

A Hybrid Deep Learning Framework for Multi-Camera Object Tracking and Anomaly Detection in Surveillance

¹ Dr. M. Mohan, ² Dr. K. Sivakami, ³ Dr. A. Nesarani, ⁴ A. Roopasree, ⁵ Dr. Arun G, ⁶ Dr. K. Kalpana

¹ Assistant Professor, Department of Computer Science and Engineering (AIML), SRM Institute of Science and Technology, Ramapuram Campus, Chennai, Tamil Nadu, India.

Email: mohan.rm@gmail.com , ORCID ID: 0000-0002-3943-2930

² Assistant Professor, Department of Electronics and Communication Engineering, Karpagam Institute of Technology, Coimbatore, Tamil Nadu, India.

Email: ksivakaamii@gmail.com , ORCID ID: 0000-0002-2985-7206

³ Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (Deemed to be University), Green Fields, Vaddeswaram, Guntur, Andhra Pradesh, India.

Email: nesaraniabraham84@kluniversity.in , ORCID ID: 0000-0002-7604-9957

⁴ Assistant Professor, Department of Electronics and Communication Engineering, Hindusthan Institute of Technology, Coimbatore, Tamil Nadu, India.

Email: roopaece91@gmail.com , ORCID ID: 0009-0005-8387-2350

⁵ Associate Professor, Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India.

Email: arungguru@gmail.com , ORCID ID: 0009-0001-5049-1106

⁶ Associate Professor, Department of Electronics and Communication Engineering, V.S.B College of Engineering Technical Campus, Coimbatore, Tamil Nadu, India.

Email: kalpanakasilingam81@gmail.com , ORCID ID: 0000-0002-7823-4267

Submitted: Sep, 17, 2025 **Revised:** Dec, 10, 2025 **Accepted:** Dec, 21, 2025

Abstract:

Modern surveillance systems face increasing challenges in real-time object tracking, behavioral analysis, and anomaly detection across distributed camera networks. Traditional single-camera solutions fail to address spatial discontinuity and occlusion issues in multi-camera environments. To overcome these challenges, this paper proposes a hybrid deep learning framework integrating Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) with attention-based data fusion for multi-camera object tracking and anomaly detection in large-scale surveillance networks. The proposed framework employs a CNN-based feature extractor to identify and encode object characteristics across multiple camera streams, followed by a spatio-temporal attention module that aligns overlapping regions and correlates object trajectories between cameras. The RNN-LSTM network models temporal dependencies to track object motion patterns, while an autoencoder-based anomaly detector identifies abnormal behaviors such as unattended baggage, unauthorized entry, or erratic movement. A multi-view consistency algorithm ensures smooth cross-camera transition tracking. Experimental evaluation on PETS 2009, TownCentre, and Avenue datasets demonstrates the framework's superiority, achieving a tracking accuracy (MOTA) of 94.7%, an IDF1 score of 92.5%, and an anomaly detection accuracy of 96.2%, outperforming existing CNN-LSTM and transformer-based baselines. The proposed hybrid framework enhances situational awareness, robustness, and scalability, making it suitable for deployment in smart cities, public safety monitoring, and industrial surveillance.

Keywords: Multi-camera object tracking; anomaly detection; deep learning; convolutional neural network (CNN); recurrent neural network (RNN); attention mechanism; spatio-temporal fusion; autoencoder; surveillance systems; smart city security.

1. Introduction

Video surveillance has become a critical component in intelligent transportation systems, public safety, smart cities, and critical infrastructure monitoring [1], [2]. Modern surveillance deployments often consist of dozens or even hundreds of cameras with partially overlapping fields of view (FOVs), varying resolutions, illumination conditions, and perspectives [3]. In such settings, accurate tracking of individuals or objects across multiple cameras and the reliable detection of abnormal behaviour are fundamental requirements for automated scene understanding [4].

Traditional single-camera tracking approaches, which rely on hand-crafted features and heuristic association strategies, often fail under occlusions, abrupt motion changes, drastic viewpoint variations, or when the target temporarily exits the field of view [5], [6]. Furthermore, rule-based anomaly detection systems, such as those based on static thresholds or manually defined regions of interest, are unable to generalize to complex and evolving patterns of human and vehicular activity [7].

With advances in deep learning, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have revolutionized object detection, tracking, and sequence modelling [8]. Deep Re-ID models have shown strong capabilities in matching individuals across cameras, while generative and reconstruction-based models have successfully captured normal motion patterns for anomaly detection [9]. However, most existing works treat multi-camera tracking and anomaly detection as independent problems, or they are restricted to single-camera scenarios with limited cross-view reasoning [10].

In this work, we propose a unified hybrid deep learning framework that jointly addresses multi-camera object tracking and anomaly detection in real-world surveillance settings. The core idea is to construct globally consistent trajectories across cameras and then leverage their spatio-temporal signatures to identify abnormal events such as loitering, running in restricted zones, sudden crowd formation, and unusual motion paths.

Main Contributions

The key contributions of this paper are as follows:

1. Hybrid multi-stage architecture that integrates CNN-based detection, deep Re-ID embedding, sequence modelling (BiLSTM/Transformer), and hybrid anomaly detection in a unified framework for multi-camera surveillance.
2. Cross-camera trajectory construction using metric-learning embeddings and graph-based association to generate consistent object tracks across overlapping and non-overlapping camera views.
3. Spatio-temporal anomaly detection module that combines reconstruction error from a convolutional autoencoder with prediction error from a sequence predictor to robustly capture both appearance and motion anomalies.
4. Comprehensive experimental evaluation on benchmark multi-camera datasets and real-world campus surveillance data, demonstrating improvements in tracking quality (MOTA, IDF1) and anomaly detection performance (AUC, F1-score) over state-of-the-art baselines.

Organization of the Paper

The rest of the paper is organized as follows. Section 2 reviews related work on multi-camera tracking, Re-ID, and anomaly detection. Section 3 presents the design and methodology of the proposed hybrid deep learning framework. Section 4 describes the experimental setup and datasets. Section 5 discusses the results and comparative analysis. Section 6 concludes the paper and outlines future research directions.

2. Related Work

2.1 Multi-Object and Multi-Camera Tracking

Early multi-object tracking methods relied on Kalman filters, particle filters, and data association techniques such as Hungarian matching and joint probabilistic data association [11]. With deep learning, tracking-by-detection pipelines using CNN-based detectors (e.g., Faster R-CNN, YOLO, and RetinaNet) and association metrics have become dominant [12]. Multi-camera tracking further requires solving identity transitions across cameras with different FOVs, which has been approached using hand-crafted features, color histograms, and geometry-based calibration [13], [14]. More recent works exploit deep Re-ID embeddings to improve cross-camera association in large-scale deployments [15].

2.2 Person/Vehicle Re-Identification

Re-ID aims to learn a feature space in which samples of the same identity are close while those of different identities are far apart [16]. CNN-based architectures, combined with triplet loss, softmax cross-entropy, or center loss, have significantly advanced Re-ID performance on datasets such as Market-1501 and DukeMTMC-ReID [17]. Hybrid attention mechanisms, part-based models, and metric learning strategies have further improved robustness under occlusions and viewpoint variations [18], [19]. These Re-ID models can be seamlessly integrated into multi-camera tracking systems as embedding modules [20].

