Tracking the Unstable: Appearance-Guided Motion Modeling for Robust Multi-Object Tracking in UAV-Captured Videos

Jianbo Ma[1,2], Hui Luo[1*], Qi Chen[3], Yuankai Qi[4], Yumei Sun[1], Amin Beheshti[4], Jianlin Zhang[1*], Ming-Hsuan Yang[5]

[1]State Key Laboratory of Optical Field Manipulation Science and Technology, Institute of Optics and Electronics, CAS
[2]School of Electronic, Electrical and Communication Engineering, UCAS
[3]University of Adelaide
[4]Macquarie University
[5]University of California, Merced
{majianbo22, luohui19, sunyumei20}@mails.ucas.ac.cn, {yuankai.qi, amin.beheshti}@mq.edu.au,
qi.chen04@adelaide.edu.au, jlin@ioe.ac.cn, mhyang@ucmerced.edu

Abstract

Multi-object tracking (MOT) aims to track multiple objects while maintaining consistent identities across frames of a given video. In unmanned aerial vehicle (UAV) recorded videos, frequent viewpoint changes and complex UAV-ground relative motion dynamics pose significant challenges, which often lead to unstable affinity measurement and ambiguous association. Existing methods typically model motion and appearance cues separately, overlooking their spatio-temporal interplay and resulting in suboptimal tracking performance. In this work, we propose AMOT, which jointly exploits appearance and motion cues through two key components: an Appearance-Motion Consistency (AMC) matrix and a Motion-aware Track Continuation (MTC) module. Specifically, the AMC matrix computes bi-directional spatial consistency under the guidance of appearance features, enabling more reliable and context-aware identity association. The MTC module complements AMC by reactivating unmatched tracks through appearance-guided predictions that align with Kalman-based predictions, thereby reducing broken trajectories caused by missed detections. Extensive experiments on three UAV benchmarks, including VisDrone2019, UAVDT, and VT-MOT-UAV, demonstrate that our AMOT outperforms current state-of-the-art methods and generalizes well in a plug-and-play and training-free manner.

Code — https://github.com/ydhcg-BoBo/AMOT

Introduction

Multi-object tracking (MOT) is a fundamental vision task with widespread applications, such as autonomous driving (Zhuang et al. 2024) and unmanned aerial vehicle (UAV) surveillance (Wu et al. 2025). A typical pipeline of MOT is to first detect multiple objects, and then assign each detection to existing tracks through data association, ensuring the continuity of track identities over time. Despite progress, robust data association remains challenging, particularly for videos captured by UAV-mounted cameras.

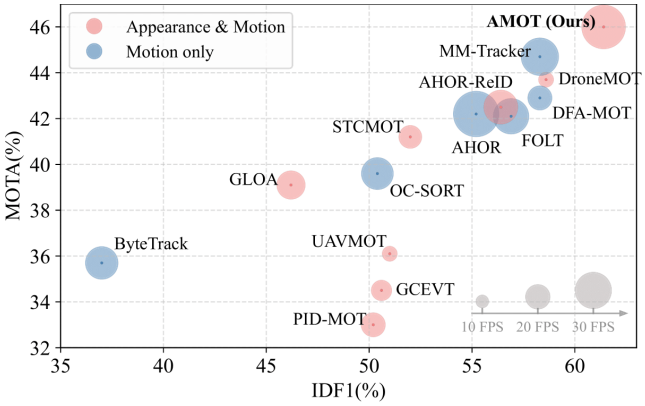


Figure 1: IDF1-MOTA-FPS comparisons of different methods on VisDrone2019. The radius of the circle denotes FPS. Our AMOT achieves the highest IDF1 of 61.4% and MOTA of 46.0%, with a real-time inference speed of 36.4 FPS.

Data association typically relies on a cost matrix that quantifies the affinity between detection–track pairs. Frequent viewpoint changes induce significant variations in object appearance and position. Compounding this, large and unpredictable object displacements caused by complex UAV-ground relative motion dynamics (e.g., varying velocities and directions) further lead to unstable affinity measurement, ultimately compromising identity assignment.

To address the above issues, prevailing methods adopt two main strategies for constructing the cost matrix. (1) Motion-based position prediction: The Kalman Filter (Zhang et al. 2022; Cao et al. 2023) is widely used as a motion model to predict current positions of tracks from their historical states. Then, a motion-based cost matrix is constructed based on the positional proximity between detections and tracks. To overcome the limitations of the Kalman Filter's linear motion predictions in UAV views, some recent works introduce complementary techniques, such as camera motion compensation (CMC) (Wang et al. 2024a;

*These authors are the corresponding authors.

Song and Lee 2024) and optical flow (Yao et al. 2023, 2025), to improve the accuracy of predicted track positions. (2) Appearance-based instance-level discrimination: Instance-level re-identification (ReID) embedding, being insensitive to position changes, offers significant advantages under substantial camera motion or object displacement. Several works (Li et al. 2024a; Wang et al. 2024a,b; Song and Lee 2024) have constructed appearance-based cost matrices using ReID embedding to facilitate accurate identity assignment. Nevertheless, these two strategies model motion and appearance cues independently to generate separate cost matrices, ignoring the intrinsic relationships between them. Specifically, motion prediction errors caused by sudden object displacements can adversely affect the construction of the motion-based cost matrix, while appearance ambiguities similarly impact the appearance-based cost matrix. Consequently, when these cost matrices produce conflicting association scores, determining a reliable matching decision becomes difficult.

To address these challenges, we propose an appearance-guided motion modeling strategy that localizes object positions across frames through dense appearance similarity measurement. Concretely, we compute a dense response map by measuring the similarity between a query ReID embedding from a reference frame and all spatial locations in the ReID feature map of the adjacent frame. The response map reflects the probability of the object's spatial position in consecutive frames. Building upon this, we introduce an appearance-motion consistency (AMC) matrix that computes forward and backward spatial distances between adjacent frames using dense response maps derived from tracks and detections. By capturing bi-directional spatial alignment, the AMC matrix reflects strong spatio-temporal correspondence, enabling the construction of a more robust cost matrix. In addition, traditional identity assignment depends on detection-to-track matching. However, missed detections may cause active tracks to lack corresponding detections, resulting in unmatched tracks. To mitigate this issue, we propose a motion-aware track continuation (MTC) module that reactivates unmatched tracks by comparing appearance-guided and Kalman-based object center predictions.

