Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Robust Continuous Person Tracking in Dense Multi-Camera Environments through Decoupled Graph Learning

Morteza Akbari, *Seyyed Mohammad Razavi*, *Sajad Mohamadzadeh*[*]

Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.

Article Info

Article History:
Received 01 July 2025
Reviewed 03 August 2025
Revised 11 October 2025
Accepted 29 October 2025

Keywords:
Person tracking
Multi-camera environment
Deep learning
Image processing
Object tracking
Spatio-temporal features
Graph neural networks

[*]Corresponding Author's Email Address:
s.mohamadzadeh@birjand.ac.ir

Abstract

Background and Objectives: Multi-object tracking in dense, multi-camera environments remains challenging due to occlusions, lighting variations, and fragmented trajectories. While existing methods rely on hierarchical two-step approaches or complex Bayesian filters, they often fail to fully exploit spatio-temporal correlations or to approach global consistency across cameras and frames. This study aims to address these limitations by proposing a novel graph-based deep learning model for continuous person tracking that independently optimizes spatial and temporal associations.

Methods: The proposed model decomposes multi-camera tracking into two tasks: temporal association (linking objects across frames using velocity and time) and spatial association (aligning objects from multiple viewpoints). A spatio-temporal graph structure is constructed, with nodes representing detected objects and edges encoding relationships. Message Passing Networks (MPNs) iteratively update node and edge features, while a graph consensus fusion module merges spatial and temporal graphs for robust tracking. The model is trained using Focal Loss and evaluated on the Wildtrack and CAMPUS datasets.

Results: The model achieves state-of-the-art performance, with a MOTA score of 85.5% on Wildtrack and 77.4–87.4% on CAMPUS subsets. Key improvements include a 100% MT (mostly tracked) rate and 0% ML (mostly lost) rate on CAMPUS, demonstrating exceptional robustness in occluded and crowded scenes. The IDF1 score (87.2%) highlights superior identity preservation. The decoupled design reduces graph size, which improves scalability.

Conclusion: By decoupling spatial and temporal associations and leveraging graph-based optimization, the proposed model significantly enhances tracking accuracy and reliability in multi-camera settings. This work provides a framework for applications like surveillance and autonomous systems, with future potential for attention mechanisms and adaptive graph integration.

How to cite this paper:
M. Akbari, S. M. Razavi, S. Mohamadzadeh, "Robust continuous person tracking in dense multi-camera environments through decoupled graph learning," J. Electr. Comput. Eng. Innovations, 14(2): 391-402, 2026.

DOI: 10.22061/jecei.2025.12094.853

URL: https://jecei.sru.ac.ir/article_2456.html

Introduction

Multi-object tracking, a primary task in computer vision, involves identifying objects and monitoring several targets throughout a series of images over time. This work benefits numerous real-world applications, including video surveillance, driverless vehicles, and sports analysis.

Despite several studies on multi-object tracking, challenges such as tracking fragmentation and identification alterations—resulting from frequent occlusions in crowded environments—continue to pose significant obstacles.

Recent trackers such as TransMOT [1] and ByteTrack [2] demonstrate impressive results by leveraging transformer-based modeling and strong appearance-driven association. However, these methods are primarily evaluated in single-camera scenarios. Our framework is complementary: transformer-based features or ByteTrack-style embeddings can be incorporated as node descriptors, while consensus fusion ensures global multi-camera consistency.

Object tracking in a multi-camera setup—commonly referred to as multi-camera multi-object tracking—represents a potential solution to this problem. Utilizing data from many cameras may yield more accurate tracking results, enabling things obscured in one perspective to be distinctly seen in another. Most 'tracking-by-detection' systems employ the Kalman filter during the data association step. This filter uses a motion model to predict the subsequent probable location and contrasts it with previous detections. Conversely, such techniques are typically predictable and struggle to adapt to dynamic environments. Moreover, outcome tracking frequently fails to ensure global consistency across multiple cameras and time frames due to variations in elements such as lighting, geometric distance, and sampling rate among datasets, conditions characteristic of real-world scenarios. Consequently, one alternative method is to reformulate the association problem as link prediction within a graph [3]-[5].

Current graph-based approaches for multi-camera person tracking have several challenges. Primarily, numerous methodologies rely on single-camera trackers to generate initial tracks [3], [4], [6]. These techniques, referred to as hierarchical two-step methods, involve single-camera tracking followed by cross-camera monitoring throughout the network. The comprehensive strategy, which addresses all detections or tracks collectively, is a choice. This method involves optimization over a batch of frames instead of merely two or a limited number of consecutive frames. This allows incorporation of global target attributes, improving overall tracking performance. This distinction is illustrated in Fig. 1. Despite numerous proposed approaches to enhance tracking systems, errors in systems reliant solely on single-camera data often remain uncorrected or lack effective solutions. Moreover, current algorithms fail to fully utilize the extensive spatial and temporal data crucial for monitoring individuals in multi-camera environments. Recent applications of spatio-temporal models have focused on acquiring representational attributes for tracking purposes. The resultant graphs, however, are occasionally complex and challenging to optimize [7]-[9].

This paper introduces a novel spatio-temporal graph model for multi-camera person tracking to address the aforementioned issues. We propose a decoupled consistency framework where spatial and temporal associations are independently optimized. This decoupling is motivated by the fact that unified spatio-temporal graphs become prohibitively dense as the number of cameras and frames grows. For C cameras, temporal window k, and average k detections per camera-frame, the joint graph includes approximately $O(C^2kN^2)$ cross-camera, cross-time edges in addition to $O(C^2N^2)$ spatial and $O(C\,kN^2)$ temporal edges. By decoupling, we remove the costly cross-term and reduce graph density, yielding lower memory and computational complexity while still retaining consistency through consensus fusion.