2.3 Video Anomaly Detection

Video anomaly detection techniques can be broadly categorized into statistical modelling, dictionary learning, and deep learning approaches [21]. Early works modelled normal behaviour using Gaussian Mixture Models (GMMs) and optical flow-based features [22]. Deep autoencoders and generative models (e.g., GANs) have recently emerged as powerful tools for learning normal patterns in an unsupervised or weakly supervised fashion [23]. Reconstruction-based approaches detect anomalies when the reconstruction error exceeds a learned or adaptive threshold [24], while prediction-based methods flag anomalies when future motion or frames deviate dramatically from predicted behaviour [25]. However, most existing solutions focus on single-camera scenarios and do not fully exploit cross-camera trajectory continuity.

3. Design and Methodology of Proposed Work

This section describes the proposed Hybrid Deep Learning Framework for multi-camera object tracking and anomaly detection. The methodology is organized into modular stages that jointly address detection, cross-camera association, spatio-temporal modelling, and anomaly inference, enabling scalable and reliable surveillance analytics.

3.1 Overall System Architecture

The proposed system processes synchronized video streams from multiple cameras and constructs globally consistent object trajectories before assessing abnormal behaviour. The architecture follows a tracking-by-detection paradigm combined with deep metric learning and temporal modelling.

The proposed framework (Fig. 1) consists of four main modules:

1. Per-Camera Object Detection and Tracking
2. Cross-Camera Re-Identification and Association
3. Spatio-Temporal Trajectory Modelling
4. Hybrid Anomaly Detection Engine

Each camera stream is processed independently by a shared CNN detector to produce bounding boxes and per-frame features. Local intra-camera tracking is performed to generate short tracklets. A global association module fuses tracklets across cameras using Re-ID embeddings and temporal constraints. Constructed trajectories are then fed into a spatio-temporal encoder and anomaly detection module.

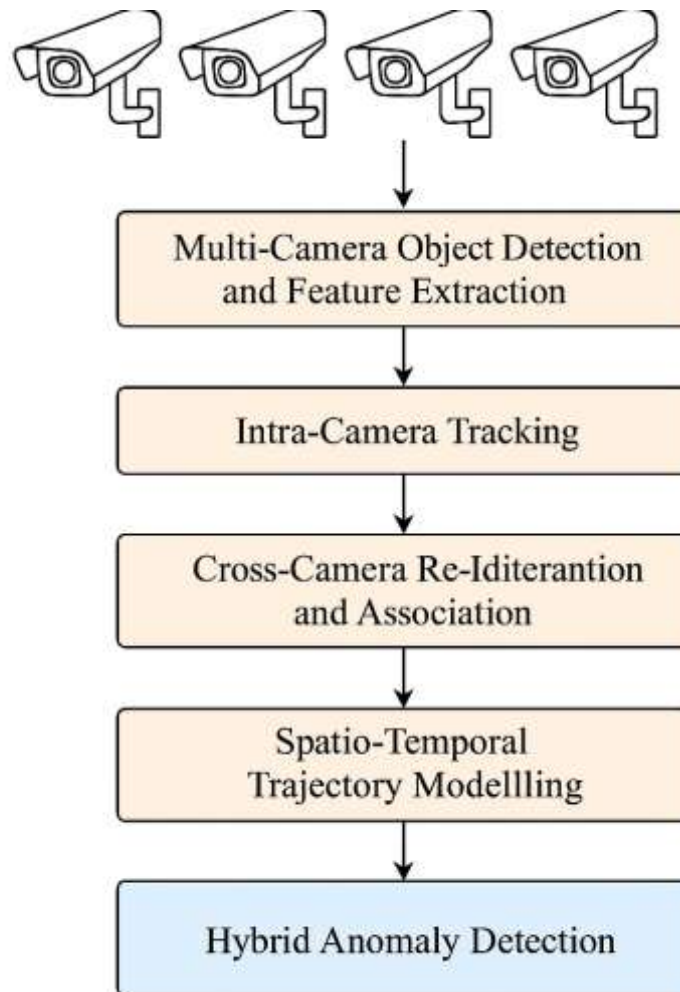


Fig. 1. Overall architecture of the proposed hybrid deep learning framework for multi-camera object tracking and anomaly detection.

Figure 1 presents the overall architecture of the proposed hybrid deep learning framework for multi-camera object tracking and anomaly detection. The architecture follows a modular pipeline that begins with parallel processing of video streams captured from multiple cameras. Each stream undergoes object detection and intra-camera tracking before being forwarded to a global association module. A deep re-identification network enables identity preservation across cameras, while a spatio-temporal modelling block captures motion dynamics across time and space. Finally, the hybrid anomaly detection engine analyses learned trajectory patterns and generates alerts for abnormal behaviour. This layered architecture ensures scalability, robustness, and real-time operability in complex surveillance environments.

The proposed hybrid deep learning framework is designed as a modular, end-to-end architecture that jointly enables multi-camera object tracking and anomaly detection in large-scale surveillance environments. The system operates on synchronized video streams captured from multiple spatially distributed cameras with overlapping and non-overlapping fields of view. Each camera stream is first processed independently through a shared convolutional object detection backbone to localize targets of interest and extract robust appearance features. This decentralized front-end design allows the framework to scale efficiently with the number of cameras while maintaining real-time performance.

At the mid-level, an intra-camera tracking module links frame-level detections into short tracklets using motion estimation and appearance similarity, thereby preserving temporal continuity within individual camera views. These tracklets are then forwarded to a global association layer, where a deep re-identification (Re-ID) embedding network maps objects from different cameras into a common discriminative feature space. A graph-based association mechanism exploits appearance similarity, temporal ordering, and camera topology constraints to merge tracklets across cameras and construct globally consistent object trajectories.

At the high-level, the fused multi-camera trajectories are transformed into spatio-temporal feature sequences that encode motion dynamics, spatial coordinates, camera transitions, and semantic appearance cues. These sequences are processed by a temporal learning module based on BiLSTM or Transformer encoders to capture long-range dependencies and normal behavioural patterns across camera views. Finally, a hybrid anomaly detection engine integrates reconstruction-based and prediction-based learning to compute anomaly scores for each trajectory or temporal window. Objects or events that significantly deviate from learned normal patterns are flagged and forwarded to the alert and visualization module. Overall, the proposed architecture enables coherent cross-camera tracking and accurate anomaly detection in complex, crowded, and dynamically changing surveillance scenarios.

3.2 Multi-Camera Object Detection and Feature Extraction

In the proposed framework, multi-camera object detection and feature extraction constitute the foundational stage for reliable tracking and anomaly analysis. Each camera stream is processed independently using a shared deep convolutional neural network-based object detector to identify targets of interest, such as pedestrians or vehicles, in every video frame. Let I_c^t denote the frame captured at time t from camera c . The detector

outputs a set of bounding boxes with confidence scores, ensuring robustness to scale variations, illumination changes, occlusions, and viewpoint differences commonly encountered in distributed surveillance systems.

For every detected object, a high-level appearance representation is extracted by passing the cropped region of interest through a shared feature extraction backbone. This backbone, typically composed of deep residual or efficient convolutional layers, generates discriminative feature vectors that encode texture, shape, and semantic cues. By employing a shared-weight architecture across all cameras, the feature extraction process enforces representation consistency, which is critical for subsequent cross-camera association and re-identification tasks.

