

Article

Low-Altitude Multi-Object Tracking via Graph Neural Networks with Cross-Attention and Reliable Neighbor Guidance

Hanxiang Qian , Xiaoyong Sun, Runze Guo, Shaojing Su *, Bing Ding and Xiaojun Guo

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China; qianhanxiang_@nudt.edu.cn (H.Q.); sunxiaoyong14@nudt.edu.cn (X.S.); guorunze14@nudt.edu.cn (R.G.); dingbing23@nudt.edu.cn (B.D.); jeanakin@nudt.edu.cn (X.G.)

* Correspondence: ssjing@nudt.edu.cn

Highlights

What are the main findings?

- A novel multi-object tracking framework, NOWA-MOT, is proposed that leverages stable group context and graph neural networks to resolve tracking ambiguities caused by occlusion and non-linear motion in UAV imagery.
- The framework introduces a cascaded association mechanism using cross-graph attention for robust feature enhancement and reliably matched neighbors as anchors to guide the matching of more difficult, ambiguous targets.

What is the implication of the main findings?

- The proposed approach achieves state-of-the-art performance on challenging UAV datasets (VisDrone, UAVDT), significantly reducing identity switches and improving tracking continuity in dense, dynamic scenes.
- This context-aware tracking paradigm provides a more reliable foundation for downstream remote sensing applications, such as traffic flow analysis and smart city monitoring, by delivering higher-fidelity trajectory data.

Abstract

In low-altitude multi-object tracking (MOT), challenges such as frequent inter-object occlusion and complex non-linear motion disrupt the appearance of individual targets and the continuity of their trajectories, leading to frequent tracking failures. We posit that the relatively stable spatio-temporal relationships within object groups (e.g., pedestrians and vehicles) offer powerful contextual cues to resolve such ambiguities. We present NOWA-MOT (Neighbors Know Who We Are), a novel tracking-by-detection framework designed to systematically exploit this principle through a multi-stage association process. We make three primary contributions. First, we introduce a Low-Confidence Occlusion Recovery (LOR) module that dynamically adjusts detection scores by integrating IoU, a novel Recovery IoU (RloU) metric, and location similarity to surrounding objects, enabling occluded targets to participate in high-priority matching. Second, for initial data association, we propose a Graph Cross-Attention (GCA) mechanism. In this module, separate graphs are constructed for detections and trajectories, and a cross-attention architecture is employed to propagate rich contextual information between them, yielding highly discriminative feature representations for robust matching. Third, to resolve the remaining ambiguities, we design a cascaded Matched Neighbor Guidance (MNG) module, which uniquely leverages the reliably matched pairs from the first stage as contextual anchors. Through MNG, star-shaped topological features are built for unmatched objects relative to their stable neighbors, enabling accurate association even when intrinsic features are weak. Our comprehensive experimental evaluation on the VisDrone2019 and UAVDT



Academic Editor: Pedro Melo-Pinto

Received: 31 August 2025

Revised: 14 October 2025

Accepted: 17 October 2025

Published: 21 October 2025

Citation: Qian, H.; Sun, X.; Guo, R.; Su, S.; Ding, B.; Guo, X. Low-Altitude Multi-Object Tracking via Graph Neural Networks with Cross-Attention and Reliable Neighbor Guidance. *Remote Sens.* **2025**, *17*, 3502. <https://doi.org/10.3390/rs17203502>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

datasets confirms the superiority of our approach, achieving state-of-the-art HOTA scores of 51.34% and 62.69%, respectively, and drastically reducing identity switches compared to previous methods.

Keywords: multi-object tracking; graph neural networks; unmanned aerial vehicle; occlusion recovery

1. Introduction

Visual multi-object tracking (MOT) is a fundamental computer vision task with the aim of detecting objects in video sequences and assigning each target a consistent, unique identity. Driven by advancements in deep learning, MOT techniques have evolved significantly over the past decade. Compared to single-object tracking (SOT), MOT presents additional challenges, including the disappearance of existing objects, the emergence of new objects, dynamic variation in target count, and unreliable cues during trajectory association. Currently, most research [1–5] focuses on tracking objects in settings with stationary or horizontally moving cameras—such as handheld or vehicle-mounted cameras—which operate with limited perception ranges.

In recent years, unmanned aerial vehicles (UAVs) have been adopted across various domains, including search and rescue, smart cities, intelligent transportation, and remote sensing. Compared to stationary or horizontally moving cameras, UAV-mounted cameras offer higher perspectives and broader fields of view. However, they also introduce more complex challenges, such as small target sizes and non-linear camera motion. Consequently, there is a growing demand for innovative techniques capable of addressing the unique complexities in UAV scenarios.

As shown in Figure 1, The following challenges are primarily encountered in UAV-based MOT:

- (1) Image blurring induced by high-speed UAV motion.
- (2) Complex motion patterns resulting from UAV maneuvers (e.g., hovering, ascending, forward movement).
- (3) Frequent occlusions caused by elevated viewing angles, where trees, buildings, and objects that lie in the same imaging plane.
- (4) Indistinct texture information for small objects at high altitudes, increasing susceptibility to confusion with similar objects.



Figure 1. Challenges in UAV-based MOT.

Previous MOT research [6–8] predominantly features the tracking-by-detection (TBD) paradigm, which involves two steps: detection and association. Algorithms typically employ detection models to identify objects of interest, followed by re-identification (Re-ID) modules to extract appearance features. Motion characteristics are estimated using Kalman filtering, and both features are combined for association. Through this method, target feature extraction and data association are primarily optimized. Considering the characteristics of UAV imagery, relying solely on features from individual objects often leads to inaccurate associations. The reliability of detected features diminishes, making it

difficult to maintain visual consistency for occluded objects. Object detection bounding boxes can overlap, particularly when they are proximate, causing the extracted appearance features to be confused across objects. Furthermore, sudden non-linear motion (e.g., abrupt platform or target direction changes) significantly alters motion features, even for objects with a high detection confidence, rendering both the motion and appearance cues ineffective for matching.

Occlusion is a major factor contributing to MOT failure [9]. It reduces the visible area of a target, hindering the extraction of sufficiently discriminative features via Re-ID for appearance-based matching methods. For motion-based matching methods, deformation of the detection bounding box due to occlusion causes the Intersection over Union (IoU) between predicted and detected boxes to fall below the matching threshold. Additionally, occlusion lowers the detection confidence. Many cascade matching methods [6,8] prioritize high-confidence detections before processing low-confidence ones, but this approach prevents occluded objects from matching their true trajectories in the first matching stage due to their low confidence. Occlusion frequently occurs for dense objects that are spatially close and visually similar, increasing the likelihood of erroneous trajectory associations.

The core contribution of this study stems from a fundamental observation of how complex scenes are often structured [10–12]. Target occlusion does not happen in a vacuum; rather, it often occurs when a smaller, individual target is obscured by surrounding, more distinct objects or groups that exhibit consistent, coordinated motion patterns. For instance, a pedestrian might be temporarily hidden by a group of people moving in unison, or a car might be occluded by a convoy of trucks. In these scenarios, while the occluded target's individual appearance and motion features become unreliable, the spatio-temporal topology of the surrounding group remains relatively stable. Our method is designed to systematically exploit this observation. We posit that by modeling and leveraging these stable group dynamics, we can infer the identity and trajectory of an occluded object with far greater accuracy than by relying on its degraded individual features alone. This principle of using contextual group information to resolve individual ambiguity forms the foundational basis of our proposed framework.

Motivated by these insights, this paper introduces a confidence adjustment and cross-attention feature enhancement modules. The former integrates multiple cues to enhance the confidence scores of low-confidence objects, partially recovering occluded and blurred objects. The latter concatenates appearance features with spatial neighbor features and employs cross-attention mechanisms alongside graph neural networks to obtain features that better discriminate between trajectories and detections.

Another key innovation of our approach lies in utilizing reliable matched pairs from the first round as neighbors in the second matching stage. By modeling the topological relationships between unmatched detections/trajectories and these neighbors, more robust features are obtained. Specifically, we design velocity and directional constraints to identify reliable neighbors around unmatched trajectories and detections, enabling object association via a star-shaped topological distance metric. This dual mechanism ensures effective secondary matching even under severe occlusion.

The contributions of this paper can be summarized as follows:

- (1) A novel target confidence adjustment method, termed Low-Confidence Occlusion Recovery (LOR), is introduced. This technique incorporates mutual occlusion relationships, positional data, and detection confidence to re-evaluate initially low-confidence detections, improving their reliability. Consequently, the anticipated errors in multi-round associations are effectively mitigated.

(2) A Graph Cross-Attention Enhancement Mechanism (GCA) is proposed to refine object Re-ID features. Complementarily, an innovative topological similarity computation function is developed to facilitate the First-Round target association.

(3) A cascaded matching mechanism named Matched Neighbor Guidance (MNG) is designed, leveraging initial matches and neighboring trajectories to associate detections via directed graph feature propagation.

(4) The integrated approach enables NOWA-MOT to achieve state-of-the-art performance on the VisDrone2019 and UAVDT benchmarks.

The remainder of this paper is structured as follows: Section 2 reviews prior work on detection-based tracking, graph neural networks, and occlusion-aware methods. Section 3 details the proposed methodology, and Section 4 presents the experimental configuration and the results of comparative analyses with state-of-art trackers. Finally, Section 5 discusses the framework's attributes, and future research directions are suggested.

2. Related Work

2.1. Tracking by Detection

Modern multi-object tracking systems predominantly employ a tracking-by-detection (TBD) framework comprising three sequential stages: object detection, feature extraction, and data association. The detection module provides per-frame bounding box coordinates, object categories, and confidence scores and appearance features are extracted using re-identification (Re-ID) models. Some methods [13,14] feature improved re-identification models to obtain more discriminative features, where motion characteristics are quantified via techniques such as Kalman filtering. OC-SORT [15] utilizes target observations to compute virtual trajectories during occlusion periods, thereby mitigating the accumulation of Kalman filter parameter errors caused by occlusion. The association module establishes cross-frame correspondences by assigning unique identifiers based on trajectory detection similarity metrics. Common metrics include IoU [16], appearance cosine similarity [17], and novel feature modalities incorporating topological structures [18] or hybrid cues [19].

