Journal of Applied Automation Technologies

Research Paper

# GCN-Enhanced Multi-Object Tracking for Video Surveillance: Adaptive Spatial-Temporal Graph Convolutional Modeling

**Agnieszka Szymanski[1,*] and Weronika Czarnecki[2]**

[1] Maria Curie-Sklodowska University, Faculty of Mathematics, Physics and Computer Science, 20-031 Lublin, Poland
[2] Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering, Gdansk University of Technology, 80-233 Gdansk, Poland
*Corresponding author: agnieszka.s@umcs.lublin.pl

**Abstract.** Multi-object tracking is still challenging since many objects in intelligent video surveillance regions move irregularly or are frequently obscured and alter owing to lighting changes. This research presents a new tracking framework that integrates graph convolutional neural networks and adaptive spatial-temporal graph creation to overcome the shortcomings of appearance-based and sequential association approaches. Using a composite attention mechanism as a guide, dynamically create a graph in the system with each node representing a detection and edges representing spatial-temporal correlations. Contextual information is spread by Hierarchical Graph Convolutional Networks, which also strengthen identity association and trajectory continuity. The suggested framework has also demonstrated good improvement based on the experiment results in the MOT17, MOT20, and DukeMTMC datasets; for MOT17 and MOT20, it achieved a MOTA of 75.2% and 69.4%, respectively, outperforming earlier approaches in both precision and identity preservation. According to ablation analysis, in situations of dense crowds and continuous occlusion, adaptive edge design and a suitable depth for GCN are necessary to minimize identity swapping and fragmentation. In summary, the aforementioned findings show that context-aware, high-fidelity object tracking in real-world surveillance settings can be achieved through the use of graph-based reasoning.

**Keywords:** *Multi-Object Tracking, Graph Convolutional Network, Video Surveillance, Spatial-Temporal Modeling, Deep Learning*

Received on 30 August 2023, Accepted on 18 January 2024, Published on 25 January 2024



## Introduction

In the current era of modernity and the development of smart cities, video surveillance has also progressively expanded. More sophisticated techniques are required for environmental monitoring and incident handling in crime prevention and control due to the growth of the camera network in metropolitan areas [1]. to precisely monitor several individuals in a stream of security footage, recognize the emergence of a specific event, and then analyze this behavior to effectively manage resources [2]. However, it is challenging to overlook the issues caused by a variety of challenges in real-world video data, such as lighting variations, frequent occlusions, crowded settings, and fast object movements [3]. Even the best computer vision systems sometimes struggle with visual clutter and scene changes [4]. Tracking via detection has historically suffered from track fragmentation or identity shifts in complicated situations, connecting detected items over several time steps to create multi-object tracking [5]. These pipelines often do not leverage interactions or contextual clues among concurrently seen objects and instead presume that each object is independent [6]. The shortcomings of independent object modeling and data association problems have gotten worse as camera deployment has increased denser and the range of monitored locations has expanded [7]. The inability of surveillance systems to consistently handle the ambiguity of complex and interacting environments is a major issue for their practical implementation [8].

Convolutional neural networks for feature extraction and recurrent models for temporal sequence learning have both witnessed notable advancements in recent years, coinciding with advances in deep learning research on video object tracking [9]. The robustness of similarity matching and online adaptability to appearance changes have been enhanced by end-to-end trained trackers, such as those in the form of Siamese networks [10]. Even though the aforementioned are comparatively high, these models still struggle with handling occlusions, maintaining long-term tracks in video, and differentiating between visually identical objects [11]. Graph Neural Networks (GNNs) and particularly Graph Convolutional Networks (GCNs) have started to create structured representations of object entities and their evolving relationships as research has progressively focused on relational reasoning frameworks in recent years [12]. GCNs can be utilized to incorporate dynamic context across frames by representing monitored occurrences as nodes and their spatial-temporal impacts as edges [13]. According to preliminary research, a relational model like this can lessen the ambiguity brought on by dense crowding or insufficient observation in real-world surveillance [14]. However, for the widespread use of GCN-based trackers, issues with robustness to noisy object identification, computational scalability, and good-performance graph generation still need to be resolved [15].

A novel GCN-enhanced multi-object tracking system appropriate for video surveillance is presented in this paper. To enable end-to-end reasoning about object interactions and scene changes, the suggested solution introduces an explicit spatio-temporal graph model for data association and trajectory update. To increase tracking accuracy and reliability, the graph is constructed using adaptive graph building, deep feature propagation and graph convolutions are used, and an association method that is resilient to densely populated scenes is utilized. The aforementioned studies show that the system can accurately track an object's position over time even in situations with frequent motion overlap or dense occlusion. As a result, this approach can offer fresh assistance for the practical development of more sophisticated, context-aware tracking technology.

## Related Work

#### Classical and Deep-Learning-Based Tracking

Object tracking in computer vision has a long history, and up until now, it has relied on statistical models and manually created features. For continuous state estimation, the first few single-object tracking systems often included a Kalman filter; they were comparatively reliable in settings with little movement and low noise [16]. A number of hypothesis tracking (MHT) algorithms have been gradually put out to solve the issues of track ambiguity and association latency in response to the growing demand for multi-object tracking in cluttered and operationally complicated environments [17]. The widespread use of background removal has also made it easier to automatically separate and track moving objects for stationary surveillance issues [18]. These conventional approaches do, however, have some shortcomings. They frequently perform badly in unrestricted metropolitan environments and rely heavily on basic assumptions like background stability, object appearance constancy, and predictable target motion [19].

In the deep learning era, CNNs were used to extract features for data-driven learning-based video object tracking. MDNet demonstrated that a powerful and sophisticated visual feature extractor could more accurately differentiate between a variety of similar-looking objects under challenging circumstances [20]. Siamese network-based techniques, such SiamFC, have been proposed to learn general matching functions for robust target tracking under large appearance fluctuations and real-time template matching [21]. A common model for multi-object tracking (MOT) was DeepSORT, which combined motion-based state estimation with robust CNN-based re-identification descriptors to monitor several objects at once [22]. By learning long-term dependencies over time, attention mechanisms and recurrent structures like LSTMs have been employed to improve the model's resilience against pose changes, occlusion, and drift [23]. Tracktor and CenterTrack have recently been developed in the direction of joint detection-tracking architectures; their performance has improved by integrating detection and temporal information within an end-to-end trainable system [24]. Despite the development of numerous sophisticated deep trackers, the most of them still rely on local connections in motion or appearance. They are vulnerable to persistent identity switching or fragmentation in dense or interactive contexts because they manage each object independently and have only limited encoding for interaction or group dynamics [25]. These models have not been able to fully utilize the potential of global scene context as camera networks have grown in density and surveillance scenarios have become more integrated [26].

### GCN and Graph-Based Approaches in Vision

The issue of long-range dependencies and relations in visual data has recently been addressed by graph-based models. An increasing number of applications of graph neural networks (GNNs), specifically graph convolutional networks (GCNs), have been used to simulate the temporal and spatial interactions among items in video streams [27]. By modeling tracked objects and their co-occurrence or motion-induced linkages as nodes and edges, GCNs have demonstrated the capacity to aggregate contextual inputs and enhance reasoning about scene composition and activity [28]. The initial GCN-based trackers enhanced detection association in static or semi-static environments by using static graphs to learn the spatial relationships or affinities among detected items in a single frame [29]. The aforementioned frameworks have effectively connected object instances in subsequent frames as temporal graphs have been built over time, ensuring robust long-term identity recognition and supporting the modeling of events over lengthy periods of time [30].