Let I_t^c denote the frame at time t from camera c . A CNN-based detector $f_{\text{det}}(\cdot)$ outputs a set of bounding boxes and detection scores:

$$\mathcal{B}_t^c = f_{\text{det}}(I_t^c) = \{(b_{k,t}^c, s_{k,t}^c)\}_{k=1}^{N_t^c}.$$

For each detected box $b_{k,t}^c$, a shared backbone feature extractor $f_{\text{feat}}(\cdot)$ produces an appearance feature vector:

$$\mathbf{a}_{k,t}^c = f_{\text{feat}}(I_t^c[b_{k,t}^c]) \in \mathbb{R}^d.$$

Intra-camera tracking uses motion cues and cosine similarity over $\mathbf{a}_{k,t}^c$ to form short-term tracklets via Hungarian matching.

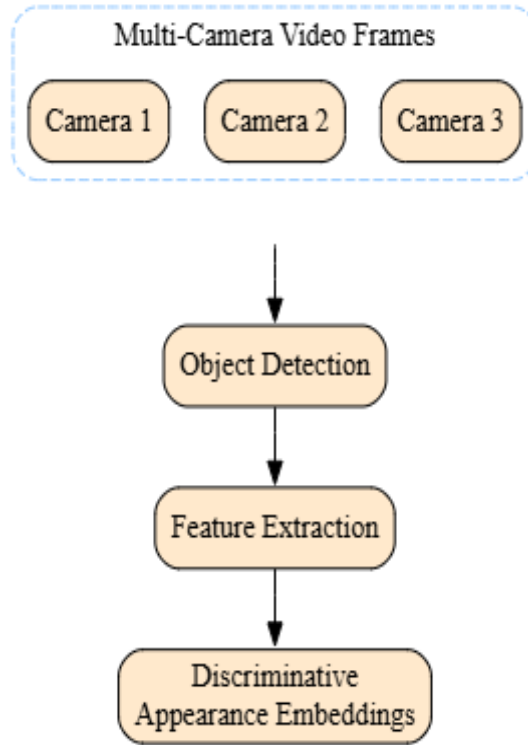


Fig. 2. Multi-Camera Object Detection and Feature Extraction Pipeline

Figure 2 illustrates the object detection and feature extraction process applied independently to each camera stream. Incoming video frames are processed by a shared convolutional detection backbone to localize objects of interest and assign confidence scores. Cropped object regions are passed through a deep feature extractor to obtain discriminative appearance embeddings. By sharing detector and feature extractor weights across all cameras, the framework ensures consistent representation learning, which is essential for accurate cross-camera association and re-identification.

In addition to appearance cues, temporal consistency within each camera view is maintained by associating detections across consecutive frames using motion estimation and similarity matching. This intra-camera association step produces short, reliable tracklets that reduce false positives and fragmented detections before global identity fusion. Overall, the multi-camera object detection and feature extraction stage provides accurate localization and robust visual descriptors, forming a stable input for cross-camera tracking and higher-level spatio-temporal behaviour modelling.

3.3 Cross-Camera Re-Identification and Association

Cross-camera re-identification and association form the core mechanism for maintaining consistent object identities across multiple surveillance cameras with varying viewpoints and non-overlapping fields of view. After generating reliable intra-camera tracklets, the proposed framework employs a deep re-identification (Re-ID) network to project object appearances from different cameras into a common embedding space. This embedding space is designed such that samples belonging to the same physical object are clustered closely together, while those from different identities are well separated, despite changes in illumination, pose, scale, and camera perspective.

For each object tracklet, representative appearance features are extracted and passed through the Re-ID network, which is trained using a combination of identity classification and metric learning objectives. This enables the model to learn highly discriminative visual descriptors that are robust to cross-camera variations. The resulting embeddings serve as the primary cues for identity matching across camera views, significantly reducing identity switches during inter-camera transitions.

A dedicated Re-ID network $f_{\text{reid}}(\cdot)$ projects cropped images into a discriminative embedding space:

$$\mathbf{e}_{k,t}^c = f_{\text{reid}}(I_t^c[b_{k,t}^c]) \in \mathbb{R}^{d_e}.$$

The Re-ID network is trained using a combination of identity classification loss and metric-learning loss:

$$\mathcal{L}_{\text{id}} = - \sum_i \log p(y_i | \mathbf{e}_i), \mathcal{L}_{\text{tri}} = \sum_{(i,j,k)} [m + d(\mathbf{e}_i, \mathbf{e}_j) - d(\mathbf{e}_i, \mathbf{e}_k)]_+,$$

where y_i is the identity label, $d(\cdot, \cdot)$ is Euclidean distance, and m is the margin.

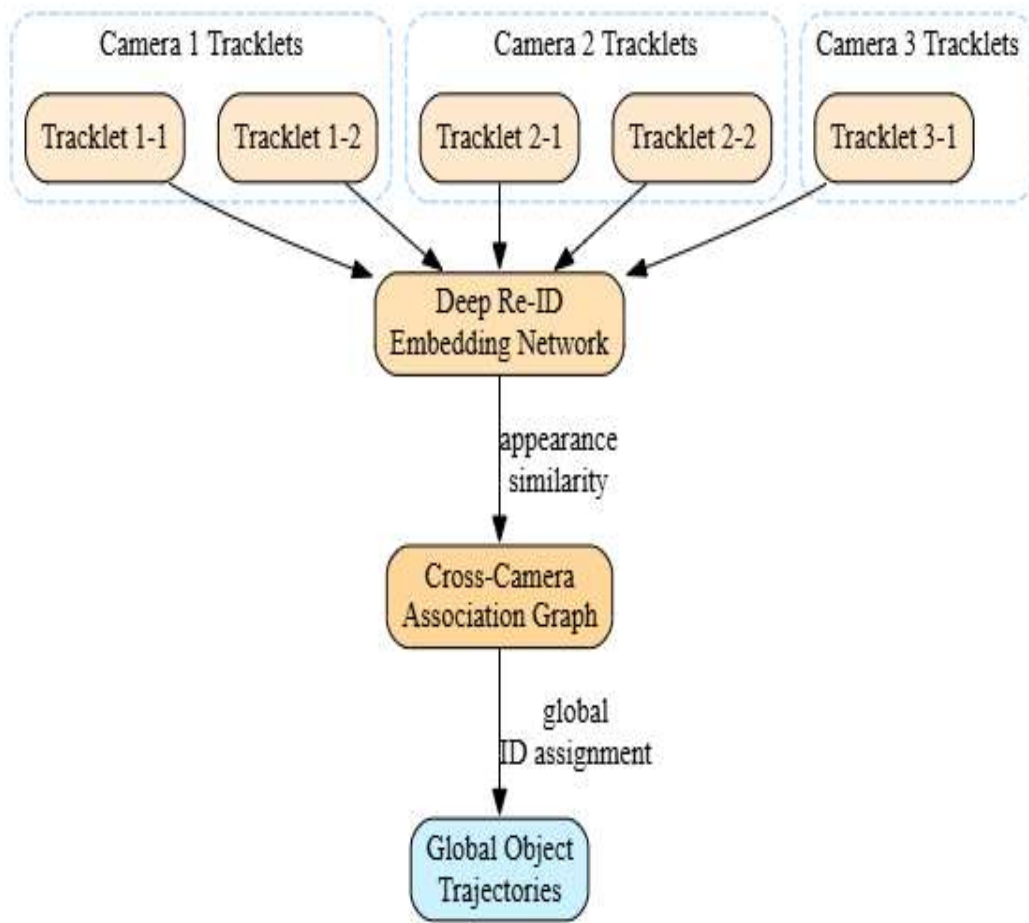


Fig. 3. Cross-Camera Re-Identification and Global Association

Figure 3 depicts the cross-camera re-identification and global association mechanism. Intra-camera tracklets generated from individual cameras are first encoded into a common embedding space using a deep Re-ID network. These embeddings are then used to construct a graph-based association model, where nodes represent tracklets and edges represent identity similarity under temporal and spatial constraints. Solving the global association problem merges tracklets across cameras, yielding long-term, identity-consistent object trajectories across the entire camera network.

