Enhancing Surveillance System through Video Anomaly Detection

Shivin Bhagwan Tarare[1] Mrs. Vrushali Shinde[2]
[1,2]Master of Computer Applications
[1,2]Modern College of Engineering, Pune, Maharashtra, India

Abstract — The rapid proliferation of surveillance infrastructure in urban environments necessitates intelligent systems capable of autonomous monitoring and real-time threat assessment. Traditional surveillance frameworks relying on manual operator oversight are susceptible to cognitive fatigue, scalability constraints, and delayed response times. This paper presents an AI-based intelligent surveillance system integrating three functional modules: deep learning-based person detection using YOLOv8, multi-object tracking via DeepSORT, and cross-camera person re-identification (ReID) employing convolutional neural network-based feature extraction. The proposed end-to-end pipeline maintains identity consistency across non-overlapping camera views under challenging conditions including occlusion, illumination variability, and significant pose changes. Experimental evaluation on a custom multi-camera dataset demonstrates a mean Average Precision (mAP@0.5) of 91.4% for detection, a Multi-Object Tracking Accuracy (MOTA) of 82.5%, and a Rank-1 re-identification accuracy of 87.2% on the Market-1501 benchmark. The complete pipeline sustains an average throughput of 22.3 FPS on mid-range GPU hardware, confirming real-time operational feasibility. The system establishes a robust, extensible foundation for next-generation intelligent surveillance with planned extensions toward behavioral anomaly detection and automated alert generation.

Keywords: Person Re-Identification; Multi-Object Tracking; Object Detection; YOLOv8; DeepSORT; Convolutional Neural Networks; Intelligent Surveillance; Anomaly Detection; Computer Vision

I. INTRODUCTION

The proliferation of closed-circuit television (CCTV) cameras in urban environments, transportation hubs, and public infrastructure has fundamentally transformed modern security practices. Global installed surveillance camera count is estimated to exceed 770 million units, generating continuous, large-volume video streams that rapidly outpace the capacity for human review [1]. Traditional surveillance systems place the interpretive burden squarely on human operators, who must simultaneously monitor multiple video feeds under sustained cognitive load. Such conditions are inherently prone to fatigue-induced lapses, introducing critical monitoring gaps precisely when vigilance is most essential [2].

Recent advances in artificial intelligence (AI) and deep learning have opened new avenues for automated video analysis. Convolutional neural networks (CNNs) [10,11] have revolutionized the field of object detection, enabling real-time pedestrian localization in complex video streams. Single-stage detectors—most notably the YOLO (You Only Look Once) family [1,2,3]—achieve state-of-the-art detection performance through a unified inference pipeline that processes entire frames in a single forward pass. Multi-object tracking (MOT) algorithms such as SORT [5] and DeepSORT [4] enable robust temporal association of detected entities across consecutive video frames. Person re-identification (ReID) extends this capability by enabling cross-camera identity association across spatially separated, non-overlapping fields of view [6,7,13].

Despite individual progress in detection, tracking, and ReID, many existing surveillance deployments treat these capabilities as isolated modules, resulting in identity discontinuities at camera handoff boundaries, operational redundancy, and reduced situational awareness. A compelling need exists for a unified, end-to-end framework that cohesively integrates these components into a single identity-aware surveillance pipeline [12]. This research addresses this gap by proposing and evaluating such an integrated system.

The primary contributions of this paper are as follows: (1) design and implementation of a real-time person detection module using YOLOv8 optimized for multi-camera surveillance; (2) robust multi-object tracking via DeepSORT incorporating motion estimation and appearance-based re-association; (3) cross-camera person ReID using deep CNN feature extraction and cosine similarity matching; (4) comprehensive quantitative evaluation on a custom multi-camera dataset using standard MOT and ReID metrics; and (5) a modular, extensible pipeline architecture suitable for future integration of anomaly detection and behavioral analysis.

II. LITERATURE REVIEW

The trajectory of automated surveillance systems spans several decades of research. Early motion detection methods relied on handcrafted image processing techniques—including Gaussian mixture model-based background subtraction, temporal frame differencing, and optical flow estimation. While computationally lightweight, these approaches exhibited brittleness under dynamic background conditions, shadow artifacts, and illumination fluctuations, leading to prohibitively high false alarm rates in practical deployments.