Together with the AMC and MTC modules, we propose a novel multi-object tracker, termed AMOT, which is built on the joint detection and embedding (JDE) architecture (Zhang et al. 2021). AMOT is specifically designed to enhance the robustness and accuracy of identity assignment under challenging UAV-captured videos. Experiments on multiple UAV benchmarks demonstrate that AMOT achieves superior performance in terms of IDF1 and MOTA, as shown in Figure 1. The main contributions are summarized as follows:

- We propose a novel appearance-motion consistency (AMC) matrix that integrates appearance similarity with bi-directional spatial distances, offering a reliable affinity measurement for robust identity assignment.
- We design a motion-aware track continuation (MTC) module that recovers unmatched tracks by aligning appearance-guided and Kalman-based predictions, reducing fragmented trajectories under detection failures.
- AMC and MTC modules are plug-and-play and training-free, allowing easy integration into JDE-based trackers to boost tracking performance.
- Extensive evaluations on VisDrone2019, UAVDT and VT-MOT-UAV benchmarks demonstrate that AMOT consistently outperforms existing trackers. For example, AMOT attains an IDF1 of 61.4% on VisDrone2019, surpassing MM-Tracker by 2.8%.

Related Works

Motion Modeling

The purpose of motion modeling in MOT is to predict the positions of tracks. Most MOT methods (Stadler and Beyerer 2023; Li et al. 2024b; Cheng et al. 2023) adopt the Kalman Filter for motion modeling, owing to its high computational efficiency for real-time applications. Predicted positions of tracks are compared with current detections via Mahalanobis distance (Du et al. 2023) or Intersection over Union (IoU) (Lv et al. 2024a), which are employed as motion-based cost metrics for data association. Some works (Cao et al. 2023; Maggiolino et al. 2023) enhance the Kalman Filter by adopting an observation-centric strategy to refine track estimation, yielding better performance under non-linear motion patterns. Despite their effectiveness, Kalman-based methods exhibit limited capability in scenarios involving substantial camera motion and object displacement. Meanwhile, learning-based motion models (Shuai et al. 2021; Qin et al. 2023; Song et al. 2025) have recently gained attention. These models leverage data-driven architectures to learn motion patterns. However, such methods often incur high computational costs and are unsuitable for real-time tracking. In contrast, we introduce a training-free motion modeling strategy that balances robustness and efficiency, specifically tailored for UAV-captured videos.

Appearance Modeling

Appearance modeling aims to extract discriminative appearance features to re-identify objects. The tracking-by-detection paradigm (Jin et al. 2024; Huang et al. 2024) first detects objects using an off-the-shelf detector, and then employs a ReID network to extract identity embedding for each object. Although this pipeline delivers impressive performance, it suffers from high computational cost due to separate detection and extraction stages. In contrast, the joint detection and embedding (JDE) paradigm (Liu et al. 2023; Meng et al. 2023) integrates object detection and ReID feature extraction into a unified framework, enabling simultaneous object localization and embedding extraction. Furthermore, several JDE-based trackers incorporate global attention (Wu et al. 2023; Shi et al. 2023) and temporal cues (Liu et al. 2022; Ma et al. 2024) to enhance the discriminability of instance-level ReID embedding. The aforementioned methods measure instance-level appearance similarity by computing the cosine distance between ReID embeddings. Differently, we reformulate the appearance similarity measurement as a dense response between the instance embedding and the global ReID feature map. This enables the joint modeling of visual similarity and spatial coherence.

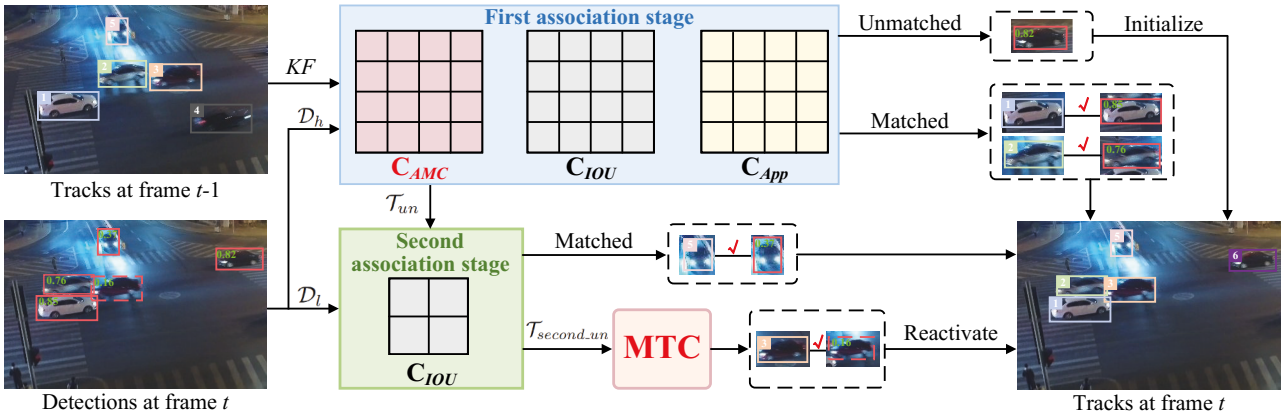


Figure 2: Tracking pipeline of AMOT. Specifically, we introduce the appearance-motion consistency (AMC) matrix $\mathbf{C}_{AMC}$ that integrates it with the appearance similarity matrix $\mathbf{C}_{App}$ and the Intersection-over-Union (IOU) matrix $\mathbf{C}_{IOU}$ to robustly associate high-confidence detections $\mathcal{D}_h$ with tracks $\mathcal{T}$ at frame $t-1$. Then, the unmatched tracks $\mathcal{T}_{un}$ are associated with low-confidence detections $\mathcal{D}_l$ in the second stage. The remaining unmatched tracks $\mathcal{T}_{second\_un}$ are further potentially reactivated through our proposed motion-aware track continuation (MTC) module. *KF* means the Kalman Filter.

