FALCONMOT: SELF-CALIBRATING EVIDENTIAL ASSOCIATION AND VISUAL-KINEMATIC RECOVERY FOR AERIAL MULTI-OBJECT TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-object tracking (MOT) aims to maintain continuous target trajectories across video frames. In unmanned aerial vehicle (UAV) recorded videos, rapid platform motion, continuous viewpoint variations, and small object scales pose severe challenges, frequently causing unreliable affinity estimation and detector dropouts. Existing methods typically rely on fixed linear weighting between motion and appearance, which fails to adapt when visual cues fluctuate, or employ unconstrained feature-map querying that triggers false recoveries from visually similar distractors. In this work, we propose FalconMOT, a robust aerial tracking framework driven by two training-free components: Self-calibrating Probabilistic Evidence Association (SPEAR) and Query Appearance-Motion (QAM) target recovery. Specifically, SPEAR dynamically balances motion and appearance by evaluating the directional consistency of historical track embeddings; when visual features fluctuate, its concentration parameter automatically decreases, down-weighting appearance affinity and relying on spatial motion for robust association. The QAM module addresses missed detections by querying dense backbone feature maps around predicted track locations, filtering out ambiguous background distractors to ensure reliable recovery. Extensive experiments on VisDrone2019 and UAVDT demonstrate that FalconMOT outperforms current state-of-the-art methods, achieving 69.1% IDF1 and 55.6% MOTA while operating at 25.0 FPS on an embedded Jetson Orin NX in a plug-and-play, training-free manner.

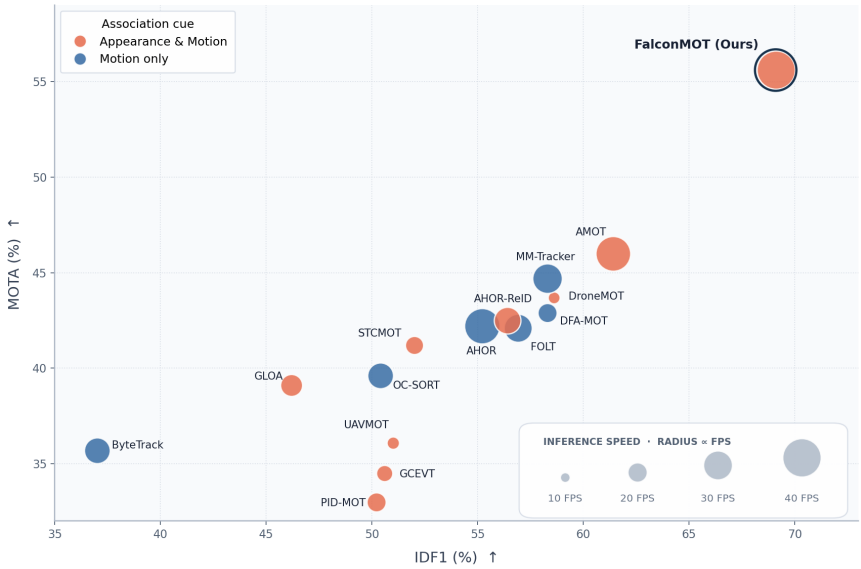

Figure 1: Tracking accuracy versus throughput on VisDrone2019-MOT. FalconMOT achieves 69.1% IDF1, 55.6% MOTA, and reduces identity switches to 702 establishing an optimal Pareto frontier over prior aerial and ground trackers.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Airborne visual perception from Unmanned Aerial Vehicles (UAVs) enables wide-area traffic monitoring, disaster response, and autonomous surveillance (Zhu et al., 2021; Du et al., 2018; Liu et al., 2022; Wang et al., 2024), with online Multi-Object Tracking (MOT) as its core perceptual foundation (Bewley et al., 2016; Wojke et al., 2017; Zhang et al., 2022). While MOT has progressed rapidly on ground-level benchmarks (Dendorfer et al., 2020), aerial tracking operates under a substantially harder sensing regime (Zhu et al., 2021; Ma et al., 2025; Zheng et al., 2025): targets are small and distant with degraded textures that blend into cluttered backgrounds (Wang et al., 2022; Liu et al., 2024), while platform ego-motion and wind disturbances induce camera jitter, abrupt accelerations, and motion blur (Liu et al., 2022; Ma et al., 2024). Detectors therefore suffer frequent confidence dropouts, causing trackers to lose spatial continuity and switch identities (Cao et al., 2023; Ma et al., 2025). A further, often overlooked bottleneck is detector complexity: ground trackers such as ByteTrack and OC-SORT (Zhang et al., 2022; Cao et al., 2023) deploy YOLOX-X ($\sim$ 99 M parameters, 281 GFLOPs), and aerial trackers such as UAVMOT, FOLT, and AMOT (Liu et al., 2022; Yao et al., 2023; Ma et al., 2025) rely on DLA-34 or ResNet-50 networks, whose latency and power demands preclude closed-loop execution on airborne edge platforms (< 20 W). Pairing a compact YOLO11s detector (Jocher & Qiu, 2024) (~ 9.4 M parameters, 21.5 GFLOPs) with a parallel ReID head enables efficient single-pass feature extraction onboard.