Deep learning fundamentally transformed the detection landscape. The foundational convolutional architectures established by LeCun et al. [11] were scaled to large-scale visual recognition tasks by Krizhevsky et al. [10], demonstrating the representational power of deep CNNs. Ren et al. [9] introduced Faster R-CNN, which employed region proposal networks for accurate two-stage object detection with competitive inference speed. The YOLO family [1,2] demonstrated that single-stage architectures could achieve competitive accuracy at dramatically higher inference speeds, unlocking practical real-time deployment. YOLOv8 [3], the latest iteration, incorporates an anchor-free detection head and CSP bottleneck architecture that yields superior accuracy-speed trade-offs compared to earlier versions.

In the multi-object tracking domain, Bewley et al. [5] proposed SORT, combining Kalman filtering for

kinematic state estimation with the Hungarian algorithm for bipartite detection-to-track assignment, achieving real-time performance at the cost of identity fragmentation under occlusion. DeepSORT [4] addressed this limitation by integrating deep appearance descriptors into the association cost matrix, substantially reducing identity switch frequency. More recently, ByteTrack [16] demonstrated that leveraging both high- and low-confidence detections during association yields further improvements in dense crowd scenarios. A comprehensive survey by Luo et al. [12] categorizes MOT approaches by their motion modeling and association strategies, providing context for the design choices adopted in this work.

Person re-identification has matured rapidly with the availability of large-scale benchmarks. Zheng et al. [6] introduced the Market-1501 dataset, establishing a standard evaluation protocol that subsequently catalyzed significant research activity. Vezzani et al. [13] surveyed early ReID methods in surveillance contexts, characterizing the challenges of viewpoint variation, occlusion, and intra-class appearance variability. Subsequent deep learning approaches employed ResNet-based [8] encoders with metric learning objectives—including triplet loss and contrastive loss—to learn compact, discriminative identity embeddings. Ye et al. [7] provide a comprehensive review and outlook covering supervised, unsupervised, and domain-adaptive ReID paradigms. Zhang et al. [14] proposed densely semantically aligned representations that improve robustness to partial occlusion, a persistent challenge in surveillance environments.

Multi-object tracking and segmentation (MOTS), introduced by Voigtlaender et al. [15], represents an emerging convergence of detection, tracking, and pixel-level segmentation for richer scene understanding. While promising in controlled settings, such approaches introduce substantial computational overhead that currently limits practical deployment on mid-range hardware. The present work focuses on balancing detection accuracy, tracking robustness, and ReID capability within a real-time operational envelope achievable on commodity GPU hardware.

III. Methodology

A. System Architecture Overview

The proposed system adopts a three-stage modular pipeline. Input video frames from one or more cameras are first processed by the Detection Module, which localizes all human subjects and produces bounding box proposals with associated confidence scores. Detected bounding boxes are forwarded to the Tracking Module, which assigns unique track identifiers and maintains temporal identity continuity across consecutive frames within a single camera view. Finally, the Identity Association Module extracts appearance feature embeddings for each active track and performs cross-camera matching to establish identity correspondence across non-overlapping camera fields of view. Annotated output frames, structured track logs, and cross-camera identity maps are produced as system outputs.

The pipeline architecture may be formally described as:

Input Video → Detection → Tracking → Identity Association → Annotated Output

Each module exposes a well-defined interface, enabling individual components to be independently upgraded, replaced, or disabled without disrupting overall pipeline operation.

B. Detection Module

The detection module employs YOLOv8 [3] as its core inference engine. YOLOv8 adopts an anchor-free detection paradigm that eliminates reliance on predefined anchor box templates, simplifying post-processing and improving generalization to atypical aspect ratios common in surveillance imagery. The backbone consists of C2f (Cross Stage Partial with two bottlenecks) modules that provide efficient multi-scale feature extraction, coupled with a PANet neck for hierarchical feature aggregation across spatial scales. The decoupled head independently regresses bounding box coordinates and predicts class probabilities, improving gradient flow during training. The model is initialized from COCO-pretrained weights and fine-tuned on the custom surveillance dataset. Non-Maximum Suppression (NMS) with a confidence threshold of 0.45 and an IoU threshold of 0.50 is applied to refine candidate detections.