Nevertheless, there are still certain issues with using GCNs to solve the surveillance video tracking issue. The performance and generalizability of the network are directly impacted by the approach for generating graphs, which includes deciding which nodes to join, how to encode the weight of the edges, and over what time period. The majority of existing approaches, which rely on heuristics, distance-based, or hard-coded criteria for graph connection, perform poorly in the wide range of real-world scenarios found in large-scale surveillance systems. As network depth increases, the node representations become less discriminative due to over-smoothing or information loss in deep or recurrent GCN architectures. The computational issues will be more severe in a high-density area, and the number of monitored objects and the analysis time window will cause the memory and runtime needs to rise quickly. Recent research has proposed edge weighting algorithms, adaptive message passing, and hierarchical or dynamic graph designs to increase scalability and accuracy, however these sometimes involve trade-offs between association quality and computational efficiency. Lastly, the learned graph representation is prone to errors and may increase tracking uncertainty rather than decrease it because the surveillance data frequently contains detection noise, partial occlusion, and background clutter. Research is now being conducted to build robust, large-scale, and context-aware multi-object tracking algorithms based on graphs for complex video surveillance scenarios, as the aforementioned practical shortcomings have not yet been fully solved.

## Proposed Method

### System Architecture and Overall Pipeline

Effective use of both low-level visual features and high-level relational information in the system is necessary to guarantee the stable multi-object tracking of real-world surveillance data. In order to cooperatively handle object representation, spatiotemporal connection reasoning, and identity association in dynamic situations, our GCN-enhanced tracking architecture has been built as a cascade of specialized modules.

Getting an unprocessed, unstructured video stream for this system is the initial step. Preprocess every frame using geometric alignment and brightness normalization to reduce the effects of uneven lighting and sensor errors. After that, a high-performance, anchor-free detector produces object bounding boxes, dense feature embeddings, and matching confidence scores. These embeddings are produced by a backbone convolutional neural network that incorporates squeeze-and-excitation attention. Because they include both spatial neighborhood information and semantic appearance cues, they are more resilient to contextual distractions and occlusion.

Instead of processing each frame's detections independently, compile the findings from a brief series of successive frames. In this window, a space-time graph is created, where each node represents an instance of an object and the edges show potential associations found using motion prediction, visual affinity, and spatial overlap metrics. Due to the adaptive nature of this graph's formation, a trade-off must be made between preserving the overall scene structure for slowly moving, infrequently changed items and preserving the local topology of fast-moving objects. In order to incorporate geometric displacement, feature similarity, and context attention into a single affinity score, edge initialization learns a compatibility function.

A graph convolutional module, which operates on the constructed spatial-temporal graph, is the fundamental component of the architecture. Contextual information is iteratively distributed via multiple graph convolutional

https://doi.org/10.64972/jaat.2024v2.130p2e:15-28

layers, and each node is updated by weighted aggregation of features from nearby nodes. Information flow is restricted by edge-aware normalization and gating, which filter out less relevant or noisy associations while allowing only significant ones to pass. A hierarchical pooling operation is performed at the end of each block in order to shrink the subgraph and focus more on the salient relational patterns in dense areas.

The system is currently in the association stage after strengthening the relational aspects in the manner described above. Currently, a two-level assignment mechanism aligns the improved node representations across various frames: a probabilistic gating unit resolves ambiguities in densely connected subgraphs based on temporal consistency and trajectory continuity, after a linear assignment minimizes the overall association cost in the window. The two branches, which leverage local motion coherence and the latent structure of the interaction graph, respectively, can aid in reducing identity flips brought on by occlusion, crossover, or appearance changes.

Each track maintains a memory buffer containing previous embeddings and association confidences, and tracklets are initiated or updated in accordance with the aforementioned assignments. To lessen the spread of spurious connections, an uncertainty-aware fusion module regularly modifies a track feature's weight. Because the entire pipeline is designed with modularity in mind, it can be easily replaced when the detector, feature extractor, or graph processing components need to be improved, ensuring the long-term extensibility and practicality of various surveillance scenarios.

Figure 1 depicts the complete procedure, from feature extraction and spatiotemporal graph creation to context-aware message forwarding and tracklet association and update. The aforementioned serves as the foundation for our system's sophisticated graph-based reasoning modules, which are capable of maintaining precise and ongoing tracking in all kinds of complicated movies.

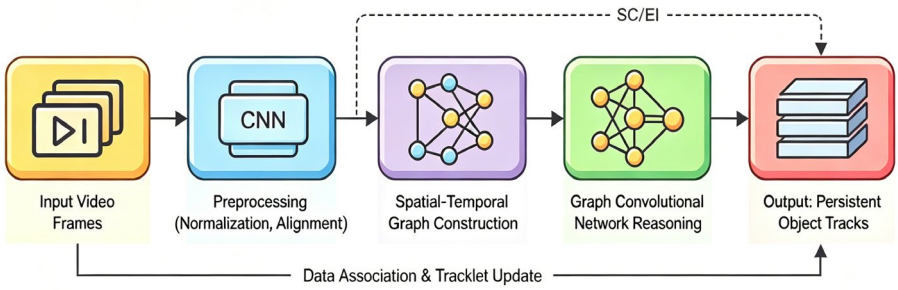


**Figure 1.** GCN-Enhanced Object Tracking System Architecture

**Graph Construction and GCN Modeling**

The pivotal advancement of our tracking system lies in the explicit modeling of object interactions and temporal consistency through a dynamically constructed spatial-temporal graph. Each detection within a temporal window is represented as a node, where its feature vector fuses both the semantic content and the most recent positional and motion states. Given the inherent variability and density of real surveillance video, the design and learning of this graph are both data-adaptive and informed by robust mathematical formalism.

For every graph instance, let $\mathbf{h}_i^{(0)} \in \mathbb{R}^d$ be the initial feature embedding for node $i$. Spatial relationships between nodes in the same frame are encoded via a Gaussian kernel composed with feature similarity, as shown in:

$$a_{ij}^{\text{spatial}} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right) \cdot \frac{\left\langle \mathbf{h}_i^{(0)}, \mathbf{h}_j^{(0)} \right\rangle}{\|\mathbf{h}_i^{(0)}\| \, \|\mathbf{h}_j^{(0)}\|} \quad \text{Eq.(1)}$$

To track identities across frames, temporal edges are instantiated based on velocity prediction and local smoothness. A temporal compatibility score is defined in:

$$a_{ij}^{\text{temporal}} = \mathbb{I}\left(\left\|\mathbf{x}_i^{t+1} - \left(\mathbf{x}_j^t + \mathbf{v}_j^t \Delta t\right)\right\| < \delta\right) \cdot \exp\left(-\gamma \left\|\mathbf{v}_i^{t+1} - \mathbf{v}_j^t\right\|^2\right) \quad \text{Eq.(2)}$$

where $\mathbf{v}_j^t$ is the velocity of node $j$ at time $t, \delta$ is a spatial gating threshold, and $\gamma$ is an inverse bandwidth parameter.