However, a lightweight detector does not remove the tracking vulnerabilities inherent to airborne flight. Tracking-by-detection methods (Bewley et al., 2016; Zhang et al., 2022; Aharon et al., 2022; Cao et al., 2023) associate discrete detections across frames using Kalman motion prediction and ReID embeddings, and multi-modal frameworks such as FairMOT (Zhang et al., 2021), BoT-SORT (Aharon et al., 2022), and DFA-MOT (Zheng et al., 2025) combine these cues through static linear mixtures or heuristic products. Yet aerial appearance reliability fluctuates dramatically: an embedding that is distinctive during stable hovering degrades sharply under viewpoint transitions, scale changes, or partial occlusion, and static weighting keeps enforcing corrupted appearance scores. Moreover, when detectors drop small targets, discrete trackers fall back on open-loop motion prediction that quickly diverges under non-linear flight dynamics. Query-based transformers (Wu et al., 2021; Zeng et al., 2022) propagate object states via attention but are too costly for edge execution, while feature-map querying methods such as AMOT (Ma et al., 2025) correlate stored embeddings across the full feature map, where an occluded target's attenuated response is easily eclipsed by a visually similar distractor, turning a transient dropout into an identity switch.

We present FalconMOT, a robust aerial tracking framework built on two training-free, plug-and-play components. *Self-Calibrating Probabilistic Evidence Association* (SPEAR) replaces fixed fusion weights with an adaptive concentration derived from the directional consistency of each track's embedding history on the unit hypersphere: consistent histories make appearance contribute strongly, whereas fluctuating features automatically shift association toward spatial proximity. *Query Appearance-Motion* (QAM) target recovery confines feature-map querying to a local window around the predicted track position and accepts the peak of a joint visual-spatial correlation surface only when a distinctiveness margin separates it from ambient clutter. With a compact YOLO11s backbone, FalconMOT establishes an optimal Pareto frontier across accuracy, latency, and power (Figure 1). Our main contributions are:

- SPEAR, a training-free, plug-and-play association module that balances motion and appearance per track and per frame based on the directional consistency of historical embeddings, replacing fixed linear weighting.
- QAM, a localized recovery module that searches for dropped targets within an uncertainty-bounded window and filters candidate peaks via a distinctiveness margin, preventing false associations with visually similar distractors.
- On VisDrone2019-MOT, FalconMOT achieves state-of-the-art tracking performance (69.1% IDF1, 55.6% MOTA, 702 IDs) at 25 FPS on an embedded Jetson Orin NX, with controlled evaluations confirming a 22.4% reduction in trajectory fragmentations.

2 Related Work

Tracking-by-Detection. Online MOT has long been dominated by tracking-by-detection, which localizes targets per frame before establishing temporal correspondences Bewley et al. (2016); Wojke et al. (2017); Wang et al. (2020); Zhang et al. (2021). These frameworks collapse continuous features into discrete boxes via score thresholds and NMS prior to association, which becomes an acute failure point for UAV imagery Zhu et al. (2021); Du et al. (2018): targets often span fewer than 20–30 pixels Wang et al. (2022); Liu et al. (2024), stride-32 feature pyramids Lin et al. (2017) dilute their activations, and motion blur makes detection scores oscillate so that targets drop out of the observation space. ByteTrack Zhang et al. (2022) mitigates moderate occlusions via a two-stage cascade but offers no recourse during complete detection voids, leaving the Kalman filter Kalman (1960) to dead-reckon; OC-SORT Cao et al. (2023) adds observation-centric momentum, yet its constant-velocity assumption breaks under the apparent non-linear accelerations induced by UAV pitch and roll.

Temporal Propagation and Edge Compute. Continuous tracking paradigms instead propagate unthresholded evidence across time, inspired by radar Track-Before-Detect Bar-Shalom et al. (2001); Boers & Driessen (2004). They span point-based tracking (CenterTrack Zhou et al. (2020)), deformable feature querying (TraDeS Wu et al. (2021)), query-based transformers (TransTrack Sun et al. (2020), TrackFormer Meinhardt et al. (2022), MOTR Zeng et al. (2022)), and selective state-space models (Samba MOT Segu et al. (2025)). CenterTrack lacks identity embeddings and is vulnerable to swaps during crossings, while TraDeS and query transformers require dense spatio-temporal cost volumes or global cross-attention whose latency violates the 20 W budget of airborne platforms. FalconMOT instead performs closed-form feature querying only within a kinematically bounded region of interest.