C. Multi-Object Tracking Module

The tracking module is constructed on the DeepSORT framework [4], which extends the foundational SORT algorithm [5] with appearance-based re-association. Each active track is modeled by a Kalman filter operating on an eight-dimensional state vector comprising bounding box centroid coordinates, aspect ratio, height, and their first-order time derivatives. Detection-to-track assignment is formulated as a linear sum assignment problem solved via the Hungarian algorithm [5], minimizing a composite cost function that integrates Mahalanobis distance (quantifying kinematic affinity) and cosine distance (quantifying appearance affinity). A lightweight CNN-based appearance encoder, pretrained on a ReID dataset, generates 128-dimensional L2-normalized embedding vectors for each detection crop. The maximum track age parameter is configured to accommodate short-term occlusions without premature track termination.

D. Person Re-Identification Module

The re-identification module employs a ResNet-50 [8] backbone, pretrained on ImageNet [10] and fine-tuned on the Market-1501 benchmark [6], to extract 2048-dimensional identity feature embeddings. Global average pooling compresses the spatial feature map of the final convolutional stage into a fixed-length person descriptor, which is subsequently L2-normalized. Cross-camera identity matching is performed by computing pairwise cosine similarity between probe embeddings (detections from a query camera) and gallery embeddings (accumulated detections from reference cameras). Identities are assigned using a nearest-neighbor retrieval strategy with a configurable similarity threshold. A rolling gallery is maintained and periodically pruned to accommodate long-term appearance changes while controlling memory consumption [7].

E. System Design Considerations

The architecture is organized around four key engineering principles: (1) *Scalability*—the modular design permits horizontal extension to accommodate additional camera streams with linear resource scaling; (2) *Real-time Processing*—each processing stage is optimized for low-latency inference, collectively sustaining greater than 22 FPS end-to-end throughput on mid-range GPU hardware; (3) *Robustness*—appearance-based re-association and deep ReID embeddings provide resilience against occlusion, pose variation, and illumination changes; and (4) *Modularity*—standardized inter-module interfaces enable independent development, testing, and deployment of individual components.

IV. Experimental Setup and Dataset

A. Hardware and Software Configuration

All experiments were performed on a workstation equipped with an Intel Core i7-10750H CPU (6 cores, 2.60 GHz base clock), 16 GB DDR4 RAM, and an NVIDIA GeForce GTX 1660 Ti GPU (6 GB GDDR6). The software stack comprised Python 3.10, PyTorch 2.0.1 with CUDA 11.8, OpenCV 4.8.0, and the Ultralytics YOLOv8 library. All models were executed in inference mode without test-time augmentation to ensure realistic real-time performance benchmarking.

B. Datasets

Two datasets were employed for evaluation. A custom multi-camera surveillance dataset was constructed comprising 1,200 annotated video clips captured across four cameras deployed in a campus environment with fully non-overlapping fields of view. The dataset encompasses 48 unique individuals under varied clothing, lighting conditions, and time periods. Ground-truth annotations for bounding boxes and track identities were generated using a semi-automated labeling pipeline with manual verification. In addition, the Market-1501 benchmark [6], containing 32,668 images of 1,501 identities across six cameras, was employed for standardized re-identification evaluation.

C. Evaluation Metrics

Detection performance was quantified using mean Average Precision at IoU threshold 0.5 (mAP@0.5), Precision, Recall, and inference speed in frames per second (FPS). Tracking performance was assessed using Multi-Object Tracking Accuracy (MOTA) and Precision (MOTP) [12], the number of Identity Switches (ID-Sw), and FPS. Re-identification performance was measured using Rank-1 and Rank-5 Cumulative Matching Characteristic (CMC) accuracies and mean Average Precision (mAP).