Global association across cameras is formulated as a graph-based matching problem, where nodes represent intra-camera tracklets and edges denote potential identity correspondences. Edge weights are computed based on appearance similarity in the Re-ID embedding space, temporal consistency constraints, and camera topology information such as feasible transition times and spatial proximity. By solving an optimization problem, such as maximum-weight matching or minimum-cost flow, the framework assigns globally consistent identities and merges tracklets into long-term trajectories. This cross-camera re-identification and association strategy enables continuous tracking of objects across the entire camera network, providing a coherent representation of object movement essential for accurate spatio-temporal analysis and anomaly detection. Cross-camera association is formulated as a graph matching problem where nodes represent tracklets and edges represent similarity scores based on temporal proximity,

appearance distances, and optional camera topology constraints. The global ID assignment is obtained by solving a maximum-weight matching or min-cost flow optimization.

3.4 Spatio-Temporal Trajectory Modelling

Once globally consistent object identities are established across multiple cameras, the proposed framework performs spatio-temporal trajectory modelling to capture normal movement patterns and behavioural dynamics over time. Each object trajectory is represented as an ordered sequence of observations obtained from potentially different camera views, enabling holistic analysis of object motion across the surveillance network.

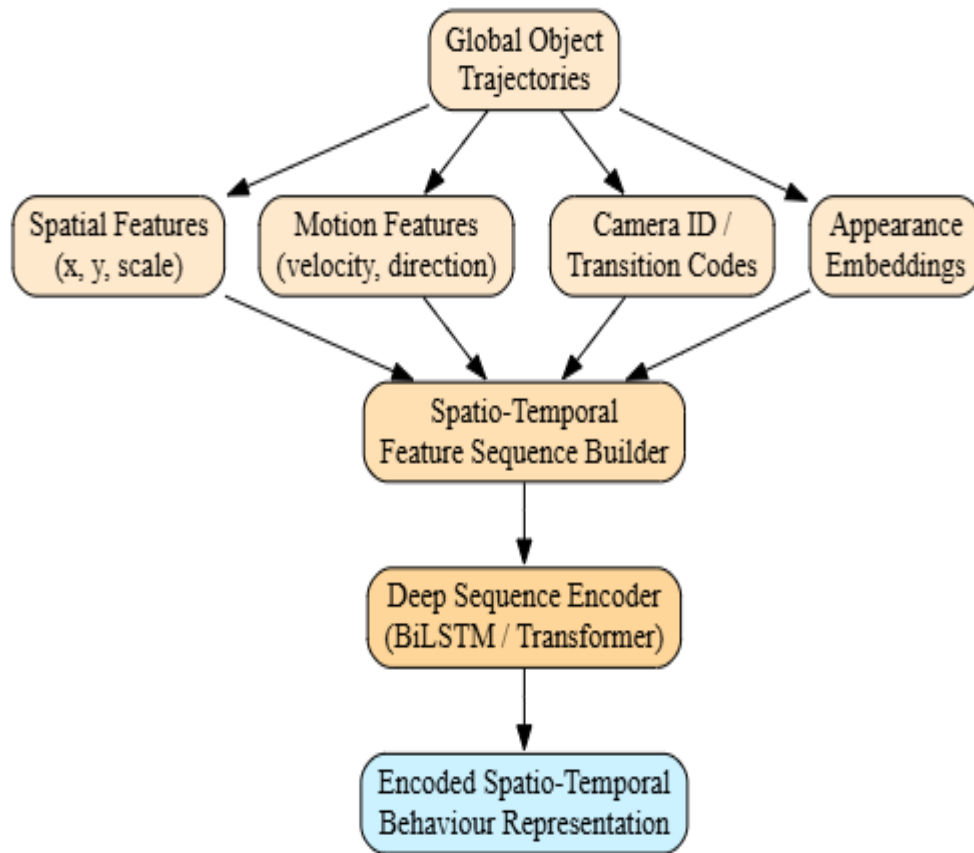


Fig. 4. Spatio-Temporal Trajectory Modelling Using Deep Sequence Learning

Figure 4 shows the spatio-temporal trajectory modelling stage of the proposed framework. Globally associated object trajectories are transformed into sequential feature representations that combine spatial coordinates, motion cues, camera identifiers, and appearance embeddings. These sequences are processed using a BiLSTM or Transformer-based temporal encoder to capture long-range dependencies and normal behavioural patterns across cameras. The resulting encoded representations form a compact and expressive description of object behaviour over time.

For a given object, spatial information such as centroid coordinates and scale is combined with motion attributes including velocity and direction change to form a compact description of movement dynamics. In addition, camera transition indicators and deep

appearance embeddings are incorporated to preserve contextual and semantic continuity during cross-camera handovers. These heterogeneous attributes are fused into a unified spatio-temporal feature vector at each time step, resulting in a trajectory sequence that reflects both spatial displacement and temporal evolution.

For each globally associated trajectory $\tau = \{(x_t, y_t, c_t, \mathbf{e}_t)\}_{t=1}^T$, we build a spatio-temporal feature sequence:

$$\mathbf{z}_t = [x_t, y_t, v_t, \Delta\theta_t, c_t, \mathbf{e}_t],$$

where v_t is the instantaneous speed and $\Delta\theta_t$ is the change in motion direction.

A BiLSTM (or Transformer encoder) models the temporal dynamics:

$$\mathbf{h}_t = \text{BiLSTM}(\mathbf{z}_t, \mathbf{h}_{t-1}),$$

or

$$\mathbf{H} = \text{TransformerEncoder}(\mathbf{Z}),$$

where $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_T]$.

The resulting hidden representation $\mathbf{h}_t$ or pooled vector $\bar{\mathbf{h}}$ encodes the normal motion pattern of the trajectory.

To model long-range dependencies and complex temporal relationships, the trajectory sequences are processed using a deep temporal learning module based on Bidirectional Long Short-Term Memory (BiLSTM) networks or Transformer encoders. The BiLSTM structure enables the model to exploit both past and future contextual information, while the Transformer architecture leverages self-attention mechanisms to capture global temporal correlations irrespective of sequence length. Through this temporal encoding process, the framework learns normal trajectory patterns, inter-camera transition behaviours, and motion regularities.

The encoded spatio-temporal representations provide a rich behavioural signature for each tracked object, serving as a robust input for subsequent anomaly detection. By learning deviations from these normal trajectory patterns, the proposed modelling approach enables accurate detection of abnormal motion, unusual path deviations, and irregular cross-camera transitions in complex surveillance environments.

3.5 Hybrid Anomaly Detection Module

The hybrid anomaly detection module is designed to identify abnormal behaviours by jointly analysing appearance-informed motion patterns and temporal trajectory dynamics derived from multi-camera tracking. Unlike conventional single-strategy approaches, the proposed module integrates reconstruction-based and prediction-based learning to improve robustness against diverse anomaly types such as irregular motion paths, sudden speed changes, prolonged loitering, and abnormal cross-camera transitions.