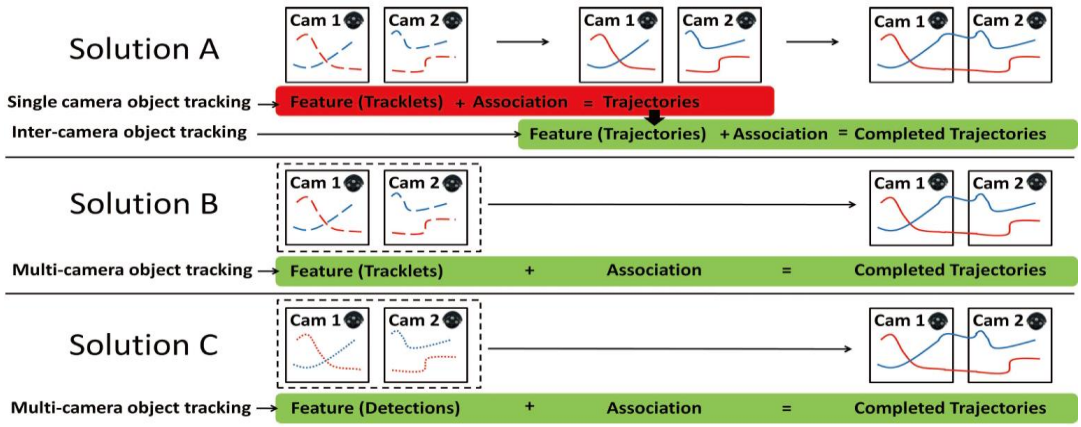


Fig. 1: illustrates two overarching approaches in visual object tracking utilizing several cameras [10].

Related Works

In single-camera scenarios, multi-object tracking has recently been the subject of substantial research. Research, such as [11]-[13], has concentrated on improving the data association phase and achieving more accurate motion estimates. Studies such as [14]-[17] have integrated object identification and tracking into a comprehensive model. Recently, multi-camera person tracking has garnered significant interest and advanced rapidly [4], [3], [18]. The model in [19] utilizes a Bayesian filter designed to monitor things in three-dimensional space with many cameras. This method uses a probabilistic model to forecast the status of each object, including its position, velocity, and direction of movement. The Bayesian filter continuously changes the state of objects by incorporating data from several cameras, utilizing probabilistic predictions and updates. Continuous state estimation via a Bayesian filter necessitates substantial computational resources, especially when monitoring numerous objects. Furthermore, the Bayesian filter may inadequately estimate object states if the initial predictions are erroneous or the input data is imprecise, resulting in tracking errors.

While multi-camera person tracking provides more comprehensive spatio-temporal information compared to single-camera tracking, challenges such as insufficient integration of spatial-temporal data and changing environmental conditions are still present in this field.

A. Spatio-temporal Representation Learning

Method [20] initially employs an occupancy map to integrate spatial correlations across several perspectives, subsequently utilizing deep learning to improve person detection in highly obstructed environments, facilitating more accurate detection through the simultaneous information from various cameras.

Method [6] characterizes the problem of multi-camera person tracking as a joint-structure optimization that connects tracklets based on appearance, geometry, and motion continuity.

Reference [21] employs a tracklet-to-target assignment approach. A "tracklet" refers to a concise, continuous sequence of a target's whereabouts recorded by a single camera. This model correlates several tracklets with actual targets through an assignment mechanism. Tracklets are initially retrieved from each camera; thereafter, comparable tracklets are allocated to the same target based on many parameters, including spatial position and temporal factors. The assignment algorithm identifies the association of tracklets through distance evaluations and feature-similarity criteria. The model incorporates re-identification and verification mechanisms to ensure that inaccurate or missing tracklets are adequately addressed, hence averting assignment errors. It employs a Kalman filter to predict the future positions of each target based on its previous condition.

Model [22] concurrently analyzes spatial and temporal data from all camera perspectives, focusing on the movement of individuals in three-dimensional space. This information enables the model to account for spatial and temporal differences that may arise from alterations in camera angle or position. The model is designed to analyze scene information jointly across multiple camera views, enabling consistent multi-camera tracking. This technique seeks to integrate spatial and temporal data from multiple cameras to facilitate person tracking in intricate environments. It presupposes that the scene remains relatively stable and that tracking precision may be influenced by complex or dynamic environmental alterations. The model is specifically designed for steady conditions with moderate movement; it may encounter difficulties in very congested environments or scenarios with rapid state transitions.

The study in paper [18] seeks to analyze and predict object behavior by tracking items across many cameras through the utilization of 3D convolutional neural networks and recurrent neural networks. Conversely, deep 3D models require the processing of intricate data and, especially in real-time tracking scenarios, demand substantial computational power that may be lacking in systems with limited processing capabilities. This model necessitates training on extensive datasets that must precisely represent real-world conditions. Alternatively, the model may exhibit suboptimal performance when confronted with novel circumstances.

B. Graph-based Methods

Method [23] introduces a traditional graph model that transmits messages through local polar feature representations. The relationship becomes complex when the graph edges are integrated within a graph and the spatial and temporal attributes are ambiguous. A candidate graph is initially constructed in [12]; thereafter, a feature encoder using the Transformer model [24]. This method, conversely, disregards spatio-temporal correlations within the graph. To enhance feature extraction for link prediction, [4] utilizes structural and temporal attention layers. This approach is suboptimal for simultaneous multi-view scenarios because it relies on data from single-camera trackers.

The computational expense of the dual-attention layers is significantly elevated. In this strategy, objects are organized in consecutive frames through the direct incorporation of new nodes and edges.

The integration of 3D geometric mapping with the Lifted Multicut optimization model in the paper [3] presents an innovative method for multi-object tracking across multiple cameras. This enhances path association precision and reduces identity switch errors. This article's weaknesses include dependence on the accuracy of input data and the capacity of the pre-processing pipeline.

Numerous issues identified in other models analyzed in other works can be addressed by our proposed model, utilizing spatio-temporal graphs. This model exhibits superior performance relative to others in conditions of noise, occlusion, rapid motion, and path crossings. Innovations in research:

- We break down the multi-object tracking problem into two parts: temporal association and spatial association, with multiple cameras.
- Our model is different from other approaches since it is not dependent on the output of individual camera trackers.
- Our methodology operates incrementally on frame-by-frame inputs, without relying on future frames.
- The CAMPUS [6] and Wildtrack [20] datasets have demonstrated that our model performs remarkably well in experimental results.

Proposed Method

Our method performs object tracking through identification, following [25]. Our methodology diverges from previous techniques by not depending on pre-trained single-camera trackers utilizing extensive person-tracking datasets [21], [3], [4]. Rather than addressing tracking and association issues, one must resolve a graph link prediction challenge. The model framework for person tracking in multi-camera scenarios includes temporal and spatial association. Spatial association captures relationships between objects from different viewpoints. The establishment of spatial connections among nodes is the principal emphasis of the spatial graph.