Data Association

Data association typically involves constructing a cost matrix that quantifies the affinity between current detections and existing tracks based on motion and appearance information. To improve association accuracy, motion and appearance cues are often modeled separately and then integrated into a unified cost matrix (Liang et al. 2022; Yang et al. 2024). Once the cost matrix is constructed, data association is formulated as an assignment problem and solved using the Hungarian algorithm (Yi et al. 2024; Qin et al. 2024). Nevertheless, existing approaches commonly fail to consider the inherent interplay between motion and appearance cues. This independent modeling can cause instability in affinity measurement, arising from prediction errors or appearance ambiguities, which ultimately leads to conflicting data associations. To this end, we propose the AMC matrix, which jointly enforces consistency in appearance, motion, and temporal domains, enabling reliable identity assignment.

Methodology

Preliminaries

Given an input frame $\mathbf{I}^t$, it is first fed into a feature extractor to obtain a feature map $\mathbf{F}^t \in \mathbb{R}^{H\times W\times 64}$. Then, the detection branch processes $\mathbf{F}^t$ to generate a heatmap $\mathbf{H}^t \in \mathbb{R}^{H\times W\times C}$ for object center localization, where C denotes the number of object classes, along with a regression map $\mathbf{B}^t \in \mathbb{R}^{H\times W\times 2}$ that predicts the height and width of objects. The ReID branch produces a ReID feature map $\mathbf{E}^t \in \mathbb{R}^{H\times W\times D}$, where $D = 128$ is embedding dimension.

We retain the locations in $\mathbf{H}^t$ whose confidence scores exceed a threshold τ as object centers, formulated as:

$$\mathcal{O}_{det} = \{[x_i, y_i] \mid \mathbf{H}^t_{(x_i,y_i)} > \tau\}_{i=1}^{N}, \tag{1}$$

where $\mathcal{O}_{det}$ denotes the set of center coordinates for detections, and N is the total number of detections. For each detection, the width and height are obtained from the regression map as $[w_i, h_i] = \mathbf{B}^t_{(x_i,y_i)}$, and the confidence score is given by $c_i = \mathbf{H}^t_{(x_i,y_i)}$. Meanwhile, the ReID embedding $\mathbf{e}_i$ for each detection is extracted from $\mathbf{E}^t$ at the corresponding center coordinate, expressed as $\mathbf{e}_i = \mathbf{E}^t_{(x_i,y_i)} \in \mathbb{R}^D$.

The set of detections is defined as $\mathcal{D} = \{\mathbf{d}_i\}_{i=1}^{N}$, where each detection is defined as $\mathbf{d}_i = [c_i, x_i, y_i, w_i, h_i, \mathbf{e}_i]$. The set of tracks can be represented as $\mathcal{T} = \{\mathbf{t}_j\}_{j=1}^{M}$, where M is the total number of existing tracks. The tracking states of a track is defined as $\mathbf{t}_j = [id, x_j, y_j, w_j, h_j, \mathbf{e}_j]$, where id denotes the track identity, and the set of center coordinates for tracks is defined as:

$$\mathcal{O}_{trk} = \{[x_j, y_j]\}_{j=1}^{M}. \tag{2}$$

In our AMOT, as shown in Figure 2, we construct an appearance-motion consistency (AMC) matrix to enable robust detection-to-track association. Additionally, we propose a motion-aware track continuation (MTC) module designed to reactivate unmatched tracks without relying on explicit detection-to-track matching.

Appearance-Motion Consistency Matrix

Most existing cost matrices are constructed using either motion or appearance cues independently. However, this separate modeling approach is insufficient under challenging UAV tracking conditions, often resulting in tracking failures. To address this issue, we propose the AMC matrix, which jointly models appearance similarity and spatio-temporal correspondence to improve the robustness of association.

Specifically, we compute the track-specific dense response maps $\mathbf{A}_{trk}$ by evaluating the similarity between each track's ReID embedding and the current ReID feature map $\mathbf{E}^t$, defined as:

$$\mathbf{A}^{(j)}_{trk}(x, y) = \text{sim}(\mathbf{E}^t(x, y), \mathbf{e}_j), \tag{3}$$

where $\mathbf{e}_j$ denotes the ReID embedding of the j-th track, and $\text{sim}(\cdot)$ is the cosine similarity function. The response map

$\mathbf{A}_{trk}^{(j)} \in \mathbb{R}^{H\times W}$ highlights regions that are most semantically correlated with the track's ReID embedding. Then, the spatial location corresponding to the maximum response on $\mathbf{A}_{trk}^{(j)}$ is defined as the center of the j-th track in the current frame. Accordingly, the set of predicted center coordinates $\mathcal{Q}_{trk}$ for tracks in the current frame is formulated as:

$$\mathcal{Q}_{trk} = \{[x_j^*, y_j^*] \,|\, \arg\max_{(x,y)\in\Omega} \mathbf{A}_{trk}^{(j)}(x,y)\}_{j=1}^{M}. \tag{4}$$

Similarly, detection-specific dense response maps $\mathbf{A}_{det}$ are computed by evaluating the similarity between each detection's ReID embedding and the ReID feature map $\mathbf{E}^{t-1}$ from the previous frame, defined as:

$$\mathbf{A}_{det}^{(i)}(x,y) = \mathrm{sim}(\mathbf{E}^{t-1}(x,y),\, \mathbf{e}_i), \tag{5}$$

where $\mathbf{e}_i$ denotes the ReID embedding of the i-th detection. The set of predicted center coordinates $\mathcal{Q}_{det}$ for detections in the previous frame is given by:

$$\mathcal{Q}_{det} = \{[x_i^*, y_i^*] \,|\, \arg\max_{(x,y)\in\Omega} \mathbf{A}_{det}^{(i)}(x,y)\}_{i=1}^{N}. \tag{6}$$