In UAV-based tracking, motion compensation modules are commonly integrated to mitigate platform-induced motion errors and camera jitter. Target occlusion constitutes a critical challenge for detection-based MOT systems, where partial or complete occlusion causes reduced detection reliability. To enhance robustness, specialized methods have been developed: Chen et al. [20] proposed Arbitrack, which employs rotated object detectors to replace traditional axis-aligned bounding boxes, capturing target orientation for improved motion estimation and feature matching. Deng et al. [21] leveraged topological features to augment target representations in aerial imagery, addressing motion feature degradation from non-linear trajectories while predicting occlusion probabilities via temporal-topological consistency, thus enabling dynamic recovery of occluded objects. Song et al. [22] proposed SFTrack, which proactively utilizes low-confidence detections as tracking initiators and exhibits an enhanced tracking accuracy through the reintroduction of traditional appearance matching algorithms for data association.

Our Matched Neighbor Guidance (MNG) module provides a theoretically stronger alternative to traditional cascaded matching paradigms. Standard cascaded approaches typically process all remaining low-confidence detections in a second stage, attempting to associate them based on weaker cues like IoU. The fundamental limitation of this is that this matching occurs in the context of high uncertainty. MNG introduces a novel principle: guidance from the certain to the uncertain. It uniquely leverages the high-confidence, reliably matched pairs from the first stage as stable contextual anchors. By constructing star-shaped topological features for unmatched objects relative to this shared, reliable context, MNG establishes a robust frame of reference to resolve ambiguity. This anchor-based

guidance provides a stronger associative signal than simply comparing two uncertain elements (an unmatched track and a low-confidence detection) in isolation.

2.2. Tracking with Graph Neural Networks

Graphs are widely used across disciplines to model relationships in complex unstructured data, representing inherent connections through nodes and edges. In multi-object tracking, this framework is formulated as a bipartite graph matching problem, where trajectories and detections serve as nodes and edges, respectively. Optimization methods including network flows [23], multi-cut [24], minimum clique graphs [25], and disjoint paths [26] have emerged from this paradigm. Although initially prevalent in offline methods [27,28], their extension to online MOT—requiring per-frame matching of heterogeneous trajectory detection features—faces core challenges regarding the development of effective association methods. Key challenges for graph-based tracking involve (1) significant growth in computational/storage requirements during extended tracking and (2) balancing feature propagation and smoothing. Excessive propagation causes blurring of target-background distinction, while insufficient propagation limits higher-order feature extraction.

The predominant method involves creating detection graphs and trajectory graphs for target matching through bipartite graph matching. GSM [29] generates a graph for each target, evaluating target similarity through graph comparison, while in the GNMOT [30] approach, appearance graphs and matching graphs are constructed for detections and trajectories, matching scores are computed for pairs, and the optimal match is determined. GMTracker [31] transforms the matching problem into a graph matching problem based on vertex mapping relationships and employs the implicit function theorem to ensure differentiability in the matching layer. SGT [32] utilizes edge classification to ascertain whether two detections belong to the same target, aiding in the recovery of low-confidence detections. The MotionTrack method introduces an interaction module to learn interaction-aware motion patterns from short-term trajectories. This module utilizes an asymmetric adjacency matrix to depict interactions between objects and integrates information through graph convolutional networks to make predictions.

2.3. Tracking Under Occlusion

Occlusion in multi-object tracking presents a significant challenge, including obstacle occlusion, inter-object occlusion, and both short-term and long-term occlusion. To address these issues, researchers have developed various methods. For instance, in [33], motion models were employed to predict the positions of occluded pedestrians over extended periods, while in [34], robust appearance models were constructed to re-identify occluded individuals.

This paper focuses on occlusion in the context of UAV-based ground observations, which are characterized by a bird's-eye view, the presence of numerous small objects, weak texture information, infrequent long-term occlusion, and frequent short-term inter-group occlusion among objects with similar appearances—all under non-linear platform motion. You et al. [35] proposed a method in which a social topology matrix is constructed based on spatio-temporal constraints, incorporating directional and velocity information to associate objects with their neighbors. This approach leverages statistical consistency in group motion to suppress false positives and recover missed detections. Additionally, BUSCA [36] introduced a transformer-based plug-and-play module that identifies missed detections by leveraging adjacent trajectories, motion cues, and learned trajectory labels.

While the occlusion recovery mechanism within our GCA module shares a high-level goal with “soft recovery” methods in the detection field, its underlying principle is fundamentally different. Traditional soft recovery techniques typically act as score

boosters, elevating the confidence of low-score detections based on simple spatio-temporal priors like IoU with predicted track locations. In contrast, our GCA module operates as a feature enhancer. Instead of merely adjusting confidence scores, it actively propagates rich contextual information between trajectory and detection graphs via a cross-attention mechanism. This process generates an enhanced, context-aware feature representation for occluded objects at the embedding level. By considering an object's relationship with its neighbors, GCA provides a more discriminative feature for the matching task, addressing the root cause of ambiguity rather than just its symptoms.

Based on this analysis, we argue that effectively handling occlusion in aerial view images hinges on the ability to recover low-confidence detections while minimizing false positives and carefully reassessing the appearance features of low-confidence objects. This is particularly critical in tracking-by-detection (TBD) frameworks, where discarding small objects during the detection phase precludes subsequent tracking. To safely enhance occlusion recovery without increasing false positives, it is essential to incorporate multiple validation mechanisms based on the spatio-temporal context of the target.

3. Proposed Method

This section details the proposed NOWA-MOT multi-object tracking framework. Section 3.1 provides an overview of the overall framework architecture and then Sections 3.2–3.4 delve into the three core components: the Low-Confidence Occlusion Recovery (LOR) module, the First-Round Association with Graph Cross-Attention (GCA) module for initial association, and the Second-Round Association with Matched Neighbor Guidance (MNG) module for secondary association.

3.1. Overview

We propose NOWA-MOT, a robust tracker designed to comprehensively address the challenges of low-altitude UAV tracking. As illustrated in Figure 2, the framework extends the classic tracking-by-detection (TBD) paradigm.

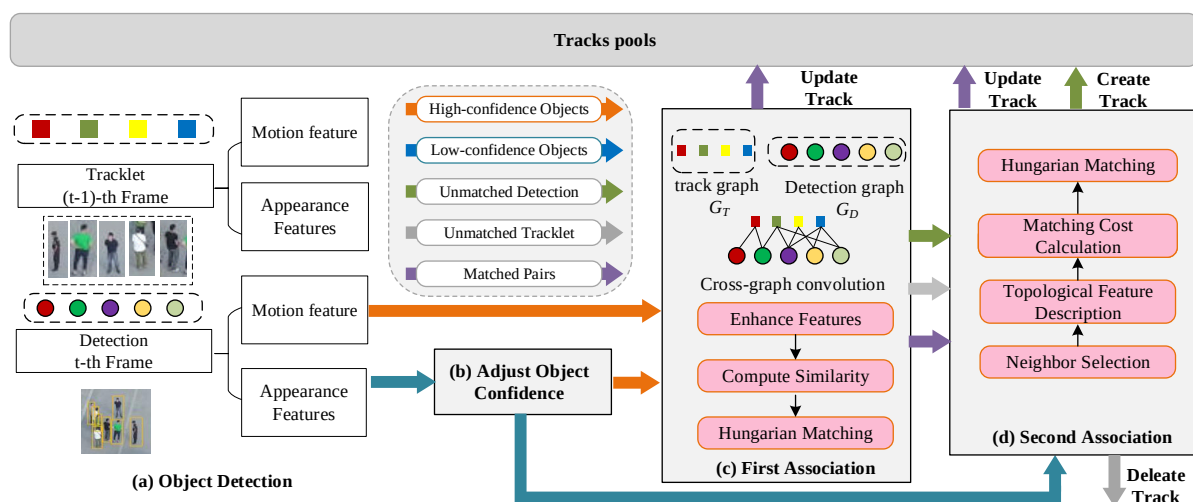


Figure 2. The overall architecture of NOWA-MOT is as follows: Tracklets and the detections at the t th frame are fed into a backbone network and a Re-ID module to extract motion and appearance features. Subsequently, the LOR module restores omitted objects with a low detection confidence. Following this, the GCA module enhances features between detections and tracklets via Graph Cross-Attention mechanisms and accomplishes first-stage association. Finally, the MNG module performs second-stage matching by leveraging reliably matched neighbors as a reference.

LOR: As shown in Figure 2b, the LOR module preprocesses the raw detector output. The detection confidence is dynamically adjusted by integrating multiple cues, including IoU, RIoU, location constraints, and initial detection scores—with the aim of recovering low-scoring true objects obscured by occlusion. This enables their participation in high-priority initial matching, mitigating adverse effects from detector failures.

GCA: Figure 2c illustrates the initial association stage. We employ a Graph Cross-Attention mechanism to enhance target feature representation. Specifically, a detection graph (current frame high-confidence detections) and a trajectory graph (active historical trajectories) are constructed, and node features are interactively enhanced via cross-attention between the graphs. The enriched features incorporate both intrinsic appearance and spatio-temporal contextual information from neighbors, yielding highly discriminative representations even under partial occlusion, thereby enabling accurate initial matching.

MNG: Leveraging initial matching results as priors (Figure 2d), the MNG module selects reliably matched neighboring trajectories (“reliable neighbors”) for each unmatched trajectory. Star-shaped topological features centered on unmatched trajectories/detections are built by aggregating the neighbors’ motion and appearance features. A cost matrix integrating multiple cues is computed, and final associations are resolved using the Hungarian algorithm.

3.2. Low-Confidence Object Recovery

Inspired by [37], the primary function of this module, as depicted in the figure, is to identify objects with a low detection confidence but a high likelihood of being present. Inputs include the current frame’s detections $D = \{d_i\}_{i=1}^M$, $d_i = \{loc_i, conf_i\}$, with M representing the number of detections, and trajectories $T = \{T_j\}_{j=1}^N$, $T_j = \{loc_j, status_j\}$, where N denotes the number of trajectories and $status_j$ indicates their current state (active or inactive). Predicted trajectory positions for the current frame, $\{\hat{D}_1, \hat{D}_2, \dots, \hat{D}_m\}$, are derived using Kalman filtering. Detections with a low confidence are defined by the threshold τ_{low} , set as $\{D_1^{low}, D_2^{low}, \dots, D_j^{low}\}$. The new confidence score $\hat{c}_{d_i}$ is calculated primarily from the detection confidence c_{d_i} and a similarity evaluation function $S(D_i, T_j)$, which measures the similarity between detections and trajectories. If this score exceeds the threshold τ_{high} for high-confidence objects, the target is included in the high-confidence target set $\{D_1^{high}, D_2^{high}, \dots, D_i^{high}\}$. The calculation formula is as follows:

$$\hat{c}_{d_i} = \max(c_{d_i}, \min(S(D_i, T_j), 1)) \quad (1)$$

where c_{d_i} indicates the prior confidence of the target and $\hat{c}_{d_i}$ signifies the adjusted confidence. If the confidence derived from the new algorithm surpasses the original confidence, the new confidence is adopted as the target confidence.

