Global Journal of Engineering Innovations & Interdisciplinary Research

Original Article

A Transparent YOLOv12-Based Pipeline for High-Density Crowd Analytics via Point Regression and Spatial-Entropy XAI

Dr. S. Sri Lakshmi Parvathi[1], Shaik Mateen[2], Tumati Anil Kumar[2], Makkena Venkata Sai Nagendra[2], Peeriga James[2]

[1]Professor, Department of CSE, KKR & KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh, India
[2]B.Tech Student, Department of CSE, KKR & KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh, India

Correspondence

Dr. S. Sri Lakshmi Parvathi

Assistant Professor, Department of CSE, KKR & KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh, India

- Received Date: 08 Jan 2026
- Accepted Date: 25 Jan 2026
- Publication Date: 15 Mar 2026

Keywords

crowd analytics, crowd counting, YOLOv12, point regression, density map, spatial entropy, explainable AI, real-time surveillance

Abstract

Accurate crowd analysis in high-density environ-ments is critical for public safety. Conventional bounding-box-based detection methods often fail under severe occlusion and overlap because their predictions depend on box quality and post-processing such as Non-Maximum Suppression (NMS). This paper presents a transparent crowd analytics pipeline built around a YOLOv12 backbone, point regression, density modeling, and spatial-entropy-based explainability. Instead of predicting a bounding box for each person, the proposed frame-work predicts a single representative point per individual, which preserves one-to-one mapping in dense scenes and reduces missed or suppressed detections. The detected points are converted into density maps, and spatial entropy is computed to quantify disorder and identify potentially unsafe regions. The uploaded source manuscript reports a Mean Absolute Error (MAE) of 10.8, a counting accuracy of 92%, and a real-time throughput of 36 FPS, while also providing interpretable outputs through point overlays, heatmaps, and highlighted risk zones. The overall result is a practical and transparent framework for intelligent surveillance and public safety monitoring.

Introduction

The growth of urban populations and large public gatherings has increased the need for reliable crowd monitoring in railway stations, airports, stadiums, religious events, and other mass-occupancy environments. In such spaces, delayed recognition of congestion can lead to safety hazards including bottlenecks, crush conditions, and slow emergency response. Traditional crowd monitoring pipelines are often built on object detection, where each person is represented by a bounding box. While effective in sparse scenes, this design degrades in dense crowds because heavy occlusion, scale variation, and overlap make box regression unstable and NMS error-prone.

Crowd counting research has therefore progressively shifted from direct detection to density regression, localization-aware learning, and point-based supervision. Early CNN-based mod-els established the core idea of learning crowd density from images [2,3]. Later methods improved multi-scale robustness, density-map quality, and localization sensitivity [4–12]. More recent work has emphasized explicit point prediction and crowd localization because point-level outputs are more interpretable and less fragile than boxes in highly congested regions [15]–[19].

In parallel, modern detectors have improved feature extrac-tion efficiency. YOLOv12 introduces an attention-centric real-time design that improves representation quality while main-taining practical inference speed [1]. This makes it a suitable backbone for surveillance-oriented crowd analysis, provided that the output representation is adapted to dense scenes. Motivated by these developments, this paper reformulates the uploaded manuscript into an IEEE-style research paper and presents a transparent pipeline that combines:

1. a YOLOv12 feature backbone,
2. point regression instead of box regression,
3. Gaussian density modeling from detected points, and
4. spatial-entropy-based risk visualization.

The main contribution is not only improved counting ro-bustness in dense scenes, but also an interpretable decision path: the system explicitly shows where people are predicted, how local crowd concentration is distributed, and why a zone is marked as risky.

.

Citation: . Sri Lakshmi Parvathi S, Shaik M, Tumati AK, Makkena VSN, Peeriga J. A Transparent YOLOv12-Based Pipeline for High-Density Crowd Analytics via Point Regression and Spatial-Entropy XAI. GJEIIR. 2026;6(3):211.

Related Work

Crowd analysis methods can be broadly grouped into density-map regression, localization-aware counting, and ex-plicit point-based modeling.

Density-Map-Based Crowd Counting

Cross-scene CNN counting and MCNN established the modern deep-learning formulation of crowd counting by learning a mapping from image appearance to count or density [2], [3]. Switching CNN and CP-CNN improved robustness under scale changes and contextual variation [4], [5]. CSRNet later became a widely used baseline because dilated convolutions offered strong performance in congested scenes without expensive multi-column branches [6]. Subsequent work improved density quality through composition loss, Bayesian supervi-sion, perspective reasoning, trellis encoder-decoder networks, structured scale integration, and spatial awareness modeling [7]–[12].