Subsequently, as shown in Figure 3, we quantify the appearance-guided spatio-temporal correspondence by measuring the forward and backward spatial distances, which can be defined as:

$$\mathbf{D}_f(j,i) = \left\| \mathcal{Q}_{trk}^{(j)} - \mathcal{O}_{det}^{(i)} \right\|_2, \tag{7}$$

$$\mathbf{D}_b(i,j) = \left\| \mathcal{Q}_{det}^{(i)} - \mathcal{O}_{trk}^{(j)} \right\|_2. \tag{8}$$

Here, $\mathbf{D}_f(j,i)$ denotes the forward spatial distance from the predicted position of the j-th track $\mathcal{Q}_{trk}^{(j)}$ to the observed center of the i-th detection $\mathcal{O}_{det}^{(i)}$. Conversely, $\mathbf{D}_b(i,j)$ represents the backward spatial distance from the predicted position of the i-th detection $\mathcal{Q}_{det}^{(i)}$ to the observed center of the j-th track $\mathcal{O}_{trk}^{(j)}$. In both cases, lower values indicate stronger spatial coherence. A detection-track pair is considered a reliable match only when both forward and backward distances are small.

Next, we construct the AMC matrix $\mathbf{C}_{AMC}$ using a Gaussian kernel that integrates the bi-directional spatial distances, defined as:

$$\mathbf{C}_{AMC}(i,j) = 1 - \exp\left(-\frac{\mathbf{D}_f^\top(i,j) + \mathbf{D}_b(i,j)}{2\sigma^2}\right), \tag{9}$$

where σ is a scale factor that controls the spatial sensitivity, and is set to 5. $\mathbf{C}_{AMC}$ intrinsically encodes the joint spatial and appearance similarities. It is designed to impose a smooth penalty on potentially ambiguous detection-track pairs, enhancing the robustness of affinity measurement.

Motion-aware Track Continuation

Recovery of short-term lost tracks is critical for maintaining track identities. To this end, we propose the MTC module, which effectively propagates unmatched tracks to mitigate association failures caused by temporary missed detections.

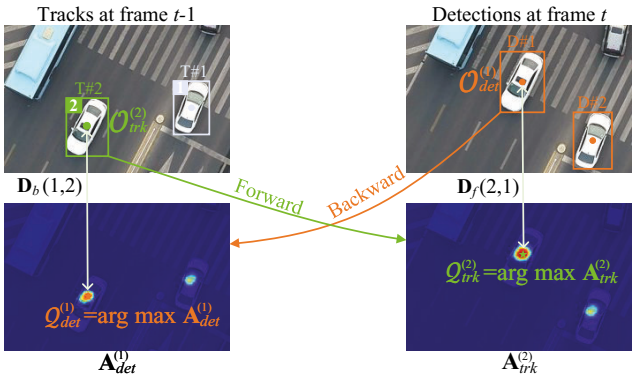


Figure 3: Overview of bi-directional spatial distances in AMC matrix. $\mathbf{A}_{trk}^{(2)}$ and $\mathbf{A}_{det}^{(1)}$ are the track-specific dense response map of T#2 and the detection-specific dense response map of D#1, respectively. $\mathbf{D}_f(2,1)$ represents the forward spatial distance from predicted center $\mathcal{Q}_{trk}^{(2)}$ to observed center $\mathcal{O}_{det}^{(1)}$, while $\mathbf{D}_b(1,2)$ denotes the backward spatial distance from predicted center $\mathcal{Q}_{det}^{(1)}$ to observed center $\mathcal{O}_{trk}^{(2)}$.

At the beginning of tracking, we initialize a buffer for each track to store its frame indices and corresponding tracking states, denoted as Buff $= \{\mathbf{t}_j^s\}_{s=t-20}^{t-1}$, and update following a first-in, first-out policy. In later frames, if a track fails to associate with any detection, while retaining recent and temporally consecutive tracking states in its buffer, we mark it as a reactivation candidate. Then, we introduce the MTC module to determine whether these candidates should be reactivated. To be specific, we first employ the Kalman Filter to predict the bounding box of all candidates, yielding a set of Kalman-based predictions: $\mathbf{Box}_{kf} = \{[\hat{x}_k, \hat{y}_k, \hat{w}_k, \hat{h}_k]\}_{k=1}^{K}$, where K is the total number of candidates. The corresponding predicted centers are denoted as $\mathbf{c}_{kf} = \{[\hat{x}_k, \hat{y}_k]\}_{k=1}^{K}$. Subsequently, we compute the similarity between the candidate's latest ReID embedding $\mathbf{e}_k^{t-1}$ and the current ReID feature map $\mathbf{E}^t$, obtaining a dense response map: $\mathbf{M}^{(k)} = \mathrm{sim}(\mathbf{E}^t, \mathbf{e}_k^{t-1}) \in \mathbb{R}^{H\times W}$. The region with the maximum response in $\mathbf{M}$ not only exhibits the highest similarity to the given ReID embedding, but also serves as an appearance-guided prediction of the candidate's center coordinates in the current frame, denoted as $\mathbf{c}_{reid} = \{[\tilde{x}_k, \tilde{y}_k]\}_{k=1}^{K}$.

Subsequently, we compute the Euclidean distance between the appearance-guided predicted center $\mathbf{c}_{reid}$ and the Kalman-based predicted center $\mathbf{c}_{kf}$ for each candidate. This distance is formulated as:

$$d_k = \left\| \mathbf{c}_{reid}^{(k)} - \mathbf{c}_{kf}^{(k)} \right\|_2, \tag{10}$$

where d_k denotes the spatial offset between the two predicted centers for the k-th candidate. If d_k is smaller than the predefined threshold λ, which is set to 3 in our experiments, and there is no significant overlap with any current detection, the candidate is considered present in the current frame and is reactivated to ensure identity consistency. Otherwise, it remains unmatched.

Tracking Pipeline

The tracking pipeline of AMOT is illustrated in Figure 2. The Kalman Filter is used to predict the current positions of tracks. We then adopt a two-stage matching strategy similar to BYTE (Zhang et al. 2022), where detections are divided into high-confidence sets $\mathcal{D}_h$ and low-confidence sets $\mathcal{D}_l$.