There are several methods for setting $S(D_i, T_j)$, with the most basic being IoU. In this study, a multi-cue combination of IoU, Recovery Intersection over Union (RIoU), and location similarity is utilized, as illustrated in Equation (2).

While the Intersection over Union (IoU) serves as a foundational metric for defining the similarity function $S(D_i, T_j)$, it often proves insufficient under challenging conditions such as occlusion. To create a more robust association score, we introduce a multi-cue formulation that synergistically combines three components: standard IoU, RIoU, and a measure of location similarity, as detailed in Equation (2). Through a grid search on the validation set, the optimal weights were determined to be $\lambda_1 = 0.4$, $\lambda_2 = 0.3$, and $\lambda_3 = 0.3$.

$$S(D_i, T_j) = \lambda_1 \text{IoU}(D_i, T_j) + \lambda_2 \text{RIoU}(D_i, T_j) + \lambda_3 S^{loc}(D_i, T_j) \quad (2)$$

$$\lambda_1 + \lambda_2 + \lambda_3 = 1 \quad (3)$$

$IoU(D_i, T_j)$ represents the ratio of the area of overlap between the detection box and the trajectory to the total area, which is defined as

$$IoU = \frac{B_1 \cap B_2}{B_1 \cup B_2} \quad (4)$$

where B_1 and B_2 represent the areas of the detection box and the tracking box, respectively.

To better address occlusion, we introduce a custom metric termed RIoU [38]. Its core idea is that an occluded object's detection box is often smaller than its true size. RIoU simulates a "recovered," non-occluded state by expanding the dimensions of both the detection and trajectory prediction boxes toward a common reference size. This preprocessing allows for the calculation of a more meaningful overlap score, even when partial occlusion results in a low direct IoU, thereby more accurately reflecting the object's true positional relationship.

The calculation of RIoU involves two key steps. The first is determining the reference width and height of the D_i and T_j :

$$\begin{cases} w^r = \max(w^T, w^D) \\ h^r = \max(h^T, h^D) \end{cases} \quad (5)$$

where w^T, h^T and w^D, h^D denote the width and height of the D_i and T_j , respectively.

$$\begin{cases} \hat{w}^T = w^T + \beta(w^r - w^T) \\ \hat{h}^T = h^T + \beta(h^r - h^T) \\ \hat{w}^D = w^D + \beta(w^r - w^D) \\ \hat{h}^D = h^D + \beta(h^r - h^D) \end{cases} \quad (6)$$

The recovery strength is controlled by the hyperparameter β , which adjusts the box dimensions. This parameter dictates the extent of box expansion: $\beta = 0$ corresponds to no recovery (i.e., standard IoU). To determine its optimal value, we conduct systematic ablation studies on the validation set. The results indicate that setting β to 2 strikes an optimal balance between effectively recovering occluded targets and avoiding erroneous associations with nearby objects due to over-expansion. A schematic diagram of RIoU is shown in Figure 3.

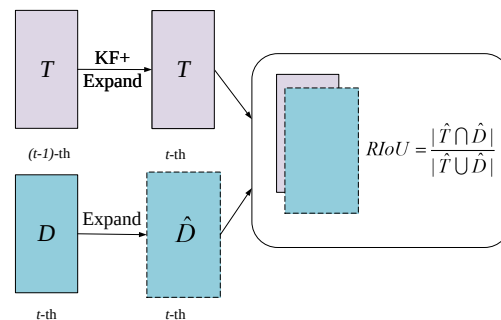


Figure 3. Visualization of the RIoU process.

The standard IoU between the adjusted boxes $\hat{T}$ and $\hat{D}$ is calculated:

$$RIoU = \frac{|\hat{T} \cap \hat{D}|}{|\hat{T} \cup \hat{D}|} \quad (7)$$

S^{loc} is designed to quantify the similarity of adjacent bounding boxes based on the core assumption that objects within the same local topology share similar spatial characteristics.

We construct this feature using the aspect ratios and relative positions of neighboring objects. Specifically, we calculate the weighted average m_i^t of the aspect ratio and distance for target i :

$$m_i^t = \frac{1}{m} \sum |\cos \theta_{i,j}^t| \sigma_j^t \quad (8)$$

where σ_j^t represents the aspect ratio of the neighbor detection box for target i in frame t and denotes the angle from neighbor node j to target i . The final location similarity is then calculated according to Equation (9).

$$S^{loc} = \frac{\min(m_i^t, m_i^{t-1})}{\max(m_i^t, m_i^{t-1})} \quad (9)$$

To ensure the reliability of the spatial context, neighbors are not just selected based on spatial proximity, but a critical filtering step is also included: we exclusively select neighbors from high-confidence detections. In practice, we select the five nearest high-confidence neighbors to construct the location feature. This strategy is designed to mitigate noise at the source, effectively preventing low-confidence false positives from corrupting the topological structure and thereby ensuring the robustness of the S^{loc} feature.

3.3. First-Round Association with GCA

Compared to conventional matching approaches that rely solely on appearance similarity and IoU metrics, graph-based feature embedding produces higher-order discriminative representations that significantly enhance the robustness of detection–trajectory associations. As depicted in Figure 4, a three-stage process is executed in this module: (1) construction of a bipartite graph and initialization of node features; (2) feature enhancement via a multi-head graph attention mechanism; and (3) final association using a robust similarity metric and optimal assignment.

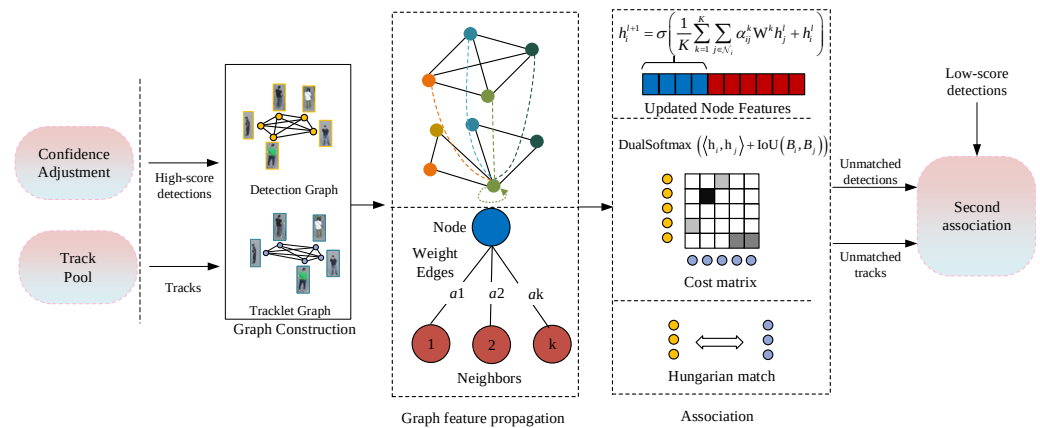


Figure 4. Framework of the First-Round Association using GCA.

3.3.1. Graph Initialization

We separately construct a detection graph $\mathcal{G}_d = (\mathcal{V}_d, \mathcal{E}_d)$ for the high-confidence detection set $\mathcal{D}_{\text{high}}$ and a trajectory graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$ for the active trajectory set $\mathcal{T}t - 1$, employing lightweight and efficient OSNet [39] as the Re-ID backbone network. Through its distinctive omni-scale residual modules, OSNet effectively captures multiscale features spanning fine to coarse granularities. This capability is particularly crucial for recognizing small objects in UAV perspectives. For conciseness, we denote the initial features as $\mathbf{h}_i^0$.

for nodes in $\mathcal{G}_d$ and as $\mathbf{h}_j^0$ for nodes in $\mathcal{G}_t$. Node features are initialized by concatenating topological and appearance features:

$$h_i^{(0)} = f_{\text{topo}}(\text{pos}_i \| \mathbf{l}_i \| \theta_i) + f_{\text{app}}(\text{app}_i) \quad (10)$$

where

$$\text{pos}_i = \left[\frac{x_c}{w_0}, \frac{y_c}{h_0}, \frac{w}{w_0}, \frac{h}{h_0} \right] \quad (11)$$

$$\mathbf{l}_i = [\text{dist}(\mathbf{d}_i, \mathbf{d}_k) / w_0], \quad \mathbf{d}_k \in \mathcal{N}_i \quad (12)$$

$$\theta_i = [\angle(\mathbf{d}_{k-1} \mathbf{d}_i \mathbf{d}_k) / 2\pi], \quad \mathbf{d}_{k-1}, \mathbf{d}_k \in \mathcal{N}_i. \quad (13)$$

$\|$ denotes concatenation, $\angle(\mathbf{d}_{k-1} \mathbf{d}_i \mathbf{d}_k)$ represents the angular relationship between three objects, and $f_{\text{topo}}(\cdot)$ is an MLP (implemented as two-layer MLPs) that encodes topological features in a higher-dimensional space to 256. This representation exhibits scale and rotation invariance, accommodating camera motion in UAV scenarios. $f_{\text{app}}(\cdot)$ projects appearance features in dimensions comparable to topological features, and $\mathcal{N}_i$ denotes the set of neighbors for the target object i . To construct this set, we select the K -nearest neighboring objects (where $K = 5$ in our experiments) based on the Euclidean distance from a pool of high-confidence detections and stable trajectories.

Edge Initialization: The edges between the detection graph nodes and the trajectory graph nodes are constructed under distance constraints, with edge features representing pairwise similarity to the weight Graph Cross-Attention. We employ multi-head attention to capture node relations from both appearance and spatial perspectives. The number of heads is set as 2, and the k th head's attention coefficient e_{ij}^k is computed as

$$e_{ij}^k = \text{LeakyReLU}\left(\omega_k^T \left[W^k h_i \| W^k h_j \right] \right) \quad (14)$$

where ω_k^T denotes learnable parameters, $\|$ denotes concatenation, and W^k denotes shared parameters across graphs.

3.3.2. Cross-Graph Feature Propagation

As illustrated in Figure 4, we employ a Graph Cross-Attention mechanism for feature enhancement between trajectory and detection graphs, allowing detection and trajectory nodes to exchange contextual information. This process enhances the node features by aggregating information from their respective neighbors in the opposing set.

$$\alpha_{ij}^k = \frac{\exp(e_{ij}^k)}{\sum_{m \in \mathcal{N}_i} \exp(e_{im}^k)} \quad (15)$$

Node feature propagation aggregates features via

$$h_i^{l+1} = \sigma \left(\frac{1}{K} \sum_{k=1}^K \sum_{j \in \mathcal{N}_i} \alpha_{ij}^k W^k h_j^l + h_i^l \right) \quad (16)$$

where $\sigma(\cdot)$ is the activation function.

