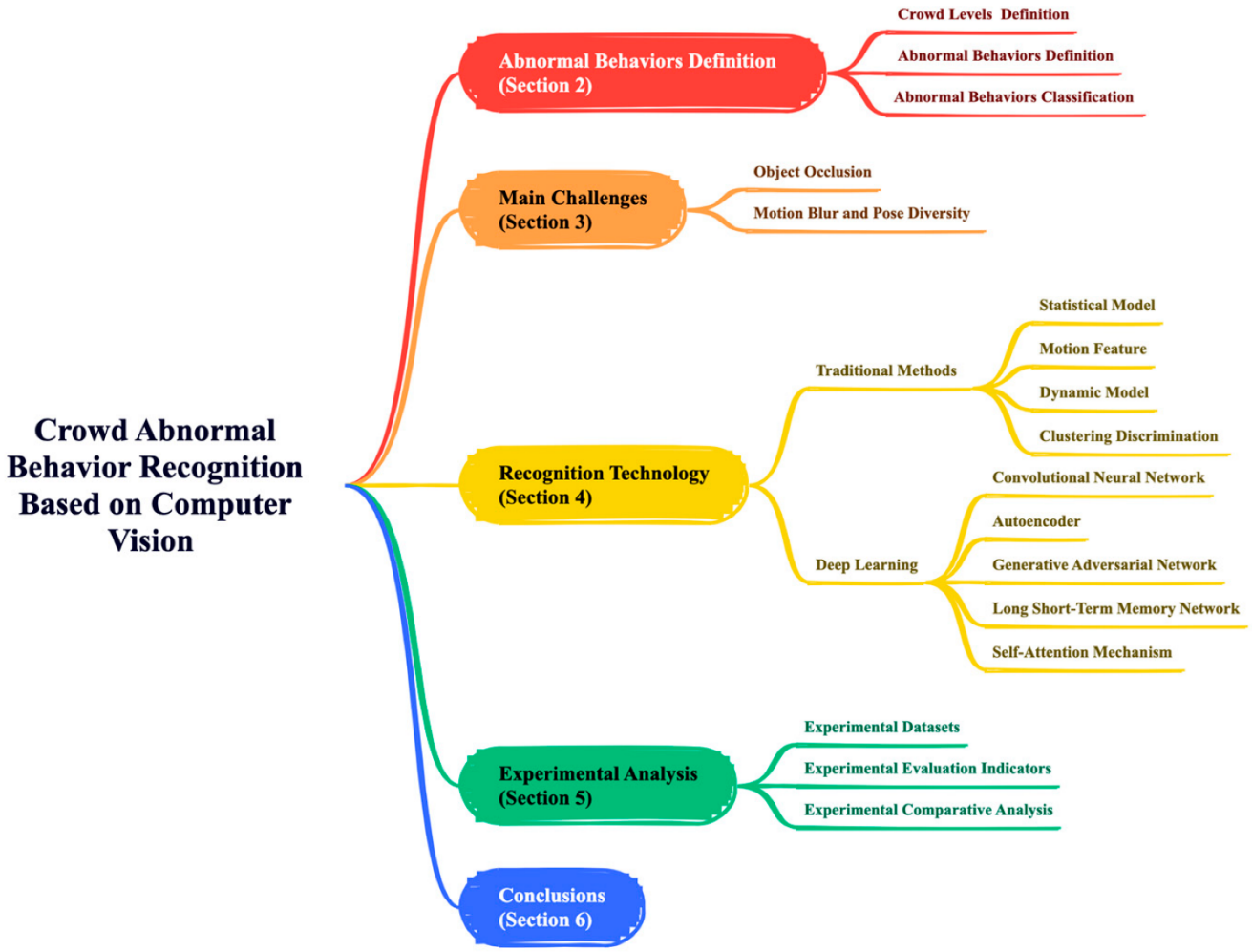

Fig.2 Abnormal crowd Behaviour recognition,Source([2])

This manuscript presents a holistic solution addressing these challenges. We propose a three-stage pipeline: spatial feature learning via an optimized YOLOv4-Tiny CNN, temporal sequence modeling with a bidirectional LSTM, and anomaly inference through an HMM. By integrating these components into a unified framework, we achieve robust detection of

defined behavior states—normal flow, mild congestion, sudden halt, and dispersal—while maintaining real-time performance. Through extensive simulation and real-world validation, we quantify the system's accuracy, latency, and scalability, demonstrating its readiness for deployment in live surveillance networks.

