GLoMOT: Efficient Online GNN-based Low-Frame-Rate Multi-Object Tracker

Yaxuan Hu[1], Jie Hua[1], Gang Wu[2*], Yuhong Yang[1], Atsushi Suzuki[3], Zhongyuan Wang[1*]

[1]National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
[2]College of Cyber Security, Tarim University
[3]Department of Mathematics, Faculty of Sciences, The University of Hong Kong
sad123yxhu@gmail.com, huajie@whu.edu.cn, 120070034@taru.edu.cn,
yangyuhong@whu.edu.cn, atsushi.suzuki.rd@outlook.com, wzy_hope@163.com

Abstract

Low-frame-rate (LFR) Multi-Object Tracking (MOT) is crucial for efficient tracking on edge devices, as it significantly reduces computational and storage demands. However, existing trackers struggle in LFR settings due to large temporal gaps, extreme appearance changes, and motion non-linearity. While Graph Neural Network (GNN)-based trackers are effective at associating objects across these gaps, most operate offline, which prevents their use for online tracking. To address these limitations, we propose GLoMOT, a novel online **G**NN-based **Lo**w-Frame-Rate **M**ulti-**O**bject **T**racker designed for robust performance in LFR videos. To bridge the large temporal gaps, we introduce a Dynamic Node Buffer Pool. This acts as a long-term memory, caching the states of absent objects to enable their robust re-association. To tackle extreme motion uncertainty, we propose an adaptive context-aware module that dynamically adjusts the weights of positional and appearance features, generating more robust features for predicting node connections. Furthermore, we propose a pseudo-depth feature calculation method. This provides the GNN with critical geometric context, which helps resolve spatial ambiguity arising from occlusions. Extensive experiments on several public MOT benchmarks, including DanceTrack, MOT17 and VisDrone, demonstrate GLoMOT's effectiveness and superiority, particularly in challenging Low-Frame-Rate conditions.

Introduction

Multi-Object Tracking (MOT) is an essential and challenging task in computer vision, involving the detection and association of multiple objects across consecutive frames in a video. It plays an important role in various real-world applications such as intelligent transportation systems (Chiu et al. 2021), surveillance and security (Elhoseny 2020), and sports events (Cioppa et al. 2022).

Recently, with the widespread adoption of edge computing and the increasing demand for data processing on resource-constrained devices, a new paradigm, Low-Frame-Rate (LFR) Multi-Object Tracking, has received considerable attention (Ganesh et al. 2023). LFR-MOT focuses on tracking objects using only sparsely sampled keyframes.

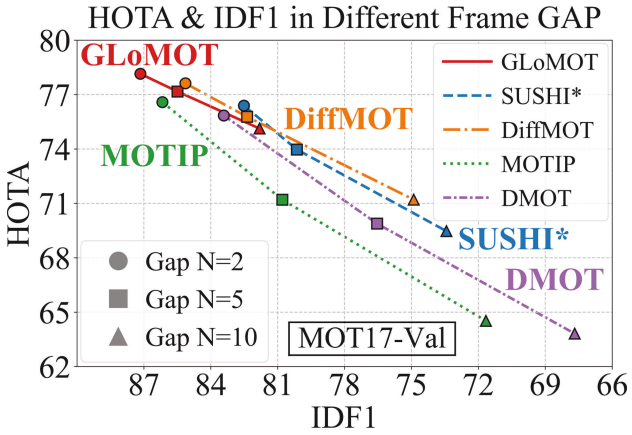


Figure 1: Comparison of tracking performance of trackers with different tracking paradigms on MOT17 val with frame gap ranging from 2 to 10. (Top left is the best). The results are normalized and the detailed results are shown in Tables 1, and the DMOT in the figure refers to DiffusionMOT.

Crucially, this paradigm not only enables tracking in scenarios where high-frame-rate video capture is physically or economically infeasible due to camera hardware or bandwidth limitations, but also drastically reduces the subsequent computational load and storage capacity. This makes it particularly suitable for real-world applications where continuous high-frame-rate processing is not practical. However, this efficiency comes at the cost of significantly increased tracking difficulty. The main challenges in LFR-MOT are as follows: (1) Large Temporal Gaps: The long intervals between frames lead to large, unpredictable object displacements, often with no spatial overlap between consecutive detections of the same object. (2) Severe Motion Non-linearity: Standard linear motion models, such as the Kalman filter, which are the foundations of many trackers, become ineffective as they cannot model the complex, non-linear motion occurring over large time gaps. (3) Extreme Appearance and Visibility Changes: An object's appearance and visibility can change dramatically between sparse frames, weakening the reliability of appearance features for re-identification and making occlusions more abrupt.

We compared the performance of several state-of-the-

*Corresponding Author.

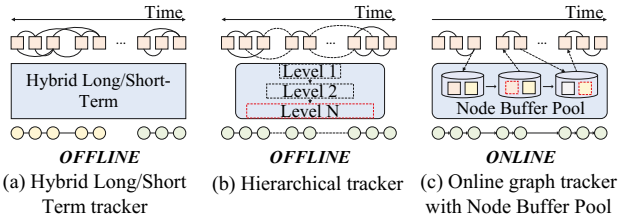
art (SOTA) trackers from mainstream paradigms under LFR conditions. These included the graph-based SUSHI (Cetintas, Brasó, and Laura 2023), DiffMOT (Lv et al. 2024), and DiffusionMOT (Hu et al. 2025) from the Tracking-by-Detection (TBD) paradigm, and MOTIP (Gao, Qi, and Wang 2025) from the Joint Detection and Tracking (JDT) paradigm. As illustrated in Figure 1, the results on the MOT17 validation set reveal a clear trend: the tracking performance for all paradigms decreases significantly as the frame gap increases from 2 to 10. Therefore, exploring methods to improve tracking performance in LFR videos is a significant research challenge.

To address this, some LFR tracking methods have been proposed. However, these methods often rely on mechanisms that do not scale well to large temporal gaps. For instance, approaches extending optical flow, such as APP-Tracker (Zhou et al. 2022, 2024), struggle when visual correspondence is lost. Other methods that focus on enhancing temporal representations (e.g., ColTrack (Liu, Wu, and Fu 2023)) or encoding frame rate information (e.g., FraMOT (Feng et al. 2024)) improve robustness but do not fundamentally redesign the association logic for managing objects that are absent for extended periods. This can cause local mismatches to amplify global tracking errors.