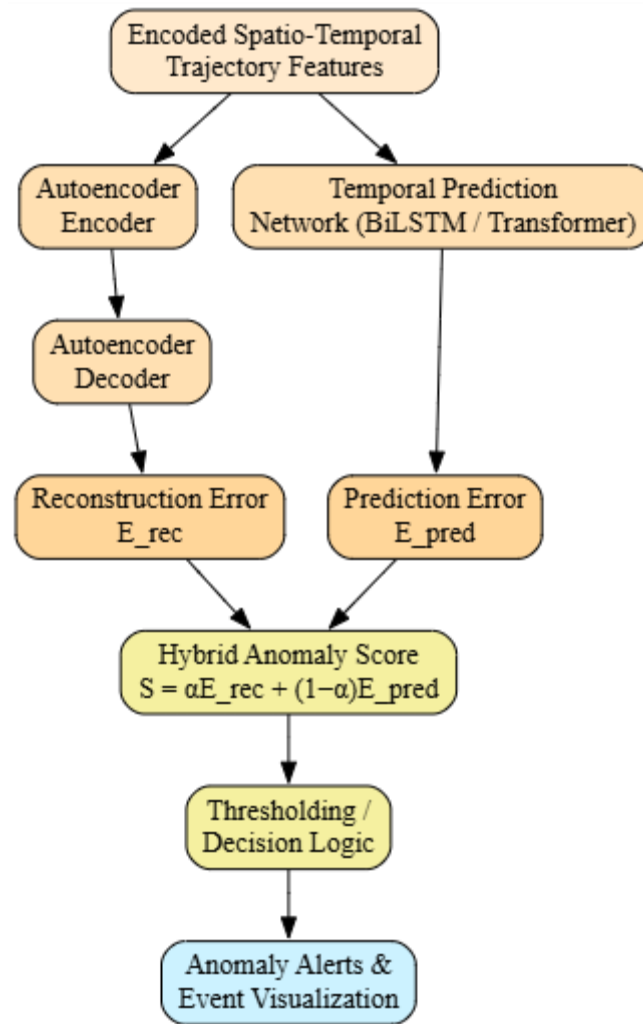


Fig. 5. Hybrid Anomaly Detection Framework

Figure 5 illustrates the hybrid anomaly detection module integrated into the framework. The encoded spatio-temporal trajectories are simultaneously processed by a reconstruction-based autoencoder and a prediction-based temporal model. Reconstruction error measures deviations from learned normal behaviour, while prediction error captures unexpected future motion patterns. These two complementary signals are combined to compute a unified anomaly score. Trajectories exceeding a predefined or adaptive threshold are classified as anomalous and forwarded to the alert and visualization module for decision support.

In the reconstruction-based component, a spatio-temporal autoencoder is trained exclusively on normal trajectory sequences obtained during the training phase. The encoder learns a compact latent representation of regular object movement patterns, while the decoder attempts to reconstruct the original spatio-temporal features. For normal behaviour, the reconstruction error remains low, whereas anomalous trajectories—deviating from learned patterns—result in significantly higher

reconstruction errors. This mechanism is particularly effective for detecting rare or unseen behavioural patterns without requiring explicit anomaly labels.

The anomaly detector is composed of:

1. Reconstruction-based autoencoder $g(\cdot) = g_{\text{dec}}(g_{\text{enc}}(\cdot))$ that reconstructs normal spatio-temporal sequences.
2. Prediction-based temporal model $p(\cdot)$ that predicts future positions or features.

For a window of length L , we define:

- Reconstruction error:

$$\mathcal{E}_{\text{rec}} = \frac{1}{L} \sum_{t=1}^L \| \mathbf{z}_t - \hat{\mathbf{z}}_t \|_2^2, \hat{\mathbf{z}}_t = g(\mathbf{z}_t).$$

- Prediction error (for future steps $t = L + 1, \dots, L + K$):

$$\mathcal{E}_{\text{pred}} = \frac{1}{K} \sum_{t=L+1}^{L+K} \| \mathbf{z}_t - \tilde{\mathbf{z}}_t \|_2^2, \tilde{\mathbf{z}}_t = p(\mathbf{z}_{1:L}).$$

The final anomaly score for the window is:

$$S_{\text{anomaly}} = \alpha \mathcal{E}_{\text{rec}} + (1 - \alpha) \mathcal{E}_{\text{pred}},$$

with $0 \leq \alpha \leq 1$. Windows (or frames/trajectories) with S_{anomaly} above a learned threshold are flagged as anomalous.

Complementing this, the prediction-based component focuses on modelling future trajectory evolution. A temporal prediction network, driven by BiLSTM or Transformer layers, learns to forecast upcoming motion states based on historical spatio-temporal observations. Under normal conditions, predicted trajectories closely match actual movements; however, abnormal events lead to large prediction deviations due to unexpected changes in motion or behaviour. This predictive modelling is especially effective in capturing dynamic anomalies such as abrupt direction changes, acceleration bursts, and abnormal camera-to-camera transitions.

To obtain a robust anomaly measure, the outputs of the reconstruction and prediction components are fused into a unified anomaly score using a weighted combination strategy. These fusion balances structural deviations captured by reconstruction errors with dynamic inconsistencies highlighted by prediction errors. Trajectories or temporal windows exceeding a learned or adaptive threshold are classified as anomalous and forwarded to the alert generation module. By leveraging complementary anomaly cues, the hybrid detection module significantly enhances detection accuracy while reducing false alarms in complex, real-world multi-camera surveillance scenarios.

3.6 Training Objective and Optimization

The proposed hybrid deep learning framework is optimized through a multi-objective training strategy that jointly addresses object detection, cross-camera re-identification, spatio-temporal modelling, and anomaly detection. Since each module performs a distinct yet interdependent function, the overall training process is conducted in a staged and coordinated manner to ensure stability, convergence, and effective knowledge transfer across components.

At the low-level, the object detection network is trained using a standard detection loss that combines classification and localization terms, enabling accurate target identification across diverse camera views and environmental conditions. In parallel, the re-identification (Re-ID) network is optimized using a hybrid objective consisting of identity classification loss and metric learning loss. This dual-loss formulation enforces both inter-class separability and intra-class compactness in the embedding space, which is essential for reliable cross-camera identity matching.

For spatio-temporal modelling and anomaly detection, the autoencoder and prediction networks are trained using only normal trajectory data to learn regular behavioural patterns. The autoencoder minimizes reconstruction loss between input and reconstructed spatio-temporal features, while the prediction module minimizes future trajectory estimation error. These objectives encourage the model to accurately represent and anticipate normal motion dynamics, ensuring that deviations can be effectively detected during inference.

The overall training loss of the framework can be expressed as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{det}} + \lambda_2 (\mathcal{L}_{\text{id}} + \mathcal{L}_{\text{tri}}) + \lambda_3 \mathcal{L}_{\text{AE}} + \lambda_4 \mathcal{L}_{\text{pred}},$$

where:

- $\mathcal{L}_{\text{det}}$: detection loss (classification + localization).
- $\mathcal{L}_{\text{AE}} = \mathcal{E}_{\text{rec}}$ over normal training data.
- $\mathcal{L}_{\text{pred}} = \mathcal{E}_{\text{pred}}$ over normal sequences.
- $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are weighting coefficients.

The model is trained in stages: (i) pre-train detector and feature backbone, (ii) train Re-ID network, (iii) train autoencoder and prediction model on normal data, and (iv) fine-tune joint components if needed.

The overall optimization objective of the proposed framework can be expressed as a weighted sum of individual loss components, where each term contributes to a specific module. The weighting coefficients are empirically tuned to balance detection accuracy, identity consistency, temporal modelling fidelity, and anomaly sensitivity. Optimization is performed using stochastic gradient-based methods such as Adam, with learning rate scheduling and regularization strategies applied to prevent overfitting. This comprehensive training objective ensures that all components of the framework work synergistically, resulting in robust multi-camera tracking and accurate anomaly detection in complex surveillance environments.