LITERATURE REVIEW

The domain of automated crowd analysis spans decades of research across computer vision, machine learning, and human behavior modeling. We categorize prior work into three principal areas: density estimation, motion and flow modeling, and anomaly detection.

2.1 Density Estimation

Early density estimation approaches relied on pixel-level operations—background subtraction to isolate moving objects and blob counting to estimate head or torso counts. While simple, these methods degrade rapidly under occlusion and require manual threshold tuning. The advent of convolutional neural networks (CNNs) enabled density map regression, where a network outputs a continuous density field indicating the expected number of people per pixel region. CSRNet and MCNN architectures achieved substantial gains by using dilated convolutions to enlarge receptive fields without downsampling. However, their heavy computational footprints limit real-time deployment.

2.2 Motion and Flow Modeling

Optical flow techniques, originating with the Lucas-Kanade and Horn-Schunck algorithms, quantify pixel displacements between consecutive frames, offering a dense motion field. Researchers have applied flow histograms and vector clustering to interpret collective movement patterns. More recent works, such as Social Force Models, simulate inter-agent repulsion and attraction forces to predict pedestrian trajectories. Nonetheless, purely physics-based models struggle to generalize across diverse crowd behaviors, motivating hybrid data-driven approaches.

2.3 Anomaly Detection

Statistical anomaly detection leverages models of "normalcy" derived from historical video. Early methods applied Gaussian Mixture Models (GMMs) to flow vectors, flagging deviations beyond learned thresholds. Deep learning introduced autoencoders that learn compact representations of normal frames, with high reconstruction error indicating anomalies. GAN-based methods pit a generator against a discriminator to detect irregular frames. Despite impressive results, these techniques often yield high false positives when the training set lacks diverse normal examples.

2.4 Hybrid Architectures

To capture both spatial detail and temporal context, hybrid CNN-RNN pipelines have gained traction. Ibrahim et al. (2016) integrated a CNN encoder with an LSTM decoder, detecting anomalies in crowd motion sequences. Liang et al. (2019) extended this by constructing scene graphs linking individual trajectories to group behaviors. However, the computational complexity and latency of these models remain barriers to real-time application.

Our approach builds on this body of work by selecting an efficient CNN backbone (YOLOv4-Tiny) for rapid spatial encoding, pairing it with a BiLSTM for bidirectional temporal context, and incorporating an HMM layer to formalize behavior transitions. This design balances the need for expressive modeling with the imperative for real-time inference.

METHODOLOGY

This section details the architecture, data preparation, and training procedures used in our framework.

3.1 Data Preprocessing

- **Frame Sampling**: Input video streams are sampled at 20 fps. Frame rate was chosen to balance temporal resolution with computational load.
- **Normalization and Resizing**: Frames resized to 416×416 pixels; pixel intensities normalized to the [0, 1] range.
- **Background Subtraction**: A running Gaussian mixture model flags static background regions, enabling the CNN to focus on active areas containing crowd motion.
- **Data Augmentation**: To enhance robustness to illumination and viewpoint variation, we apply random flips, brightness shifts (±20%), and affine transformations during training.

3.2 Spatial Feature Extraction

- **Backbone Network**: We adopt YOLOv4-Tiny—featuring CSPDarknet53-Tiny—due to its favorable accuracy-throughput trade-off. The network outputs three detection heads corresponding to small, medium, and large pedestrian scales.
- **Region-level Embedding**: For each detection, we extract the feature map region and pass it through a 512-unit dense layer, producing a fixed-length embedding representing local appearance and context.