Aerial Motion Modeling and Multi-Modal Fusion. Aerial trackers Liu et al. (2022); Ma et al. (2024; 2025); Wang et al. (2025) stabilize state estimation against camera displacement with Global Motion Compensation (GMC) Aharon et al. (2022), but typically fuse motion and appearance via linear mixtures (Cost $= \alpha d_{\text{mot}} + \beta d_{\text{app}}$). This blends Euclidean innovation distances in $[0, \infty)$ with cosine distances on the hypersphere $\mathbb{S}^{D-1}$ in $[0, 2]$, and a static weight cannot adapt to asymmetric degradation: rotor-induced blur disperses embeddings while kinematics remain informative, whereas abrupt yaw invalidates Kalman predictions while textures stay sharp. Building on multi-target tracking theory Bar-Shalom et al. (2001) and directional statistics Mardia & Jupp (2000); Banerjee et al. (2005), FalconMOT replaces such mixtures with Bayesian log-likelihood ratios against spatial Poisson clutter and a self-calibrated appearance concentration.

3 Methodology

FalconMOT addresses two central challenges of aerial tracking: association instability under fluctuating visual features, and trajectory fragmentation when miniature targets are missed by the detector. It combines an evidential association framework (SPEAR, Section 3.2) that regulates appearance reliability from temporal feature consistency with a localized feature-map recovery module (QAM, Section 3.3) that recovers lost targets within an uncertainty-bounded window verified by a distinctiveness margin; the full pipeline is given in Section 3.4 and Figure 2.

3.1 Preliminaries

At each time step t, the tracker receives a frame $I_t \in \mathbb{R}^{H \times W \times 3}$ from a moving UAV. To satisfy the latency and power budgets (< 20 W) of onboard systems (Appendix D.1), I_t is processed by a single-pass Joint Detection and Embedding (JDE) network built on a compact YOLO11s backbone (Jocher & Qiu, 2024). Its FPN-PAN neck prunes the coarse stride-32 level (P_5) and adds a high-resolution stride-4 level (P_2, $s = 4$) for miniature targets; a parallel ReID head on P_2 produces a dense embedding map $E_t \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times D}$ with $D = 128$, trained with an identity classification loss $\mathcal{L}_{\text{id}}$ (architecture in Appendix C, Figure 4).

Each candidate proposal j is a tuple $d_j = [\mathbf{c}_j, w_j, h_j, s_j, \mathbf{e}_j]$ with box center $\mathbf{c}_j = [x_j, y_j]^\top$, size (w_j, h_j), confidence $s_j \in [0, 1]$, and L_2-normalized embedding $\mathbf{e}_j \in \mathbb{R}^D$ sampled from E_t at the

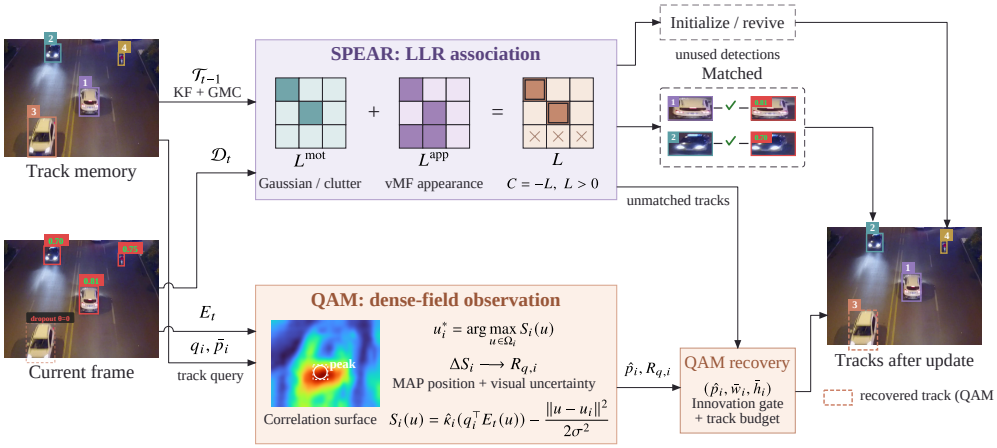

Figure 2: Overview of FalconMOT. SPEAR constructs an evidential cost matrix combining spatial Gaussian log-likelihood ratios (L^{mot}) and directional vMF appearance evidence (L^{app}). For tracks left unmatched after low-score IoU rescue, QAM queries dense feature map E_t within localized window $\Omega_i(R_i)$, verifying candidate peaks via distinctiveness margin ΔS_i to update Kalman covariance $\mathbf{R}_{q,i}$ without distractor confusion.