The similarity between the augmented detection node $\mathbf{h}_i$ and the trajectory node $\mathbf{h}_j$ is

$$s_{i,j} = \begin{cases} \text{DualSoftmax}(\langle \mathbf{h}_i, \mathbf{h}_j \rangle + \text{IoU}(B_i, B_j)), & \text{if } d_E(B_i, B_j) < 2(w_i + w_j) \\ 0, & \text{otherwise} \end{cases} \quad (17)$$

where $d_E(B_i, B_j)$ is the distance between B_i and B_j . The dual-softmax operation normalizes the raw score matrix S' along both its rows and columns and computes their geometric

mean. This enforces both row- and column-wise competition, ensuring that a high score for a pair (i, j) indicates that detection i is a strong candidate for trajectory j and that trajectory j is a strong candidate for detection i .

Finally, the matching matrix $\mathbf{M}$ is estimated by the Hungarian algorithm, based on the similarity matrix $\mathbf{S}$.

$$\mathbf{M} = \text{Hungarian}(\mathbf{S}) \quad (18)$$

3.3.3. Training Strategy

The primary objective of the training process is to optimize the GCA module to produce powerful node embeddings that lead to accurate matching. A significant challenge arises due to the Hungarian algorithm, used during inference to obtain the final discrete assignments, being non-differentiable. This property prevents the flow of gradients, making it unsuitable for end-to-end training.

To overcome this, we adopt a strategy inspired by SuperGlue [40], which formulates the assignment as a differentiable optimal transport problem. We utilize the Sinkhorn algorithm, a well-established method for approximating optimal transport, to convert the raw similarity matrix $\mathbf{S}$ into a doubly stochastic matrix. This matrix can be interpreted as a “soft assignment” matrix, where each element represents the probability of a match between a detection and a trajectory. Crucially, the Sinkhorn operation is differentiable, allowing gradients to be back-propagated through it.

$$\bar{\mathbf{S}} = \text{Sinkhorn}(\mathbf{S}) \quad (19)$$

With this differentiable pipeline, we can train the GCA module by supervising its output against the ground-truth matching assignments. Our total loss function is composed of two synergistic components: matching loss to supervise the correctness of the match predictions and Node Loss to enforce the quality of the underlying feature embeddings.

Matching Loss: To address the severe class imbalance inherent in the matching task (where non-matches vastly outnumber true matches), we replace the standard cross-entropy loss with a more powerful combination of Focal Loss and Dice Loss.

The Focal Loss component alleviates class imbalance by down-weighting the loss assigned to well-classified examples, allowing the model to focus on hard-to-classify ambiguous pairs. The Dice Loss component directly optimizes the intersection between predicted and ground-truth matches and is particularly effective at forcing the model to produce higher confidence scores for positive samples. The combined matching loss is defined as

$$L_M = \beta_1 \cdot \frac{1}{N_+} \sum_{e_{i,j} \in E'} (-\alpha(1 - p_t)^\gamma \log(p_t)) + \beta_2 \cdot \left(1 - \frac{2TP + 1}{\Sigma + 1}\right) \quad (20)$$

where p_t is the predicted probability for a given match, N_+ is the number of positive samples, and α and γ are the focusing parameters for the Focal Loss. For the Dice Loss term, TP is the true positive count (sum of probabilities for ground-truth matches) and Σ is the sum of probabilities over all predicted and ground-truth matches.

Node Loss: The goal of Node Loss is to structure the embedding space by ensuring that features from nodes corresponding to the same object are pulled closer together (intra-class compactness), while features from different objects are pushed apart (inter-class separability). We retain the effective margin-based contrastive loss for this purpose:

$$L_N = \sum_{(i,j) \in P} d(h_i, h_j)^2 + \sum_{(i,k) \in N} \max(0, m - d(h_i, h_k))^2 \quad (21)$$

where $d(\cdot, \cdot)$ is the Euclidean distance between the final node embeddings, P is the set of positive pairs (correct matches), N is the set of negative pairs (incorrect matches), and m is the margin hyperparameter.

Total Loss: The final loss function is a weighted sum of the Edge and Node Losses, balanced by a hyperparameter:

$$L_{\text{total}} = \lambda_{\text{loss}} L_M + (1 - \lambda_{\text{loss}}) L_N \quad (22)$$

To ensure a principled approach, the loss weights in Equations (20) and (22) are not set heuristically. Instead, they are determined through a systematic grid search on the VisDrone2019 validation set, with HOTA as the primary optimization metric.

3.3.4. First-Round Association Algorithm

Algorithm 1 delineates our primary matching pipeline, which consists of four main stages. **(1) Graph Initialization:** Given the current frame's high-confidence detections D_{high} and active trajectories from the previous frame, two separate graphs are constructed: a detection graph $\mathcal{G}_d$ and a trajectory graph $\mathcal{G}_t$. The initial feature for each node in both graphs is then computed using Equation (10). **(2) Cross-Graph Feature Propagation:** A multi-head graph attention mechanism is employed to iteratively update and enhance the node features. This allows contextual information to be exchanged between the detection and trajectory graphs. **(3) Similarity Computation:** After feature enhancement, a final similarity matrix S is computed between all detection and trajectory nodes as per Equation (18). **(4) Matching:** The optimal assignments are resolved using the Hungarian algorithm, which outputs the final matched pairs.

Algorithm 1 First-Round Association with Graph Cross-Attention

Require: High-confidence detections $\mathcal{D}_{\text{high}}$; Active trajectories $\mathcal{T}_{t-1}$; Confidence threshold η

Ensure: Matched pairs $\mathcal{M}$; Unmatched detections $\mathcal{U}_d$; Unmatched trajectories $\mathcal{U}_t$

- 1: ▷ Graph Initialization
- 2: Construct detection graph $\mathcal{G}_d(\mathcal{V}_d, \mathcal{E}_d)$ from $\mathcal{D}_{\text{high}}$
- 3: Construct trajectory graph $\mathcal{G}_t(\mathcal{V}_t, \mathcal{E}_t)$ from $\mathcal{T}_{t-1}$
- 4: **for** each node $v_i \in \mathcal{V}_d \cup \mathcal{V}_t$ **do**
- 5: Initialize node feature $\mathbf{h}_i^{(0)}$
- 6: **end for**
- 7: ▷ Cross-Graph Feature Propagation
- 8: **for** $l \leftarrow 0$ **to** $L - 1$ **do**
- 9: Update all node features to $\mathbf{h}^{l+1}$ via cross-graph attention
- 10: **end for**
- 11: ▷ Similarity Computation
- 12: Compute similarity matrix $\mathbf{S}$ between $\mathcal{V}_d$ and $\mathcal{V}_t$
- 13: ▷ Matching
- 14: $\mathcal{M}, \mathcal{U}_d, \mathcal{U}_t \leftarrow \text{Hungarian}(\mathbf{S})$
- 15: **return** $\mathcal{M}, \mathcal{U}_d, \mathcal{U}_t$

3.4. Second-Round Association Based on Reliable Neighbor Guidance

Building upon the reliable matches from the initial round, our framework introduces a key strategic departure from methods like ByteTrack: we retain these matched pairs to serve as contextual anchors during the second-stage association, rather than processing only the unmatched elements. We posit that these existing matches, which exhibit stable trajectory links, provide critical contextual cues for resolving more ambiguous cases. The effectiveness of this principle is empirically validated in our ablation study.

Our tracking method proceeds in four stages. First, we input the state from the previous frame: matched tracks T_{match} and detections D_{match} , unmatched tracks T_{unmatch} and detections D_{unmatch} , and low-confidence detections D_{low} . Second, for each unmatched tracklet, we select its four nearest matched neighbors based on proximity, velocity, and direction. Third, we search the predicted area using IoU matching and topological feature similarity. Finally, during track management, new tracks are initiated from remaining high-confidence detections and low-confidence ones are discarded. Tracks that remain unmatched for 30 consecutive frames are terminated. These stages are described in detail in the following sections.

3.4.1. Spatio-Temporally Constrained Neighbor Selection

Intuitively, neighboring objects (e.g., pedestrian groups or vehicle clusters) exhibit coherent motion patterns. We therefore select adjacent trajectories for each unmatched trajectory by enforcing three constraints: distance thresholds, velocity consistency, and directional alignment.

Distance Constraint: We let (x_i, y_i) and (x_j, y_j) denote the center positions of the target trajectory and candidate neighbor trajectory, respectively, with w_i representing the bounding box width. For high-confidence matches established during initial association, the detection box size inherently encodes camera perspective information and implicitly conveys target depth. Consequently, a normalized Euclidean distance metric is constructed as follows:

$$S_{i,j}^d = \frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{w_i} \quad (23)$$

Directional Constraint: Adjacent trajectories inherently exhibit compatible motion patterns with the target. We thus employ a cosine similarity-based distance metric to select trajectories with directional consistency:

$$S_{i,j}^{\text{dir}} = 1 - \frac{\Delta x_t^i \cdot \Delta x_t^j + \Delta y_t^i \cdot \Delta y_t^j}{\sqrt{(\Delta x_t^i)^2 + (\Delta y_t^i)^2} \cdot \sqrt{(\Delta x_t^j)^2 + (\Delta y_t^j)^2}} \quad (24)$$

where $(\Delta x_t^i, \Delta y_t^i)$ and $(\Delta x_t^j, \Delta y_t^j)$ denote the velocity vectors of the trajectories T_i and T_j , respectively. A smaller $S_{i,j}^{\text{dir}}$ value indicates a higher degree of directional alignment.

Velocity Constraint: Within correlated groups, object velocities are constrained by neighboring objects, resulting in comparable motion magnitudes. The velocity similarity $S_{i,j}^v$ between trajectory T_i and T_j is consequently quantified as

$$S_{i,j}^v = \left(\Delta x_t^i - \Delta x_t^j \right)^2 + \left(\Delta y_t^i - \Delta y_t^j \right)^2 \quad (25)$$

A smaller $S_{i,j}^v$ indicates a higher probability of cohesive group movement.

The similarity between matched trajectories T_j and target trajectories T_i in the current frame is ultimately calculated as

$$S(T_i, T_j) = S_{i,j}^d + S_{i,j}^{\text{dir}} + S_{i,j}^v \quad (26)$$

The matched trajectories are ranked based on this composite score, and the top four trajectories with the lowest aggregate score are selected as the reliable neighbors for T_i . Subsequently, matching attempts are performed between the current frame's unmatched detections and the unmatched trajectory to form temporary pairs. The formation of a valid match pair is determined by comparing their group similarity.