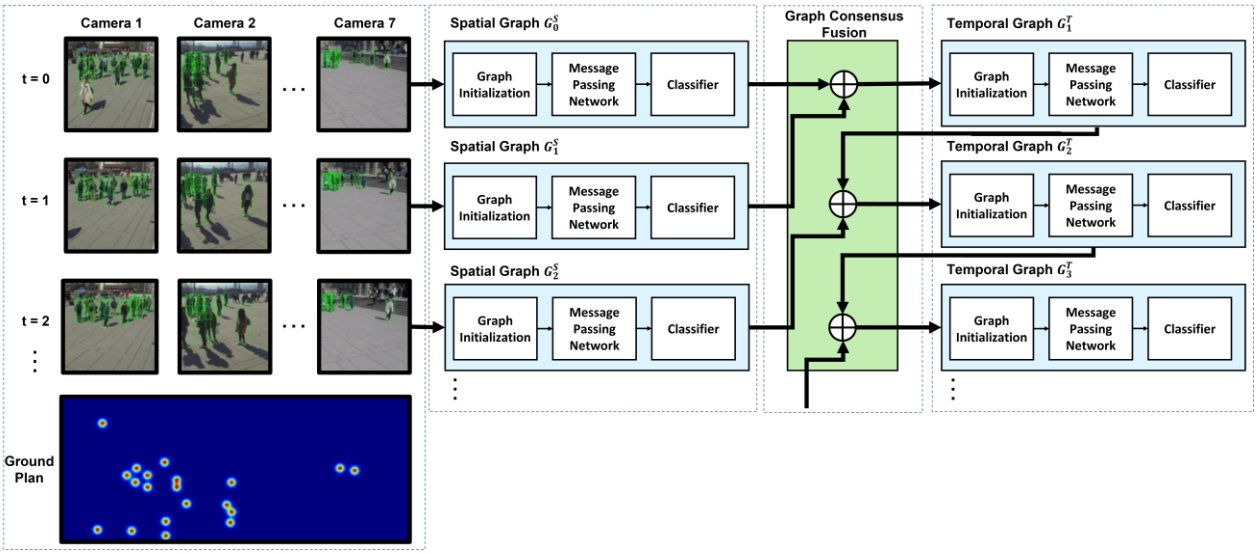


Fig. 2: Architecture of the proposed decoupled spatio-temporal graph learning model. The framework consists of three modules: (i) spatial graph construction, (ii) temporal graph construction, and (iii) consensus fusion, which merges spatial and temporal consistency into a unified temporal graph for identity tracking.

The temporal graph utilizes time-related information, including duration and velocity, to link objects across frames. Both modules systematically produce geographical and temporal graphs for each frame. Our framework introduces modular graph training, where spatial and temporal graphs are trained independently with specialized feature encoders and loss functions. We simplify optimization through consensus fusion, averaging node features within connected components to reduce dimensionality while preserving topology. This design avoids the computational overhead of graph reconfiguration and ensures compatibility with low-resource systems.

Fig. 2 illustrates the proposed model's system architecture.

A. Problem Formulation

The objective of in-person tracking in multi-camera environments is to monitor several objects across frames and diverse viewpoints. Assume that there are C cameras with fields of view that overlapped. We create a graph structure consisting of nodes V_t and edges E_t at time t $(\mathrm{G}_t(V_t, E_t))$.

Suppose H_c is the homography matrix that transforms

the coordinates of the bottom point of the bounding box (x, y, w, h) from the image plane of the camera view to coordinates in the common ground plane. The spatial location $p_{(v_i)}$ and velocity information $s_{(v_i)}$ for the reference node v_i are computed as follows:

$$p_{v_i} = H_{c_{v_i}}\left(x + \frac{w}{2}, y\right) \tag{1}$$

$$s_{v_i} = \frac{p_{v_i} - p_{v_j}}{t_{v_i} - t_{v_j}} \tag{2}$$

Hence, we compute the relative distance of the appearance feature $d_{(v_i)}$ by using cosine similarity and L1 distance. Cosine emphasizes orientation/appearance direction alignment, while L1 norm provides robustness to illumination noise [26].

$$\Delta d_{ij} = \left[||d_{v_i} - d_{v_j}||_1, 1 - \text{cosine similarity}\left(d_{v_i}, d_{v_j}\right)\right] \tag{3}$$

where $[.,.]$ denotes the concatenation of two terms. Spatial location, and velocity between two nodes computed by combining L1 and L2 distances. L1 is more robust to outliers (e.g., occlusions), while L2 preserves geometric smoothness [21].

$$\Delta p_{ij} = \left[||p_{v_i} - p_{v_j}||_1, ||p_{v_i} - p_{v_j}||_2\right] \tag{4}$$

$$\Delta s_{ij} = \left[||s_{v_i} - s_{v_j}||_1, ||s_{v_i} - s_{v_j}||_2\right] \tag{5}$$

B. Spatio-temporal Graph

To achieve object tracking and association, our model adheres to a sequential process: first, at time t − 1, we establish the temporal graph through temporal association; subsequently, at time t, we create the spatial graph via spatial association, and ultimately, at time t, we develop a new temporal graph by consolidating these features.

The spatial graph referred to as G^S and constructed with spatial information.

The set of nodes of the spatial graph G_t^S at time t includes all the detected objects from every camera at a time t .

After constructing the spatial graph, the initial features of the nodes $h_{v_i}^0$ and the initial features of the edges $h_{(v_i,v_j)}^0$ are defined as follows:

$$h_{v_i}^0 = \mathcal{E}_v(d_{v_i}) \tag{6}$$

$$h_{(v_i,v_j)}^0 = \mathcal{E}_e([\Delta p_{ij}, \Delta d_{ij}]) \tag{7}$$

where $\mathcal{E}_v$ is a node feature encoder, and $\mathcal{E}_e$ is an edge feature encoder. The result of $[\Delta p_{ij}, \Delta d_{ij}]$ is the concatenation of two expressions.