downsampled center:

$$\mathbf{e}_j = E_t(\mathbf{c}_j/s), \quad \|\mathbf{e}_j\|_2 = 1. \tag{1}$$

The detections at frame t form $\mathcal{D}_t = \{d_j\}_{j=1}^N$. Let $\mathcal{T}_{t-1} = \{\tau_i\}_{i=1}^M$ denote the M active trajectories up to frame $t-1$. Each τ_i maintains a state $\mathbf{x}_i = [\mathbf{p}_i^\top, w_i, h_i, \dot{\mathbf{p}}_i^\top, \dot{w}_i, \dot{h}_i]^\top \in \mathbb{R}^8$ with center $\mathbf{p}_i \in \mathbb{R}^2$, covariance $\mathbf{P}_i \in \mathbb{R}^{8\times 8}$, and a history buffer $\mathcal{H}_i = \{\mathbf{e}_{i,1}, \ldots, \mathbf{e}_{i,|\mathcal{H}_i|}\}$ of unit-norm ReID embeddings. At frame t, ego-motion is compensated via GMC (Aharon et al., 2022) and a Kalman filter (Kalman, 1960) yields the predicted center $\bar{\mathbf{p}}_i$ and position covariance $\bar{\mathbf{\Sigma}}_{p,i} \in \mathbb{R}^{2\times 2}$ (top-left block of $\bar{\mathbf{P}}_i$).

3.2 SPEAR: Self-Calibrating Probabilistic Evidence Association

Conventional multi-modal trackers (Wojke et al., 2017; Aharon et al., 2022; Zheng et al., 2025) fuse a spatial distance d_{mot} (Mahalanobis or IoU) and an appearance cosine distance d_{app} through a static linear combination:

$$C_{ij} = \alpha \cdot d_{\text{mot}}(i,j) + (1-\alpha) \cdot d_{\text{app}}(i,j), \tag{2}$$

where $\alpha \in [0,1]$ is a fixed global hyperparameter. Under motion blur, scale change, or partial occlusion, a fixed appearance weight keeps enforcing temporarily corrupted affinities, while during aggressive maneuvers spatial predictability drops even though appearance remains distinctive; a single global weight cannot adapt to either case per track and per frame.

SPEAR instead formulates association as adaptive probabilistic evidence evaluation. Since embeddings lie on the unit hypersphere $\mathbb{S}^{D-1}$, their temporal stability is quantified by the directional dispersion of $\mathcal{H}_i$ (Mardia & Jupp, 2000). We compute the sample mean resultant vector $\bar{\mathbf{R}}_i$, the unit mean direction vector $\mathbf{q}_i$, and the mean resultant length $\bar{r}_i$:

$$\bar{\mathbf{R}}_i = \frac{1}{|\mathcal{H}_i|}\sum_{k=1}^{|\mathcal{H}_i|} \mathbf{e}_{i,k}, \quad \mathbf{q}_i = \frac{\bar{\mathbf{R}}_i}{\|\bar{\mathbf{R}}_i\|_2}, \quad \bar{r}_i = \min\left(\|\bar{\mathbf{R}}_i\|_2,\ 1-10^{-6}\right) \in [0,1). \tag{3}$$

The mean resultant length $\bar{r}_i$ directly reflects directional consistency over time. Under stable observation, successive embeddings align tightly on the hypersphere, yielding $\bar{r}_i \to 1$. When the target undergoes visual corruption, embeddings scatter across the sphere, causing $\bar{r}_i \to 0$. Grounding this behavior in directional statistics and the von Mises-Fisher (vMF) distribution (Mardia & Jupp, 2000; Banerjee et al., 2005), the maximum likelihood estimate of the concentration parameter evaluates in closed form (Banerjee et al., 2005) (see Appendix A.2 for full derivation from Bessel function asymptotics) as:

$$\hat{\kappa}_i = \min\left(\frac{\bar{r}_i(D-\bar{r}_i^2)}{1-\bar{r}_i^2},\ \kappa_{\max}\right), \tag{4}$$