V. Results and Analysis

A. Detection Performance

The YOLOv8 detection module was evaluated on a 20% held-out test split of the custom dataset. As shown in Table I, the model achieves a mAP@0.5 of 91.4%, demonstrating reliable pedestrian localization across diverse lighting conditions and camera viewpoints. The precision of 89.7% and recall of 87.3% reflect a well-calibrated detection threshold, with limited false positives from static background objects. The inference speed of 28.6 FPS on the GTX 1660 Ti confirms real-time operational capacity for single-stream surveillance.

Metric	mAP@0.5	Prec.	Recall	FPS
YOLOv8	91.4%	89.7%	87.3%	28.6

Table I: Detection Module Performance (YOLOv8 on Custom Dataset)

B. Tracking Performance

Table II reports tracking performance of DeepSORT evaluated on the custom multi-camera dataset, with SORT included as a baseline. DeepSORT achieves a MOTA of 82.5%, representing an 8.3 percentage point improvement over SORT (74.2%). Most significantly, the identity switch count is reduced from 41 (SORT) to 14 (DeepSORT)—a 65.9% reduction—attributable to the appearance descriptor's disambiguation capability under short-term occlusion. The MOTP of 79.3% indicates accurate bounding box localization throughout tracking. The tracking pipeline operates at 24.1 FPS, maintaining real-time performance despite the additional appearance encoding overhead.

Method	MOTA	MOTP	ID-Sw	FPS
SORT [5]	74.2%	76.1%	41	31.4
DeepSORT [4]	82.5%	79.3%	14	24.1

Table II: Tracking Performance Comparison on Custom Dataset

C. Re-Identification Performance

The ReID module was evaluated on the Market-1501 benchmark following the standard evaluation protocol. Table III presents results alongside a published ResNet-50 baseline from Zhang et al. [14]. The proposed implementation achieves Rank-1 accuracy of 87.2% and Rank-5 accuracy of 94.6%, with an mAP of 76.8%—exceeding the published ResNet-50 baseline. This improvement is attributed to augmented training data preprocessing and cosine annealing learning rate scheduling employed during fine-tuning. The lower mAP relative to Rank-1 accuracy reflects the inherent challenge of exhaustively retrieving all positive gallery instances, a well-documented characteristic of the Market-1501 evaluation protocol [6].

Backbone	Rank-1	Rank-5	mAP
ResNet-50 [14]	85.9%	93.4%	74.5%
ResNet-50 (Ours)	87.2%	94.6%	76.8%

Table III: Re-Identification Performance on Market-1501

D. System-Level Performance Summary

Table IV consolidates the end-to-end pipeline performance. The integrated system sustains an average throughput of 22.3 FPS on the evaluation hardware, exceeding the 20 FPS threshold commonly accepted for real-time surveillance applications. The modest throughput reduction relative to the standalone detection rate (28.6 FPS) reflects the additional latency introduced by the tracking and ReID modules. The pipeline incurs no measurable accuracy penalty from module integration, as each component operates on its own dedicated processing thread.

Module	Metric	Value
Detection (YOLOv8)	mAP@0.5	91.4%
Detection (YOLOv8)	FPS	28.6

Tracking (DeepSORT)	MOTA	82.5%
Tracking (DeepSORT)	ID-Sw	14
Re-ID (ResNet-50)	Rank-1	87.2%
Re-ID (ResNet-50)	Rank-5	94.6%
Full Pipeline	Avg. FPS	22.3

Table IV: End-to-End System Performance Summary

VI. Discussion

The experimental results confirm that the integrated pipeline delivers competitive performance across all three functional modules while maintaining real-time operational throughput. The YOLOv8 detector's anchor-free design proves particularly advantageous in surveillance contexts, where pedestrian bounding boxes exhibit high scale variability due to camera placement and subject distance. Fine-tuning on the domain-specific dataset yielded measurable precision and recall improvements over direct deployment of COCO-pretrained weights, validating the value of domain adaptation even with limited labeled data.