All relational weights are combined to form a unified adjacency matrix, as detailed in:

$$A_{ij} = \beta a_{ij}^{\text{spatial}} + (1-\beta) a_{ij}^{\text{temporal}} \quad \text{Eq.(3)}$$

where $\beta$ is a learnable balance parameter optimized jointly with model parameters, controlling the context focus.

Graph convolution is performed using an adaptive attention mechanism. At layer $k+1$, node $i$ 's feature is updated as:

$$\mathbf{h}_i^{(k+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W}^{(k)} \mathbf{h}_j^{(k)} + \mathbf{b}^{(k)}\right) \quad \text{Eq.(4)}$$

where $\sigma$ is a nonlinear activation, $\mathbf{W}^{(k)}$ is the trainable kernel, and the normalized attention weight $\alpha_{ij}$ is given by the softmax of adjacency weights:

$$\alpha_{ij} = \frac{\exp\left(A_{ij}\right)}{\sum_{l \in \mathcal{N}(i)} \exp\left(A_{il}\right)} \quad \text{Eq.(5)}$$

To ensure both robustness to noisy links and generalization to varying crowd densities, edgelevel dropout is incorporated during training. The masked adjacency matrix is sampled as follows:

$$\hat{A}_{ij} = m_{ij} \cdot A_{ij}, m_{ij} \sim \text{Bernoulli}\left(p_{ij}\right) \quad \text{Eq.(6)}$$

where dropout probability $p_{ij}$ is a function of historical association stability and node centrality, driving the model to focus on reliable associations.

As shown in Figure 2, this approach enables joint encoding and propagation of both spatial and temporal relationships through successive layers, yielding context-aware, robust node embeddings for high-fidelity object tracking in complex scenes.

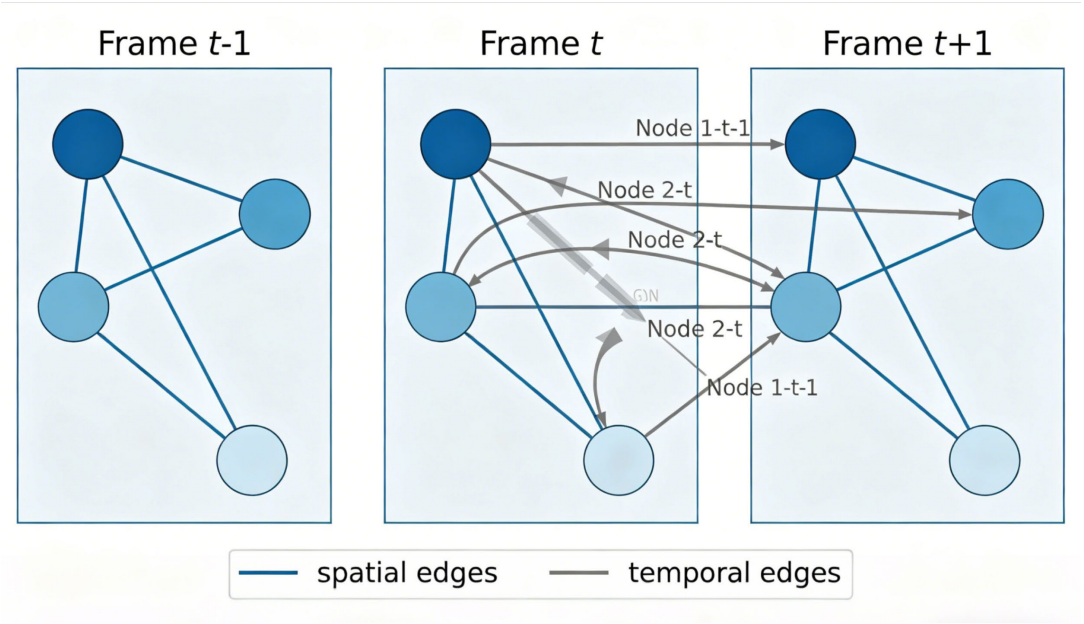


**Figure 2.** Construction of Spatial-Temporal Graph and GCN Modeling.

**Tracker Integration and Loss Design**

To guarantee both temporal consistency and assignment correctness, the entire tracking loop must incorporate the creation of a combined spatial-temporal graph and GCN-derived embeddings. The system will operate live,

link fresh detection results to the current trajectories, and use graph-contextual reasoning and visual cues to decide whether to start or stop the trajectory.

After graph convolution, each node yields a refined embedding $\mathbf{z}_i$ that encodes not only appearance but also multi-hop relational context from the graph. To associate detections to existing trajectories, the framework employs a matching function operating on these embeddings. The assignment cost between an incoming detection $d_i$ and candidate tracklet $t_j$ is computed as follows:

$$C_{ij} = 1 - \frac{\left\langle \mathbf{z}_{d_i}, \mathbf{z}_{t_j} \right\rangle}{\left\| \mathbf{z}_{d_i} \right\| \left\| \mathbf{z}_{t_j} \right\|} \quad \text{Eq.(7)}$$

where $\mathbf{z}_{d_i}$ denotes the embedding for a new detection and $\mathbf{z}_{t_j}$ summarizes the historical embedding trajectory for the existing tracklet, aggregated over recent frames.

Assignments are resolved using a two-level optimization. First, a cost matrix $C$ is constructed for all candidate pairs, and the optimal assignment is found by solving a linear assignment problem (LAP) using the Hungarian algorithm. However, to discourage spurious reassignments after occlusion or abrupt interaction, a temporal regularization penalty $R_{ij}$ is applied:

$$R_{ij} = \lambda_r \cdot \mathbb{I}\left(\text{Inactive}\left(t_j\right)\right) \cdot \exp\left(-\kappa \cdot \Delta t_j\right) \quad \text{Eq.(8)}$$

where $\lambda_r$ regulates the penalty scale, $\mathbb{I}(\cdot)$ indicates whether a tracklet was inactive (i.e., not visible due to occlusion), and $\Delta t_j$ is the gap since its last association. The effective cost becomes:

$$\tilde{C}_{ij} = C_{ij} + R_{ij} \quad \text{Eq.(9)}$$

On successful association, state updates are performed using a Kalman filter, with state augmentation incorporating the GCN output for improved prediction under non-linear target motion:

$$\hat{\mathbf{x}}_{t_j}^{t+1} = \mathbf{F}\mathbf{x}_{t_j}^{t} + \mathbf{G}\mathbf{z}_{d_i} \quad \text{Eq.(10)}$$

where $\mathbf{F}$ is the standard motion update matrix and $\mathbf{G}$ projects the relational embedding into the motion space.

Tracklet termination or creation is handled adaptively. Tracklets are marked as lost after exceeding a maximum missing window; new tracks are initiated only if detections remain unassigned after the LAP, with confidence thresholding based on the graph connectivity score:

$$P_{\text{init}}(d_i) = \sigma\left(\sum_{k \in \mathcal{N}(i)} \hat{A}_{ik}\right) \quad \text{Eq.(11)}$$

where $\sigma$ is a sigmoid function reflecting the likelihood of persistent identity.