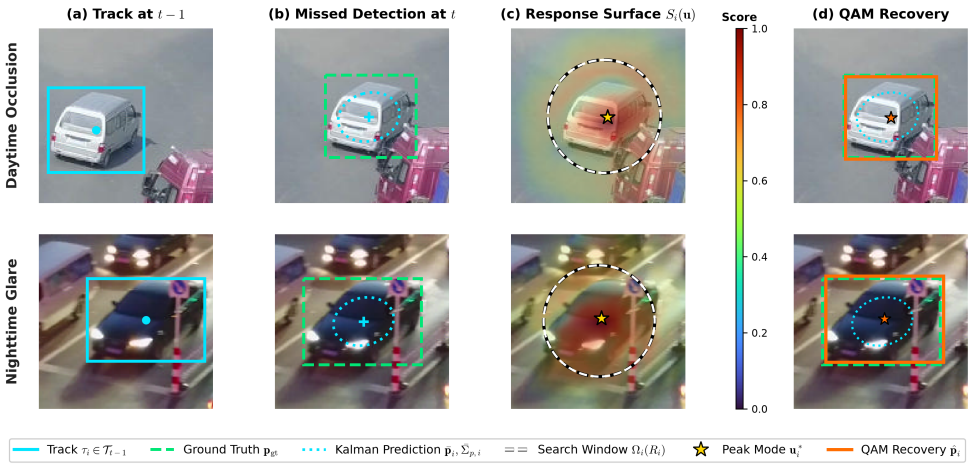

Figure 3: Visual-kinematic target recovery via QAM during detector dropouts (top: daytime tree canopy occlusion; bottom: nighttime vehicle headlight glare). (a) Active track appearance query $\mathbf{q}_i$; (b) Complete detection dropout; (c) Correlation surface $S_i(\mathbf{u})$ within localized window $\Omega_i(R_i)$; (d) Kalman recovery with distinctiveness margin verification ($\Delta S_i > \gamma_{\text{margin}}$).

where $D = 128$ is the embedding dimension, and $\kappa_{\max} = 3000.0$ bounds numerical stability. The concentration $\hat{\kappa}_i$ serves as an automatic, per-track reliability weight: it scales up when visual features are stable and automatically contracts toward zero when appearance is ambiguous.

Equipped with track mean direction $\mathbf{q}_i$, predicted center $\bar{\mathbf{p}}_i$, spatial covariance $\bar{\mathbf{\Sigma}}_{p,i}$, and reliability $\hat{\kappa}_i$, the joint evidence between track τ_i and candidate detection d_j is evaluated under a Bayesian hypothesis testing framework. We test the hypothesis that d_j and τ_i originate from the same physical target ($\mathcal{H}_1$) against ambient clutter ($\mathcal{H}_0$). The total Log-Likelihood Ratio $L_{ij} \triangleq \ln \frac{p(d_j|\tau_i,\mathcal{H}_1)}{p(d_j|\mathcal{H}_0)}$ decomposes into spatial motion and directional appearance components:

$$L_{ij} = L_{\text{mot}}(i,j) + L_{\text{app}}(i,j). \tag{5}$$

For spatial motion, under $\mathcal{H}_1$ the detection center $\mathbf{c}_j$ follows $\mathcal{N}(\mathbf{c}_j; \bar{\mathbf{p}}_i, \mathbf{S}_i)$ with measurement innovation covariance $\mathbf{S}_i = \bar{\mathbf{\Sigma}}_{p,i} + \sigma_{\text{det}}^2 \mathbf{I}_2$ ($\sigma_{\text{det}} = 2.0$ px). Under $\mathcal{H}_0$, candidate detections follow a spatial Poisson clutter process with density $\lambda_0 = \frac{\max(|\mathcal{D}_t|,1)}{A_{\text{img}}}$ ($A_{\text{img}} = H \times W$). The spatial log-likelihood ratio evaluates as (derivation in Appendix A.1):

$$L_{\text{mot}}(i,j) \triangleq \ln \frac{p(\mathbf{c}_j \mid \tau_i, \mathcal{H}_1)}{p(\mathbf{c}_j \mid \mathcal{H}_0)} = -\tfrac{1}{2} d_{\text{M}}^2(i,j) - \tfrac{1}{2} \ln \det(2\pi \mathbf{S}_i) - \ln \lambda_0, \tag{6}$$

where $d_{\text{M}}^2(i,j) = (\mathbf{c}_j - \bar{\mathbf{p}}_i)^\top \mathbf{S}_i^{-1} (\mathbf{c}_j - \bar{\mathbf{p}}_i)$ is the squared Mahalanobis distance. Pairs exceeding spatial gating ($d_{\text{M}}^2 > 13.82$, 99.9% quantile of χ_2^2) are masked with $L_{\text{mot}} = -\infty$. In parallel, under $\mathcal{H}_1$ the embedding $\mathbf{e}_j \in \mathbb{S}^{D-1}$ follows vMF likelihood $p(\mathbf{e}_j \mid \mathbf{q}_i, \hat{\kappa}_i, \mathcal{H}_1) \propto \exp(\hat{\kappa}_i \mathbf{q}_i^\top \mathbf{e}_j)$, while ambient clutter under $\mathcal{H}_0$ exhibits baseline cosine similarity $m_0 \triangleq \text{median}_{(i,j)}(\mathbf{q}_i^\top \mathbf{e}_j)$, yielding:

$$L_{\text{app}}(i,j) \triangleq \ln \frac{p(\mathbf{e}_j \mid \tau_i, \mathcal{H}_1)}{p(\mathbf{e}_j \mid \mathcal{H}_0)} = \hat{\kappa}_i \cdot \left(\mathbf{q}_i^\top \mathbf{e}_j - m_0\right). \tag{7}$$