3.4.2. Star-Shaped Topological Feature Description

As illustrated in Figure 5, in this stage, a robust, context-aware representation is constructed for each remaining object by introducing a directed graph topology. A key distinction from the first association round is our strategy to leverage the reliably matched pairs as contextual anchors. Inspired by the work of You et al. [35], we find that topological features, which characterize the precise geometric relationships between a target and its neighboring nodes, provide a highly discriminative representation. This approach is particularly effective for distinguishing between objects that exhibit similar appearances and positions.

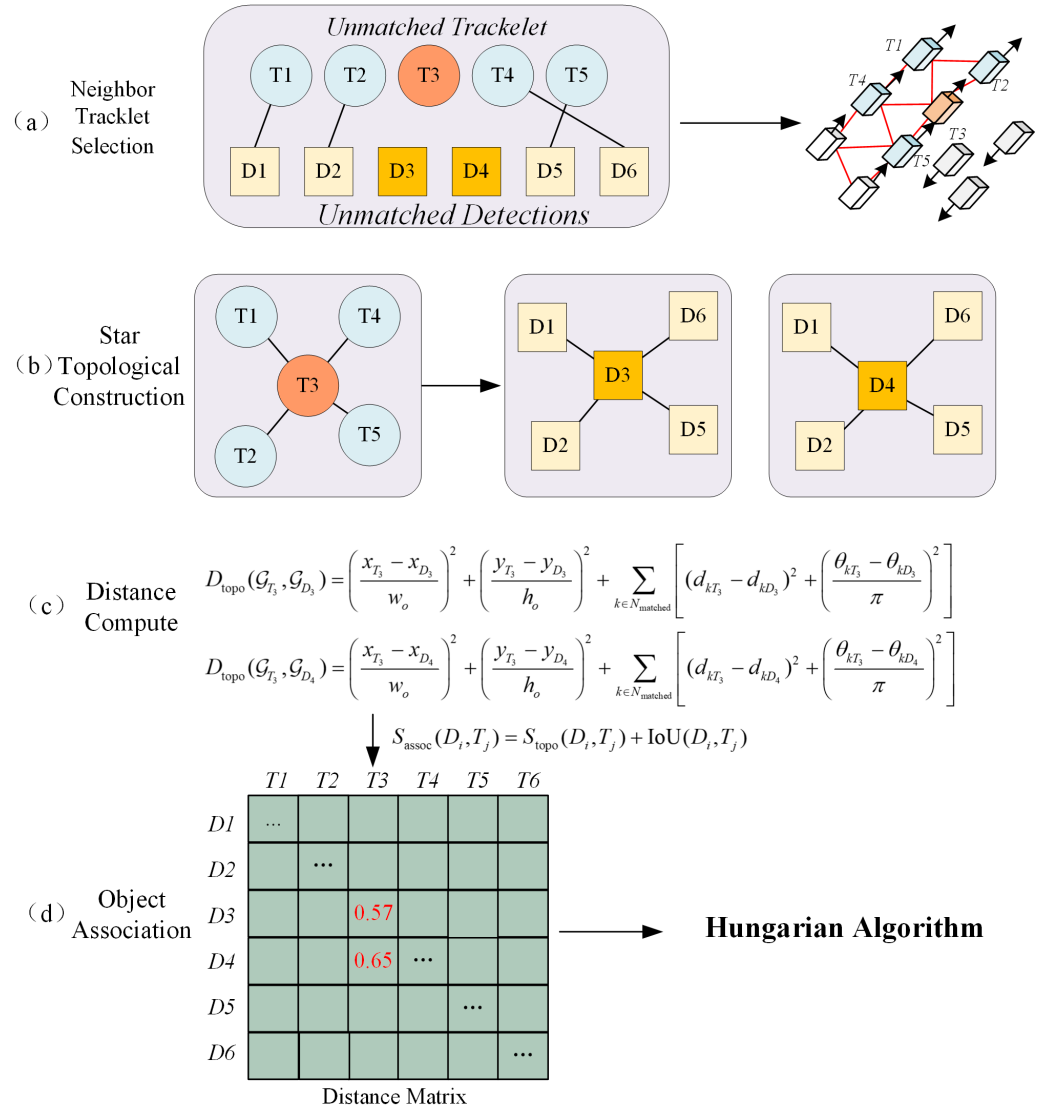


Figure 5. Star-shaped topological feature description.

To implement this strategy, we formally define a star-shaped topological graph for each unmatched trajectory T_j denoted by G_{T_j} and for each unmatched detection D_i denoted by G_{D_i} . Both graphs are defined by their respective central node (the unmatched object) and a shared set of reliable matched neighbors N_{matched} . By comparing the structure of these two graphs, we can measure the similarity of the unmatched objects based on their relationship with the same stable context. The topological distance between these two subgraphs is then calculated using Equation (27).

$$D_{\text{topo}}(\mathcal{G}_{T_j}, \mathcal{G}_{D_i}) = \left(\frac{x_j - x_i}{w_o} \right)^2 + \left(\frac{y_j - y_i}{h_o} \right)^2 + \sum_{k \in N_{\text{matched}}} \left[(d_{kj} - d_{ki})^2 + \left(\frac{\theta_{kj} - \theta_{ki}}{\pi} \right)^2 \right] \quad (27)$$

The topological similarity $S_{\text{topo}}(D_i, T_j)$ between the detection D_i and tracking boxes T_j is normalized using Equation (28):

$$S_{\text{topo}}(D_i, T_j) = \exp\left(-D_{\text{topo}}(\mathcal{G}_{T_j}, \mathcal{G}_{D_i})\right) \quad (28)$$

Subsequently, Hungarian matching is performed between residual detections and trajectories using $S_{\text{topo}}(D_i, T_j)$.

$$S_{\text{assoc}}(D_i, T_j) = S_{\text{topo}}(D_i, T_j) + \text{IoU}(D_i, T_j) \quad (29)$$

3.4.3. Second-Round Association Algorithm

As shown in Algorithm 2, given initial matched pairs and unmatched detections/trajectories, we first select reliable neighbors around unmatched detections using velocity–spatial consistency constraints. Star-shaped topologies are then constructed centering on unmatched detections and trajectories. Topological similarity metrics compare unmatched trajectories, resolved via Hungarian matching. Unlike conventional two-stage methods, this approach leverages prior matched knowledge and physical geometric constraints to reduce identity switches, ensuring reliable tracking continuity in dynamic UAV videos. The lightweight design does not require additional networks, guaranteeing deployability on UAV platforms. Unmatched detections initialize new trajectories in $\mathcal{T}^t$, while unmatched trajectories are marked as inactive and added to $\mathcal{T}^t$. Using Kalman filtering, we predict their positions for 30 frames, purging persistently unassociated trajectories.

Algorithm 2 Reliable Neighbor-Guided Second Association

Require: $\mathcal{M}_1$: Matched pairs from first round $\mathcal{U}_d$: Unmatched detections $\mathcal{U}_t$: Unmatched tracks

Ensure: $\mathcal{M}_2$: Matched pairs from second round $\mathcal{U}'_d, \mathcal{U}'_t$: Remaining unmatched sets

```

1:  $\mathcal{M}_2 \leftarrow \emptyset$ 
2: Initialize similarity matrix  $\mathbf{S}_{\text{assoc}}$ 
3: for each  $T_j \in \mathcal{U}_t$  do                                     ▷ Select reliable neighbors
4:   Compute  $S(T_j, T_k)$  for all  $T_k \in \mathcal{M}_1$ 
5:    $\mathcal{N}_j \leftarrow$  Top-4 matched neighbors for  $T_j$ 
6: end for
7: for each pair  $(T_j, D_i) \in \mathcal{U}_t \times \mathcal{U}_d$  do                 ▷ Compute association similarity
8:   Construct subgraphs  $\mathcal{G}_{T_j}$  and  $\mathcal{G}_{D_i}$  using  $\mathcal{N}_j$ 
9:   Compute  $\mathbf{S}_{\text{assoc}}(j, i)$  using Equations (27)–(29)
10: end for
11:                                     ▷ Perform matching
12:  $\mathcal{M}_2, \mathcal{U}'_d, \mathcal{U}'_t \leftarrow \text{Hungarian}(\mathbf{S}_{\text{assoc}})$ 
13:                                     ▷ Track Management
14: for each  $T_j \in \mathcal{U}'_t$  do
15:   Update  $T_j$  state using motion from  $\mathcal{N}_j$ 
16:   Terminate if  $\text{age}(T_j) > 30$ 
17: end for
18: return  $\mathcal{M}_2, \mathcal{U}'_d, \mathcal{U}'_t$ 

```

4. Experiments

4.1. Implementation Details

4.1.1. Datasets and Metrics

To verify the effectiveness of the proposed method, experiments were conducted on two commonly used unmanned aerial vehicle (UAV) multi-target tracking datasets: VisDrone2019 [41] and UAVDT [42]. VisDrone2019 contains 56 training video sequences, 7 validation sequences, and 17 test evaluation sequences. These videos cover various scenes, including sports stadiums, commercial streets, highways, and suburban areas. When evaluating multi-target tracking tasks, the official evaluation tool for this dataset only considers five object types: people, cars, vans, trucks and buses. In contrast, UAVDT only tracks a single target category: vehicles. This dataset contains 50 sequences (30 for training and 20 for testing), mainly covering scenarios such as squares, intersections, and highways under different lighting conditions (such as sunny, nighttime, and foggy). It should be noted that we refer to the true value labels provided by SFtrack [22] for the revised test set of UAVDT.

This study features common multi-object tracking evaluation criteria used in previous work. The evaluation metrics are as follows: high-order tracking accuracy (HOTA), multi-target tracking accuracy (MOTA), IDF1, false positive (FN), false positive (FP), and ID switching (IDs). HOTA is a high-order metric used to measure tracking performance, and can better reflect the performance of tracking algorithms, while MOTA is a widely used MOT evaluation index, mainly related to detection performance (FP and FN) and ID stability.

4.1.2. Training Details

We use PyTorch Geometric 3.12 to construct graph neural networks for NOWA-MOT and train it on one NVIDIA RTX 4090 GPU for 20 epochs, with a batch size of 1. The Adam optimizer is applied with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the initial learning rate is 3×10^{-5} and weight decay is 1×10^{-5} . For the Re-ID module, we use an OSNet network for appearance feature extraction and pre-trained it on VisDrone and UAVDT datasets. The normalized crop size for the Re-ID module is set as (256, 128).