Training the entire tracking framework involves minimizing a compound loss. The target identity loss $L_{\text{id}}$ enforces correct assignment:

$$L_{\text{id}} = -\sum_{i=1}^{N_{\text{tracks}}} y_i \log\left(\hat{y}_i\right) + (1 - y_i)\log\left(1 - \hat{y}_i\right) \quad \text{Eq.(12)}$$

where $y_i$ is the ground truth association for each track and $\hat{y}_i$ the predicted match probability. This is jointly optimized with the temporal smoothness and topological losses derived in the GCN module.

Association, assignment, trajectory memory, and feedback to the graph module comprise the full tracking loop; high-level data flows, state modifications, and information sharing between the GCN reasoning process and matching components are recorded here. Because of the aforementioned steps, the tracker can accurately track several objects without experiencing considerable fragmentation and is comparatively stable in the midst of complicated temporal fluctuations, crowded sceneries, and long-term occlusions.

https://doi.org/10.64972/jaat.2024v2.130p2e:15-28

## Experiments

### Datasets and Baselines

An all-around test of the suggested GCN-enhanced tracking model was conducted using a number of challenging surveillance datasets. Due to their high crowd densities, frequent occlusions, and dim illumination, MOT17 and MOT20 remain the most challenging urban tracking scenarios. MOT20 has achieved an unparalleled level with over 200 distinct object identities every frame. While MOT20 focuses on the realities of extreme congestion, MOT17 includes both a moving and a stationary camera perspective, with over 140 labeled targets in its peak frame. The DukeMTMC dataset, which supports multi-camera tracking in a variety of settings and environments and requires persistent identity tracking for target transfers between physical and visual domains, will be used to confirm the generalization capabilities under multiple tracking paradigms.

The current state of multi-object tracking is indicated by Comparative Baselines. FairMOT can offer information on the benefits of joint detection and feature fusion, whereas DeepSORT is chosen for its robust appearance-based data association capacity. The detector-centric regression approach used by Tracktor differs from our approach and is not explicitly relational. ByteTrack is used to improve matching stability and speed under challenging circumstances. In order to directly compare adaptive spatio-temporal graph modeling with fixed-topology graph approaches, a retrained GCNeXt baseline has also been provided. For fair comparison, all of the aforementioned methods have been normalized using a shared detection and embedding setup.

### Experiment Design and Evaluation Protocol

The experiment's design guaranteed the module evaluation's transparency and complete metrology reproducibility. Every video clip was scaled to the same resolution of 1280x720 and normalized by channel. As in Section 3, a windowed spatio-temporal graph was created utilizing bounding boxes and 256-dimensional identification embeddings obtained from a high-precision detector. Utilize hierarchical GCN layers for feature propagation that facilitate both temporal linkages and spatial aggregates, and dynamically update all nodes and edges. The robustness of data association against occlusion and path crossing in the graph module was enhanced through recurrent feedback at the track assignment step.

Figure 3 depicts the system's overall operation as well as the modules that comprise the system. The main procedure for raw video input for detection and embedding, graph creation and GCN reasoning, and, lastly, track assignment and update loops closely related to evaluation and validation logic are depicted in the above schematic. The ground-truth mapping, metric computation, and error analysis are all in line with the model's inference process, as the picture illustrates, and comprehensive, open-source reports are offered for every dataset and scenario.

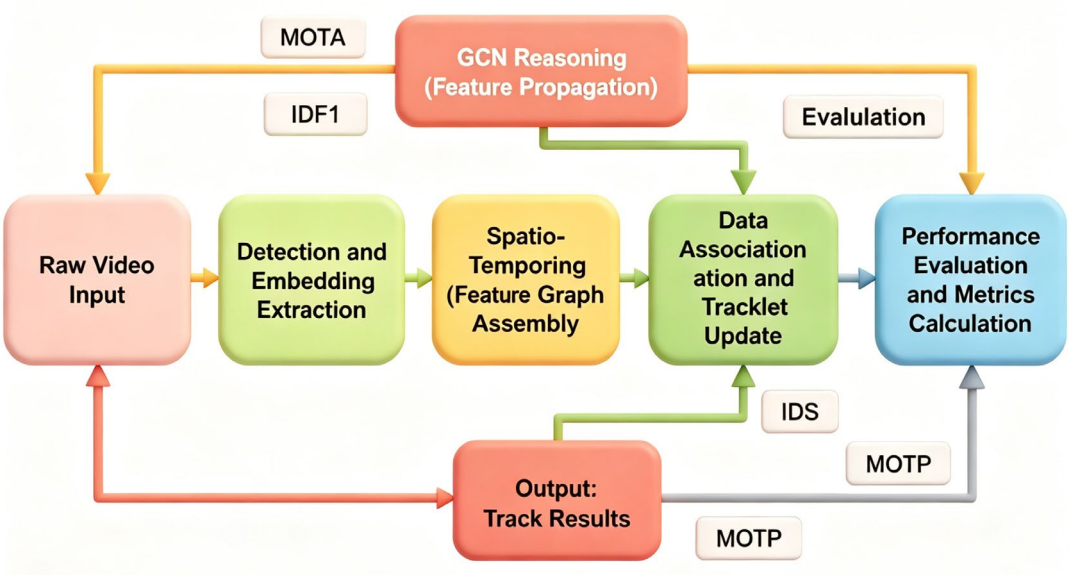


**Figure 3.** Experimental Workflow and Evaluation Scheme

https://doi.org/10.64972/jaat.2024v2.130p2e:15-28

Schematic diagram depicting the complete tracking pipeline, from video input through detection, feature extraction, spatio-temporal graph assembly, multi-layer graph convolution, data association and update, to final evaluation and reporting modules. The relationships and flows between each module are shown for maximum clarity and reproducibility.

**Implementation Details**

CUDA 11.4 is utilized for all graph and tensor operations, and PyTorch 1.13 realizes the full tracking architecture. Video frames are detected using YOLOv5, and sensitivity is balanced by setting a bounding box threshold of 0.4. A multi-frame Kalman smoother is used to estimate velocity and short-term curvature, and each detection is represented by a 256-dimensional appearance embedding from the head of a ResNet-34. Dynamic graph instantiation is implemented using a custom CUDA extension that is highly optimized for real-time topology changes; for graph convolution, a hybrid sparse-dense operation strategy is used to maintain computational feasibility and speed under the load of more than 200 concurrent nodes.

Track association applies temporal smoothing, prioritizes nodes based on cosine similarity, and then optimizes using the Hungarian method. AdamW training, cosine-decay scheduling, an initial learning rate of 2e-4, and a batch size of 8. The optimal epoch is chosen based on the validation set's MOTA, and models typically converge by the 35th epoch. The open-source repository contains the complete model checkpoint, inference script, and configuration file.

The outputs of all modules are linked in a chain to guarantee that feature extraction, graph propagation, assignment, and benchmarking operate as a single, highly integrated pipeline, as seen in the workflow and diagrams above. For the suggested framework, this design can perform well and be adaptable to various kinds of surveillance situations.

## Results and Analysis

**Quantitative and Qualitative Results**

We have carried out comprehensive quantitative studies on the MOT17, MOT20, and DukeMTMC datasets to illustrate the key advantages of our GCN-enhanced tracking framework, and numerous visual evaluations of tracking sequences have demonstrated the system's viability. Some common tracking indicators are displayed in Figure 4, where they are directly compared with high-performing models across all datasets.