3.3 Temporal Sequence Modeling

- **Sequence Buffer**: We buffer embeddings across a sliding window of 30 frames (~1.5 s). Overlapping windows (stride = 10 frames) ensure continuity and reduce latency.
- **BiLSTM Architecture**: A two-layer bidirectional LSTM (256 units per direction) processes each sequence, capturing forward and backward temporal dependencies.
- **Regularization**: Dropout (0.4) applied between LSTM layers to prevent overfitting, along with L2 weight decay (1e-4).

3.4 Anomaly Inference

- **Hidden Markov Model (HMM)**: We fit an HMM on the BiLSTM output distributions for four behavior states (normal, congestion, halt, dispersal). Transition probabilities encode expected temporal progressions (e.g., normal→congestion, congestion→halt).
- **Anomaly Scoring**: For each window, we compute the negative log-likelihood of the observed state sequence under the HMM. Scores above a threshold (tuned on validation set at 0.65) trigger an alert.

3.5 Training and Implementation

- **Loss Functions**: We jointly optimize the YOLO detection loss (bounding box + classification + objectness) and a cross-entropy loss for sequence-level behavior classification.
- **Optimization**: Stochastic gradient descent with a cosine annealing learning-rate schedule, initial LR = 1e-3, batch size = 16 sequences.
- **Deployment**: Implemented in PyTorch with TorchScript export for C++ inference. On an NVIDIA RTX 2080, single-stream latency averages 40 ms/frame.

STATISTICAL ANALYSIS

We conducted quantitative comparisons on five synthetic scenarios covering densities from 0.5 to 3.0 persons/m² and motion types: steady flow, stop-and-go, and dispersal. Each scenario comprises 1,000 annotated frames. Models evaluated:

1. **Density-Only CNN** (CSRNet variant): density map regression + heuristic thresholding
2. **LSTM-Autoencoder**: CNN encoder + LSTM decoder reconstruction error

3. **Proposed CNN+BiLSTM+HMM**

Performance metrics (mean ± SD over five runs):

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	False Alarm Rate (%)
Density-Only CNN	78.2 ± 2.4	75.4 ± 3.1	80.1 ± 2.7	77.7 ± 2.9	18.5 ± 2.0
LSTM-Autoencoder	85.6 ± 1.8	83.2 ± 2.0	87.4 ± 1.5	85.3 ± 1.7	12.3 ± 1.4
Proposed CNN+BiLSTM+HMM	**92.3 ± 1.2**	**90.8 ± 1.5**	**93.7 ± 1.0**	**92.2 ± 1.2**	**6.9 ± 1.0**

Table 1. Performance comparison across models in simulated scenarios.

An ANOVA test confirms that differences across models are statistically significant for all metrics ($p < 0.001$). Post-hoc Tukey tests reveal our method significantly outperforms both baselines ($p < 0.01$), with the most pronounced gains in reducing false alarms.

SIMULATION RESEARCH

To emulate real-world surveillance challenges, we built a synthetic environment in Unity3D:

1. **Scene Configuration**: A 20 × 20 m plaza with four dynamic entry/exit zones and variable obstacle placement.
2. **Agent Dynamics**: 500 agents governed by a Social Force Model, programmed to execute normal flow, evacuation drills, and random dispersal patterns.
3. **Camera Network**: Four 1080p virtual cameras positioned to provide 90% scene coverage, streaming at 25 fps.

We evaluated throughput and latency under single-GPU and multi-GPU settings:

- **Single GPU (RTX 2080)**: Processes one stream at 25 fps; end-to-end latency 120 ms/frame.
- **Quad-Stream (×4)**: Sustains 18 fps per stream; latency 160 ms/frame.
- **Scale-Out**: On a four-GPU cluster, achieves ≥20 fps for 16 concurrent streams, confirming the framework's scalability.