4.1.3. Inference Details

Due to the TBD structure of the proposed algorithm, the detection performance was adopted as the primary evaluation metric. For fairness, YoloX-m was retrained as the detector on both VisDrone and UAVDT datasets, with identical detection results provided to all comparison algorithms. Detection score thresholds were set to 0.6 (high) and 0.1 (low) and motion compensation (CMC) was employed to enhance tracking. Following the robust configuration established in BoT-SORT, we utilized the Enhanced Correlation Coefficient (ECC) algorithm to register images between consecutive frames. Specifically, the ECC method estimates the optimal affine transformation matrix that aligns the previous frame with the current one. This transformation is then applied to the trajectory locations predicted by the Kalman filter, effectively warping them into the current frame's coordinate system before the association step. This process significantly mitigates the tracking errors caused by camera motion rather than true object movement. Target trajectories are initialized directly from the first frame's detections. For subsequent frames, the tracker assigns detections to corresponding trajectories.

4.2. Comparison with State of the Art

NOWA-MOT was compared with other multi-object tracking (MOT) methods on UAVDT and VisDrone2019-MOT datasets. MOT methods fall into two paradigms: tracking by detection (TBD) (e.g., ByteTrack [8], BoT-SORT [6], UAVMOT [43]) and joint detection and tracking (JDT) (e.g., FairMOT [44], DroneMOT [45], GLOA [46]). Since the detector

performance significantly impacts MOT metrics, identical detector weights were used for TBD methods on the same dataset. For JDT methods, retraining was performed on trackers with open-source code; results from original papers were adopted for non-open-source methods. The OSNet appearance feature extraction network was utilized for trackers requiring appearance features.

4.2.1. Results on UAVDT

NOWA-MOT was trained on the UAVDT training set and evaluated on its test set. As shown in Table 1, NOWA-MOT achieves superior performance in terms of HOTA, MOTA, and IDF1, reaching 62.69%, 69.01%, and 80.58%, indicating significant improvements in tracking accuracy and identity consistency. Compared with the two-stage tracker ByteTrack, NOWA-MOT substantially reduced ID switching (IDs: 305 $\rightarrow$ 131) and false positives (FPs: 43,056 $\rightarrow$ 36,098), achieving a 1.44% higher HOTA and demonstrating its ability to enhance tracking accuracy via multi-cue fusion. Compared with the UAV-specific tracker UAVMOT, NOWA-MOT achieved 2.5% and 8.6% gains in MOTA and IDF1, respectively. These results demonstrate that NOWA-MOT efficiently maintains tracking accuracy and identity consistency under non-linear UAV motion, outperforming competing methods.

Table 1. Comparison with state-of-the-art online MOT methods on UAVDT. Metrics with $\uparrow$ ($\downarrow$) indicate higher (lower) values are better. The best results are indicated in bold. Results marked with * are taken from the original paper. This convention applies to all subsequent tables.

Method	HOTA $\uparrow$	MOTA $\uparrow$	IDF1 $\uparrow$	MOTP $\uparrow$	MT $\uparrow$	ML $\downarrow$	FP $\downarrow$	FN $\downarrow$	IDs $\downarrow$
SORT [47]	41.24	38.99	43.71	74.28	493	410	33,037	172,628	2350
DeepSORT[17]	61.48	68.45	78.38	80.03	817	114	36,670	70,044	850
ByteTrack [8]	61.25	67.19	78.86	79.93	830	109	43,056	68,509	305
UAVMOT [43]	57.74	66.51	71.98	80.04	822	113	43,191	68,869	2100
StrongSORT [7]	61.37	68.32	78.17	80.12	808	113	36,517	70,485	985
FLOT * [48]	—	48.5	68.3	80.1	—	—	24,105	107,630	800
DroneMOT * [45]	—	50.10	69.60	74.50	638	178	57,411	112,548	129
GLOA * [46]	—	49.60	68.90	79.80	626	220	55,822	115,567	433
BOT-SORT [6]	61.99	67.86	79.73	80.45	763	127	31,035	78,439	184
SuperMOT * [49]	—	50.60	70.60	75.70	674	227	60,587	107,538	407
MM-tracker [50]	58.87	62.48	76.13	79.99	723	168	31,649	95,958	295
Tracktrack [51]	57.62	58.16	70.96	77.68	723	168	30,547	101,132	781
Ours	62.69	69.01	80.58	80.13	814	114	36,098	69,407	131

4.2.2. Results on VisDrone

NOWA-MOT was trained on the VisDrone2019 training set and evaluated on its test development set, with the metrics, including HOTA, MOTA, IDF1, and IDs, compared against those of state-of-the-art methods. As shown in Table 2, NOWA-MOT achieved an HOTA of 51.34% and an IDF1 of 67.33%, surpassing all existing methods. VisDrone contains five target categories, with numerous small, irregularly moving objects (e.g., pedestrians, non-motor vehicles), leading to more frequent trajectory switches and association failures than for UAVDT. Compared with ByteTrack, the HOTA and MOTA of NOWA-MOT were improved by 3.96% and 4.79%, respectively, and compared to BOT-SORT, its HOTA and IDF1 were increased by 1.05% and 1.61%, respectively. It also achieved the highest IDF1 (highlighting robust identity preservation) and the lowest IDs (trajectory switches), demonstrating effectiveness in generating coherent trajectories—especially in complex environments. Compared to Unigraph (which also uses topological features), the HOTA and MOTA of NOWA-MOT were improved by 3.68% and 2.86%, respectively. These results

confirm NOWA-MOT's significant advantages in balancing detection accuracy and identity preservation in challenging UAV tracking scenarios.

Table 2. Comparison with state-of-the-art online MOT methods on VisDrone2019.

Method	HOTA ↑	MOTA ↑	IDF1 ↑	MOTP ↑	MT ↑	ML ↓	FP ↓	FN ↓	IDs ↓
SORT [47]	33.29	32.18	21.61	68.51	318	511	64,548	85,453	5728
DeepSORT [17]	45.75	43.27	55.91	75.06	849	249	50,118	60,436	4616
ByteTrack [8]	47.38	45.68	60.37	74.89	872	285	45,211	62,800	2274
UAVMOT [43]	39.43	36.10	51.00	74.20	520	574	27,983	115,925	2775
StrongSORT [7]	46.03	43.16	56.40	75.07	846	251	49,716	60,672	5013
Unigraph [21]	47.66	47.61	60.01	75.62	895	290	42,527	61,532	2306
FOLT * [48]	–	42.1	56.9	77.6	–	–	24,105	107,630	800
DroneMOT* [45]	–	43.7	58.6	71.4	689	397	41,998	86,177	1112
GLOA * [46]	–	39.1	46.2	76.1	581	824	18,715	158,043	4426
BOT-SORT [6]	50.29	49.55	65.72	76.90	716	471	14,260	88,377	483
SuperMOT * [49]	–	51.70	66.70	77.2	892	407	30528	79579	1105
MM-tracker [50]	50.54	50.61	65.59	75.12	715	443	15,934	83,781	558
Tracktrack [51]	50.75	51.54	67.03	75.86	810	337	27,608	69,558	1202
Ours	51.34	50.47	67.33	76.40	883	289	23,239	76,524	793

4.3. Ablation Studies

In this section, we demonstrate the importance of each module and design strategy in NOWA-MOT through ablation experiments. Subsequently, we explore the sensitivity of the parameters to graph initialization and association. All models in this section were trained on the VisDrone2019 training set and evaluated on the VisDrone2019 test-dev set.

4.3.1. Effectiveness of Each Module

Using ByteTrack as the baseline (first row), we evaluated the effectiveness of each module; the results are shown in Table 3.

Table 3. Ablation study for each module of NOWA-MOT.

LOR	GCA	MNG	HOTA ↑	MOTA ↑	IDF1 ↑	IDs ↓	FPS ↑
			47.38	45.68	60.37	2274	37.25
✓			47.68	47.53	59.89	2140	35.44
	✓		49.62	48.63	61.81	1105	21.71
		✓	48.13	47.56	61.84	1564	34.09
✓	✓		49.97	49.23	65.87	862	19.38
✓		✓	48.34	47.63	62.25	1477	32.68
	✓	✓	50.88	49.56	65.88	937	19.76
✓	✓	✓	51.34	50.47	67.33	793	18.05

LOR: Without LOR, detections with confidence below 0.6 are excluded from the first matching step, as in ByteTrack. Incorporating LOR improves the HOTA and MOTA by 0.3% and 1.85%, respectively. This is because in low-altitude scenarios, the confidence scores fluctuate significantly due to occlusion and motion. Directly excluding low-confidence detections may lead to erroneous matches in crowded scenes. However, IDF1 slightly decreases by 0.48%, indicating that the module may introduce false positives (FPs) and cause track fragmentation. When combined with GCA, the performance is further improved, with HOTA, MOTA, and IDF1 increasing by 2.59%, 3.55%, and 5.5%, respectively, over the baseline. The combination of LOR and MNG also brings gains, though less pronounced, suggesting functional overlap in handling low-confidence and occluded objects.

GCA: GCA improves the HOTA and MOTA by 2.24% and 2.95%, respectively, demonstrating that graph convolution with spatio-temporal feature enhancement effectively aggregates neighborhood information and improves feature discriminability. The IDF1 also increases by 1.44%, indicating that topological features provide more stable association in dynamic drone scenarios. The combination of GCA and MNG yields the second best performance, highlighting their complementarity.

MNG: Using MNG alone improves the HOTA and IDF1 by 0.75% and 1.47%, respectively, showing its effectiveness in improving second-stage matching and trajectory stability.

The ablation study demonstrates that the GCA module is the primary computational bottleneck, reducing the FPS from 37.25 to 21.71. This is attributed to its computationally intensive graph convolution and cross-attention mechanisms. In contrast, the LOR and MNG modules are highly efficient, introducing minimal overhead as they rely on lightweight geometric calculations; MNG's efficiency is further enhanced by reusing cached motion variables.

These results demonstrate the importance of each module in enhancing tracking performance and maintaining identity consistency.

4.3.2. Composition Analysis of the LOR Module

To further analyze the contribution of each component in LOR, we conducted fine-grained experiments, the results of which are shown in Table 4. Adding RIoU to IoU improves the MOTA by 0.44% but decreases the IDF1 by 1.2%, suggesting increased fragmentation. Incorporating S^{loc} leads to improvements of 0.63% in HOTA, 1.41% in MOTA, and 0.73% in IDF1 (compared to the second row). Overall, the results indicate that combining multiple similarity metrics improves tracking, while RIoU should be carefully integrated to avoid adverse effects.

Table 4. Ablation study on the components of the LOR module.