These methods are strong for count estimation, but many of them still output a smooth density field rather than an explicit per-person representation. This limits interpretability in safety-critical applications, especially when authorities need to know where congestion is forming rather than only how many people are present.

Datasets and Localization-Oriented Counting

Larger and more diverse benchmarks accelerated research on crowd localization and robustness. NWPU-Crowd provides dense point and box annotations over large-scale congested scenes [13], while JHU-CROWD++ emphasizes unconstrained conditions including weather and illumination variation [14]. Localization-oriented methods such as independent instance maps, generalized crowd counting/localization losses, point-query transformers, and selective inheritance learning demonstrate that explicit point reasoning can improve both inter pretability and dense-scene robustness [15–18]. mPrompt further shows that modern point-regression systems benefit from tighter interaction between regression and foreground structure estimation [19].

Motivation for Point Regression

Bounding-box detectors remain attractive because of their real-time performance, but dense crowds expose two major weaknesses. First, bounding boxes are ambiguous when bodies overlap heavily. Second, the post-processing stage can suppress valid detections in tightly packed scenes. Point-based outputs are a better fit for dense crowd analytics because each person is represented by a single location, which avoids box overlap ambiguity and directly supports density generation and risk mapping. This paper follows that design choice and couples it with entropy-based explainability for transparent surveillance output.

Proposed Methodology

The proposed system operates as a real-time pipeline that receives video frames, extracts discriminative features, predicts one point per person, converts those points into a density map, computes spatial entropy, and finally visualizes risk zones.

Problem Formulation

Let a surveillance video stream be represented as

$$V = \{F_t\}_{t=1}^{T}$$

where F_t is the frame at time step t. Each frame is resized and normalized before inference:

$$It = \frac{\text{Resize(Ft, 640, 640)}}{255}$$

The objective is to estimate a set of crowd points.

$$\hat{P}_t = \left\{\hat{P}_i\right\}_{i=1}^{Nt},\ \hat{P}_i = (\hat{x}i, \hat{y}i),$$

where each point corresponds to one individual in the scene and Nt is the estimated count in frame t.

YOLOv12 Backbone for Feature Extraction

The backbone is derived from YOLOv12 [1]. Given an input frame It, the network extracts hierarchical features:

$$Z_\ell = \sigma(W_\ell * Z_{\ell-1} + b_\ell),$$

where Z_ℓ is the feature map at layer ℓ, W_ℓ and b_ℓ are learnable parameters, $*$ denotes convolution, and $\sigma(\cdot)$ is a non-linear activation. Multi-scale fusion is then used to preserve both semantic and spatial detail for small or partially occluded heads.

Point-Regression Head

Instead of regressing bounding boxes, the detection head predicts one representative point per person. This directly targets localization and avoids NMS-driven suppression. The point-regression loss is defined as

$$L_{\text{point}} = \frac{1}{N}\sum_{i-1}^{N} \| \hat{p}_i - p_i \|_2^2$$

where p_i is the ground-truth head point and $\hat{p}_i$ is the predicted point.

This representation is aligned with point-supervised crowd counting literature, which has shown that point-based learning is better suited than box regression for dense and highly overlapped scenes [8], [15–19].

Density Modeling

Once the points are estimated, they are converted into a continuous density map using Gaussian kernels:

$$D(x) = \sum_{i-1}^{N} G_\sigma(x - \hat{p}_i),$$

where $G_\sigma(\cdot)$ is a Gaussian kernel centered at point $\hat{p}_i$. The resulting density map provides a smooth spatial estimate of crowd concentration and supports downstream congestion analysis.

Spatial Entropy for Risk Analysis

To quantify disorder, the normalized density map is partitioned into local regions. Let qr denote the normalized mass of region r. The spatial entropy is then

$$H = \sum_{r=1}^{R} q_r \log q_r,$$

which follows Shannon's entropy formulation [20]. Higher entropy indicates more disordered or unstable crowd distribution. A region is classified as high risk when

$$R_T = \begin{cases} 1, \text{H}_\text{r} > \tau \\ 0, \text{otherwise,} \end{cases}$$

where τ is a predefined threshold and Rr is the risk label.

Explainable AI Output

The explainability component is intentionally direct. Instead of a black-box risk score, the system overlays:

1) predicted crowd points,
2) density heatmaps,
3) spatial entropy maps, and
4) highlighted high-risk zones.