The DeepSORT tracker's 65.9% reduction in identity switches relative to SORT demonstrates the substantial benefit of incorporating appearance information into the data association stage. Qualitative inspection reveals that remaining identity switches occur predominantly in high-density crowd regions, where spatial proximity renders Mahalanobis distance-based association unreliable. Integration of graph neural network-based or transformer-based global association mechanisms [16] may address this limitation in future iterations.

The ReID module achieves Rank-1 accuracy that modestly surpasses the published ResNet-50 baseline on Market-1501, lending confidence to the training pipeline implementation. Failure analysis reveals that the predominant error modes involve extreme appearance change (e.g., significant clothing modification) and inter-camera illumination discontinuities exceeding the model's learned invariance. The rolling gallery update strategy partially mitigates long-term appearance drift, but temporal modeling of identity evolution remains an active research frontier [7].

Several limitations merit consideration. First, the detection module's accuracy directly bounds downstream tracking and ReID performance—missed detections introduce irrecoverable gaps in track continuity. Second, the system does not currently implement higher-level behavioral interpretation, limiting autonomous threat identification. Third, deployment in large-scale environments with many simultaneous cameras will require hardware acceleration strategies, model quantization, or edge-cloud hybrid architectures to maintain acceptable latency.

VII. Conclusion and Future Work

This paper presented an AI-based intelligent surveillance system integrating YOLOv8-based person detection, DeepSORT multi-object tracking, and ResNet-50-based person re-identification into a unified real-time processing pipeline. The system achieves 91.4% mAP@0.5 for detection, 82.5% MOTA for tracking, and 87.2% Rank-1 ReID accuracy on the Market-1501 benchmark, while sustaining 22.3 FPS end-to-end throughput on mid-range GPU hardware. These results confirm the practical feasibility of deploying integrated intelligent surveillance without high-end infrastructure investment.

The modular pipeline provides a robust foundation for future enhancements, including: (1) transformer-based detection and tracking architectures for improved performance in dense crowd scenes; (2) spatiotemporal behavioral anomaly detection for autonomous identification of loitering, intrusion, and crowd surge events; (3) automated alert generation and integration with security management platforms; (4) model compression, pruning, and quantization for embedded edge deployment; and (5) large-scale evaluation on public multi-camera benchmarks such as DukeMTMC-reID and MSMT17 to assess broader generalizability.

References

[1] J. 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), Las Vegas, NV, USA, 2016, pp. 779–788.

[2] A. Hampapur, L. Brown, J. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, and S. Pankanti, "Smart Video Surveillance: Exploring the Concept of Multiscale Spatiotemporal Tracking," IEEE Signal Process. Mag., vol. 22, no. 2, pp. 38–51, Mar. 2005.

[3] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," version 8.0.0, Ultralytics, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[4] N. Wojke, A. Bewley, and D. Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric," in Proc. IEEE Int. Conf. Image Process. (ICIP), Beijing, China, 2017, pp. 3645–3649.

[5] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple Online and Realtime Tracking," in Proc. IEEE Int. Conf. Image Process. (ICASSP), Shanghai, China, 2016, pp. 3464–3468.

[6] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable Person Re-Identification: A Benchmark," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, 2015, pp. 1116–1124.

[7] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep Learning for Person Re-Identification: A Survey and Outlook," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 28, 2015, pp. 91–99.

[10] A. 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.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document

Recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[12] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T.-K. Kim, "Multiple Object Tracking: A Literature Review," Artif. Intell., vol. 293, p. 103448, Apr. 2021.
[13] R. Vezzani, D. Baltieri, and R. Cucchiara, "People Reidentification in Surveillance and Forensics: A Survey," ACM Comput. Surv., vol. 46, no. 2, pp. 1–37, Nov. 2013.
[14] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Densely Semantically Aligned Person Re-Identification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, 2019, pp. 667–676.
[15] P. Voigtlaender et al., "MOTS: Multi-Object Tracking and Segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, 2019, pp. 7942–7951.
[16] Y. Zhang et al., "ByteTrack: Multi-Object Tracking by Associating Every Detection Box," in Proc. Eur. Conf. Comput. Vis. (ECCV), Tel Aviv, Israel, 2022, pp. 1–21.