In this stage, connections can be established between objects from a variety of periods without the need to employ camera-specific data. The temporal graph prioritizes temporal correlation and is hence independent of perspective. The subsequent calculations are executed on the initial attributes of the nodes and edges of the temporal graph G_t^T at time t:

$$h_{v_i}^0 = \mathcal{E}_v([d_{v_i}, p_{v_i}]) \tag{8}$$

$$h_{(v_i,v_j)}^0 = \mathcal{E}_e([\Delta p_{ij}, \Delta d_{ij}, \Delta s_{ij}]) \tag{9}$$

C. Graph Consensus Fusion

After the association stage, several connected components representing the same object may be obtained. To bridge two consecutive association stages, our model reconfigures the graphs into a new temporal graph G_t^T. Specifically, G_t^T is constructed from the spatial graph G_t^S at frame t and the temporal graph $G_{(t-1)}^T$ from the previous frame.

We first compute connected components in G_t^S and $G_{(t-1)}^T$ using a standard Union–Find (disjoint-set) algorithm. Two detections are assigned to the same component if their association confidence exceeds a threshold τ). This ensures each component corresponds to a single identity hypothesis.

From G_t^S and $G_{(t-1)}^T$, all nodes in the same connected component are aggregated into a single node in G_t^T. Let $M = M_1, \ldots, M_n$ denote the set of connected components in graph GGG. The initial node features of G_t^T are then defined by averaging the features of all detections in the component:

$$d_v = \frac{1}{|H_i|}\sum_{v \in H_i} d_v \tag{10}$$

$$p_v = \frac{1}{|H_i|}\sum_{v \in H_i} p_v \tag{11}$$

where $H_i \in H(\boldsymbol{G_t^S}) \cup H(\boldsymbol{G_{(t-1)}^T})$, and d_v, p_v represent the appearance and position features, respectively. After node aggregation, we define edges in $\boldsymbol{G_t^T}$ by connecting nodes that belong to different time steps. In other words, an edge exists between two nodes if they originate from different frames; otherwise, no edge is created.

When a node belongs to multiple potential components (for instance, due to overlapping spatial and temporal evidence), we resolve ambiguity by a hierarchical rule:

1. Keep the link with the highest edge confidence score,
2. If tied, select the match with minimal appearance distance,

3. If still tied, prefer the match with the smallest velocity discontinuity on the ground plane.

This ensures that ambiguous detections are consistently merged while reducing the number of identity switches.

The pseudocode for executing the model, along with the messaging network, is provided in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1: *Inference Algorithm*

Input: *Temporal graph* G_{t-1}^{T}
Output: *Temporal graph* G_t^T, *tracking result at time t*
Construct spatial graph G_t^S *and Compute initial feature* $h_{v_i}^0$ *and* $h_{(v_i,v_j)}^0$
for $l = 1$ *to* L *do*
  $h_{v_i}^l, h_{(v_i,v_j)}^l = MPN\left(G_t^S, h_{v_i}^{l-1}, h_{(v_i,v_j)}^{l-1}\right)$
end for
$G_t^T = consensus\ fusion(G_t^S, G_{t-1}^T)$
Compute initial feature $h_{v_i}^0, h_{(v_i,v_j)}^0$ *for* G_t^T
for $l = 1$ *to* L *do*
  $h_{v_i}^l, h_{(v_i,v_j)}^l = MPN\left(G_t^T, h_{v_i}^{l-1}, h_{(v_i,v_j)}^{l-1}\right)$
end for
$\hat{y}_{(v_i,v_j)}^l = \mathcal{C}\left(h_{(v_i,v_j)}^l\right)$
assigning tracklet ID

Algorithm 2: *Message Passing Network* (*MPN*)

Input: *graph* G_t *at time* t, $h_{v_i}^{l-1}, h_{(v_i,v_j)}^{l-1}$ *node feature and edge feature of iteration* $l-1$
Output: $h_{v_i}^l, h_{(v_i,v_j)}^l$ *node feature and edge feature of iteration* l
$h_{(v_i,v_j)}^l = \mathcal{E}_e\left(\left[h_{v_i}^{l-1}, h_{(v_i,v_j)}^{l-1}, h_{v_j}^{l-1}\right]\right)$
$m_{ij}^l = \mathcal{E}_v\left(\left[h_{v_j}^{l-1}, h_{(v_i,v_j)}^l\right]\right)$
$h_{v_i}^l = \sum_{j\in N(v_i)} m_{ij}^l$

D. Message Passing Network

We utilize the standard message passing network framework [5], [27], [23], using the graph's initial node and edge features as input. Each node and edge computes its outgoing messages at each stage of propagation, then aggregates its received messages, and ultimately updates its representation by fusing the new data with its prior representation.

Each message passing phase consists of two primary stages: edge update and node update. Both updates are iteratively executed across L iterations, or L message passing steps.

The edge feature is updated by aggregating the source node's feature and the destination node's feature:

$$\boldsymbol{h}_{\boldsymbol{e_{ij}}}^{\boldsymbol{l}} = \boldsymbol{\mathcal{F}_e}\left(\left[\boldsymbol{h}_{\boldsymbol{v_i}}^{\boldsymbol{l-1}}, \boldsymbol{h}_{\boldsymbol{(v_i,v_j)}}^{\boldsymbol{l-1}}, \boldsymbol{h}_{\boldsymbol{v_j}}^{\boldsymbol{l-1}}\right]\right) \quad (12)$$

where $\boldsymbol{\mathcal{F}_e}$ is a learnable MLP encoder. After updating the edge features, each node is updated with the messages sent by its nearby nodes:

$$\boldsymbol{h}_{\boldsymbol{v_i}}^{\boldsymbol{l}} = \sum_{\boldsymbol{j\in N(v_i)}} \boldsymbol{m}_{\boldsymbol{ij}}^{\boldsymbol{l}} \quad (13)$$

where $\boldsymbol{N}(.)$ denotes the neighboring nodes, and:

$$\boldsymbol{m}_{\boldsymbol{ij}}^{\boldsymbol{l}} = \boldsymbol{\mathcal{F}_v}\left(\left[\boldsymbol{h}_{\boldsymbol{v_j}}^{\boldsymbol{l-1}}, \boldsymbol{h}_{\boldsymbol{(v_i,v_j)}}^{\boldsymbol{l}}\right]\right) \quad (14)$$

where $\boldsymbol{\mathcal{F}_v}$ is another learnable MLP encoder.