When visual features degrade due to motion blur or occlusion, $\bar{r}_i \to 0$ causes $\hat{\kappa}_i \to 0$, naturally extinguishing the appearance term ($L_{\text{app}} \to 0$). Association smoothly falls back on spatial proximity (L_{mot}), preventing corrupted visual features from causing identity switches. The pairwise cost for Hungarian assignment (Kuhn, 1955) is formed as $C_{\text{SPEAR}}(i,j) = 1 - \sigma(L_{ij})$, where $\sigma(z) = \frac{1}{1+e^{-z}}$, restricting valid matching support to pairs exhibiting positive joint evidence ($L_{ij} > 0$). Matched tracks update their kinematic states and append observed embeddings $\mathbf{e}_j$ to history buffer $\mathcal{H}_i$.

Algorithm 1 FalconMOT Online Tracking Pipeline

Require: Video frame I_t, active tracks $\mathcal{T}_{t-1}$, appearance gallery $\mathcal{G}_c$.
Ensure: Updated active tracks $\mathcal{T}_t$, verified output detections $\mathcal{Y}_t$.
1: **Motion Update:** Predict states via Kalman filter and align camera motion: $\bar{\mathcal{T}}_{t-1} \leftarrow$ Kalman + GMC$(\mathcal{T}_{t-1}, I_t)$
2: **Feature Extraction:** Extract proposals $\mathcal{D}_t^{\text{high}}, \mathcal{D}_t^{\text{low}}$ and dense ReID map $E_t \leftarrow$ Detector(I_t)
3: **Stage 1 (SPEAR):** Associate $\bar{\mathcal{T}}_{t-1}$ and $\mathcal{D}_t^{\text{high}}$ via evidential cost (Eq. equation 5) $\rightarrow$ matched $\mathcal{M}_1$, unmatched $\mathcal{T}_{\text{rem}}, \mathcal{D}_{\text{rem}}^{\text{high}}$
4: **Stage 2 (IoU):** Associate remaining tracks $\mathcal{T}_{\text{rem}}$ with $\mathcal{D}_t^{\text{low}}$ via spatial IoU $\rightarrow$ matched $\mathcal{M}_2$, lost tracks $\mathcal{T}_{\text{lost}}$
5: **Stage 3 (QAM):** Recover lost tracks $\mathcal{T}_{\text{lost}}$ on feature map E_t with margin $\Delta S_i > \gamma_{\text{margin}}$ (Eq. equation 10) $\rightarrow$ recovered $\mathcal{M}_3$
6: **Stage 4 (Lifecycle Management):** Update active tracks $\mathcal{T}_t \leftarrow \mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3$; retire lost tracks exceeding $N_{\text{lost}} = 30$ into $\mathcal{G}_c$
7: **Track Management:** Initialize new tracks from unassigned $\mathcal{D}_{\text{rem}}^{\text{high}}$ or revive from gallery $\mathcal{G}_c$
8: **return** Active tracks $\mathcal{T}_t$ and verified output detections $\mathcal{Y}_t$

3.3 QAM: Query Appearance-Motion Target Recovery

Tracks left unassociated after discrete matching form the lost set $\mathcal{T}_{\text{lost}}$. Although E_t still carries the visual evidence of targets that fell below the detection threshold, correlating across the entire map is prone to visually similar distractors. QAM therefore restricts querying to a kinematically bounded window $\Omega_i(R_i)$ anchored at $\mathbf{u}_i = \bar{\mathbf{p}}_i/s$ on E_t. The search radius R_i dynamically scales with 3σ spatial positional uncertainty:

$$R_i = \max\left(r_{\min},\ 3\sqrt{\text{Tr}(\bar{\mathbf{\Sigma}}_{p,i})}/s\right), \tag{8}$$

where $r_{\min} = 4$ feature cells (16 px). Within window $\Omega_i(R_i) = \{\mathbf{u} \in \mathbb{Z}^2 \mid \|\mathbf{u} - \mathbf{u}_i\|_2 \le R_i\}$, QAM evaluates a scalar joint visual-spatial response score $S_i(\mathbf{u}) \in \mathbb{R}$ at each grid location $\mathbf{u} = [u, v]^\top$ (derivation in Appendix A.3):

$$S_i(\mathbf{u}) = \hat{\kappa}_i\left(\mathbf{q}_i^\top E_t(\mathbf{u})\right) - \tfrac{1}{2}(\mathbf{u} - \mathbf{u}_i)^\top\left(\tfrac{1}{s^2}\bar{\mathbf{\Sigma}}_{p,i}\right)^{-1}(\mathbf{u} - \mathbf{u}_i), \quad \mathbf{u} \in \Omega_i(R_i). \tag{9}$$