IoU	RIoU	S^{loc}	HOTA ↑	MOTA ↑	IDF1 ↑	IDs ↓
✓			47.38	45.68	60.37	2274
✓	✓		47.05	46.12	59.16	2377
✓	✓	✓	47.68	47.53	59.89	2140

4.3.3. Analysis of Message Passing Steps

The number of graph neural network layers (L) impacts feature aggregation: too few layers limit neighborhood aggregation, while too many leads to over-smoothing. We tested the performance with different numbers of layers, as shown in Table 5. The results in Table 5 show that the model performs best when $L = 2$, achieving a balance between aggregating contextual semantic information and avoiding feature over-smoothing.

Table 5. Analysis of the impact of different numbers of message passing layers.

L	HOTA ↑	MOTA ↑	IDF1 ↑
1	45.54	47.56	57.78
2	49.62	48.63	61.81
3	46.17	48.24	58.48

4.3.4. Impact of Data Augmentation on the Graph Network Performance.

Data augmentation exerts a significant impact on the graph network. As shown in the table, we incorporate data augmentation during the training of the graph neural network. To simulate the input of a real detector, in our approach, detector outputs are compared

with the ground truth, Hungarian matching based on IoU is applied, and ground truth IDs are assigned to the corresponding detector results. This process generates a detector-based training set that more closely resembles the actual output distribution of the detector. The results are shown in Table 6 and show that our method exhibits improvements in the HOTA and MOTA of 1.33% and 1.47%, respectively.

Table 6. Impact of data augmentation on the graph network performance.

Augmentation	HOTA ↑	MOTA ↑	IDF1 ↑
w/o	48.29	47.16	60.01
w/	49.62	48.63	61.81

4.3.5. Impact of Spatio-Temporal Geometric Constraints on Secondary Association

To validate the necessity of the spatio-temporal geometric constraints in MNG, we compare our method with a nearest neighbor approach. As shown in Table 7, the HOTA and IDF1 decrease by 0.89% and 3.48%, respectively, without constraints. These constraints prevent the selection of inversely moving targets, which is crucial for accurately computing the topological similarity.

Table 7. Ablation study on spatio-temporal geometric constraints in secondary association.

Constraints	HOTA ↑	MOTA ↑	IDF1 ↑
Nearest Neighbor	47.24	47.54	58.36
Proposed (w/ Constraints)	48.13	47.56	61.84

4.3.6. Computational Efficiency Analysis

To evaluate the computational efficiency of our proposed framework, we compare its inference speed (Frames Per Second, FPS) against several representative state-of-the-art trackers. As shown in Table 8, our NOWA-MOT achieves the highest tracking accuracy with a HOTA score of 51.34, surpassing all compared methods.

Table 8. Algorithm speed ablation study.

Methods	HOTA ↑	MOTA ↑	IDF1 ↑	FPS ↑
ByteTrack [8]	47.38	45.68	60.37	37.25
BOT-SORT [6]	50.29	49.55	65.72	24.16
MM-tracker [50]	50.54	50.61	65.59	31.89
NOWA-MOT	51.34	50.47	67.33	18.05

This superior performance is achieved with an inference speed of 18.05 FPS. While trackers like ByteTrack (37.25 FPS) and MM-tracker (31.89 FPS) offer faster processing speeds, they exhibit lower tracking accuracy. Specifically, compared to the high-speed baseline ByteTrack, NOWA-MOT provides a substantial 3.96% absolute improvement in HOTA. Although BOT-SORT and MM-tracker are closer in accuracy, our method still holds a clear advantage. This analysis highlights a deliberate trade-off in our design, prioritizing tracking robustness and accuracy over raw speed. The resulting inference rate remains practical for many real-world deployment scenarios where high performance is the primary concern.

4.3.7. Parameter Sensitivity Analysis

We conducted a hyperparameter sensitivity analysis to evaluate the impact of the number of neighbors (N) on the performance of the GCA and MNG modules, with the

results presented in Table 9. For the GCA module, we varied N from 3 to 9. The performance, measured by HOTA, peaks at our chosen value of $N = 5$ (51.34 HOTA). Using fewer neighbors ($N = 3$) provides insufficient contextual information, leading to a drop in accuracy, while using too many ($N = 9$) introduces noise from irrelevant objects, which also degrades performance. This confirms that $N = 5$ offers the best balance of contextual richness and noise for feature enhancement.

Similarly, for the MNG module, we tested N from 2 to 8. The model achieves optimal performance at $N = 4$, which provides the most stable set of contextual anchors for resolving ambiguous associations. Using too few neighbors ($N = 2$) makes the topological feature less reliable, while using too many ($N = 6$ or 8) can incorporate misleading motion patterns from more distant objects, harming accuracy. This analysis validates our choice of hyperparameters and demonstrates that the model's performance is robust around these optimal values.

Table 9. Hyperparameter sensitivity analysis for the number of neighbors (N) in GCA and MNG modules on the VisDrone dataset.

Module	N	HOTA $\uparrow$	MOTA $\uparrow$	IDF1 $\uparrow$
GCA	3	49.13	50.08	66.53
	5	51.34	50.47	67.33
	7	50.72	50.35	66.84
	9	47.29	48.02	65.21
MNG	2	49.51	48.24	62.75
	4	51.34	50.47	67.33
	6	49.64	49.39	64.91
	8	48.42	49.11	63.48

4.4. Qualitative Comparisons with State-of-the-Art Trackers

Figure 6 presents a qualitative comparison between our method and other state-of-the-art trackers. Red circles indicate the occurrence of ID switches, while green circles denote missed detections. In this video sequence, the upper part of the image contains dense small objects with similar appearances, while targets moving around the roundabout exhibit significant non-linear motion.

As observed, both ByteTrack and UAVMOT exhibit a noticeable number of ID switches, whereas our method maintains correct identities. This is because motion-based features often fail to meet the matching threshold under non-linear motion. Moreover, due to the small size of the targets, the cosine similarity of appearance features also falls short of the matching threshold, leading the tracker to mistakenly treat them as new objects. In contrast, our approach exhibits a relatively stable topological similarity. Based on the knowledge that objects do not appear out of nowhere, we associate targets with similar neighborhood contexts, thus achieving more robust data association.

Multiple trackers also exhibited missed detections of small objects. This occurred because it is difficult to associate detection boxes with existing tracks or initialize them as new trajectories due to occlusions and small object sizes. Our method addresses this issue through the LOR module, which enhances the confidence scores of low-confidence detections using contextual cues, allowing them to be associated with existing tracks or initialized as new ones.



Figure 6. Qualitative results of UGT. The frames in this figure are from uav0000355_00001_v in the VisDrone2019-MOT-test-dev set.

4.5. Discussion

NOWA-MOT demonstrates good adaptability to the specific requirements of multi-object tracking in low-altitude scenarios; however, there are areas that warrant improvement. In particular, a targeted long-term occlusion strategy was not developed. This decision was made based on the enhanced field of view at low altitudes, where prolonged occlusion between objects is less frequent. Consequently, specific enhancements to streamline the algorithm and enhance its real-time performance have not been pursued. Moreover, our approach employs feature aggregation through graph neural networks in the second step of trajectory association. While graph neural networks are effective in capturing point relationships, the interactions between trajectories are intricate and diverse. Future enhancements could involve utilizing mechanisms like hypergraphs to link multiple nodes through hyperedges, facilitating exploration of feature transfers within groups and enhancing motion information extraction. This approach could lead to more precise trajectory predictions and enable progressive matching strategies from the group to the individual level using hypergraph matching methods. Furthermore, it is important to note that our methodology relies on the YOLOX detector. Should a different detector be employed, adjustments to certain hyperparameters may be necessary to align with the detector's performance.

Our approach is limited by its core assumption that topological neighbors share similar spatial features. This assumption fails in complex scenarios such as “stereo traffic” overpasses, where vehicles on different levels appear adjacent in the 2D projection but are in reality independent. As shown in Figure 7, other challenging cases include low-light conditions that degrade detection, scenes with highly non-linear motion and occlusion like

basketball courts, and situations with too few observable neighbors. These conditions can lead to tracking failures, evidenced by an increase in ID switches. Addressing these issues requires future work, such as incorporating targeted image enhancement and leveraging visual-language models to integrate scene priors.



Figure 7. Examples of failure cases. IDSW indicates that the target is assigned a different identity compared to its track in the previous frame. Each target with IDSW is annotated accordingly.

5. Conclusions

This paper introduces a novel multi-object tracking algorithm to address challenges that arise due to a low-altitude perspective, such as small target sizes, similar appearances, and frequent mutual occlusion. This method leverages multiple cues to recover low-confidence occluded objects. Subsequently, Graph Cross-Attention is applied to enhance the features of both detection and trajectory nodes, enabling first-stage matching based on the similarity between detection and trajectory nodes. Furthermore, a novel cascaded second-stage matching strategy is introduced. Unlike previous, relatively isolated cascaded matching methods, the proposed approach utilizes already matched detection–trajectory pairs to identify the neighboring trajectories of unmatched ones based on motion characteristics such as velocity and direction. These neighboring trajectories are then used to construct star-shaped topological features, leading to more robust matching results. Thanks to these improvements, NOWA-MOT’s HOTA on the VisDrone test-dev set reached 51.34%, an increase of 3.96% compared to Bytetrack’s 47.38%.