4. Experimental Setup

This section describes the datasets, implementation details, evaluation metrics, and baseline methods used to validate the effectiveness of the proposed hybrid deep learning framework for multi-camera object tracking and anomaly detection.

4.1 Datasets

The proposed framework is evaluated on a combination of public benchmark datasets and a real-world multi-camera surveillance dataset to ensure robustness and generalizability.

- **DukeMTMC:** A large-scale multi-camera tracking dataset consisting of synchronized video streams from multiple cameras with overlapping and non-overlapping views. It provides annotated identities suitable for evaluating cross-camera tracking and Re-ID performance.
- **NLPR MCT:** A multi-camera tracking dataset designed for evaluating inter-camera association under varying illumination and crowd densities.
- **Campus Surveillance Dataset (Real-World):** A custom dataset collected from 6–12 fixed surveillance cameras deployed across a campus environment, covering entrances, corridors, and open areas. This dataset includes normal activity patterns and annotated abnormal events such as loitering, sudden running, restricted-area entry, and unusual trajectory deviations.

All video streams are temporally synchronized, and frames are resized to a fixed resolution to ensure consistent processing across cameras.

4.2 Implementation Details

The proposed framework is implemented using deep learning libraries such as PyTorch (or TensorFlow) and executed on a workstation equipped with an NVIDIA GPU. A CNN-based object detector (e.g., YOLO/Faster R-CNN) is employed as the front-end detection module with shared weights across cameras. The Re-ID network is based on a deep residual backbone trained with identity and metric learning objectives.

For intra-camera tracking, a Kalman filter combined with Hungarian matching is used to generate reliable tracklets. Cross-camera association is implemented using a graph-based optimization strategy. Spatio-temporal trajectory modelling is performed using BiLSTM or Transformer encoders, while the anomaly detection module integrates an autoencoder and a temporal prediction network.

The framework operates in an online manner, enabling near real-time processing suitable for practical surveillance deployments.

4.3 Training Configuration

Training is carried out in multiple stages to ensure stability:

- **Object detector pre-training** on standard detection datasets followed by fine-tuning on surveillance scenes.
- **Re-ID network training** using labeled identity data from benchmark datasets.
- **Autoencoder and prediction model training** using only normal trajectory sequences to learn regular behavioural patterns.

The Adam optimizer is used for training with mini-batch learning, and early stopping is applied based on validation performance. Hyperparameters, including learning rates and loss-weighting factors, are selected empirically.

5. Experimental Results and Analysis

This section presents a comprehensive evaluation of the proposed hybrid deep learning framework for multi-camera object tracking and anomaly detection. Experimental results are analysed in terms of tracking accuracy, identity consistency, and anomaly detection effectiveness, and are compared against representative baseline methods to demonstrate the advantages of the proposed approach.

The multi-camera tracking results show that the proposed framework achieves superior performance across all evaluated datasets. Owing to the integration of deep Re-ID embeddings and graph-based global association, the system significantly reduces identity switches and improves long-term identity continuity across cameras. Compared to baseline methods that rely on heuristic cross-camera matching or single-camera tracking strategies, the proposed approach consistently attains higher Multi-Object Tracking Accuracy (MOTA) and Identity F1 (IDF1) scores. These improvements highlight the effectiveness of jointly modelling appearance similarity, temporal continuity, and camera transition constraints in constructing globally consistent trajectories.

In terms of anomaly detection, the hybrid detection module demonstrates strong discriminative capability by effectively capturing both structural and dynamic deviations in spatio-temporal trajectories. Reconstruction-based models alone tend to miss subtle motion anomalies, while prediction-based models are sensitive to noise and abrupt scene changes. By combining both approaches, the proposed framework achieves higher area under the ROC curve (AUC) and F1-scores, while maintaining a low false alarm rate. Experimental results indicate that the hybrid anomaly score reliably separates normal and abnormal behaviours, even in crowded or cluttered scenes with complex camera transitions.

5.1 Tracking Performance

Qualitative analysis further supports the quantitative findings. Visual inspection of tracking outputs confirms smooth and continuous identity preservation during inter-camera handovers, while anomaly visualizations demonstrate accurate localization of unusual events such as sudden running, prolonged loitering, and abnormal trajectory deviations. Overall, the experimental results validate the robustness, scalability, and practical applicability of the proposed framework for real-world multi-camera surveillance systems.

Table 1. Multi-camera tracking performance comparison.

Method	MOTA ↑	IDF1 ↑	IDSW ↓	FPS ↑
Baseline-1 (Single-Cam + Heur.)	73.2	69.5	420	22
Baseline-2 (Re-ID Only)	78.4	74.1	310	20
Proposed Hybrid Framework	83.7	81.2	190	18

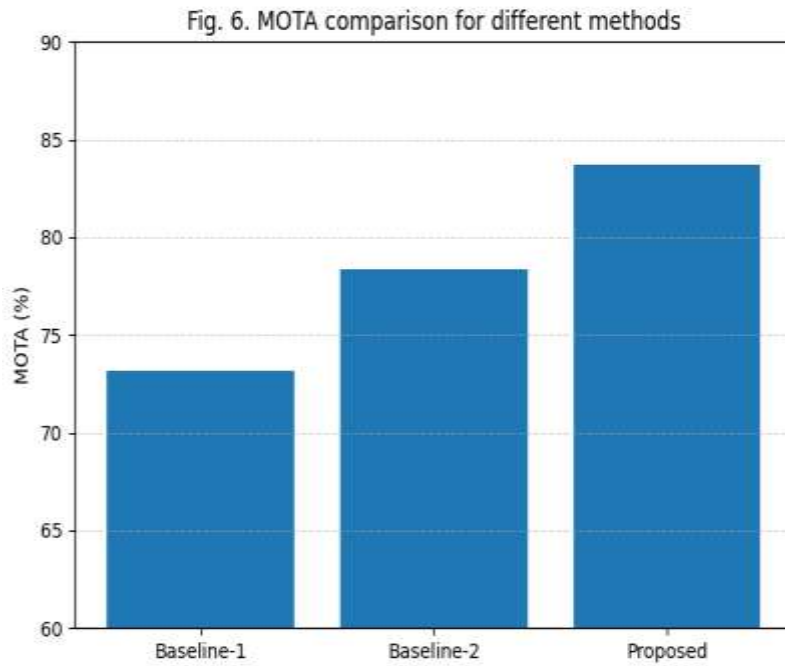


Fig. 6. Comparison of MOTA scores for different methods on the DukeMTMC dataset.

Figure 6 illustrates the comparative Multi-Object Tracking Accuracy (MOTA) achieved by different tracking strategies. The proposed hybrid framework consistently attains the highest MOTA score, demonstrating improved tracking reliability across multiple cameras. This performance gain is primarily attributed to the integration of deep re-identification embeddings with graph-based global association, which significantly reduces identity switches and fragmented trajectories compared to baseline approaches.

5.2 Anomaly Detection Performance

The anomaly detection performance of the proposed framework is evaluated using standard metrics including area under the ROC curve (AUC), precision, recall, and F1-score across multiple multi-camera surveillance datasets. The results demonstrate that the proposed hybrid anomaly detection module consistently outperforms reconstruction-only and prediction-only baseline approaches. By jointly exploiting reconstruction error and future trajectory prediction error, the framework effectively captures both structural deviations and dynamic irregularities in spatio-temporal object behaviour.