Qualitative analysis of flagged anomalies shows precise localization of congestion pockets and timely alerts for abrupt dispersal episodes, with an average detection lead time of 1.2 s before manual annotation.

RESULTS

We validated on two real-world datasets:

1. **University Campus CCTV** (10 h footage, annotated for congestion): 90.1% accuracy, 88.7% precision, 91.5% recall. False alarm rate: 7.8%.
2. **Public Marathon Coverage** (3 h high-density segments): During peak density (>1.5 persons/m²), the system identified crowd surges with 93.4% F1-score. Alerts aligned within ±2 s of ground-truth events.

Robustness tests under varying illumination (day/night cycles) and occlusion scenarios (>60% overlap) indicate stable performance (<5% degradation). Operator feedback in a live pilot rated the system's alerts as actionable and reduced manual monitoring load by 70%.

CONCLUSION

This work presents a scalable, AI-driven framework for automated crowd behavior analysis in surveillance video streams. The integration of a YOLOv4-Tiny CNN backbone, BiLSTM temporal modeling, and HMM-based anomaly inference yields high accuracy (92.3%), low false alarm rates (6.9%), and real-time processing (≥20 fps on multi-GPU clusters). Extensive

simulation and real-world evaluations confirm the system's efficacy in diverse settings—from controlled synthetic plazas to crowded marathon routes.

Key contributions include:

- A unified pipeline balancing expressive power with computational efficiency.
- A probabilistic HMM layer that formalizes behavior transitions, reducing spurious alerts.
- Demonstrated scalability to multi-camera deployments.

Future enhancements will explore self-supervised pre-training to address labeled-data scarcity, multi-view depth fusion for 3D crowd reconstruction, and integration with edge-AI devices for on-site inference. By transforming reactive surveillance into proactive crowd management, this framework paves the way for smarter cities and safer public spaces.

REFERENCES

- *Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.*
- *Chan, A. B., Liang, Z.-S. J., & Vasconcelos, N. (2008). Privacy preserving crowd monitoring: Counting people without people models or tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1–8).*
- *Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Vol. 1, pp. 886–893).*
- *Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … & Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (Vol. 27, pp. 2672–2680).*
- *Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning temporal regularity in video sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 733–742).*
- *Helbing, D., & Molnár, P. (1995). Social force model for pedestrian dynamics. Physical Review E, 51(5), 4282–4286. https://doi.org/10.1103/PhysRevE.51.4282*
- *Ibrahim, M. S., Muralidharan, V., Chang, K., Ranganathan, A., & Ryoo, M. S. (2016). A hierarchical deep temporal model for group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1971–1980).*
- *Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR).*
- *Li, Y., Zhang, X., & Chen, D. (2018). CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1091–1100).*
- *Liang, X., Zheng, Y., Zhao, Y., Li, C., & Li, W. (2019). Crowd behavior analysis with spatio-temporal graph convolutional network. In Proceedings of the 27th ACM International Conference on Multimedia (pp. 1178–1186). https://doi.org/10.1145/3343031.3351129*
- *Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision (pp. 21–37).*
- *Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94*
- *Mahadevan, V., Li, W., Bhalodia, V., & Vasconcelos, N. (2010). Anomaly detection in crowded scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1975–1981).*
- *Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 779–788).*
- *Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.*
- *Rodriguez, M., Laptev, I., Sivic, J., & Audibert, J.-Y. (2011). Density-aware person detection and tracking in crowds. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2423–2430).*
- *Sabokrou, M., Fathy, M., Hoseini, M., & Klette, R. (2018). Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Computer Vision and Image Understanding, 172, 88–97. https://doi.org/10.1016/j.cviu.2018.05.009*
- *Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations (ICLR).*
- *Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2022). Scaled-YOLOv4: Scaling cross stage partial network. arXiv preprint arXiv:2011.08036.*

- *Zhang, C., Li, H., Wang, X., & Yang, X. (2016). Single-image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 589–597).*