Formally, each message-passing update has complexity $O(|E|\,.\,d)$, where $|E|$ is the number of edges and d the hidden feature dimension. In a joint graph, $|E|$ grows quadratically with cameras × frames due to cross-camera cross-time links. Our decoupled formulation instead builds (i) per-frame spatial graphs and (ii) per-camera temporal graphs. The total number of edges is reduced to

$$\left|E_{decoupled}\right| = O(C^2N^2) + O(CkN^2) \quad (15)$$

Versus

$$\left|E_{joint}\right| = O(C^2N^2) + O(CkN^2) + O(C^2kN^2) \quad (16)$$

Thus, the decoupled model removes the dominant cross-term and ensures better scalability. This design aligns with prior findings that denser graphs can hinder learning efficiency [5], [4], [12].

E. Link Prediction

After the message passing network is executed, the updated edge features are prepared for link prediction. Our goal is to learn to predict which edges in the graph should be kept or eliminated. One may consider it to be an edge classification problem. The classification of a particular edge at a given iteration is computed using:

$$\boldsymbol{\hat{y}}_{\boldsymbol{e_{ij}}}^{\boldsymbol{l}} = \boldsymbol{\mathcal{C}}\left(\boldsymbol{h}_{\boldsymbol{(v_i,v_j)}}^{\boldsymbol{l}}\right) \quad (17)$$

Being $\boldsymbol{\mathcal{C}}$ is a binary classifier implemented using a multilayer perceptron network and a Softmax layer.

F. Loss Function

To improve the model's spatial and temporal feature representations, the two graphs are trained separately during model training. For every message that goes through iteration l, the loss between the predicted and

true labels is calculated using the Focal Loss function [28].

Experiments

A. Dataset

Our experiments were conducted on the Wildtrack dataset [20] for multi-object multi-camera tracking, which includes various environmental conditions, including lighting changes, detection quality, and crowd density. All videos are captured using fixed cameras with a certain overlap ratio in their fields of view. The Wildtrack dataset contains 7 calibrated cameras at 1920×1080 resolution, capturing up to 25 individuals per frame in a dense pedestrian zone. Severe occlusions are frequent, with more than 70% of individuals partially occluded in several frames. The dataset is particularly challenging for cross-camera identity matching because multiple people overlap in the ground-plane projection. Each frame contains about 25 individuals, either standing or moving. We trained using the first 360 frames and tested using the last 40 frames, following the typical parameters [29]-[31].

The CAMPUS dataset [6] was also used to address additional challenges in multi-object tracking from multiple viewpoints. The CAMPUS dataset provides 3 camera views at lower resolution, with more dynamic movement and less structured pedestrian flow. While crowd density is lower than Wildtrack, occlusions and frame drops introduce temporal discontinuities, requiring robust temporal association to maintain identity consistency. This dataset includes sequences named Garden 1, Garden 2, and ParkingLot, each filmed by 3 to 4 high-quality cameras installed at a height of 1.5 to 2 meters above the ground. Each camera covers overlapping and non-overlapping areas with other cameras. The videos were recorded at a frame rate of 30 frames per second and for a duration of approximately 3 to 4 minutes, with image resolution maintained at 1080x1920.

The ability to capture diverse movements is emphasized in Garden 1, where people are performing various sports. Garden 2 presents a sequence characterized by a lower density. Due to the frequent occurrence of cars blocking players' paths in the ParkingLot sequence, it becomes increasingly difficult to reconstruct their paths from various vantage points.

Together, these datasets complement each other: Wildtrack stresses spatial consistency under heavy occlusion, while CAMPUS highlights temporal continuity under challenging motion and missing frames.

B. Evaluation Metrics

In this part, the proposed model's performance is assessed using the standard datasets Wildtrack [20] and CAMPUS [6]. To evaluate the methods for multi-camera person tracking, the CLEAR MOT metrics [32] and the ID score [33] (including IDF1) are used to provide a fair comparison.

The CLEAR MOT metrics include:

- MOTA: The multi-object tracking accuracy metric that measures the number of false positives, false negatives, and identity switches, focusing on the coverage of detections.
- MOTP: The multi-object tracking accuracy metric that addresses the overall mismatch between true positives and ground truth objects.
- IDF1: A metric that measures the agreement between the computed tracks and the ground truth tracks through the F1 score.
- MT: The number of objects that have been completely tracked for at least 80% of their lifespan.
- ML: The number of objects that have been tracked for at most 20% of their lifespan.

C. Implementation details

The OSNet model [34] is used for extracting appearance features, providing a feature vector of dimension 512. The graph model's node feature dimensions are set to 32, while the edge feature dimensions are set to 6. In all experiments, 4 iterations of the messaging network are performed, and finally, a 6-dimensional feature vector is provided as the output for classifying edges.

Following prior graph-based multi-camera tracking works (e.g., LMGP; DyGLIP; DMCT), we used ground-truth bounding boxes to isolate the association problem and ensure fair comparison. This choice factors out detector variability and highlights the contribution of our graph formulation.

In consensus fusion, we set the confidence threshold to $\boldsymbol{\tau = 0.5}$, following common practice in graph-based multi-object tracking [4], [5]. We also observed stable results when varying $\boldsymbol{\tau}$ in the range 0.4–0.6.

The temporal graph's time window size is set to 3. The following parameters are used to train the model:

- One hundred training epochs are utilized by the Adam optimizer [35].
- During the initial ten epochs, the learning rate is fine-tuned from 0 to 0.01.

D. Evaluation Results

The evaluation results presented in Table 1 and Table 2 show that the proposed model has achieved significant improvements in the accuracy and efficiency of multi-camera tracking. For the Wildtrack dataset, as shown in Table 1 and Fig. 3, the GLMB-DO method [19] achieves moderate MOTA but suffers from a high ML rate, indicating weaker robustness in crowded scenes.

Similarly, Fig. 3 illustrates that DMCT [18] improves identity preservation compared to GLMB but still falls short in maintaining consistency across cameras.

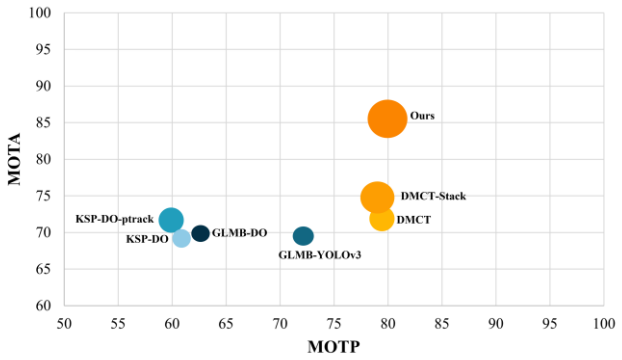


Fig. 3: Visual diagram of the evaluation results on the Wildtrack dataset. (The size of each circle is determined based on the IDF1 criterion)

In contrast, our proposed method achieves the highest balance between MOTA and IDF1, as highlighted in Fig. 3, where the larger circle for our method demonstrates both improved accuracy and robustness.