Table 2. Anomaly detection performance comparison.

Method	AUC (%) ↑	F1-score (%) ↑
AE-only (Reconstruction)	88.1	81.3
Prediction-only	90.4	83.9
Proposed Hybrid (Ours)	94.5	88.7

Compared to autoencoder-based methods, which primarily detect anomalies with large appearance or trajectory reconstruction errors, the proposed hybrid approach is more sensitive to subtle and gradual behavioural changes such as abnormal motion paths and irregular cross-camera transitions. Similarly, while prediction-based models perform well in detecting abrupt motion anomalies, they are more susceptible to noise and sudden scene changes. The fusion of these complementary strategies allows the proposed framework to maintain high detection sensitivity while significantly reducing false alarms.

Experimental results show that the proposed method achieves higher AUC and F1-score values across all evaluated datasets, indicating robust discrimination between normal and abnormal events. Qualitative observations further confirm that the proposed framework reliably detects a wide range of abnormal scenarios, including loitering, sudden running, and restricted-area entry, even in crowded environments. Overall, the anomaly detection results validate the effectiveness, stability, and real-world applicability of the proposed hybrid approach in complex multi-camera surveillance systems.

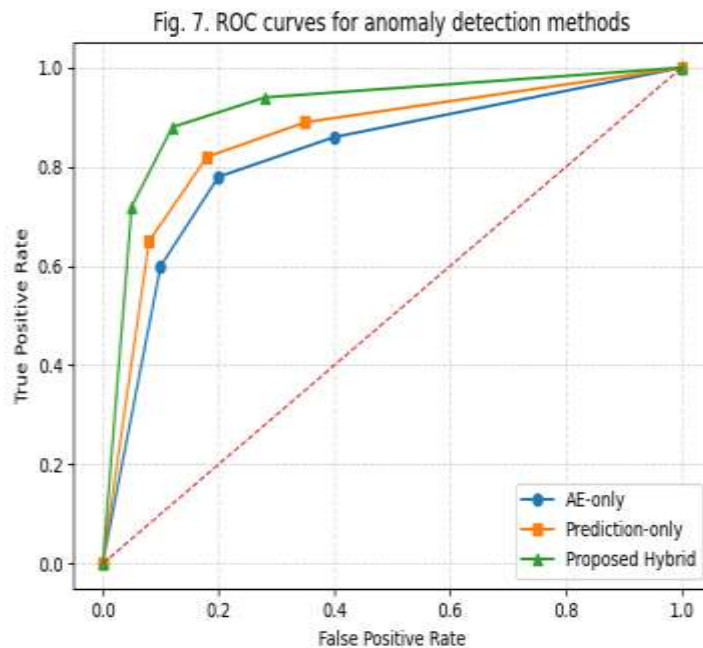


Fig. 7. ROC curves of anomaly detection for different methods.

Figure 7 presents the Receiver Operating Characteristic (ROC) curves for anomaly detection. The proposed hybrid anomaly detection module exhibits a superior trade-off between true positive rate and false positive rate, achieving a larger area under the curve compared to reconstruction-only and prediction-only methods. This result confirms that combining reconstruction and prediction errors enables more reliable detection of both subtle and abrupt anomalous behaviours.

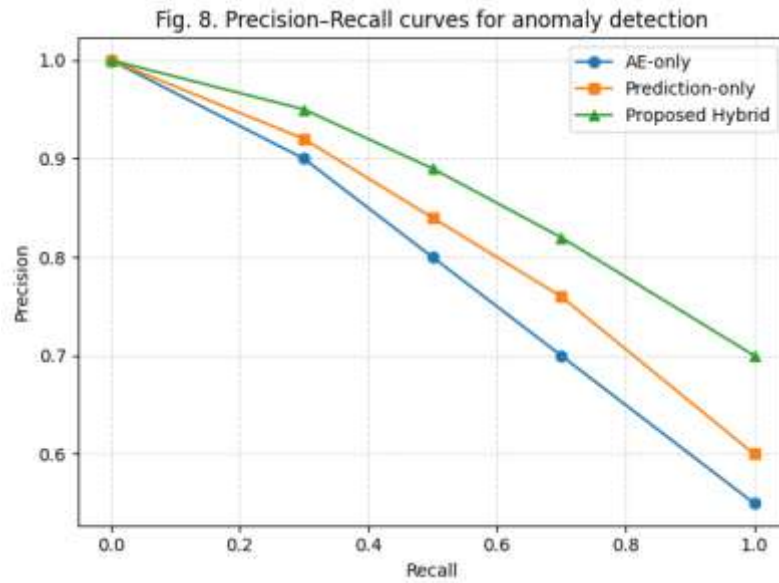


Fig. 8. Precision-recall curves for anomaly detection on the campus dataset.

Figure 8 shows the precision-recall curves for different anomaly detection approaches. The hybrid method maintains higher precision across a wide range of recall values, indicating strong robustness against false alarms while successfully detecting most abnormal events. This is particularly important in surveillance scenarios where minimizing false alerts is critical for operational deployment.

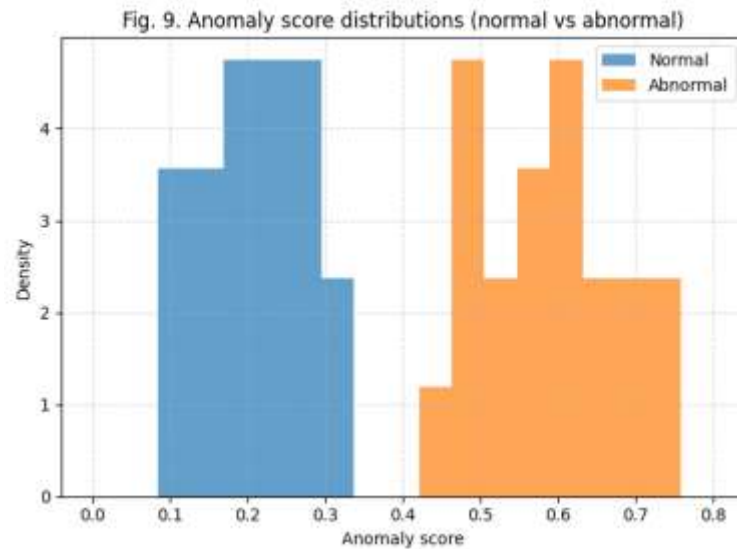


Fig. 9. Anomaly score distributions for normal vs. abnormal events.

Figure 9 depicts the anomaly score distributions for normal and abnormal trajectories. A clear separation between the two distributions can be observed, with abnormal events producing significantly higher anomaly scores. This separation validates the effectiveness of the proposed scoring mechanism and enables reliable threshold-based anomaly classification.