Author Contributions: Conceptualization, H.Q.; methodology, H.Q.; software, H.Q. and X.G.; validation, H.Q., R.G. and B.D.; formal analysis, H.Q.; investigation, S.S.; resources, S.S.; data curation, H.Q.; writing—original draft preparation, H.Q.; writing—review and editing, H.Q. and X.S.; visualization, H.Q. and B.D.; supervision, H.Q.; project administration, H.Q.; funding acquisition, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Graduate innovation program of National University of Defense Technology, grant number: XJZH2024016.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Guo, Y.; Li, Y.; Wang, S.; Sun, K.; Liu, M.; Wang, Z. Pedestrian multi-object tracking combining appearance and spatial characteristics. *Expert Syst. Appl.* **2025**, *272*, 126772. [\[CrossRef\]](#)
- Li, M.; Zhang, Y.; Jia, Y.; Yang, Y. Advancing multi-object tracking through occlusion-awareness and trajectory optimization. *Knowl.-Based Syst.* **2025**, *310*, 112930. [\[CrossRef\]](#)
- Liu, Y.; Liu, X.; Jiang, Z.; Liu, J. Co-MOT: Exploring the Collaborative Relations in Traffic Flow for 3D Multi-Object Tracking. *IEEE Trans. Intell. Transport. Syst.* **2025**, *26*, 4744–4756. [\[CrossRef\]](#)
- Seidenschwarz, J.; Brasó, G.; Serrano, V.C.; Elezi, I.; Leal-Taixé, L. Simple Cues Lead to a Strong Multi-Object Tracker. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 13813–13823. [\[CrossRef\]](#)
- Tang, Z.; Naphade, M.; Liu, M.Y.; Yang, X.; Birchfield, S.; Wang, S.; Kumar, R.; Anastasiu, D.; Hwang, J.N. CityFlow: A City-Scale Benchmark for Multi-Target Multi-Camera Vehicle Tracking and Re-Identification. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 8789–8798. [\[CrossRef\]](#)
- Aharon, N.; Orfaig, R.; Bobrovsky, B.Z. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv* **2022**, arXiv:2206.14651.
- Du, Y.; Zhao, Z.; Song, Y.; Zhao, Y.; Su, F.; Gong, T.; Meng, H. StrongSORT: Make DeepSORT Great Again. *IEEE Trans. Multimed.* **2023**, *25*, 8725–8737. [\[CrossRef\]](#)
- Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. ByteTrack: Multi-object Tracking by Associating Every Detection Box. In Proceedings of the Computer Vision—ECCV 2022, Tel Aviv, Israel, 23–27 October 2022; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer: Cham, Switzerland, 2022; pp. 1–21. [\[CrossRef\]](#)
- Sun, Z.; Wei, G.; Fu, W.; Ye, M.; Jiang, K.; Liang, C.; Zhu, T.; He, T.; Mukherjee, M. Multiple Pedestrian Tracking Under Occlusion: A Survey and Outlook. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 1009–1027. [\[CrossRef\]](#)
- Zhang, Y.; Liang, Y.; Leng, J.; Wang, Z. SCGTracker: Spatio-temporal correlation and graph neural networks for multiple object tracking. *Pattern Recognit.* **2024**, *149*, 110249. [\[CrossRef\]](#)
- Qin, Z.; Zhou, S.; Wang, L.; Duan, J.; Hua, G.; Tang, W. MotionTrack: Learning Robust Short-Term and Long-Term Motions for Multi-Object Tracking. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 17939–17948. [\[CrossRef\]](#)
- Xu, X.; Ren, W.; Sun, G.; Ji, H.; Gao, Y.; Liu, H. GroupTrack: Multi-Object Tracking by Using Group Motion Patterns. In Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 14–18 October 2024; pp. 4896–4903. [\[CrossRef\]](#)
- Liu, S.; Shen, X.; Xiao, S.; Li, H.; Tao, H. A Multi-Scale Feature-Fusion Multi-Object Tracking Algorithm for Scale-Variant Vehicle Tracking in UAV Videos. *Remote Sens.* **2025**, *17*, 1014. [\[CrossRef\]](#)
- Fu, H.; Guan, J.; Jing, F.; Wang, C.; Ma, H. A real-time multi-vehicle tracking framework in intelligent vehicular networks. *China Commun.* **2021**, *18*, 89–99. [\[CrossRef\]](#)
- Cao, J.; Pang, J.; Weng, X.; Khiroudkar, R.; Kitani, K. Observation-centric sort: Rethinking sort for robust multi-object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 9686–9696.
- Bochinski, E.; Eiselein, V.; Sikora, T. High-Speed tracking-by-detection without using image information. In Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; pp. 1–6. [\[CrossRef\]](#)
- Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649. [\[CrossRef\]](#)
- Zhang, J.; Wang, M.; Jiang, H.; Zhang, X.; Yan, C.; Zeng, D. STAT: Multi-Object Tracking Based on Spatio-Temporal Topological Constraints. *IEEE Trans. Multimed.* **2024**, *26*, 4445–4457. [\[CrossRef\]](#)
- Yang, M.; Han, G.; Yan, B.; Zhang, W.; Qi, J.; Lu, H.; Wang, D. Hybrid-sort: Weak cues matter for online multi-object tracking. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 6504–6512.
- Chen, Y.; Wang, J.; Zhou, Q.; Hu, H. ArbiTrack: A Novel Multi-Object Tracking Framework for a moving UAV to Detect and Track Arbitrarily Oriented Targets. *IEEE Trans. Multimed.* **2025**, *27*, 5387–5397. [\[CrossRef\]](#)
- Deng, C.; Wu, J.; Han, Y.; Wang, W.; Chanussot, J. Learning a robust topological relationship for online multi-object tracking in uav scenarios. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5628615. [\[CrossRef\]](#)
- Song, I.; Lee, J. SFTrack: A Robust Scale and Motion Adaptive Algorithm for Tracking Small and Fast Moving Objects. In Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Abu Dhabi, United Arab Emirates, 14–18 October 2024; pp. 10870–10877. [\[CrossRef\]](#)
- Zhang, L.; Li, Y.; Nevatia, R. Global data association for multi-object tracking using network flows. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8. [\[CrossRef\]](#)

24. Tang, S.; Andriluka, M.; Andres, B.; Schiele, B. Multiple People Tracking by Lifted Multicut and Person Re-identification. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3701–3710. [\[CrossRef\]](#)
25. Roshan Zamir, A.; Dehghan, A.; Shah, M. GMCP-Tracker: Global Multi-object Tracking Using Generalized Minimum Clique Graphs. In *Computer Vision—ECCV 2012*; Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7573, pp. 343–356. [\[CrossRef\]](#)
26. Tang, S.; Andres, B.; Andriluka, M.; Schiele, B. Subgraph decomposition for multi-target tracking. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5033–5041. [\[CrossRef\]](#)
27. Cetintas, O.; Brasó, G.; Leal-Taixé, L. Unifying Short and Long-Term Tracking with Graph Hierarchies. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 22877–22887. [\[CrossRef\]](#)
28. Brasó, G.; Leal-Taixé, L. Learning a Neural Solver for Multiple Object Tracking. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 6246–6256. [\[CrossRef\]](#)
29. Liu, Q.; Chu, Q.; Liu, B.; Yu, N. GSM: Graph Similarity Model for Multi-Object Tracking. In Proceedings of the International Joint Conference on Artificial Intelligence, Yokohama, Japan, 11–17 July 2020.
30. Li, J.; Gao, X.; Jiang, T. Graph Networks for Multiple Object Tracking. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020.
31. He, J.; Huang, Z.; Wang, N.; Zhang, Z. Learnable Graph Matching: Incorporating Graph Partitioning with Deep Feature Learning for Multiple Object Tracking. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 5295–5305. [\[CrossRef\]](#)
32. Hyun, J.; Kang, M.; Wee, D.; Yeung, D.Y. Detection Recovery in Online Multi-Object Tracking with Sparse Graph Tracker. In Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2–7 January 2023; pp. 4839–4848. [\[CrossRef\]](#)
33. Lin, J.; Liang, G.; Zhang, R. LTTrack: Rethinking the Tracking Framework for Long-Term Multi-Object Tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 9866–9881. [\[CrossRef\]](#)
34. Gao, R.; Wang, L. MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 9901–9910.
35. You, S.; Yao, H.; Xu, C. Multi-Object Tracking With Spatial-Temporal Topology-Based Detector. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 3023–3035. [\[CrossRef\]](#)
36. Vaquero, L.; Xu, Y.; Alameda-Pineda, X.; Brea, V.M.; Mucientes, M. Lost and Found: Overcoming Detector Failures in Online Multi-object Tracking. In *Computer Vision—ECCV 2024*; Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G., Eds.; Springer Nature: Cham, Switzerland, 2025; Volume 15131, pp. 448–466. [\[CrossRef\]](#)
37. Stanojević, V.; Todorović, B. BoostTrack++: Using tracklet information to detect more objects in multiple object tracking. *arXiv* **2025**, arXiv:2408.13003. [\[CrossRef\]](#)
38. Jin, H.; Nie, X.; Yan, Y.; Chen, X.; Zhu, Z.; Qi, D. AHOR: Online Multi-Object Tracking With Authenticity Hierarchizing and Occlusion Recovery. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 8253–8265. [\[CrossRef\]](#)
39. Zhou, K.; Yang, Y.; Cavallaro, A.; Xiang, T. Omni-Scale Feature Learning for Person Re-Identification. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3701–3711. [\[CrossRef\]](#)
40. Sarlin, P.E.; DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperGlue: Learning Feature Matching With Graph Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
41. Chen, G.; Wang, W.; He, Z.; Wang, L.; Yuan, Y.; Zhang, D.; Zhang, J.; Zhu, P.; Van Gool, L.; Han, J.; et al. VisDrone-MOT2021: The Vision Meets Drone Multiple Object Tracking Challenge Results. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 2839–2846. [\[CrossRef\]](#)
42. Yu, H.; Li, G.; Zhang, W.; Huang, Q.; Du, D.; Tian, Q.; Sebe, N. The Unmanned Aerial Vehicle Benchmark: Object Detection, Tracking and Baseline. *Int. J. Comput. Vis.* **2020**, *128*, 1141–1159. [\[CrossRef\]](#)
43. Liu, S.; Li, X.; Lu, H.; He, Y. Multi-Object Tracking Meets Moving UAV. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 8866–8875. [\[CrossRef\]](#)
44. Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; Liu, W. FairMOT: On the Fairness of Detection and Re-identification in Multiple Object Tracking. *Int. J. Comput. Vis.* **2021**, *129*, 3069–3087. [\[CrossRef\]](#)
45. Wang, P.; Wang, Y.; Li, D. DroneMOT: Drone-based Multi-Object Tracking Considering Detection Difficulties and Simultaneous Moving of Drones and Objects. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024; pp. 7397–7404. [\[CrossRef\]](#)

46. Shi, L.; Zhang, Q.; Pan, B.; Zhang, J.; Su, Y. Global-Local and Occlusion Awareness Network for Object Tracking in UAVs. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 8834–8844. [[CrossRef](#)]
47. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468. [[CrossRef](#)]
48. Yao, M.; Wang, J.; Peng, J.; Chi, M.; Liu, C. FOLT: Fast Multiple Object Tracking from UAV-captured Videos Based on Optical Flow. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 3375–3383. [[CrossRef](#)]
49. Ren, L.; Yin, W.; Diao, W.; Fu, K.; Sun, X. SuperMOT: Decoupling Motion and Fusing Temporal Pyramid Features for UAV Multiobject Tracking. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 14188–14202. [[CrossRef](#)]
50. Yao, M.; Peng, J.; He, Q.; Peng, B.; Chen, H.; Chi, M.; Liu, C.; Benediktsson, J.A. MM-Tracker: Motion Mamba for UAV-platform Multiple Object Tracking. *AAAI* **2025**, *39*, 9409–9417. [[CrossRef](#)]
51. Shim, K.; Ko, K.; Yang, Y.; Kim, C. Focusing on Tracks for Online Multi-Object Tracking. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025; pp. 11687–11696. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.