Across search window $\Omega_i(R_i)$, the primary candidate recovery location is identified as the peak correlation coordinate $\mathbf{u}_i^* = \arg\max_{\mathbf{u}\in\Omega_i(R_i)} S_i(\mathbf{u})$. To ensure that peak $\mathbf{u}_i^*$ represents the genuine target rather than an ambiguous background artifact, we verify its distinctiveness relative to the surrounding response landscape. Defining $\mathcal{N}_3(\mathbf{u}_i^*) = \{\mathbf{u} \in \Omega_i(R_i) \mid \|\mathbf{u} - \mathbf{u}_i^*\|_\infty \le 1\}$ as the 3×3 grid neighborhood centered at the primary peak, the distinctiveness margin ΔS_i is computed as:

$$\Delta S_i = S_i(\mathbf{u}_i^*) - \max_{\mathbf{u}\in\Omega_i(R_i)\setminus\mathcal{N}_3(\mathbf{u}_i^*)} S_i(\mathbf{u}). \tag{10}$$

A candidate peak is accepted as a valid target recovery only when its distinctiveness exceeds margin threshold $\gamma_{\text{margin}} = 0.25$ ($\Delta S_i > \gamma_{\text{margin}}$). Upon acceptance, the grid coordinate is mapped back to the image plane as $\hat{\mathbf{p}}_i = s \cdot \mathbf{u}_i^* \in \mathbb{R}^2$, and the track state is updated via a Kalman measurement update using an uncertainty-calibrated observation covariance $\mathbf{R}_{q,i} \in \mathbb{R}^{2\times 2}$ (proof in Appendix A.3):

$$\mathbf{R}_{q,i} = (w_i h_i) \cdot \exp(-\beta \Delta S_i)\mathbf{I}_2, \quad \beta = 5.0, \tag{11}$$

where $(w_i h_i)$ scales covariance with target bounding box area, and $\beta = 5.0$ controls exponential uncertainty tightening for sharp peaks. If competing peaks produce an ambiguous response landscape ($\Delta S_i \le \gamma_{\text{margin}}$), the recovery proposal is rejected, and the trajectory continues coasting under motion prediction.

3.4 Unified Tracking Pipeline and Lifecycle Management

Algorithm 1 summarizes the per-frame procedure. High-confidence proposals are first matched via SPEAR (Eq. equation 5); unmatched tracks are then matched to low-confidence proposals by IoU (≥ 0.8) (Zhang et al., 2022) to rescue occluded detections without corrupting appearance histories; tracks still unassociated undergo QAM recovery, and lifecycle management handles initialization, survival, and retirement.

Track Lifecycle and Identity Revival. Following established online tracking conventions (Zhang et al., 2022; Aharon et al., 2022), lost tracks are retained in a temporal buffer for up to $N_{\text{lost}} = 30$

Table 1: **Quantitative tracking comparison on the VisDrone2019-MOT benchmark** (12,000 frames across 17 video sequences, evaluated across 5 object categories via the official TrackEval toolkit). All methods are evaluated under standardized protocols. Bold and underlined entries denote the best and second-best results, respectively.