Despite these efforts, the aforementioned methods do not adequately address the three main challenges of LFR tracking. As we can see from Figure 1 and Table 1, while the performance of all trackers degrades with increasing frame gap, the GNN-based tracker, SUSHI, demonstrates the best performance retention (ΔH, the value of performance reduction from frame gap = 2 to frame gap = 10). Its advantage is that it can model complex non-local relationships and maintain robust tracking even across wide temporal gaps. This indicates that GNNs have unique potential in the domain of LFR-MOT.

However, this powerful relational reasoning is typically realized in an offline manner, processing the entire detection graph at once, as seen in methods like SUSHI, CoNo-link (Gao et al. 2024), and RTAT (Guo, Liu, and Abe 2024). This offline nature fundamentally conflicts with the real-time requirements of most tracking applications. This leads to a critical question that forms the core **Motivation** of our work: **How can we leverage the powerful relational reasoning ability of GNNs within an online framework tailored to the challenges of LFR tracking?**

To this end, we propose GLoMOT, a novel online GNN-based Low-Frame-Rate Multi-Object Tracker. Our approach contrasts with previous offline GNN paradigms, as illustrated in Figure 2. First, to bridge large temporal gaps, we introduce a Dynamic Node Buffer Pool. This new long-term memory mechanism caches the state information of missing objects, ensuring they can be robustly re-associated. Second, to handle the extreme motion and appearance uncertainty, we propose an adaptive context-aware module (ACAM). This module dynamically adjusts the weights of positional and appearance features, generating a more robust feature representation for predicting node connections. Finally, to solve spatial ambiguity from occlusions, we develop a pseudo-depth feature calculation method. This method provides the GNN with critical geometric context, enabling it to distinguish between overlapping objects.



(a) Hybrid Long/Short Term tracker (b) Hierarchical tracker (c) Online graph tracker with Node Buffer Pool

Figure 2: Comparison of tracking paradigms of different GNN trackers.

To summarize, our contributions are as follows:

- We propose GLoMOT, a novel online GNN-based framework for LFR tracking. It features a Dynamic Node Buffer Pool that acts as a long-term memory to enable robust object re-association across significant frame gaps, a key distinction from previous offline GNN trackers.
- We propose an Adaptive Context-Aware Module (ACAM) to mitigate motion and appearance ambiguity by dynamically weighting features. This is complemented by a pseudo-depth feature that resolves spatial ambiguity by providing critical geometric context during occlusions.
- Extensive experiments on public MOT benchmarks, including DanceTrack, MOT17, and VisDrone, demonstrate that our proposed GLoMOT achieves superior performance, particularly under low-frame-rate conditions.

Method

Problem Formulation

Our tracking methodology is founded upon the TBD paradigm, which first detects objects in each frame and then associates them over time. We formulate the core association task as an edge classification problem on an undirected graph. For any pair of consecutive frames at times t and $t+1$, we are given two sets of detections (detected bounding boxes), D_t and D_{t+1}. We construct an undirected graph $G_t = (V_t, E_t)$. Formally, G_t is the complete bipartite graph with the two node sets D_t and D_{t+1}. The node set is thus $V_t = D_t \cup D_{t+1}$, and the edge set E_t represents all potential associations between detections in D_t and D_{t+1}. The task of our GNN is to predict a confidence score for each edge, where a high score indicates that the two connected nodes belong to the same object.

Online Tracking Pipeline

Our online tracking pipeline extends this basic graph association framework by introducing a temporal memory component to handle the object occlusions and disappearances prevalent in LFR scenarios. The overall architecture of GLoMOT is shown in Figure 3. After performing association on the graph G_t for frames t and $t+1$, any node from D_t that remains unassociated is considered temporarily lost and is added to our novel memory component, the Dynamic Node Buffer Pool.