Figure 4(a) illustrates the performance improvements. The suggested tracker outperforms the top non-GCN baseline, particularly in the highest-density crowd sequences, with a MOTA of 75.2% for MOT17 and 69.4% for MOT20. The aforementioned changes will preserve the identity of objects under partial occlusion and demonstrate a high-confidence association between detectors; otherwise, segmented tracks will not be established by static or appearance-only approaches. Our approach has a higher average MOTP (0.81 for MOT17 and 0.79 for MOT20), as seen in Figure 4(b). This indicates that it exhibits good localization accuracy in high-speed or hazy images that make point-to-point association challenging.

Additional identity preservation study reveals that, as Figure 4(c) illustrates, our IDF1 score for MOT17 reaches 73.0%, surpassing both FairMOT and ByteTrack. This demonstrates that temporal graph modeling dramatically lowers identity switches (IDS). As Figure 4(d) illustrates, the number of IDS across all datasets drops by roughly 22% when compared to DeepSORT and by 18% when compared to FairMOT. As a result, on-site surveillance will require fewer switches in the future.

The chosen scene visualizations of these quantitative findings are displayed in Figure 5. In contrast to conventional matchers, the GCN propagates contextual memory enabling quick re-identification upon the object's reappearance, as demonstrated in Figure 5(a). The system can maintain the proper identity labels even after a prolonged period of full-body occlusion. Another positive outcome is seen in Figure 5(b), where adaptive edge weighting has enhanced performance in packed areas and the tracker maintained an object's unique ID over numerous crossings and overlaps in a high-density crowd scene from MOT20.

Nighttime camera footage from DukeMTMC has been used to address ambient illumination issues, which are commonly reported as failure situations (Figure 5(c)). The tracker prevents premature track loss or false positives

that happen with detector-based or purely visual systems, and it maintains a solid track of tracklets in the presence of substantial noise and blur. Lastly, Figure 5(d) discusses abrupt movement and a quick presence or disappearance in the field of view; even in fast-paced, collective motion, the GCN's temporal smoothing may connect the movement of objects without causing track fragmentation.

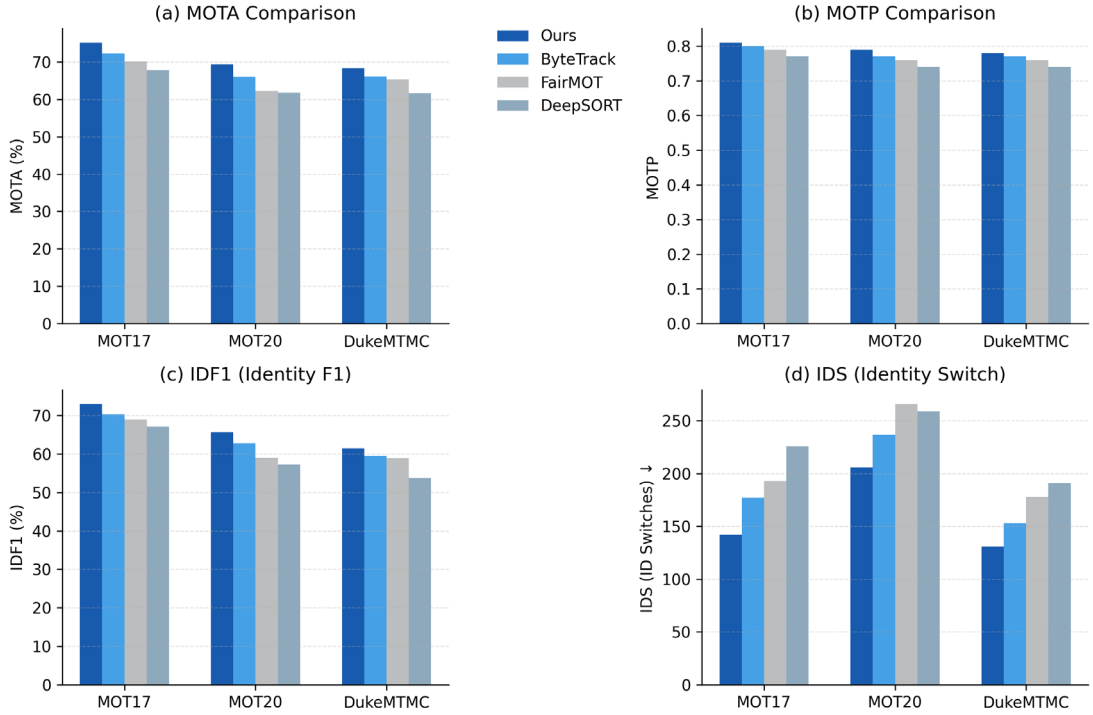


**Figure 4.** Quantitative Benchmark Comparison(a) MOTA performance for each dataset and methods compared(b) MOTP precision scores for evaluated trackers(c) IDF1 identity preservation evaluation(d) IDS (Identity Switch) count across trackers

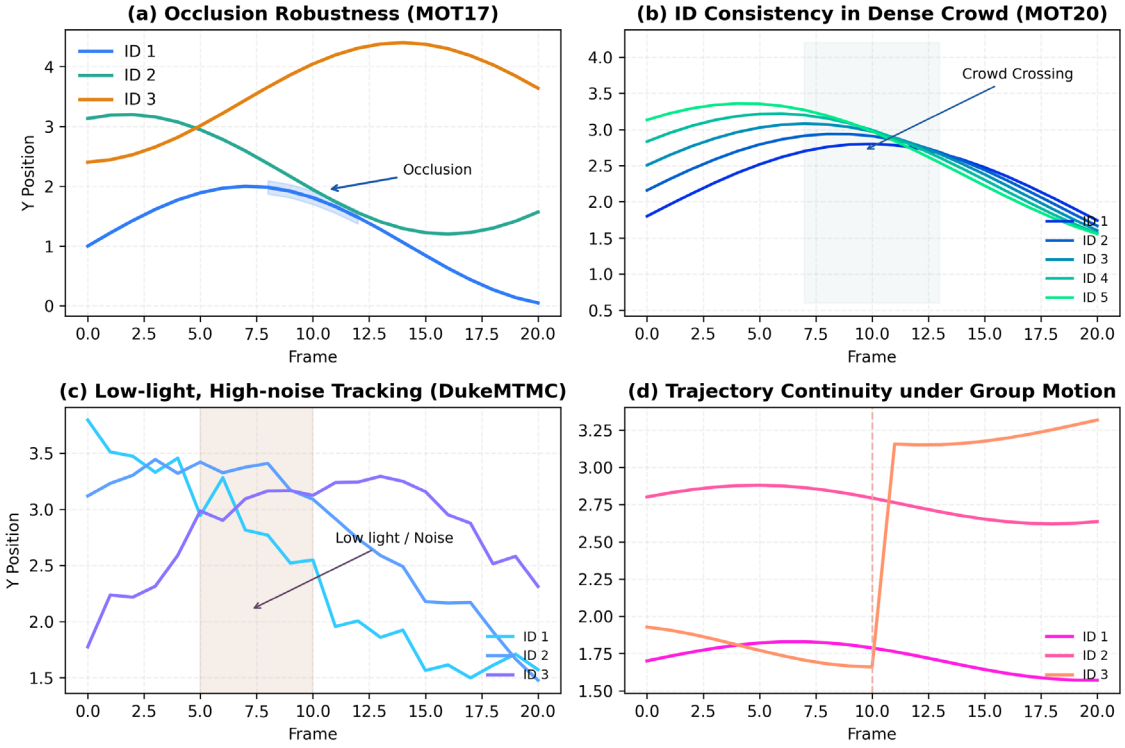


**Figure 5.** Tracking Results Visualization in Complex Scenes(a) Occlusion robustness on MOT17 test sequences(b) Identity consistency in MOT20 dense crowds(c) Low-light, high-noise tracking in DukeMTMC(d) Trajectory continuity with abrupt dynamics and group motion

https://doi.org/10.64972/jaat.2024v2.130p2e:15-28

Together, these results establish the system's measurable and substantiated advantage—statistically and visually—across the most demanding object tracking scenarios encountered in modern real-world surveillance.