Specifically, in the first stage, we construct three pairwise cost matrices between $\mathcal{D}_h$ and the tracks $\mathcal{T}$, including the appearance similarity matrix $\mathbf{C}_{App}$ (Zhang et al. 2021), the Intersection-over-Union (IOU) matrix $\mathbf{C}_{IOU}$, and the proposed AMC matrix $\mathbf{C}_{AMC}$. These matrices are integrated into a unified cost matrix $\mathbf{C}_{uni}$ as follows:

$$\mathbf{C}_{uni} = 1 - (1 - \mathbf{C}_{AMC} \cdot \mathbf{C}_{IOU}) \cdot (1 - \mathbf{C}_{App}). \quad (11)$$

Here, the cost matrix $\mathbf{C}_{uni}$ is used for bipartite matching via the Hungarian algorithm. In the second stage, unmatched tracks $\mathcal{T}_{un}$ are associated with $\mathcal{D}_l$ solely by the IOU matrix. For tracks that remain unmatched after this second association, denoted as $\mathcal{T}_{second\_un}$, we employ the proposed MTC module to determine whether they can be reactivated as matched tracks.

After that, tracks that remain unmatched for more than 30 frames are removed. New tracks are initialized from the remaining high-confidence detections that are not associated with any existing tracks, while the tracking states of the matched tracks are updated based on current observations.

Experiments

Settings

Datasets and Evaluation Metrics. We evaluated our method on various MOT datasets from UAV perspectives, including VisDrone2019 (Du et al. 2019), UAVDT (Du et al. 2018), and VT-MOT-UAV (Zhu et al. 2025).

We adopt the widely-used CLEAR metrics (Liu et al. 2022), including MOTA, IDF1, the number of mostly tracked (MT) objects, and mostly lost (ML) objects, to comprehensively evaluate the tracking performance. MOTA emphasizes the detection quality and is computed based on false positives (FP), false negatives (FN), and identity switches (IDs). IDF1 measures the tracker's ability to maintain consistent object identities over time, reflecting the accuracy of identity association across frames. Additionally, we report FPS to assess the inference speed of the tracker.

Implementation Details. We adopt DLA-34 (Zhang et al. 2021) as the default backbone network and initialize its parameters with pre-trained weights from the COCO dataset. The input image is resized to 608×1088. The corresponding feature map undergoes a $4\times$ downsampling, resulting in a resolution of 152×272, where $H = 152$ and $W = 272$. We employ the Adam optimizer with an initial learning rate of $7e^{-5}$ to train our model for 30 epochs, reducing the learning rate by a factor of 10 at 20 epochs. For the objective functions, focal loss is employed to supervise the object heatmap, while L1 loss is used to supervise the predicted object width and height. In addition, both cross-entropy loss and triplet loss are applied to guide the learning of the ReID embedding. To further enhance the model's perception of object instances, we introduce an additional mask branch (Tian, Shen, and Chen 2020) during training, which is trained using the Dice loss.

Method	IDF1↑	MOTA↑	MT↑	ML↓	IDs↓
ByteTrack (Zhang et al. 2022)	37.0	35.7	-	-	2168
UAVMOT (Liu et al. 2022)	51.0	36.1	520	574	2775
OC-SORT (Cao et al. 2023)	50.4	39.6	-	-	986
GCEVT (Wu et al. 2023)	50.6	34.5	520	612	841
GLOA (Shi et al. 2023)	46.2	39.1	581	824	4426
FOLT (Yao et al. 2023)	56.9	42.1	-	-	<u>800</u>
PID-MOT (Lv et al. 2024b)	50.2	33.0	686	424	3529
AHOR-ReID (Jin et al. 2024)	56.4	42.5	-	-	810
STCMOT (Ma et al. 2024)	52.0	41.2	667	453	3984
DroneMOT (Wang et al. 2024a)	<u>58.6</u>	43.7	<u>689</u>	**397**	1112
DFA-MOT (Zheng et al. 2025)	58.3	42.9	518	523	**792**
MM-Tracker (Yao et al. 2025)	58.3	<u>44.7</u>	-	-	-
AMOT (Ours)	**61.4**	**46.0**	**716**	<u>413</u>	1063

Table 1: Tracking performance comparison with state-of-the-art methods on VisDrone2019 test set.

Method	IDF1↑	MOTA↑	MT↑	ML↓	IDs↓
ByteTrack (Zhang et al. 2022)	59.1	41.6	-	-	296
UAVMOT (Liu et al. 2022)	67.3	46.4	624	221	456
OC-SORT (Cao et al. 2023)	64.9	47.5	-	-	288
GCEVT (Wu et al. 2023)	68.6	47.6	618	363	1801
GLOA (Shi et al. 2023)	68.9	49.6	626	220	433
FOLT (Yao et al. 2023)	68.3	48.5	-	-	338
STCMOT (Ma et al. 2024)	<u>69.8</u>	49.2	664	203	665
DroneMOT (Wang et al. 2024a)	69.6	50.1	638	<u>178</u>	**129**
DFA-MOT (Zheng et al. 2025)	69.3	49.9	<u>684</u>	230	396
MM-Tracker (Yao et al. 2025)	68.9	<u>51.4</u>	-	-	-
AMOT (Ours)	**74.7**	**55.1**	**794**	**142**	<u>272</u>

Table 2: Tracking performance comparison with state-of-the-art methods on UAVDT test set.

Benchmark Evaluation

We present the tracking results for multiple UAV datasets. ↑/↓ indicate that higher/lower is better, respectively. The best scores for each metric are marked in **bold**, and the second-best scores are marked in <u>underline</u>.

Results on VisDrone2019. VisDrone2019 serves as a fundamental benchmark for multi-object tracking in dynamic UAV-captured videos, which involve scale variations and long-range object movements. As shown in Table 1, our AMOT exhibits the highest IDF1 of 61.4% and MOTA of 46.0%, outperforming current state-of-the-art methods. Concretely, compared with DroneMOT and MM-Tracker, AMOT achieves absolute improvements in both IDF1 and MOTA, demonstrating superior association accuracy and enhanced overall tracking performance.