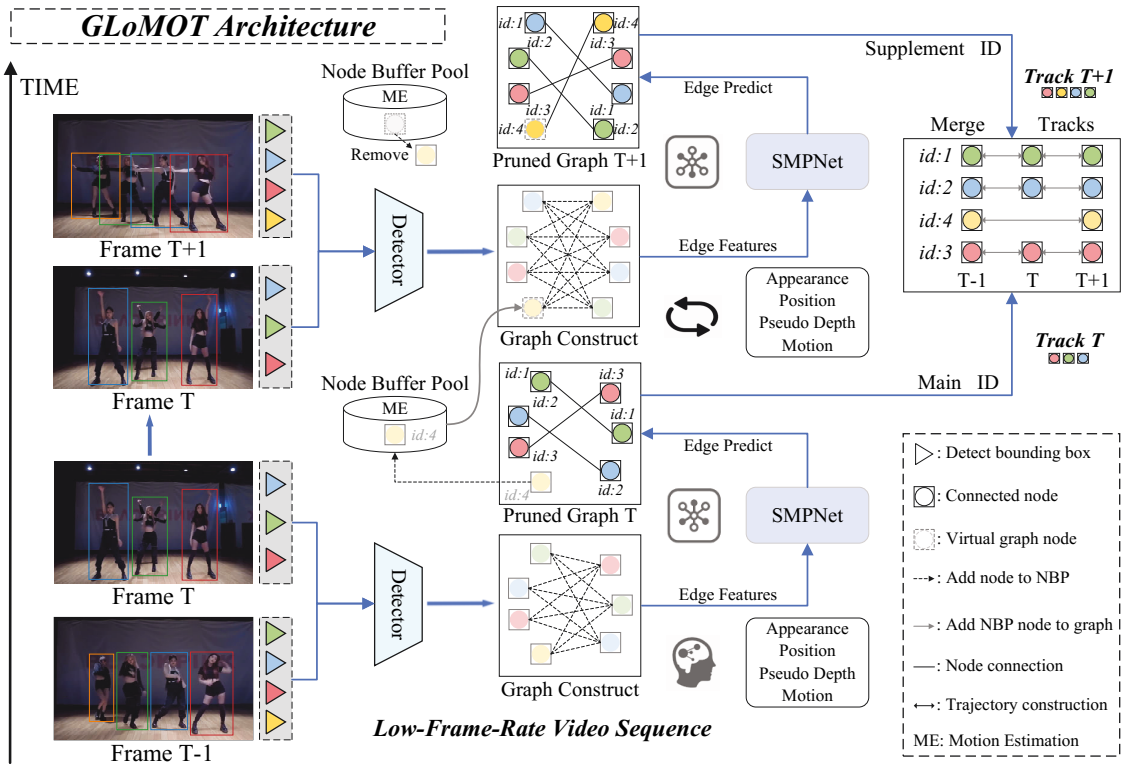


Figure 3: The overall architecture of GLoMOT. This architecture performs online tracking by iteratively constructing a graph between adjacent frames to perform association. A key component is the Dynamic Node Buffer Pool, which caches the unmatched nodes from a previous step and re-introduces them as "virtual nodes" into the current graph construction to handle long-term occlusions.

At the subsequent time step (from $t+1$ to $t+2$), the pipeline begins by updating the state of all nodes currently held in the buffer (including appearance and position). The graph construction for this new step is then fundamentally different. The source node set for the new graph G_{t+1} becomes an augmented set, composed of both the current detections D_{t+1} and the updated "virtual nodes" from the buffer. This augmented set is then associated with the new detections from the detection set D_{t+2}.

These virtual nodes, representing previously lost objects, fully participate in the GNN-based tracker's association process. A successful association between a buffered node and a new detection constitutes a successful re-identification. After re-identification, the node is removed from the buffer, and its original identity is restored to the trajectory. Unmatched buffered nodes may be retained for future frames or removed if their age exceeds a predefined threshold. This online framework, which dynamically integrates a memory of occluded objects into the graph-based association process, allows GLoMOT to maintain robust trajectory continuity in challenging LFR environments.

Strong Message Passing Network (SMPNet)

To generate robust features for predicting connections between object nodes in a graph, we propose a novel GNN architecture, the Strong Message Passing Network (SMPNet). The structure of SMPNet is illustrated in Figure 4. It is composed of an Adaptive Context-Aware Module (ACAM), parallel node and edge encoding streams, a Time-Aware Attention Module (TAM), and various MLP blocks for feature transformation and classification.

The process begins with a pre-pruned graph G'_t. This graph is derived from the complete bipartite graph G_t by removing any edge e_{ij} where the Euclidean distance between the centers of the connected nodes (d_i and d_j) exceeds a predefined threshold. Our network then iterates through each remaining edge e_{ij} in G'_t. For each edge, we compute a comprehensive feature vector by concatenating four distinct vector cues: appearance similarity ($\mathbf{f}_{\text{app}}$), positional distance ($\mathbf{f}_{\text{pos}}$), motion consistency ($\mathbf{f}_{\text{mo}}$), and our proposed pseudo-depth feature ($\mathbf{f}_{\text{pd}}$). This initial edge feature vector $\mathbf{x}_{ij}$ is defined as the concatenation of these four vectors:

$$\mathbf{x}_{ij} = \mathbf{Concat}[\mathbf{f}_{\text{app}}, \mathbf{f}_{\text{pos}}, \mathbf{f}_{\text{mo}}, \mathbf{f}_{\text{pd}}] \quad (1)$$

The feature vector first passes through our proposed ACAM. This module functions as a dynamic gating mechanism designed to intelligently re-weight the input features. Specifically, for a batch of edges, the ACAM computes a low-dimensional context vector derived from the normalized spatial distance and appearance difference. This context vector is then processed by a small neural network to generate a set of weights via a softmax function. These weights are subsequently used to scale the positional and appearance feature

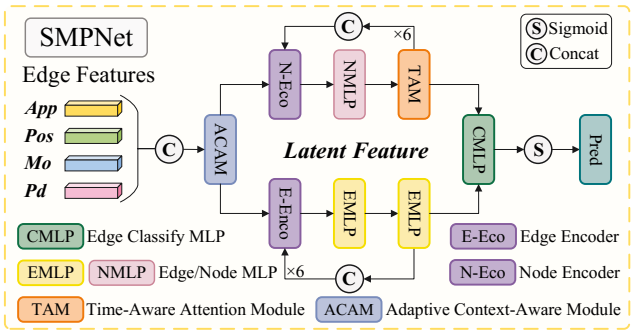


Figure 4: The structure of SMPNet.

groups. This mechanism allows SMPNet to dynamically increase the influence of appearance features when objects are spatially close but visually distinct, or conversely, to emphasize positional features when appearance is ambiguous, thereby adapting to the varying challenges of LFR tracking.

Following the ACAM, the re-weighted features are bifurcated and processed by two parallel streams: a node-centric branch and an edge-centric branch. The edge-centric stream processes features directly through a series of Edge MLPs (EMLPs). In parallel, the node-centric stream first aggregates features at the node level using a Node Encoder and then further transforms using a Node MLP (NMLP). A key component in this branch is the TAM, which is designed to capture and leverage temporal dependencies in the data.

Finally, the latent features produced by both the node-centric and edge-centric streams are concatenated. This fused representation, which now contains both refined edge attributes and node-level context, is passed to a final Edge Classify MLP (CMLP). The output is then passed through a Sigmoid activation function to produce the final prediction.

Training Object Consistency Loss

The training of SMPNet is guided by a composite objective function designed for the challenges of LFR tracking, where models are trained on frame pairs with a variable temporal gap, ΔT. Let $n = \Delta T$ denote this temporal gap in frames. The total loss $\mathcal{L}_{\text{total}}$ combines a primary classification loss $\mathcal{L}_{\text{cls}}$ with a temporal consistency regularizer $\mathcal{L}_{\text{temp}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{temp}} \tag{2}$$

The classification loss $\mathcal{L}_{\text{cls}}$ is the mean Focal Loss over all edges $e_{ij} \in E$, where E represents the set of all edges in the pruned graph G'_t. For each edge e_{ij}, $y_{ij} \in \{0, 1\}$ is its ground-truth label (where 1 indicates a true match and 0 indicates a false match), and $p_{ij} \in [0, 1]$ is the prediction probability output by the SMPNet's final sigmoid function.

$$\mathcal{L}_{\text{cls}} = \frac{1}{|E|} \sum_{e_{ij} \in E} \mathcal{L}_{\text{FL}}(p_{ij}, y_{ij}) \tag{3}$$

where $\mathcal{L}_{\text{FL}}$ is the standard Focal Loss with focusing parameter γ.