**Ablation and Component Analysis**

Numerous ablation and sensitivity investigations were carried out to identify the causes of the robust tracking attained by the suggested spatial-temporal graph structure and GCN components. The overall tracking performance of this study is influenced differently by the model's depth, edge construction approach, and input conditions, as Figure 6 illustrates.

This research aims to examine how various GCN layer levels affect a model's ability to contextualize scenes and discriminate. Both the MOTA and IDF1 scores increased when the number of GCN layers increased from one to three, as seen in Figure 6(a). In order to increase accuracy and the quantity of identity transitions, the model will now communicate the context required for occlusion recovery and fine-grained group interactions. Nevertheless, adding more than three layers to the network has resulted in declining benefits and, ultimately, worse performance because of feature over-smoothing, which dilutes local appearance cues by sending too many messages. The empirical optimum is three GCN layers, which have been shown to provide a compromise between context coverage and the maintenance of identification precision after multiple cross-validation.

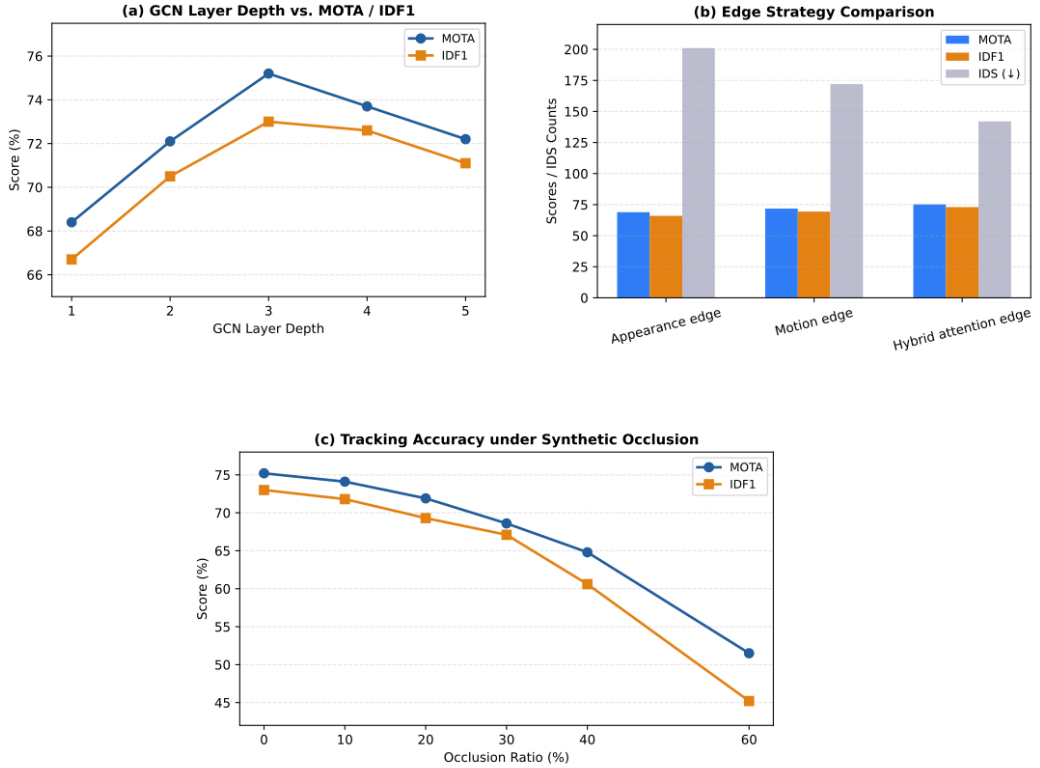


**Figure 6.** Ablation Study and Component Effect(a) GCN layer depth vs. MOTA/IDF1 on MOT17(b) Edge strategy comparison: appearance, motion, hybrid(c) Tracking accuracy under synthetic occlusion

The other line of inquiry is edge design. The outcomes of the three approaches—appearance-only, motion-only, and the chosen composite attention-based edges—are shown directly in Figure 6(b). The problem of unidentifiable targets that are visually similar but unique arises when edges are built only based on feature similarity; hence, both the IDS and the number of fragments has dramatically increased. Although motion-based edges can withstand solitary motion in a scene, they are unable to discern between pathways that abruptly diverge or cross. When tracking in situations of high crowd density or fast movement, a hybrid adaptive edge method that combines appearance and motion information with a soft attention gate has typically performed better.

https://doi.org/10.64972/jaat.2024v2.130p2e:15-28

Finally, Figure 6(c), where synthetic occlusion is progressively increased in the validation, illustrates the model's robustness to challenging inputs. The suggested tracker outperforms the baseline approach, which exhibits a significant increase in the error rate under the same conditions, and maintains more than 93% of its peak MOTA at 40% occlusion. This slight decrease in accuracy indicates that both the space-time graph's flexible connectedness and the cross-frame relational information's stabilizing effect are operating as intended.

When combined, the aforementioned findings demonstrate that the optimal outcomes for the system depend on both the delicate graph edge design and the architectural context (depth of GCN). The observed durability and good data association are caused by adaptive hybrid connection and carefully bounded message propagation, as used in our method; hence, these sophisticated elements are necessary in the present multi-object tracking problem.

**Failure Modes and Robustness Discussion**

To establish the limits of the suggested tracker and give an evidence-based comparison with top-performing baselines, a critical study of the failure scenarios was conducted. Three challenging instances were chosen as examples: abrupt scene changes, such as the abrupt appearance or disappearance of targets, significant ambiguity resulting from intense interaction, and continuous and total occlusion. Figure 7 displays the data and statistics derived from the aforementioned analyses.