Table 1: Evaluation results on the Wildtrack dataset

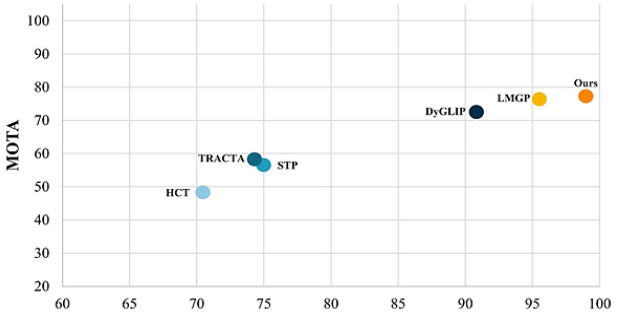
Method	IDF1 ↑	MOTA ↑	MOTP ↑	MT ↑	ML ↓
GLMB-YOLOv3 [19]	74.3	69.7	73.2	79.5	21.6
GLMB-Faster RCNN [19]	76.5	65.3	71.9	68.9	27.4
GLMB-DO [19]	72.5	70.1	63.1	93.6	22.8
KSP-DO [20]	73.2	69.6	61.5	28.7	25.1
KSP-DO-ptrack [20]	78.4	72.2	60.3	42.1	14.6
DMCT [18]	77.8	72.8	79.1	61.0	4.9
DMCT Stack [18]	81.9	74.6	78.9	65.9	4.9
Ours	**87.2**	**85.5**	**80.2**	**78.6**	**4.9**

One notable point in this evaluation is the reduction of ML to 4.9% in the proposed model, indicating an increase in the robustness of the algorithm in the complex conditions of multi-camera environments. In contrast, methods such as GLMB-DO and DMCT, while performing well, were not as accurate in terms of ML and MT as the proposed model. This superiority is due to the utilization of graph structures and spatial and information, which optimize the processing of multi-camera data and better handle common challenges such as occlusion and sudden changes. The experimental results of our proposed model on the CAMPUS dataset also outperform other methods in most metrics (Table 2 and Fig. 4). For both the ML and MT metrics, our approach attains values of 100 and 0, respectively.

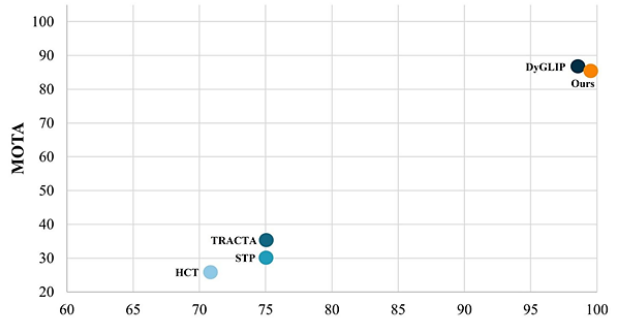
These outcomes reflect the method's ability to exploit spatio-temporal consistency for robust and reliable tracking.

Table 2: Evaluation results on the CAMPUS dataset

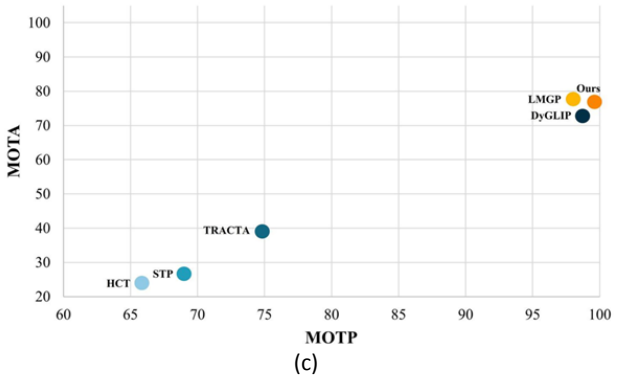
Sequence	Method	MOTA ↑	MOTP ↑	MT ↑	ML ↓
Garden 1	HCT [6]	49	71.9	31.3	6.3
	STP [22]	57	75	-	-
	TRACTA [21]	58.5	74.3	30.6	1.6
	DyGLIP [4]	71.2	91.6	31.3	0.0
	LMGP [3]	76.9	95.9	62.9	1.6
	Ours	**77.4**	**98.1**	**100.0**	**0.0**
Garden 2	HCT [6]	25.8	71.6	33.3	11.1
	STP [22]	30	75	-	-
	TRACTA [21]	35.5	75.3	16.9	11.3
	DyGLIP [4]	87.0	98.4	66.7	0.0
	Ours	**87.4**	**99.0**	**100.0**	**0.0**
Parkinglot	HCT [6]	24.1	66.2	6.7	26.6
	STP [22]	28	68	-	-
	TRACTA [21]	39.4	74.9	15.5	10.3
	DyGLIP [4]	72.8	98.6	26.7	0.0
	LMGP [3]	78.1	97.3	62.1	0.0
	Ours	**78.6**	**98.8**	**100.0**	**0.0**



(a)



(b)



(c)

Fig. 4: Visual diagram of the evaluation results on the CAMPUS dataset. (a) Garden 1 subset, (b) Garden 2 subset, and (c) ParkingLot subset.

Our qualitative tracking results are shown in Fig. 5 for the WILDTRACK dataset, with three overlapping camera views. Single-camera pedestrian tracking is insufficient for capturing all objects due to the high crowd density and clutter in the scenes. The superiority of multi-camera setups is evident in the following two scenarios. First, even if an ID, such as ID 13 in frame 372 from camera 5, is fully obscured, our tracker can still locate the object by establishing correspondences between cameras 6 and 7. Second, in frames 370 and 372 from camera 6, a misidentification mistake occurs between IDs 13 and 20 due to their ambiguous visual similarity; fortunately, we were able to avoid this error by using the matching detections for those objects in camera 7, where they are distinguishable.