The temporal consistency term $\mathcal{L}_{\text{temp}}$, regularizes the feature representations for the same object across the temporal gap. It is applied exclusively to the set of ground-truth positive edges E^+, which is the subset of E containing only true matches (i.e., $E^+ = \{e_{ij} \in E \mid y_{ij} = 1\}$). $\mathcal{L}_{\text{temp}}$ is a weighted sum of three distinct losses:

$$\mathcal{L}_{\text{temp}} = \lambda_{\text{app}}\mathcal{L}_{\text{app}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{feet}}\mathcal{L}_{\text{feet}} \tag{4}$$

where λ_{app}, λ_{depth}, and λ_{feet} are scalar weights. The components are defined as follows:

- **Appearance Consistency ($\mathcal{L}_{\text{app}}$):** This loss penalizes the appearance distance d_{reid} between matched pairs. Here, d_{reid} is a function (Cosine distance) that computes the dissimilarity between the appearance feature vectors of the two connected detections d_i and d_j.

$$\mathcal{L}_{\text{app}} = \frac{1}{|E^+|} \sum_{e_{ij} \in E^+} d_{\text{reid}}(d_i, d_j) \tag{5}$$

- **Depth Consistency ($\mathcal{L}_{\text{depth}}$):** This term penalizes the absolute distance between scalar pseudo-depth values. The function $\mathcal{D}_{\text{pseudo}}(d_k)$ returns the estimated scalar pseudo-depth value for a given detection d_k.

$$\mathcal{L}_{\text{depth}} = \frac{1}{|E^+|} \sum_{e_{ij} \in E^+} |\mathcal{D}_{\text{pseudo}}(d_i) - \mathcal{D}_{\text{pseudo}}(d_j)| \tag{6}$$

- **Feet Velocity Regularization ($\mathcal{L}_{\text{feet}}$):** This loss penalizes the average foot-point velocity to enforce physically plausible motion, which is particularly effective for LFR where simple position consistency is insufficient. It is formulated as the mean Euclidean norm of the foot-point velocity, where $c_{\text{feet}}(d_k)$ is a function that returns the 2D pixel coordinates of the bounding box's bottom-center (the "foot-point") for detection d_k, and $n = \Delta T$ is the temporal gap defined earlier:

$$\mathcal{L}_{\text{feet}} = \frac{1}{|E^+|} \sum_{e_{ij} \in E^+} \frac{1}{n} \|c_{\text{feet}}(d_i) - c_{\text{feet}}(d_j)\|_2 \tag{7}$$

By combining these losses, our final training objective equips the model to not only perform accurate association but also to learn temporally stable and physically plausible feature representations, which are critical for robust tracking in Low-Frame-Rate video.

Pseudo-Depth Feature for Geometric Context

To effectively address the spatial ambiguity caused by occlusions in LFR scenarios, we introduce a novel pseudo-depth feature calculation method. While prior works such as SparseTrack (Liu et al. 2025) and PD-SORT (Wang et al. 2025) have explored using pseudo-depth to aid in matching, their calculations are often based on a simplistic heuristic, such as the vertical position of an object's bounding box in the frame. To overcome this limitation, we propose a more comprehensive pseudo-depth calculation that integrates multiple geometric and relational cues. Furthermore, we are the first to incorporate this rich pseudo-depth information as a dynamic edge feature within a GNN framework for multi-object tracking.

Our pseudo-depth score for each detection d_i in a frame's detection set D is formulated as a weighted fusion of three distinct components. Let a detection be $d_i = (x_i, y_i, w_i, h_i)$, where x_i and y_i are the center points of the bounding box, and the image dimensions be (W, H).

Size-Based Depth. This component is based on the perspective assumption that objects appearing larger are closer. Let the area of detection d_i be $A_i = w_i \cdot h_i$. The score is calculated by normalizing the area of each detection with respect to the minimum and maximum areas, $A_{\min}$ and $A_{\max}$, within the current frame's detection set D:

$$\mathcal{D}_{size}(d_i) = \frac{A_i - A_{\min}}{A_{\max} - A_{\min} + \epsilon} \tag{8}$$

where ϵ is a small constant to prevent division by zero.

Position-Based Depth. We leverage the heuristic that objects lower in the image are typically nearer. This is computed from the normalized vertical position of the object's foot-point (bottom-center) $c_{y,i} = y_i + h_i/2$:

$$\mathcal{D}_{pos}(d_i) = 1 - \frac{c_{y,i}}{H} \tag{9}$$

Occlusion-Based Depth. This component infers depth from inter-object relationships, based on the principle that heavily occluded objects are likely farther away. We first quantify a total occlusion ratio O_i for detection d_i by summing its intersection areas with all other detections $d_j \in D$:

$$O_i = \sum_{j \neq i, d_j \in D} \frac{Area(d_i \cap d_j)}{A_i + \epsilon} \tag{10}$$

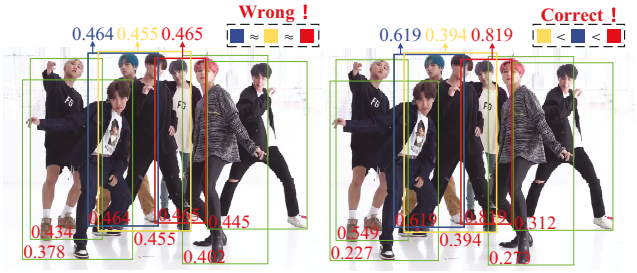
The final occlusion-based depth is then given by:

$$\mathcal{D}_{occ}(d_i) = \frac{1}{1 + O_i} \tag{11}$$

Finally, these three components are combined via a weighted sum to produce the final comprehensive pseudo-depth score.

$$\begin{aligned}\mathcal{D}_{pseudo}(d_i) = & w_{size}\mathcal{D}_{size}(d_i) + w_{pos}\mathcal{D}_{pos}(d_i) + \\ & w_{occ}\mathcal{D}_{occ}(d_i)\end{aligned} \tag{12}$$

where w_{size}, w_{pos}, and w_{occ} are the weights for each component. This calculated score is then incorporated into the feature vector for each edge in the graph.