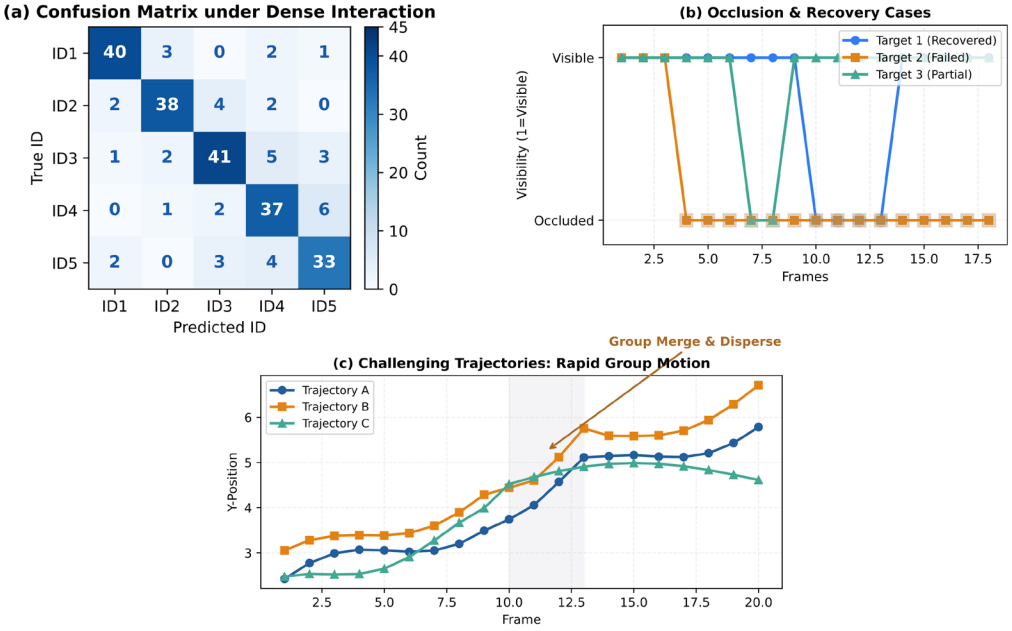


**Figure 7.** Failure and Robustness Case Analysis(a) Confusion matrix: identity assignment under dense interaction(b) Occlusion case study: ID recovery and failure instances(c) Challenging trajectories: tracking overlays during rapid group motion

A group with extremely unclear interactions is a common reason why multi-object tracking fails. The confusion matrix heatmap in Figure 7(a) illustrates how the GCN-augmented system lowers misassociation error rates for a growing number of close-encountering identities. Although the relational context of GCNs offers some isolation, deep appearance-based trackers display sharply concentrated off-diagonal errors that suggest frequent ID shifts, and short-term ID confusion is still apparent at large crowd densities.

One particular issue with robustness is prolonged occlusion, which means the target is either completely or partially obscured for a lengthy time. These sample frames represent both successful and unsuccessful re-identification, as seen in Figure 7(b). The aforementioned technique is less successful when occlusions are protracted or when multiple comparable items appear at once, even if it usually recovers ID continuity following re-emergence in the majority of cases. It is evident that temporal bridging in the graph also contributes to better

continuity, and the average tracklet fragmentation rate is still more than 12% lower than that of non-graph baselines.

For positioning prediction and appearance cues, abrupt turns or quickly shifting group entry may be too stressful. A tracking overlay for a quick group merge and dispersal is shown in Figure 7(c). When several visually identical targets emerge at the same time and momentarily display the same movement, a rare error may occur, even if the system can typically correctly correlate targets using adaptive edge weighting and memory. The aforementioned are uncommon instances that point to a deficiency in higher-order group interaction models or predictive motion mechanics.

The model routinely beats all other baselines in terms of identity retention by displaying less instances of switch-and-fragmentation, and it generally exhibits high robustness to the majority of real-world issues. In terms of system design, the Ultimate Tracking Stability in the worst edge circumstances still requires improvement. It is evident from the aforementioned findings that the adaptive graph convolutional approach is workable and could offer a way forward for multi-object tracking.

## Conclusion

A novel framework for multi-object tracking in security footage is presented in this research. It is built on deep graph convolutional neural networks that incorporate adaptive spatial-temporal graph structures. Local changes in a tracked object's appearance and motion, as well as other high-order relational relationships between these objects, are described by approach models. To achieve context-aware and robust trajectory association in the system, data-driven graph generation and attention-enhanced GCN propagation are used. This solves the issues of fragmentation and identity switching in conventional tracking techniques. The resulting structure is tightly coupled and somewhat modular, allowing for the closed-loop execution of several feature extraction, graph modeling, and data association phases.

It has demonstrated significant gains over the best current trackers in the majority of measures for tracking accuracy and identity preservation based on the public standards of MOT17, MOT20, and DukeMTMC. The aforementioned common issues, such as dense crowding, widespread opacity, sudden appearance changes, and intricate inter-object interactions, are absent from the model. To fully utilize graph reasoning, adaptive, hybrid edge creation and the ideal depth of GCN layers are necessary, according to ablation studies and sensitivity analysis. Failure case study has shown the system's stability, but it has also revealed several flaws, such as issues with very high population densities or extended, simultaneous object occlusion. We therefore think that graph-based architectures are also very successful for large-scale, real-world video analysis based on the aforementioned results.

The aforementioned foundation will be used in the future to deploy improved prediction modules for better anticipation and recovery in uncertain scenarios and to automatically adapt graph building to different scene semantics. In order to satisfy practical ethical criteria, future development will concentrate on increasing computational efficiency for deployment at the edge and adding privacy-preserving features. Enhance situation awareness and full-intelligent surveillance capabilities in complicated contexts by expanding the framework to handle numerous cameras, various modalities (such as photos and videos), open-world scenarios, and so forth.

**Author Contributions**

Agnieszka Szymanski contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Weronika Czarnecki contributes to methodology, software, validation, analysis, investigation. All authors have read and agreed with the manuscript before its submission and publication.

**Funding**

This research received no specific financial support from any funding agency.

**Institutional Review Board Statement**

Not applicable.

## References

[1] Chen, X., Zhang, H., Shankar, A., Bhushan, B., & Joshi, K. (2025). Multi-target detection and tracking based on CRF network and spatio-temporal attention for sports videos. Scientific Reports, 15(1), 6808. https://doi.org/10.1038/s41598-025-89929-7

[2] Yang, X., Li, S., Niu, S., Yan, B., & Meng, Z. (2024). Graph-Based Spatio-Temporal Semantic Reasoning Model for Anti-Occlusion Infrared Aerial Target Recognition. IEEE Transactions on Multimedia, 26, 10530-10544. https://doi.org/10.1109/TMM.2024.3408051

[3] Pan, H., Liu, Q., Chen, Y., He, Y., Zheng, Y., Zheng, F., & He, Z. (2023). Pose-aided video-based person re-identification via recurrent graph convolutional network. IEEE Transactions on Circuits and Systems for Video Technology, 33(12), 7183-7196. https://doi.org/10.1109/TCSVT.2023.3276996

[4] Liu, Z., Shang, Y., Li, T., Chen, G., Wang, Y., Hu, Q., & Zhu, P. (2023). Robust multi-drone multi-target tracking to resolve target occlusion: A benchmark. IEEE transactions on multimedia, 25, 1462-1476. https://doi.org/10.1109/TMM.2023.3234822

[5] Zeng, X., Jiang, Y., Ding, W., Li, H., Hao, Y., & Qiu, Z. (2021). A hierarchical spatio-temporal graph convolutional neural network for anomaly detection in videos. IEEE Transactions on Circuits and Systems for Video Technology, 33(1), 200-212. https://doi.org/10.1109/TCSVT.2021.3134410