Results on UAVDT. The tracking scenarios in UAVDT are derived from bird's-eye views at different outdoor scenes. Table 2 shows that our AMOT achieves the best tracking

Method	IDF1↑	MOTA↑	MT↑	ML↓	IDs↓
SORT (Bewley et al. 2016)	48.1	28.6	103	340	520
FairMOT (Zhang et al. 2021)	39.9	17.9	94	322	901
ByteTrack (Zhang et al. 2022)	50.2	28.5	129	324	415
UAVMOT (Liu et al. 2022)	47.9	22.1	**175**	248	1421
OC-SORT (Cao et al. 2023)	48.2	28.5	102	342	509
STCMOT (Ma et al. 2024)	47.4	27.9	153	277	518
AMOT (Ours)	**52.7**	**31.8**	168	**247**	**412**

Table 3: Tracking performance comparison with state-of-the-art methods on VT-MOT-UAV test set.

AMC	MTC	VisDrone2019			UAVDT			
		IDF1↑	MOTA↑	IDs↓	IDF1↑	MOTA↑	IDs↓	FPS ↑
		54.4	43.3	3847	72.4	54.6	886	37.1
✓		60.5	44.1	1078	74.4	54.9	307	36.3
	✓	57.3	44.2	1961	73.4	54.6	440	**37.7**
✓	✓	**61.4**	**46.0**	**1063**	**74.7**	**55.1**	**272**	36.4

Table 4: Ablation of various components.

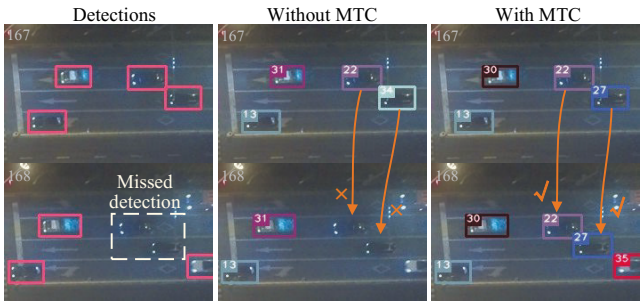


Figure 4: Visualization of tracking results with the MTC module. Despite missed detections, MTC effectively propagates tracks and maintains their correct identities.

performance, with an IDF1 of 74.7% and a MOTA of 55.1%. Specifically, AMOT outperforms STCMOT by +4.9% in IDF1 and surpasses MM-Tracker by +3.7% in MOTA, highlighting its superiority in tracking effectiveness. Moreover, AMOT has the highest MT and the lowest ML, showing its robustness in preserving identity over long-term tracking.

Results on VT-MOT-UAV. VT-MOT-UAV is dedicated to tracking multiple categories, including pedestrians and vehicles, and presents challenges such as varying illumination conditions and cluttered backgrounds. The results are summarized in Table 3. Our AMOT exhibits superior performance with an IDF1 of 52.7% and a MOTA of 31.8%, surpassing the previous advanced approaches.

Ablation Studies

Baseline model. The baseline model uses the network architecture of FairMOT (Zhang et al. 2021) combined with a two-stage association strategy (Zhang et al. 2022).

Component Ablation. We conduct a comprehensive evaluation of the proposed components, as presented in Table 4.

AMC		VisDrone2019			UAVDT		
For.	Back.	IDF1↑	MOTA↑	IDs↓	IDF1↑	MOTA↑	IDs↓
		54.4	43.3	3847	72.4	54.6	886
✓		60.4	43.4	1319	73.3	54.9	389
	✓	60.0	43.9	1132	73.0	54.9	407
✓	✓	**60.5**	**44.1**	**1078**	**74.4**	**54.9**	**307**

Table 5: Impact of the forward and backward spatial distance in AMC matrix, where "For." and "Back." denote forward and backward spatial distances, respectively.

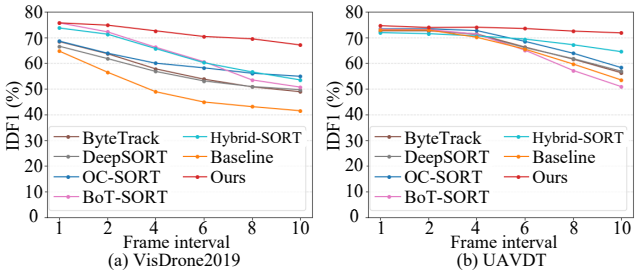
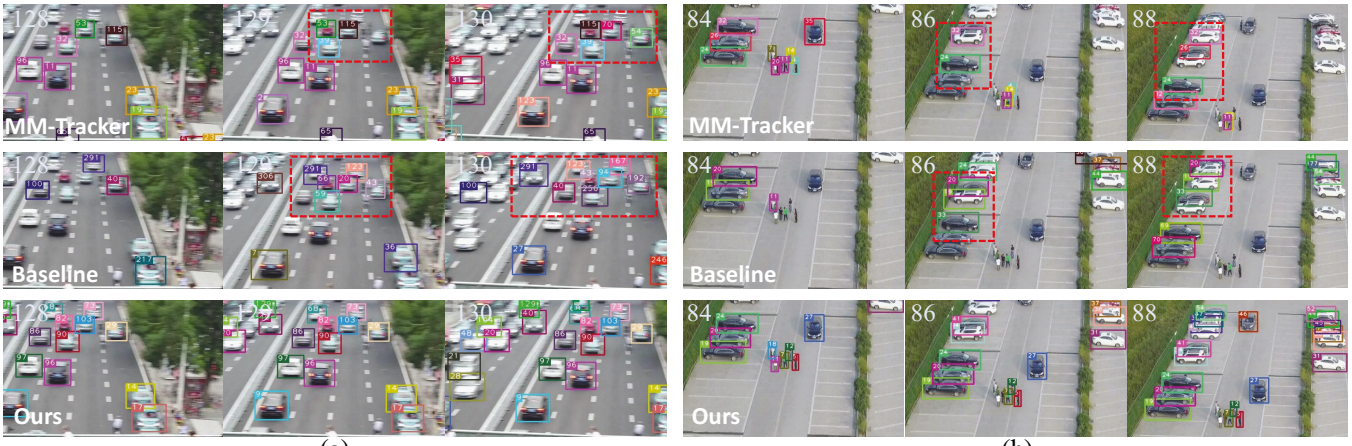


Figure 5: Association performance comparison under different frame intervals for car category. Object displacement increases with frame interval.