(a) SparseTrack (b) Ours

Figure 5: Comparison of visualization results with SparseTrack's pseudo-depth calculation method.

As illustrated in Figure 5, our approach provides a more accurate depth estimation compared to methods relying solely on vertical positioning.

Experiments

Datasets and Metrics

We evaluate our proposed GLoMOT on several diverse public benchmarks: MOT17 (Milan et al. 2016), DanceTrack (Sun et al. 2022), SportsMOT (Cui et al. 2023), and the multi-class Visdrone-MOT (Cao et al. 2021). Following standard protocols, Low-Frame-Rate (LFR) videos are generated by downsampling the original sequences at various fixed intervals, or frame gaps. Our evaluation relies on three widely used metrics: HOTA (Luiten et al. 2021), IDF1 (Ristani et al. 2016), and MOTA (Kasturi et al. 2008).

Implementation Details

Our model was trained for 5 epochs using the AdamW optimizer with an initial learning rate set to 1.5e-4. For the ACAM module, the softmax temperature is set to 0.5. For the pseudo-depth feature, the weights for the position, occlusion, and size components w_{pos}, w_{occ}, and w_{size}, are set to 0.2, 0.4, and 0.4, respectively. In our composite loss function, the weights for the temporal consistency terms λ_{app}, λ_{depth}, λ_{feet} are set to 0.05, 0.02, and 0.01. Our code is available at https://github.com/sad123-yx/GLoMOT.

Method	Venue	MOT17				DanceTrack			
		n=2	*n=5*	*n=10*	ΔH	*n=2*	*n=5*	*n=10*	ΔH
HOTA									
Bytetrack	ECCV'22	66.5	61.3	57.9	8.6	41.7	34.2	29.4	12.3
OC-SORT	CVPR'23	66.1	—	56.1	10.0	52.2	35.9	30.3	21.9
SUSHI*	CVPR'23	76.4	74.0	69.5	6.9	51.7	46.5	39.5	12.2
DiffusionTrack	AAAI'24	66.0	59.6	50.4	15.6	55.0	40.9	31.7	23.3
DiffMOT	CVPR'24	77.6	75.8	71.2	6.4	57.4	53.5	43.5	13.9
DiffusionMOT	TNNLS'25	75.8	69.9	63.8	12.0	**63.9**	48.7	37.9	26.0
MOTIP	CVPR'25	76.6	71.2	64.5	12.1	63.2	**57.5**	**48.6**	14.6
GLoMOT(Ours)	—	**78.1**	**77.2**	**75.0**	**3.1**	53.8	50.9	44.2	**9.6**
IDF1									
Bytetrack	ECCV'22	76.7	71.1	67.1	9.6	48.6	39.6	33.7	14.9
SUSHI*	CVPR'23	82.5	81.1	75.4	7.1	49.8	42.9	33.5	16.3
DiffusionTrack	AAAI'24	74.2	66.2	55.9	18.3	55.6	38.4	29.8	25.8
DiffMOT	CVPR'24	85.1	82.4	74.9	10.2	58.8	53.8	40.2	18.6
DiffusionMOT	TNNLS'25	83.4	76.5	67.7	15.7	64.3	47.4	36.1	28.2
MOTIP	CVPR'25	86.2	80.8	71.7	14.5	**68.5**	**61.4**	**51.8**	16.7
GLoMOT(Ours)	—	**87.1**	**85.5**	**81.8**	**5.3**	53.1	48.6	40.1	**13.0**
MOTA									
Bytetrack	ECCV'22	75.8	70.2	64.8	11.0	83.6	73.8	60.5	23.1
SUSHI*	CVPR'23	88.7	83.3	78.8	9.9	86.8	81.4	70.4	16.4
DiffusionTrack	AAAI'24	81.9	74.7	60.9	21.0	85.7	76.3	60.9	24.8
DiffMOT	CVPR'24	**89.2**	84.7	75.4	13.8	88.9	**84.4**	71.5	17.4
DiffusionMOT	TNNLS'25	86.8	79.9	70.3	16.5	**91.1**	81.9	66.2	24.9
MOTIP	CVPR'25	87.1	86.2	83.2	3.9	82.4	79.8	74.7	**7.7**
GLoMOT(Ours)	—	88.5	**86.6**	**85.1**	**3.4**	87.6	**84.4**	**76.1**	11.5

Table 1: Performance comparison of different methods at different frame rates on MOT17 and DanceTrack. Best results are in **bold**, second best are underlined.

Performance at Different Low-Frame-Rates

To evaluate GLoMOT's robustness in LFR conditions, we compare it with state-of-the-art (SOTA) methods (Zhang et al. 2022; Cao et al. 2023) on the MOT17 and DanceTrack valset, with results shown in Table 1. The experiments are

conducted at various frame gaps (n=2, 5, 10), and we use the performance drop (ΔH from n=2 to n=10) to measure stability. On MOT17, our proposed GLoMOT demonstrates remarkable performance. It not only achieves the highest HOTA and IDF1 scores across all frame gaps but also exhibits the greatest stability, with the smallest performance drop (ΔH=3.1 on HOTA) of all compared methods. This validates the effectiveness and robustness of our framework.

On the more challenging DanceTrack dataset, characterized by non-linear motion, GLoMOT's key advantage in robustness is even more pronounced. While some methods like MOTIP achieve higher absolute scores, GLoMOT obtains the best stability on HOTA and IDF1 metrics. This performance drop is significantly smaller than other SOTA trackers, demonstrating the superior robustness of our proposed method for tracking in challenging LFR videos.