Fig. 5: Tracking quality results in WILDTRACK in three overlapping camera view.

E. Computational Complexity

The computational complexity of each message passing step in our framework is $O(|E|\,.\,d)$, where $|E|$ is the number of edges and d is the hidden feature dimension. This complexity is comparable to other graph-based tracking methods such as [5], [4]. By decoupling spatial and temporal graphs, the number of edges is reduced compared to unified spatio-temporal formulations, which decreases memory usage and improves efficiency. While runtime performance is not the focus of this study, this design choice highlights the scalability of the approach for multi-camera environments.

F. Ablation Study

As shown in Fig. 1, training a unified spatio-temporal graph without decoupling reduces IDF1 by 4.2% on Wildtrack compared to our decoupled formulation. This confirms that separating spatial and temporal reasoning generalizes better across datasets.

Finally, the evaluations emphasize that the proposed model has not only improved tracking accuracy but also optimized the connections between different time frames. The use of spatiotemporal graph-based techniques has enabled the system to track individuals with high accuracy, even in challenging surveillance conditions.

Results and Discussion

The experimental results presented in Table 1 and Table 2 demonstrate that the proposed model achieves significant improvements in tracking accuracy and robustness compared to existing methods.

The decoupled graph design effectively reduces computational complexity while maintaining high identity consistency across multiple camera views. On the Wildtrack dataset, the model achieves the highest MOTA and IDF1 scores among all compared methods, demonstrating its superior robustness under crowded and occluded scenarios. Similarly, on the CAMPUS dataset, our method achieves the best results, indicating excellent continuity and stability in dynamic environments. These results highlight the strength of independent spatial and temporal optimization combined with consensus fusion for enhanced global consistency. The proposed framework provides a scalable and accurate foundation for real-world applications such as surveillance and autonomous systems.

Conclusion

This study presents an innovative graph-based methodology for monitoring individuals in multi-camera environments. By decoupling spatial and temporal consistency into modular graphs trained with Focal Loss, our framework achieves state-of-the-art accuracy with a modular design that is scalable and can be extended toward real-time systems in future research, though this work explicitly focuses on accuracy rather than runtime. This framework comprises two phases. In a spatial network, initial items across several perspectives at a specific time frame are interconnected. A graph consensus fusion module is employed to merge nodes within similar connected IDs, resulting in a new graph for temporal association. Ultimately, temporal association is employed to align things over several frames, thereby completing the sequential tracking process across multiple cameras.

Although we did not reimplement TransMOT or ByteTrack for direct comparison, these methods are complementary to our framework. Their advanced feature extractors can serve as inputs to our graph-based formulation, while our decoupled design and consensus fusion address the unique challenges of multi-camera identity maintenance. The results indicate that the proposed model is a significant and efficient method for individual tracking in multi-camera environments and establishes a foundation for further research in this domain.

We did not conduct an evaluation that included detector noise. A practical next step is to couple our association module with strong detectors (e.g., YOLOv8 or Faster R-CNN) and report end-to-end MOTA/IDF1 under realistic false positives/negatives and missed detections. In addition, future work will explicitly benchmark frame-per-second (FPS) throughput and deployment latency, enabling a more complete evaluation of real-time applicability in surveillance and robotics scenarios.

Author Contributions

M. Akbari, S. M. Razavi, and S. Mohamadzadeh designed the experiments. M. Akbari collected the data. M. Akbari carried out the data analysis. M. Akbari, S. M. Razavi, and S. Mohamadzadeh interpreted the results and wrote the manuscript.

Acknowledgment

The authors gratefully acknowledges the supports provided by Electrical and Computer Engineering department from University of Birjand, Birjand, Iran.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Funding

This research received no external funding. The authors declare that the study was conducted without any financial support from funding agencies or organizations.

Abbreviations

MOT	Multiple object tracking
MOTA	Multiple object tracker accuracy
MOTP	Multiple object tracker precision
MT	Mostly tracked
ML	Mostly lost
IDF1	ID measures: global min-cost F1 score
MPNs	Message Passing Networks
GNNs	Graph Neural Networks

References

[1] P. Chu, J. Wang, Q. You, H. Ling, Z. Liu, "Transmot: Spatial-temporal graph transformer for multiple object tracking," in Proc. the IEEE/CVF Winter Conference on applications of computer vision: 4870-4880, 2023.

[2] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," in Proc. European conference on computer vision: 1-121, 2022.

[3] D. M. Nguyen, R. Henschel, B. Rosenhahn, D. Sonntag, P. Swoboda, "LMGP: Lifted multicut meets geometry projections for multi-camera multi-object tracking," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8866-8875, 2022.

[4] K. G. Quach, P. Nguyen, H. Le, T. D. Truong, C. N. Duong, M. T. Tran, K. Luu, "DyGLIP: A dynamic graph model with link prediction for accurate multi-camera multiple object tracking," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 13784-13793, 2021.

[5] G. Braso, L. Leal-Taixé, "Learning a neural solver for multiple object tracking," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[6] Y. Xu, X. Liu, Y. Liu, S. C. Zhu, "Multi-view people tracking via hierarchical trajectory composition," in Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 4256–4265, 2016.

[7] H. Wang, Z. Li, Y. Li, K. Nai, M. Wen, "Sture: Spatial–temporal mutual representation learning for robust data association in online multi-object tracking," Comput. Vision Image Understanding, 220: 103433, 2022.

[8] Q. Xie, W. Zhou, G. L. Qi, Q. Tian, H. Li, "Progressive unsupervised person re-identification by tracklet association with spatio-temporal regularization," IEEE Trans. Multimedia, 23: 597-610, 2020.

[9] J. Xu, Y. Cao, Z. Zhang, H. Hu, "Spatial-temporal relation networks for multi-object tracking," in Proc. the IEEE/CVF International Conference on Computer Vision: 3988-3998, 2019.

[10] W. Chen, L. Cao, X. Chen, K. Huang, "A novel solution for multi-camera object tracking," in Proc. 2014 IEEE International Conference on Image Processing (ICIP): 2329-2333, 2014.

[11] J. Cao, X. Weng, R. Khirodkar, J. Pang, K. Kitani, "Observation-centric SORT: Rethinking SORT for robust multi-object tracking," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 9686-9696, 2023.