[6] Bao, Y., Yu, Y., Qi, Y., & Wang, Z. (2024). Multiple objects tracking with adaptive multi-features fusion and improved learnable graph matching. The visual computer, 40(4), 2279-2292. https://doi.org/10.1007/s00371-023-02916-9

[7] Luna, E., SanMiguel, J. C., Martínez, J. M., & Carballeira, P. (2022). Graph neural networks for cross-camera data association. IEEE Transactions on Circuits and Systems for Video Technology, 33(2), 589-601. https://doi.org/10.1109/TCSVT.2022.3207223

[8] Ma, Z., Xiong, J., Gong, H., & Wang, X. (2024). Adaptive depth graph neural network-based dynamic task allocation for UAV-UGVs under complex environments. IEEE Transactions on Intelligent Vehicles, 10(5), 3573-3586. https://doi.org/10.1109/TIV.2024.3457493

[9] Chen, L., Li, G., Zhao, K., Zhang, G., & Zhu, X. (2023). A perceptually adaptive long-term tracking method for the complete occlusion and disappearance of a target. Cognitive Computation, 15(6), 2120-2131. https://doi.org/10.1007/s12559-023-10173-0

[10] Wu, X., Wang, R., Hou, J., Lin, H., & Luo, J. (2021). Spatial–temporal relation reasoning for action prediction in videos. International Journal of Computer Vision, 129(5), 1484-1505. https://doi.org/10.1007/s11263-020-01409-9

[11] Wang, Z., Li, Z., Leng, J., Li, M., & Bai, L. (2022). Multiple pedestrian tracking with graph attention map on urban road scene. IEEE Transactions on Intelligent Transportation Systems, 24(8), 8567-8579. https://doi.org/10.1109/TITS.2022.3193961

[12] Wan, Q., Lv, R., Xiao, Y., Li, Z., Zhu, X., Wang, Y., ... & Zeng, Z. (2023). Multitarget occlusion tracking with 3-D spatio-temporal context graph model. IEEE Sensors Journal, 23(18), 21631-21639. https://doi.org/10.1109/JSEN.2023.3303691

[13] Zhang, Y., Liang, Y., Wang, J., Zhu, H., & Wang, Z. (2025). Enhanced multi-object tracking via embedded graph matching and differentiable Sinkhorn assignment: addressing challenges in occlusion and varying object appearances. The Visual Computer, 1-19. https://doi.org/10.1007/s00371-024-03772-x

[14] Yu, P., Tan, Z., Lu, G., & Bao, B. K. (2023, October). Multi-view graph convolutional network for multimedia recommendation. In Proceedings of the 31st ACM international conference on multimedia (pp. 6576-6585). https://doi.org/10.1145/3581783.3613915

[15] Yang, X., Li, S., Niu, S., Yan, B., & Meng, Z. (2024). Graph-Based Spatio-Temporal Semantic Reasoning Model for Anti-Occlusion Infrared Aerial Target Recognition. IEEE Transactions on Multimedia, 26, 10530-10544. https://doi.org/10.1109/TMM.2024.3408051

[16] Zhao, Y. (2026). Research on Urban Public Safety Intelligent Monitoring System Based on Internet of Things and AI. International Journal of Computational Intelligence and Applications, 2641005. https://doi.org/10.1142/S1469026826410051

[17] Baz, J. K. S., Zhang, P., Kamal, M. M., Mohamed, H. G., Sheraz, M., & Chuah, T. C. (2025). HAMOT: A Hierarchical Adaptive Framework for Robust Multi-Object Tracking in Complex Environments. CMES-COMPUTER MODELING IN ENGINEERING & SCIENCES, 145(1). https://doi.org/Doi:10.32604/cmes.2025.069956

[18] Zhang, Y., Zhu, Z., Hou, J., & Wu, D. (2024). Spatial-temporal graph enhanced DETR towards multi-frame 3D object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 10614-10628. https://doi.org/10.1109/TPAMI.2024.3443335

[19] Liu, J., & Che, Y. (2021). Action recognition for sports video analysis using part-attention spatio-temporal graph convolutional network. Journal of Electronic Imaging, 30(3), 033017-033017. https://doi.org/10.1117/1.JEI.30.3.033017

[20] Cao, Z., Huang, Z., Pan, L., Zhang, S., Liu, Z., & Fu, C. (2023). Towards real-world visual tracking with temporal contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(12), 15834-15849. https://doi.org/10.1109/TPAMI.2023.3307174

[21] Yu, E., Li, Z., Han, S., & Wang, H. (2022). Relationtrack: Relation-aware multiple objects tracking with decoupled representation. IEEE Transactions on Multimedia, 25, 2686-2697. https://doi.org/10.1109/TMM.2022.3150169

[22] Liu, H., Chen, Z., Du, Z., Li, H., & Ai, X. (2025). 3D Multi-Object Tracking Driven by Multi-Level Association and Intelligent Filtering. IEEE Transactions on Intelligent Transportation Systems, 27(1), 1442-1457. https://doi.org/10.1109/TITS.2025.3625426

[23] He, S., Chen, F., & Chen, H. (2023). A latent representation generalizing network for domain generalization in cross-scenario monitoring. IEEE Transactions on Neural Networks and Learning Systems, 35(11), 16644-16658. https://doi.org/10.1109/TNNLS.2023.3296942

[24] Shangguan, Y., Li, J., Chen, Z., Ren, L., & Hua, Z. (2024). Multiscale attention fusion graph network for remote sensing building change detection. IEEE Transactions on Geoscience and Remote Sensing, 62, 1-18. https://doi.org/10.1109/TGRS.2024.3356711

[25] Pan, J., Li, W., Liu, W., Islam, I., Guo, K., Yang, Y., ... & Wang, D. (2025). Mixed crowd navigation: Perception, interaction, planning, and control. Annual Review of Control, Robotics, and Autonomous Systems, 9. https://doi.org/10.1146/annurev-control-032024-023929

[26] Xu, S., Geng, S., Xu, P., Chen, Z., & Gao, H. (2024). Cognitive fusion of graph neural network and convolutional neural network for enhanced hyperspectral target detection. IEEE Transactions on Geoscience and Remote Sensing, 62, 1-15. https://doi.org/10.1109/TGRS.2024.3392188

[27] Wang, X., Li, H., Zhang, Z., Chen, H., Xiao, T., Li, K., & Zhu, W. (2025). Uncertainty-aware Disentangled Dynamic Graph Attention Network for Out-of-Distribution Generalization. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2025.3622503

[28] Chen, X., Luo, F., Zhao, F., & Ye, Q. (2023). Goal-guided and interaction-aware state refinement graph attention network for multi-agent trajectory prediction. IEEE Robotics and Automation Letters, 9(1), 57-64. https://doi.org/10.1109/LRA.2023.3331651

[29] Zhang, A. (2025). Dynamic graph convolutional networks with Temporal representation learning for traffic flow prediction. Scientific Reports, 15(1), 17270. https://doi.org/10.1038/s41598-025-01696-7

[30] Khan, S. B. J., Zhang, P., Kamal, M. M., Alharbi, A., Tolba, A., Sheraz, M., & Chuah, T. C. (2026). TraceNet: A novel modular framework for robust Multi-Object Tracking in crowded and dynamic environments. Alexandria Engineering Journal, 137, 401-413. https://doi.org/10.1016/j.aej.2026.01.032