GLoMOT's Component Ablation Study

The results of our component ablation study are presented in Table 2. This study validates the effectiveness of our three components: the Node Motion Status Update (NM), the Strong Message Passing Network (SMP) and the Pseudo-Depth (PD) feature calculation method. On the MOT17 dataset, the combination of all three modules consistently yields the best performance across all metrics, demonstrating their positive and complementary contributions. The results on the more challenging DanceTrack dataset underscore the importance of our full design for LFR scenarios. While the SMP+PD configuration excels at a small frame gap (n=2), the complete model including the NM module demonstrates superior robustness, achieving the top HOTA and IDF1 scores at the largest gap (n=10). This confirms that the Node Motion Status Update is critical for maintaining tracking performance in extreme LFR conditions.

Methods			MOT17				DanceTrack			
			n=2		*n=10*		*n=2*		*n=10*	
NM	SMP	PD	HOTA	IDF1	HOTA	IDF1	HOTA	IDF1	HOTA	IDF1
			77.18	85.51	73.47	79.43	53.39	53.01	38.91	34.41
✓			77.43	85.96	74.44	80.76	53.60	52.52	38.17	33.52
	✓		77.63	86.15	74.56	81.08	53.56	52.45	43.63	39.24
		✓	77.44	85.98	74.19	80.45	52.80	51.08	40.31	36.18
✓	✓		78.06	86.83	74.93	81.67	53.76	52.66	44.13	39.96
✓		✓	77.57	86.21	74.45	80.97	52.55	50.20	39.86	35.18
	✓	✓	77.82	86.53	74.75	81.23	**53.95**	**53.33**	43.24	39.20
✓	✓	✓	**78.12**	**87.15**	**75.02**	**81.81**	53.83	53.10	**44.22**	**40.11**

Table 2: Ablation study of the proposed modules at different frame rates on the validation set. Best results are in **bold**.

Pseudo-Depth Feature Components

To validate the effectiveness of our proposed pseudo-depth feature calculation method, we conducted a comprehensive ablation study; the results are shown in Table 3. We compare five configurations: a baseline without any pseudo-depth feature (None); using only the position-based method in SparseTrack (Pos); using only our proposed occlusion-based depth calculation method (Occ); using only our proposed size-based depth calculation method (Size); and our full, fused method (All).

From the Table 3, we can notice that the position-based method (Pos) offers only marginal gains on MOT17 and slightly degrades performance on the more challenging DanceTrack dataset, highlighting the need for a more robust feature. Our proposed Occ and Size components individually outperform the Pos baseline, and the fused feature achieves the best HOTA and IDF1 scores on both datasets, particularly at the more challenging n=10 frame gap. This validates that our comprehensive calculation, which combines size, position, and occlusion cues, provides a richer and more effective geometric context to the GNN, leading to more accurate and robust associations in LFR scenarios.

Depth	MOT17				DanceTrack			
	n=2		*n=10*		*n=2*		*n=10*	
	HOTA	IDF1	HOTA	IDF1	HOTA	IDF1	HOTA	IDF1
None	77.63	86.15	74.56	81.08	53.56	52.45	43.63	39.24
+ *Pos*	77.72	86.38	74.52	80.96	53.48	51.63	43.64	39.63
+ *Occ*	77.52	86.08	74.68	81.13	53.97	52.29	43.56	39.33
+ *Size*	77.80	86.49	74.64	81.09	**54.14**	52.98	43.65	39.74
All	**77.82**	**86.53**	**74.75**	**81.23**	53.95	**53.33**	**43.68**	**39.81**

Table 3: Ablation study of the pseudo-depth feature components on MOT17 and DanceTrack.

Cross-Domain Evaluation

To evaluate the generalization ability of GLoMOT, we conducted a cross-dataset evaluation; the results are summarized in Table 4. The results show that GLoMOT has excellent cross-domain transfer capabilities. Notably, models trained on the DanceTrack and SportsMOT datasets achieve SOTA performance when evaluated on MOT17, even outperforming models trained natively on MOT17 itself. Moreover, on the more challenging DanceTrack and SportsMOT datasets, models trained on other domains also achieve highly competitive results, proving the strong generalization ability of our model. We believe that training on a diverse and challenging set of scenarios is crucial to building a truly general multi-object tracking model.

Train Set	**Test on MOT17**		**Test on DanceTrack**		**Test on SportsMOT**	
	HOTA ↑	IDF1 ↑	HOTA ↑	IDF1 ↑	HOTA ↑	IDF1 ↑
MOT17	75.91	83.51	53.03	51.71	91.85	88.85
DanceTrack	76.01	83.43	**56.79**	**55.03**	90.21	87.72
SportsMOT	**76.30**	**84.33**	53.32	50.73	**93.05**	**90.87**

Table 4: Cross-dataset evaluation of GLoMOT. The model is trained on one dataset (rows) and evaluated on another (columns).

Efficiency for Edge Deployment

For LFR tracking to be practical on resource-constrained edge devices, both computational efficiency (speed) and a small model footprint (size) are critical. Table 5 provides a

comparison of GLoMOT's efficiency against SOTA methods. Performance is reported in HOTA. A primary advantage of our GLoMOT is its exceptionally lightweight design. With an association model size of only 1.11MB, it is orders of magnitude smaller than other trackers like DiffMOT (168MB) and MOTIP (677MB). Furthermore, unlike offline methods such as SUSHI, GLoMOT operates in a fully online manner, a prerequisite for real-time applications. Despite its compact size, our method maintains highly competitive tracking performance on both MOT17 and DanceTrack. In terms of speed, GLoMOT achieves 15 FPS on MOT17 and 19 FPS on DanceTrack. This combination of a minimal model footprint and strong tracking accuracy makes GLoMOT a practical solution for deployment on edge devices, where keyframes are typically at a rate of 10 FPS or less.

Method	Det	Asso	Model Size	Online	MOT17	DanceTrack
ByteTrack	Y	N	—	Y	63.1	47.3
DiffMOT	Y	Y	168MB	Y	64.5	63.4
DiffusionTrack	N	Y	3077MB	Y	60.8	52.4
SUSHI	Y	Y	2.27MB	N	66.5	63.3
MOTIP	N	Y	677MB	Y	59.2	67.5
GLoMOT (Ours)	Y	Y	**1.11MB**	Y	63.3	63.8

Table 5: Efficiency and performance comparison of GLoMOT with SOTA methods.