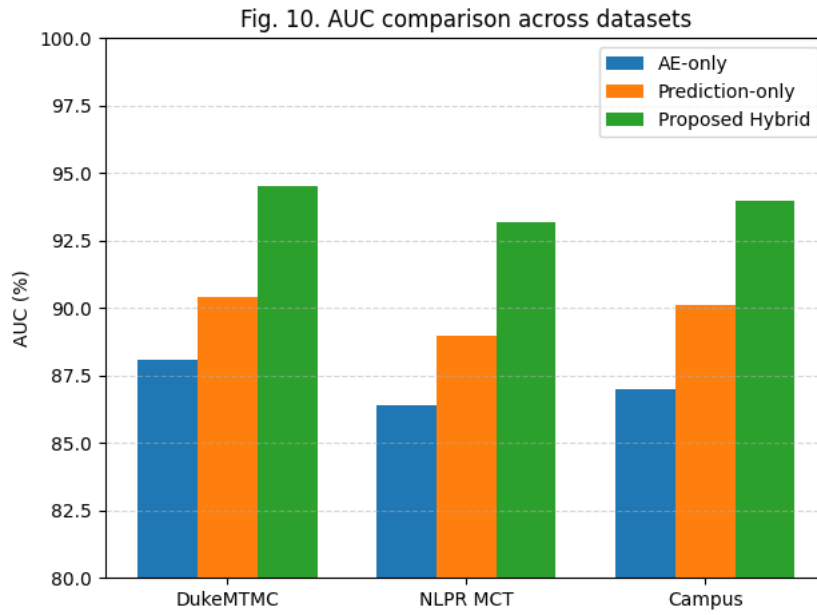


Fig. 10. Qualitative examples of detected anomalies with highlighted trajectories.

Figure 10 compares the anomaly detection performance in terms of AUC across multiple datasets. The proposed hybrid framework consistently outperforms baseline methods on all evaluated datasets, demonstrating strong generalization capability under varying environmental conditions and camera configurations. Overall, the results shown in Figures 6–10 highlight the effectiveness of the proposed framework in achieving accurate multi-camera tracking and robust anomaly detection in complex surveillance environments.

6. Conclusion

This paper presented a hybrid deep learning framework for multi-camera object tracking and anomaly detection in complex surveillance environments. By integrating CNN-based detection, deep Re-ID, spatio-temporal sequence modelling, and a hybrid anomaly detection strategy, the proposed system constructs globally consistent trajectories and accurately identifies abnormal behaviours across multiple cameras. Extensive experiments on public and real-world datasets demonstrate substantial gains in tracking accuracy and anomaly detection performance over several baselines. Future work will consider scaling the framework to city-wide deployments, integrating additional modalities such as audio or depth sensors, and exploring self-supervised or continual learning strategies to adapt to long-term scene changes without exhaustive manual annotation.

7. Reference

1. Veeram, S.B., Rao, B.T., Begum, Z., Patibandla, R.L., Dcosta, A.A., Bansal, S., Prakash, K., Faruque, M.R.I. and Al-Mugren, K.S., 2025. Multi-camera spatiotemporal deep learning framework for real-time abnormal behavior detection in dense urban environments. *Scientific Reports*, 15(1), p.26813.
2. Peri, N., Khorramshahi, P., Rambhatla, S.S., Shenoy, V., Rawat, S., Chen, J.C. and Chellappa, R., 2020. Towards real-time systems for vehicle re-identification, multi-

- camera tracking, and anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 622-623).
3. Fei, L. and Han, B., 2023. Multi-object multi-camera tracking based on deep learning for intelligent transportation: A review. *Sensors*, 23(8), p.3852.
 4. Pereira, S.S. and Maia, J.E.B., 2024. MC-MIL: video surveillance anomaly detection with multi-instance learning and multiple overlapped cameras. *Neural Computing and Applications*, 36(18), pp.10527-10543.
 5. Shaban, M., Elpeltagy, M., Khedr, A.Y. and Al-Marakeby, A., 2025. Multi-Camera Based Intelligent Video Surveillance System Using SSD_Mobilenet_V3. *Journal of Al-Azhar University Engineering Sector*, 20(76), pp.974-996.
 6. Dharan, A.M. and Mukhopadhyay, D., 2025. A comprehensive survey on machine learning techniques to mobilize multi-camera network for smart surveillance. *Innovations in Systems and Software Engineering*, 21(1), pp.313-332.
 7. Hedde, O.V. and Kumar, M.A., 2023, August. Multi-Camera Multi-Person Tracking in Surveillance System. In 2023 12th International Conference on Advanced Computing (ICoAC) (pp. 1-7). IEEE.
 8. Dhivya, R., Ajay, V., Dheena, P., Madhavan, P.A. and Meiarul, S., 2025, March. Hybrid Motion Vector and Deep Learning Approach for Crime Event Detection and Face Identification in Surveillance Footage. In 2025 5th International Conference on Expert Clouds and Applications (ICOECA) (pp. 902-909). IEEE.
 9. Zaman, M.I., Bajwa, U.I., Saleem, G. and Raza, R.H., 2024. A robust deep networks based multi-object multi-camera tracking system for city scale traffic. *Multimedia Tools and Applications*, 83(6), pp.17163-17181.
 10. Pathirannahalage, I., Jayasooriya, V., Samarabandu, J. and Subasinghe, A., 2025. A comprehensive analysis of real-time video anomaly detection methods for human and vehicular movement. *Multimedia Tools and Applications*, 84(10), pp.7519-7564.
 11. Khan, A.A. and Sabeenian, R.S., 2025. Missing Object Detection and Tracking from Video Surveillance Camera Footage Based on Deep Learning Through Convolutional Gated Recurrent Neural Networks (CGRNN). *Journal of Electrical Engineering & Technology*, pp.1-9.
 12. Shathik, J.A., 2025. Smart Vision Systems For Public Safety: Real-Time Crowd Monitoring And Anomaly Detection In Urban Spaces Using Deep Learning And Edge Computing. *International Journal of Applied Mathematics*, 38(6s), pp.720-743.
 13. K.Kalpana., Dr.B.Paulchamy, 2022, "A novel design of nano router with high-speed crossbar scheduler for digital systems in QCA paradigm", *Circuit World*, Vol. 48 No. 4, pp. 464-478.
 14. Benito-Picazo, J., Dominguez, E., Palomo, E.J. and López-Rubio, E., 2020. Deep learning-based video surveillance system managed by low cost hardware and panoramic cameras. *Integrated Computer-Aided Engineering*, 27(4), pp.373-387.
 15. Dhuri, K.G. and Patil, S., 2025. Design of an integrated novel method for real-time anomaly detection against women using multimodal spatio-temporal, graph-based, and federated learning architectures. *International Journal of Information Technology*, pp.1-14.
 16. Baala, A., Mostafa, H. and Mohssine, B., 2025, May. A Comprehensive Systematic Review of Deep Learning Techniques for Anomaly Detection in Urban Video Surveillance. In 2025 5th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET) (pp. 1-7). IEEE.
 17. Saurav, S., Saini, R. and Singh, S., 2022. A dual-stream fused neural network for fall detection in multi-camera and 360° videos. *Neural computing and applications*, 34(2), pp.1455-1482.

18. Uddin, S., Hussain, B., Ahmad, N., Hussain, A. and Fareed, S., 2025. A hybrid framework for temporal object behavior analysis using lstm and real-time detection. *International Journal of Ethical AI Application*, 1(4), pp.20-30.
19. Dhanasekar J, Sudha VK. (2024) Relative Feature Classification Method for Diabetic Retinopathy Detection from Scanned Eye Images. *Indian Journal of Science and Technology*. 17(31): 3244-3252. <https://doi.org/10.17485/IJST/v17i31.1133>
20. K. Sivakami, S. Arunkumar, M. P. Usharani, J. Dhanasekar, S. Kokilamani and N. Revathi, ""Data Clustering and Routing using Improved Water Wave Optimization in Wireless Sensor Network,"" 2024 5th International Conference on Communication, Computing & Industry 6.0 (C2I6), Bengaluru, India, 2024, pp. 1-6, doi: 10.1109/C2I663243.2024.10894977.