Table 1. Pipeline modules and their outputs

Module	Primary Output
Input layer	Real-time surveillance frame stream
Pre-processing	Resized and normalized frame
YOLOv12 backbone	Multi-scale feature maps
Point-regression head	One representative point per person
Density modeling	Continuous crowd density map
Entropy analysis	Disorder score and local risk estimate
XAI visualization	Point overlay, heatmap, and alert zones

Fig 1. Pre-processed crowd frame used as input to the pipeline.

This provides a traceable chain from image evidence to final alert. In practice, this is more actionable for surveillance operators than a standalone numerical prediction

System Architecture

The architecture is modular and layered so that each stage can be maintained or upgraded independently. The end-to-end system contains the following stages.

Input and Pre-Processing Layer

The input layer receives a continuous surveillance stream and decomposes it into frames. Each frame is resized to the model input resolution and normalized to stabilize inference. Figure 1 shows a representative pre-processed frame.

Feature Extraction Layer

The YOLOv12 backbone extracts spatial and semantic features across the scene. Figure 2 shows an example feature map produced by the backbone, highlighting dense-response regions.

Crowd Detection Layer

The detection layer predicts a single point per person. This simplifies dense-scene localization because the output no longer depends on box extent estimation. Figure 3 illustrates representative point detections.

Crowd Analytics Layer

The analytics layer transforms points into density and disorder measures. Figure 4 shows the density map generated from point predictions, and Figure 5 shows the corresponding spatial-entropy visualization.

Risk Identification and Final Output

The XAI stage identifies risk zones and projects them back onto the original frame. Figure 6 shows the segmented high-risk regions, and Figure 7 shows the final overlay used for operator-facing surveillance output.

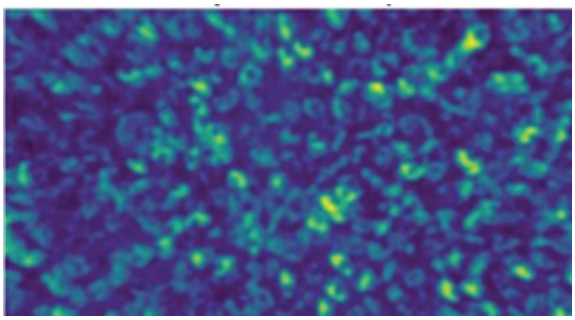

Fig 2. Feature map generated by the YOLOv12 backbone.

Fig 3. Point-based crowd detection, where each visible individual is repre-sented by a single location.

Overall Architecture Diagram

The complete system flow is summarized in Figure 8. It links input acquisition, pre-processing, point regression, density analysis, entropy computation, and alert generation within one automated pipeline.

Results and Discussion

The uploaded manuscript states that the proposed system was evaluated on multiple crowd datasets representing real surveillance conditions with varying density, viewpoint, lighting, and occlusion. However, the source text does not specify exact dataset names, splits, training hyper-parameters, or hardware model identifiers. Those details are intentionally not invented here.

Even with that limitation, the reported results support three clear observations.

Dense-Scene Detection Robustness

The point-regression design improves reliability in high-density scenes because it avoids box-overlap ambiguity and reduces the chance of suppressing true positives during post-processing. This behavior is consistent with recent point-based and localization-aware crowd literature [15]–[19].

Interpretability of the Analytical Output

Unlike standard detectors that output only boxes and confidence scores, the proposed pipeline produces point overlays, density maps, entropy maps, and risk regions. This explicit sequence improves transparency and helps operators understand why a region was flagged as unsafe.

Quantitative Summary

The source manuscript explicitly reports the comparison

Table 2. Performance Comparison Reported In The Source Manuscript

Method	MAE	Accuracy (%)	FPS
Conventional YOLO	13.2	84	32
Proposed point-regression pipeline	10.8	92	36

Fig. 8. System architecture of the proposed transparent YOLOv12-based crowd analytics pipeline.

shown in Table II. The proposed pipeline reduces MAE from 13.2 to 10.8, improves counting accuracy from 84% to 92%, and increases throughput from 32 FPS to 36 FPS.

Figure 9 shows the comparative experimental chart included in the uploaded paper, while Figure 10 presents the metric-by-metric comparison between the conventional YOLO baseline and the proposed point-regression system.