Benchmark Results

To demonstrate GLoMOT's robustness beyond LFR scenarios, we evaluated its performance on high-frame-rate benchmarks. The results of SUSHI* in the table are the online version we implemented, and GLoMOT* is the result with linear interpolation. SportsMOT benchmark results and more experimental details can be found in supplementary materials (uploaded to GitHub).

Method	Venue	HOTA ↑	IDF1 ↑	MOTA ↑
Transformer based:				
ColTrack	ICCV'23	72.6	74.0	92.1
CO-MOT	ICLR'25	65.3	66.5	89.3
MOTIP	CVPR'25	67.5	72.2	90.3
TGFormer	AAAI'25	69.1	71.7	91.9
CNN & Graph Based:				
OC-SORT	CVPR'23	55.1	54.2	89.4
SUSHI*	CVPR'23	60.8	60.5	91.6
DiffusionTrack	AAAI'24	52.4	47.5	89.3
DiffMOT	CVPR'24	63.4	64.0	92.7
CoNo-Link	AAAI'24	63.8	64.1	89.7
DiffusionMOT	TNNLS'25	63.6	64.2	90.0
TOPICTrack	TIP'25	58.3	58.4	90.9
OFTrack	AAAI'25	60.9	61.7	90.9
GLoMOT (Ours)	—	63.8	63.2	**92.8**
GLoMOT* (Ours)	—	**64.8**	**64.5**	92.7

Table 6: Performance comparison on the DanceTrack test set.

On the MOT17 and DanceTrack benchmark, as shown in Table 6 and 7, GLoMOT achieves competitive performance, demonstrating the strong baseline performance of our architecture (Liu, Wu, and Fu 2023; Luo et al. 2025; Zeng, Huang, and Pei 2025; Gao et al. 2024; Cao et al. 2025; Song et al. 2025). To further validate the suitability of GLoMOT for deployment on edge devices, we evaluated it on the VisDrone-MOT dataset, a benchmark for multi-class object tracking from aerial (UAV) perspectives (Liu et al. 2022; Du et al. 2023; Yao et al. 2023; Yi et al. 2024; Deng et al. 2024; Lv et al. 2023; Yao et al. 2025). This scenario is a practical use-case for LFR tracking on computationally constrained platforms. As shown in Table 8, our proposed GLoMOT achieves SOTA performance. This superior performance on complex aerial data underscores the robustness and efficiency of our lightweight GNN tracking framework. It demonstrates that GLoMOT is not only effective for pedestrian tracking but is also particularly well-suited for real-world deployment on platforms like drones.

Method	Venue	HOTA ↑	IDF1 ↑	MOTA ↑
ColTrack	ICCV'23	61.0	73.9	78.8
SUSHI*	CVPR'23	62.4	76.9	77.6
DiffusionTrack	AAAI'24	60.8	73.8	77.9
DiffMOT	CVPR'24	**64.5**	**79.3**	79.8
DiffusionMOT	TNNLS'25	63.2	77.8	78.2
MOTIP	CVPR'25	59.2	71.2	75.5
CO-MOT	ICLR'25	60.1	72.7	72.6
TGFormer	AAAI'25	60.3	72.0	74.9
OFTrack	AAAI'25	63.1	77.5	79.9
GLoMOT (Ours)	—	63.3	77.4	**80.6**
GLoMOT* (Ours)	—	63.4	77.6	80.5

Table 7: Performance comparison on the MOT17 test set.

Method	Venue	HOTA ↑	IDF1 ↑	MOTA ↑
UAVMOT	CVPR'22	38.1	45.1	38.6
StrongSORT	TMM'23	36.5	42.5	33.3
OC-SORT	CVPR'23	42.9	52.0	41.5
FOLT	MM'23	44.8	56.9	42.1
UCMCTrack	AAAI'24	37.1	48.6	38.4
PID-MOT	TCSVT'24	—	50.2	33.0
UGT	TGRS'24	45.5	57.7	41.8
MM-Tracker	AAAI'25	—	58.3	**44.7**
GLoMOT (Ours)	—	**47.2**	**58.7**	42.8

Table 8: Performance comparison on the VisDrone test set.

Conclusion

In this paper, we addressed the significant challenges of online Multi-Object Tracking in LFR videos. We proposed GLoMOT, a novel online GNN framework designed specifically to maintain robust performance despite large frame gaps, severe motion uncertainty, and spatial ambiguity. Our framework introduces three key contributions: a Dynamic Node Buffer Pool to bridge long temporal intervals via a long-term memory mechanism; an Adaptive Context-Aware Module (ACAM) to dynamically handle feature uncertainty; and a novel pseudo-depth feature to resolve occlusions using geometric context. Extensive experiments demonstrated that GLoMOT is not only lightweight and efficient but also achieves highly competitive performance on several benchmarks, showing exceptional robustness under a wide variety of challenging Low-Frame-Rate conditions.

Acknowledgments

This research was funded by National Natural Science Foundation of China (62371350, 62171326), Key Science and Technology Research Project of Xinjiang Production and Construction Corps (2025AB029), Alar City Science and Technology Plan Project (No.2024ZB02) and Humanities and Social Sciences Fund of the Ministry of Education of China (No.24YJAZH042).

References

Cao, J.; Pang, J.; Weng, X.; Khirodkar, R.; and Kitani, K. 2023. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9686–9696.

Cao, X.; Zheng, Y.; Yao, Y.; Qin, H.; Cao, X.; and Guo, S. 2025. TOPIC: a parallel association paradigm for multi-object tracking under complex motions and diverse scenes. *IEEE Transactions on Image Processing*, 34: 743–758.

Cao, Y.; He, Z.; Wang, L.; Wang, W.; Yuan, Y.; Zhang, D.; Zhang, J.; Zhu, P.; Van Gool, L.; Han, J.; et al. 2021. VisDrone-DET2021: The vision meets drone object detection challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2847–2854.