Integrating either AMC or MTC component into the baseline model individually leads to notable improvements in tracking performance. Compared with the baseline model, the joint application of AMC and MTC improves IDF1 and MOTA by +7.0% and +2.7%, respectively, while reducing identity switches by 2784 on VisDrone2019. On UAVDT, it achieves gains of +2.3% in IDF1 and +0.5% in MOTA, with 614 fewer identity switches. Furthermore, the AMC and MTC introduce only a minimal computational overhead and have a negligible impact on the overall inference speed. Specifically, the baseline model achieves an inference speed of 37.1 FPS, while our AMOT achieves a comparable speed of 36.4 FPS. These results highlight the effectiveness and computational efficiency of the proposed components.

Impact of Bi-directional Spatial Distances. As shown in Table 5, we assess the effect of forward and backward spatial distances in the AMC matrix. Both forward and backward spatial distances positively contribute to overall tracking performance, while their joint integration yields further gains. This indicates that modeling bi-directional spatial consistency strengthens the spatio-temporal interplay between motion and appearance cues, thereby improving the discriminative power of the affinity measurement.

Effect of MTC. Figure 4 presents a qualitative comparison of tracking results with and without the proposed MTC module. At frame 168, the baseline model fails to associate the object due to missed detections, leading to a track discontinuity. In contrast, the baseline model with MTC module successfully propagates the tracks from frame 167 to frame 168 with consistent identity assignment, highlighting the effectiveness of MTC in preserving track continuity un-



(a) (b)

Figure 6: Visualization of tracking results under challenging UAV-captured videos: (a) lateral viewpoint changes from left to right and (b) viewpoint changes from close-up to distant view.

Method	AMC	MTC	VisDrone2019			UAVDT		
			IDF1↑	MOTA↑	IDs↓	IDF1↑	MOTA↑	IDs↓
UAVMOT			50.6	40.0	3843	69.0	47.6	2079
	✓		53.7	40.2	1589	69.5	47.6	846
	✓	✓	55.0	42.4	1600	69.7	47.8	729
	Gain		**+4.4**	**+2.4**	**-2243**	**+0.7**	**+0.2**	**-1350**
STCMOT			52.7	41.0	4040	69.1	50.4	688
	✓		58.9	41.2	1004	69.7	50.6	352
	✓	✓	59.7	43.1	1043	70.0	50.8	308
	Gain		**+7.0**	**+2.1**	**-2997**	**+0.9**	**+0.4**	**-380**

Table 6: Results of applying AMC and MTC components to various JDE-based methods.

der unreliable detection conditions.

Robustness to Large Displacements. In Figure 5, we compare the association performance of our AMOT against several advanced data association methods using the same detections under varying frame intervals. As the frame interval increases, inter-frame object displacements become larger, resulting in a significant decline in association accuracy for most methods. In contrast, our AMOT demonstrates strong robustness under these challenging conditions. Although recent competitive methods such as BoT-SORT (Aharon et al. 2022) and Hybrid-SORT (Yang et al. 2024) exploit both motion and appearance cues for association, they model these cues independently, which limits their ability to handle large object displacements. Instead, our method jointly and adaptively models motion and appearance cues, enabling more consistent and accurate data association in highly dynamic tracking scenarios.

Generality on Other Trackers. We apply our design to the representative JDE-based trackers, including UAVMOT (Liu et al. 2022) and STCMOT (Ma et al. 2024). As presented in Table 6, all of these trackers benefit from the integration of AMC and MTC. For example, STCMOT achieves a notable gain of +7.0% IDF1 and +2.1% MOTA on VisDrone2019. Similar performance improvements are observed for UAVMOT, with IDF1 and MOTA increasing by +4.4% and +2.4%, respectively. Such improvements across diverse trackers and scenarios further demonstrate the generality and effectiveness of our approach.

Visualization

Frequent viewpoint changes in UAV-captured videos often induce substantial variations in object appearance and position, posing significant challenges to multi-object trackers. As illustrated in Figure 6, we evaluate the tracking performance of our method under these adverse conditions. Both the baseline model and MM-tracker (Yao et al. 2025) suffer from frequent identity switches and tracking failures, reflecting limited capability in coping with dynamic viewpoint changes. In contrast, the proposed AMOT demonstrates robust performance and maintains track identity consistency under such challenging conditions.

Conclusion

In this work, we explore the joint modeling of appearance and motion cues, and introduce two plug-and-play components for data association, namely the AMC matrix and the MTC module. AMC captures appearance, motion, and temporal coherence to ensure accurate identity assignment, while MTC mitigates broken trajectories by reactivating unmatched tracks after missed detections. Building upon these components, we develop AMOT, a simple yet effective joint detection and embedding framework tailored for real-time UAV tracking. Experiments demonstrate that AMOT exhibits strong robustness in dynamic scenarios captured by UAV-mounted cameras. It also achieves outstanding performance on several UAV benchmarks, including VisDrone2019, UAVDT, and VT-MOT-UAV.

Acknowledgments

We gratefully acknowledge the support of Prof. Haorui Zuo. The work was partially supported by Frontier Research Fund of the Institute of Optics and Electronics, Chinese Academy of Sciences (C24K003).

References

Aharon, N.; Orfaig, R.; Bobrovsky, B.-Z.; et al. 2022. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*.

Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; and Upcroft, B. 2016. Simple online and realtime tracking. In *IEEE international conference on image processing*, 3464–3468.

Cao, J.; Pang, J.; Weng, X.; Khirodkar, R.; and Kitani, K. 2023. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *IEEE/CVF conference on computer vision and pattern recognition*, 9686–9696.