[12] P. Chu, J. Wang, Q. You, H. Ling, Z. Liu, "TransMOT: Spatial-temporal graph transformer for multiple object tracking," in Proc. the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV): 4870-4880, 2023.

[13] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," in Proc. European conference on computer vision: 1-21, 2022.

[14] S. Sun, N. Akhtar, X. Song, H. Song, A. Mian, M. Shah, "Simultaneous detection and tracking with motion modelling for multiple object tracking," in Proc. the European Conference on Computer Vision (ECCV): 626-643, 2020.

[15] Z. Wang, L. Zheng, Y. Liu, Y. Li, S. Wang, "Towards real-time multi-object tracking," in Proc. the European Conference on Computer Vision (ECCV): 107-122, 2020.

[16] Y. Xu, A. Ošep,Y. Ban, R. Horaud, L. Leal-Taixé, X. Alameda-Pineda, "How to train your deep multi-object tracker," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[17] X. Zhou, V. Koltun, P. Krähenbühl, "Tracking objects as points," in Proc. the European Conference on Computer Vision (ECCV): 474-490, 2020.

[18] Q. You, H. Jiang, "Real-time 3D deep multi-camera tracking," arXiv preprint arXiv:2003.11753, 2020.

[19] J. Ong, B. T. Vo, B. N., Vo, D. Y. Kim, S. Nordholm, "A Bayesian filter for multi-view 3D multi-object tracking with occlusion handling," IEEE Trans.Pattern Anal. Mach. Intell. (TPAMI), 44(5): 2246-2263, 2020.

[20] T. Chavdarova, P. Baqué, S. Bouquet, A. Maksai, C. Jose, T. Bagautdinov, F. Fleuret, "Wildtrack: A multi-camera HD dataset for dense unscripted pedestrian detection," in Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 5030–5039, 2018.

[21] Y. He, X. Wei, X. Hong, W. Shi, Y. Gong, "Multi-target multi-camera tracking by tracklet-to-target assignment," IEEE Trans. Image Process. (TIP), 29: 5191-5205, 2020.

[22] Y. Xu, X. Liu, L. Qin, S. C. Zhu, "Cross-view people tracking by scene-centered spatio-temporal parsing," in Proc. the AAAI Conference on Artificial Intelligence (AAAI), 2017.

[23] A. Kim, G. Braso, A. Ošep, L. Leal-Taixé, "PolarMOT: How far can geometric relations take us in 3D multi-object tracking?," in Proc. the European Conference on Computer Vision (ECCV): 41-58, 2022.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need," Adv. Neural Inf. Process. Syst., 30, 2017.

[25] M. Andriluka, S. Roth, B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 1-8, 2008.

[26] Q. Xie, W. Zhou, G. J. Qi, Q. Tian, H. Li, "Progressive unsupervised person re-identification by tracklet association with spatio-temporal regularization," IEEE Trans. Multimedia, 23: 597-610, 2020.

[27] E. Luna, J. C. SanMiguel, J. M. Martínez, P. Carballeira, "Graph neural networks for cross-camera data association," IEEE Trans. Circuits Syst. Video Technol., 33(2): 589-601, 2022.

[28] T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, "Focal loss for dense object detection," in Proc. the IEEE International Conference on Computer Vision (ICCV): 2999-3007, 2017.

[29] M. Engilberge, W. Liu, P. Fua, "Multi-view tracking using weakly supervised human motion prediction," in Proc. the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV): 1582-1592, 2023.

[30] Y. Hou, L. Zheng, "Multiview detection with shadow transformer (and view-coherent data augmentation)," in Proc. the 29th ACM International Conference on Multimedia (MM '21), 2021.

[31] Y. Hou, L. Zheng, S. Gould, "Multiview detection with feature perspective transformation," in Proc. the European Conference on Computer Vision (ECCV), 2020.

[32] K. Bernardin, R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," EURASIP J. Image Video Process., 2008: 1-10, 2008.

[33] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in European Conference on Computer Vision: 17-35, 2016.

[34] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, "Omni-scale feature learning for person re-identification," in Proc. the IEEE/CVF International Conference on Computer Vision: 3702-3712, 2019.

[35] D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization," in Proc. International Conference on Learning Representations (ICLR), 2015.

Biographies

Morteza Akbari received the B.Sc. degree in Computer Science from the University of Birjand, Iran, in 2021. He received the M.Sc. degree in Computer Science from Ferdowsi University of Mashhad, Iran, in 2023. He is currently pursuing the Ph.D. degree in Electrical Engineering with a specialization in artificial intelligence and robotics at the University of Birjand, Iran. His research interests include image processing, computer vision, graph neural networks, deep learning, multi-target tracking, and medical image analysis.

- Email: morteza.akbari@birjand.ac.ir
- ORCID: 0009-0005-9407-8478
- Web of Science Researcher ID: NIS-5419-2025
- Scopus Author ID: NA
- Homepage: NA

Seyyed Mohammad Razavi received his B.Sc. degree in Electrical Engineering from Amirkabir University of Technology, in 1994 and his M.Sc. and Ph.D. degree in Electrical Engineering from the Tarbiat Modares University, Iran, in 1996 and 2006 respectively. Now, he is a Full Professor in University of Birjand. His research interests include Computer Vision, Pattern Recognition and Artificial Intelligence.

- Email: smrazavi@birjand.ac.ir
- ORCID: 0000-0002-3493-7614
- Web of Science Researcher ID: AAF-7386-2021
- Scopus Author ID: 56214431100
- Homepage: https://cv.birjand.ac.ir/mrazavi/fa

Sajad Mohamadzadeh received the B.Sc. degree in Electrical Engineering from Sistan & Baloochestan, University of Zahedan, Iran, in 2010. He received the M.Sc. degree in Telecommunication Engineering from university of Birjand, Birjand, Iran, in 2012. He received the Ph.D. degree in Telecommunication Engineering from university of Birjand, Birjand, Iran, in 2016. He is currently an academic staff in Department of electrical and computer engineering, university of Birjand, Birjand, Iran. His area research includes image processing and retrieval, pattern recognition, digital signal processing and sparse representation.

- Email: s.mohamadzadeh@birjand.ac.ir
- ORCID: 0000-0002-9096-8626
- Web of Science Researcher ID: AAF-4605-2021
- Scopus Author ID: 57056477500
- Homepage: https://cv.birjand.ac.ir/mohamadzadeh