Cetintas, O.; Brasó, G.; and Laura, L.-T. 2023. Unifying short and long-term tracking with graph hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22877–22887.

Chiu, H.-k.; Li, J.; Ambruş, R.; and Bohg, J. 2021. Probabilistic 3D multi-modal, multi-object tracking for autonomous driving. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 14227–14233. IEEE.

Cioppa, A.; Giancola, S.; Deliege, A.; Kang, L.; Zhou, X.; Cheng, Z.; Ghanem, B.; and Van Droogenbroeck, M. 2022. Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3491–3502.

Cui, Y.; Zeng, C.; Zhao, X.; Yang, Y.; Wu, G.; and Wang, L. 2023. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9921–9931.

Deng, C.; Wu, J.; Han, Y.; Wang, W.; and Chanussot, J. 2024. Learning a robust topological relationship for online multiobject tracking in UAV scenarios. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–15.

Du, Y.; Zhao, Z.; Song, Y.; Zhao, Y.; Su, F.; Gong, T.; and Meng, H. 2023. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 25: 8725–8737.

Elhoseny, M. 2020. Multi-object detection and tracking (MODT) machine learning model for real-time video surveillance systems. *Circuits, Systems, and Signal Processing*, 39(2): 611–630.

Feng, W.; Bai, L.; Yao, Y.; Yu, F.; and Ouyang, W. 2024. Towards frame rate agnostic multi-object tracking. *International Journal of Computer Vision*, 132(5): 1443–1462.

Ganesh, S. V.; Wu, Y.; Liu, G.; Kompella, R.; and Liu, L. 2023. Fast and resource-efficient object tracking on edge devices: A measurement study. *arXiv preprint arXiv:2309.02666*.

Gao, R.; Qi, J.; and Wang, L. 2025. Multiple object tracking as id prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 27883–27893.

Gao, Y.; Xu, H.; Li, J.; Wang, N.; and Gao, X. 2024. Multi-scene generalized trajectory global graph solver with composite nodes for multiple object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1842–1850.

Guo, S.; Liu, R.; and Abe, N. 2024. RTAT: A Robust Two-Stage Association Tracker for Multi-object Tracking. In *International Conference on Pattern Recognition*, 432–447. Springer.

Hu, Y.; Hua, J.; Han, Z.; Zou, H.; Wu, G.; and Wang, Z. 2025. DiffusionMOT: A Diffusion-Based Multiple Object Tracker. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10): 18203–18217.

Kasturi, R.; Goldgof, D.; Soundararajan, P.; Manohar, V.; Garofolo, J.; Bowers, R.; Boonstra, M.; Korzhova, V.; and Zhang, J. 2008. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2): 319–336.

Liu, S.; Li, X.; Lu, H.; and He, Y. 2022. Multi-object tracking meets moving UAV. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8876–8885.

Liu, Y.; Wu, J.; and Fu, Y. 2023. Collaborative tracking learning for frame-rate-insensitive multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9964–9973.

Liu, Z.; Wang, X.; Wang, C.; Liu, W.; and Bai, X. 2025. Sparsetrack: Multi-object tracking by performing scene decomposition based on pseudo-depth. *IEEE Transactions on Circuits and Systems for Video Technology*.

Luiten, J.; Osep, A.; Dendorfer, P.; Torr, P.; Geiger, A.; Leal-Taixé, L.; and Leibe, B. 2021. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129(2): 548–578.

Luo, W.; Zhong, Y.; Gan, Y.; Ma, L.; et al. 2025. CO-MOT: Boosting End-to-end Transformer-based Multi-Object Tracking via Coopetition Label Assignment and Shadow Sets. In *The Thirteenth International Conference on Learning Representations*.

Lv, W.; Huang, Y.; Zhang, N.; Lin, R.-S.; Han, M.; and Zeng, D. 2024. Diffmot: A real-time diffusion-based multiple object tracker with non-linear prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19321–19330.

Lv, W.; Zhang, N.; Zhang, J.; and Zeng, D. 2023. One-shot multiple object tracking with robust ID preservation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6): 4473–4488.

Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; and Schindler, K. 2016. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, 17–35. Springer.

Song, Z.; Luo, R.; Ma, L.; Tang, Y.; Chen, Y.-P. P.; Yu, J.; and Yang, W. 2025. Temporal Coherent Object Flow for Multi-Object Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6978–6986.

Sun, P.; Cao, J.; Jiang, Y.; Yuan, Z.; Bai, S.; Kitani, K.; and Luo, P. 2022. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20993–21002.

Wang, Y.; Zhang, D.; Li, R.; Zheng, Z.; and Li, M. 2025. PD-SORT: Occlusion-Robust Multi-Object Tracking Using Pseudo-Depth Cues. *IEEE Transactions on Consumer Electronics*.

Yao, M.; Peng, J.; He, Q.; Peng, B.; Chen, H.; Chi, M.; Liu, C.; and Benediktsson, J. A. 2025. MM-Tracker: Motion Mamba for UAV-platform Multiple Object Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9409–9417.

Yao, M.; Wang, J.; Peng, J.; Chi, M.; and Liu, C. 2023. Folt: Fast multiple object tracking from uav-captured videos based on optical flow. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3375–3383.

Yi, K.; Luo, K.; Luo, X.; Huang, J.; Wu, H.; Hu, R.; and Hao, W. 2024. Ucmctrack: Multi-object tracking with uniform camera motion compensation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6702–6710.

Zeng, R.; Huang, Y.; and Pei, S. 2025. TGFormer: Transformer with Track Query Group for Multi-Object Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9824–9832.

Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; and Wang, X. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, 1–21. Springer.

Zhou, T.; Luo, W.; Shi, Z.; Chen, J.; and Ye, Q. 2022. Apptracker: Improving tracking multiple objects in low-frame-rate videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, 6664–6674.

Zhou, T.; Ye, Q.; Luo, W.; Ran, H.; Shi, Z.; and Chen, J. 2024. Apptracker+: Displacement uncertainty for occlusion handling in low-frame-rate multiple object tracking. *International Journal of Computer Vision*, 1–26.