Cheng, S.; Yao, M.; Xiao, X.; and other. 2023. Dc-mot: Motion deblurring and compensation for multi-object tracking in uav videos. In *IEEE International Conference on Robotics and Automation*, 789–795.

Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; and Tian, Q. 2018. The unmanned aerial vehicle benchmark: Object detection and tracking. In *European conference on computer vision*, 370–386.

Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. 2019. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In *IEEE/CVF international conference on computer vision workshops*, 213–226.

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.

Huang, C.; Han, S.; He, M.; Zheng, W.; and Wei, Y. 2024. Deconfusetrack: Dealing with confusion for multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19290–19299.

Jin, H.; Nie, X.; Yan, Y.; Chen, X.; Zhu, Z.; and Qi, D. 2024. AHOR: Online Multi-Object Tracking With Authenticity Hierarchizing and Occlusion Recovery. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9): 8253–8265.

Li, S.; Ke, L.; Danelljan, M.; Piccinelli, L.; Segu, M.; Van Gool, L.; and Yu, F. 2024a. Matching anything by segmenting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18963–18973.

Li, Z.; Zhang, D.; Wu, S.; Song, M.; and Chen, G. 2024b. Sampling-resilient multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3297–3305.

Liang, C.; Zhang, Z.; Zhou, X.; Li, B.; and Hu, W. 2022. One more check: making "fake background" be tracked again. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1546–1554.

Liu, K.; Jin, S.; Fu, Z.; Chen, Z.; Jiang, R.; and Ye, J. 2023. Uncertainty-aware unsupervised multi-object tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9996–10005.

Liu, S.; Li, X.; Lu, H.; and He, Y. 2022. Multi-Object Tracking Meets Moving UAV. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8866–8875.

Lv, W.; Huang, Y.; Zhang, N.; Lin, R.-S.; Han, M.; and Zeng, D. 2024a. 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. 2024b. One-shot multiple object tracking with robust id preservation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6): 4473–4488.

Ma, J.; Tang, C.; Wu, F.; Zhao, C.; Zhang, J.; and Xu, Z. 2024. STCMOT: Spatio-Temporal Cohesion Learning for UAV-Based Multiple Object Tracking. In *IEEE International Conference on Multimedia and Expo*, 1–6.

Maggiolino, G.; Ahmad, A.; Cao, J.; and Kitani, K. 2023. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. In *IEEE International conference on image processing*, 3025–3029.

Meng, S.; Shao, D.; Guo, J.; and Gao, S. 2023. Tracking without label: Unsupervised multiple object tracking via contrastive similarity learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 16264–16273.

Qin, Z.; Wang, L.; Zhou, S.; Fu, P.; Hua, G.; and Tang, W. 2024. Towards generalizable multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18995–19004.

Qin, Z.; Zhou, S.; Wang, L.; Duan, J.; Hua, G.; and Tang, W. 2023. Motiontrack: Learning robust short-term and long-term motions for multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17939–17948.

Shi, L.; Zhang, Q.; Pan, B.; Zhang, J.; and Su, Y. 2023. Global-Local and Occlusion Awareness Network for Object Tracking in UAVs. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16: 8834–8844.

Shuai, B.; Berneshawi, A.; Li, X.; Modolo, D.; and Tighe, J. 2021. Siammot: Siamese multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12372–12382.

Song, I.; and Lee, J. 2024. SFTrack: A Robust Scale and Motion Adaptive Algorithm for Tracking Small and Fast Moving Objects. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 10870–10877.

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*, 6978–6986.

Stadler, D.; and Beyerer, J. 2023. An improved association pipeline for multi-person tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3170–3179.

Tian, Z.; Shen, C.; and Chen, H. 2020. Conditional convolutions for instance segmentation. In *European conference on computer vision*, 282–298.

Wang, P.; Wang, Y.; Li, D.; et al. 2024a. DroneMOT: Drone-based Multi-Object Tracking Considering Detection Difficulties and Simultaneous Moving of Drones and Objects. In *IEEE International Conference on Robotics and Automation*, 7397–7404.

Wang, Y.-H.; Hsieh, J.-W.; Chen, P.-Y.; Chang, M.-C.; So, H.-H.; and Li, X. 2024b. Smiletrack: Similarity learning for occlusion-aware multiple object tracking. In *Proceedings of the AAAI conference on artificial intelligence*, 5740–5748.

Wu, H.; He, Z.; Gao, M.; et al. 2023. GCEVT: Learning Global Context Embedding for Vehicle Tracking in Unmanned Aerial Vehicle Videos. *IEEE Geoscience and Remote Sensing Letters*, 20: 1–5.

Wu, H.; Sun, H.; Ji, K.; and Kuang, G. 2025. Temporal-Spatial Feature Interaction Network for Multi-Drone Multi-Object Tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2): 1165–1179.

Yang, M.; Han, G.; Yan, B.; Zhang, W.; Qi, J.; Lu, H.; and Wang, D. 2024. Hybrid-sort: Weak cues matter for online multi-object tracking. In *Proceedings of the AAAI conference on artificial intelligence*, 6504–6512.

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*, 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 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*, 6702–6710.

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.

Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; and Liu, W. 2021. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International journal of computer vision*, 129: 3069–3087.

Zheng, Y.; He, C.; Chen, X.; Zhang, H.; Qu, T.; and Wang, D. 2025. DFA-MOT: A Dynamic Field-Aware Multi-Object Tracking Framework for Unmanned Aerial Vehicles. *IEEE Transactions on Circuits and Systems for Video Technology*, Early Access: 1–13.

Zhu, Y.; Wang, Q.; Li, C.; Tang, J.; Gu, C.; and Huang, Z. 2025. Visible–thermal multiple object tracking: Large-scale video dataset and progressive fusion approach. *Pattern Recognition*, 161: 111330.

Zhuang, Z.; Wang, Z.; Chen, S.; Liu, L.; Luo, H.; and Tan, M. 2024. Robust 3d semantic occupancy prediction with calibration-free spatial transformation. *arXiv preprint arXiv:2411.12177*.