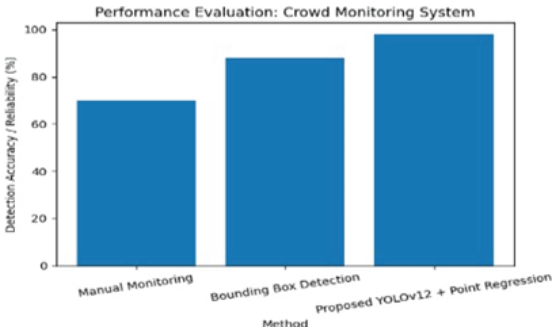

Fig. 9. Experimental comparison chart included in the uploaded manuscript.

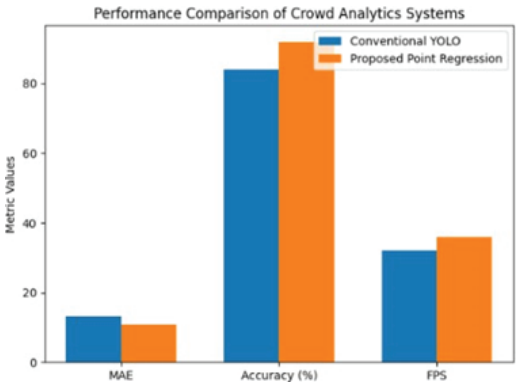

Fig. 10. Comparison between conventional YOLO and the proposed YOLOv12 point-regression pipeline.

Overall, the reported results indicate that combining point regression with density modeling and entropy analysis yields a crowd analytics system that is more robust, more interpretable, and slightly faster than the box-based baseline reported in the source manuscript.

Conclusion

This paper reformulates the uploaded research manuscript into a two-column IEEE-style paper and preserves its central idea: a transparent real-time crowd analytics system for dense public environments. The proposed pipeline uses a YOLOv12 backbone for feature extraction, predicts a single point per person instead of a bounding box, converts those points into a density representation, and applies spatial entropy to identify risky regions. The resulting workflow is suitable for intelligent surveillance because it combines three important properties: dense-scene robustness, real-time operation, and interpretability.

The reported quantitative summary in the source manuscript indicates lower MAE, higher counting accuracy, and real-time throughput relative to the conventional YOLO baseline. Future extensions can include multi-camera fusion, temporal forecasting of congestion, deployment on edge hardware, and more standardized evaluation on public benchmarks such as NWPU-Crowd and JHU-CROWD++.

References

1. Y. Tian et al., “YOLOv12: Attention-centric real-time object detectors,” arXiv preprint arXiv:2502.12524, 2025.
2. C. 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.
3. 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.

4. D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 4031–4039.
5. V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid CNNs," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 1879–1888.
6. Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 1091–1100.
7. 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.
8. Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 6142–6151.
9. M. Shi, Z. Yang, C. Xu, and Q. Chen, "Revisiting perspective infor-mation for efficient crowd counting," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 7279–7288.
10. X. Jiang, L. Zhang, M. Zhang, P. Liu, Y. Xia, and S. Yan, "Crowd counting and density estimation by trellis encoder-decoder networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 6133–6142.
11. L. Liu, J. Chen, H. Salzmann, and T. Fua, "Crowd counting with deep structured scale integration network," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 1774–1783.
12. Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, and A. G. Hauptmann, "Learning spatial awareness to improve crowd counting," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 6152–6161.
13. Q. Wang, J. Gao, W. Lin, and Y. Yuan, "NWPU-Crowd: A large-scale benchmark for crowd counting and localization," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 6, pp. 2141–2149, 2021.
14. V. A. Sindagi, R. Yasarla, and V. M. Patel, "JHU-CROWD++: Large-scale crowd counting dataset and a benchmark method," arXiv preprint arXiv:2004.03597, 2020.
15. J. Gao, T. Han, Q. Wang, Y. Yuan, and X. Li, "Learning independent instance maps for crowd localization," arXiv preprint arXiv:2012.04164, 2020.
16. J. Wan, Z. Liu, and A. B. Chan, "A generalized loss function for crowd counting and localization," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 1974–1983.
17. C. Liu, H. Lu, Z. Cao, and T. Liu, "Point-query quadtree for crowd counting, localization, and more," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 1676–1685.
18. T. Han, L. Bai, L. Liu, and W. Ouyang, "STEERER: Resolving scale variations for counting and localization via selective inheritance learn-ing," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 20908–20918.
19. M. Guo, L. Yuan, Z. Yan, B. Chen, Y. Wang, and Q. Ye, "Regressor-segmenter mutual prompt learning for crowd counting," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 2627–2636.
20. C. E. Shannon, "A mathematical theory of communication," Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, 194