Method	Backbone	IDF1 ↑	MOTA ↑	HOTA ↑	AssA ↑	MT ↑	ML ↓	IDs ↓	Frag ↓
ByteTrack (Zhang et al., 2022)	YOLOX-X	37.0	35.7	32.4	38.6	–	–	2,168	4,077
UAVMOT (Liu et al., 2022)	Custom DLA-34	51.0	36.1	41.5	46.2	520	574	2,775	3,892
OC-SORT (Cao et al., 2023)	YOLOX-X	50.4	39.6	42.1	48.9	–	–	986	2,913
GCEVT (Wu et al., 2023)	ResNet-50	50.6	34.5	39.8	44.7	520	612	841	3,420
GLOA (Shi et al., 2023)	CenterNet	46.2	39.1	38.7	42.1	581	824	4,426	5,110
FOLT (Yao et al., 2023)	ResNet-50	56.9	42.1	45.8	51.2	–	–	<u>800</u>	3,250
PID-MOT (Lv et al., 2024)	DLA-34	50.2	33.0	39.1	43.8	686	424	3,529	4,890
AHOR-ReID (Jin et al., 2024)	YOLOv5s	56.4	42.5	46.2	52.0	–	–	810	**3,110**
STCMOT (Ma et al., 2024)	YOLOv5s	52.0	41.2	43.9	47.8	667	453	3,984	4,720
DroneMOT (Wang et al., 2024)	YOLOv7-Tiny	58.6	43.7	48.1	54.3	689	<u>397</u>	1,112	3,450
DFA-MOT (Zheng et al., 2025)	ResNet-50	58.3	42.9	47.9	53.6	518	523	<u>792</u>	3,190
MM-Tracker (Yao et al., 2025)	ResNet-50	58.3	44.7	48.5	55.1	–	–	–	–
AMOT (Ma et al., 2025)	DLA-34	<u>61.4</u>	<u>46.0</u>	<u>49.8</u>	<u>57.2</u>	<u>716</u>	413	1,063	<u>3,150</u>
FalconMOT (Ours)	YOLO11s-P2	**69.1**	**55.6**	**53.6**	**63.1**	**914**	**316**	**702**	4,146

frames (1.0 s at 30 FPS). When an unassociated track exceeds N_{lost} frames, it is retired from active tracking into the long-term appearance gallery $\mathcal{G}_c$, preserving its historical embedding $\mathbf{q}_i$. If an unassigned proposal re-emerges with high cosine similarity ($\mathbf{q}_i^\top \mathbf{e}_j > 0.75$), the trajectory is revived under its original identity, preventing fragmentation across extended dropouts. Detailed mathematical derivations and proofs are provided in Appendix A, and complete hyperparameter calibrations are cataloged in Appendix F.1.

4 Experiments

4.1 Experimental Settings

Datasets and Metrics. We evaluate on **VisDrone2019-MOT** (Zhu et al., 2021) (test-dev, 17 sequences, 12,000 frames from 14 cities), which features severe ego-motion, steep tilt angles, day/night illumination changes, dense urban congestion, and miniature targets (often $< 16 \times 16$ pixels) over five classes (*pedestrian*, *car*, *van*, *truck*, *bus*), and on **UAVDT** (Du et al., 2018) (20 sequences, 16,592 frames) under direct *zero-shot transfer*: models trained on VisDrone are evaluated without fine-tuning or re-calibration. Following the official `TrackEval` toolkit, we report HOTA (Luiten et al., 2021) with DetA, AssA, and LocA, IDF1 (Ristani et al., 2016), and CLEAR-MOT metrics (Bernardin & Stiefelhagen, 2008) (MOTA, MT, ML, IDs, Frag).

Implementation Details. The YOLO11s backbone (Jocher & Qiu, 2024) (7.73 M parameters, 54.45 GFLOPs at 960×544) uses pyramid levels $[P_2, P_3, P_4]$ ($s \in \{4, 8, 16\}$) with P_5 pruned, and a ReID head on P_2 producing 128-d embeddings. The joint network is trained on the VisDrone training split with AdamW (learning rate 1×10^{-3}, cosine schedule, weight decay 5×10^{-4}) for 30 epochs at batch size 16, using CIoU and DFL for box regression, BCE for classification, and a supervised directional contrastive loss on $\mathbb{S}^{D-1}$ for embeddings. All hyperparameters are fixed across benchmarks: $\tau_{\text{high}} = 0.40$, $\tau_{\text{low}} = 0.10$, $\kappa_{\max} = 3000.0$, $R_i \geq 4.0$ cells, $\gamma_{\text{margin}} = 0.25$, $\beta = 5.0$, and $N_{\text{lost}} = 30$ (Appendix F.1). End-to-end latency (preprocessing, detection, ReID, GMC, SPEAR, and QAM) is measured at batch size 1 on an NVIDIA RTX 3060 GPU (PyTorch FP32 and TensorRT FP16) and on an embedded NVIDIA Jetson Orin NX (16 GB, 20 W).

4.2 Benchmark Evaluation

Results on VisDrone2019-MOT. As shown in Table 1, FalconMOT achieves state-of-the-art performance with 69.1% IDF1, 55.6% MOTA, 53.6 HOTA, and 63.1% AssA. Compared with the previous best aerial tracker AMOT (Ma et al., 2025) (61.4% IDF1, 46.0% MOTA, 49.8 HOTA, 57.2% AssA), it gains +7.7% IDF1, +9.6% MOTA, +3.8 HOTA, and +5.9% AssA, reduces identity switches

Table 2: **Zero-shot cross-dataset generalization on the UAVDT benchmark** (16,592 frames across 20 video sequences). All methods are evaluated under direct zero-shot transfer without domain fine-tuning. Bold and underlined entries denote best and second-best results, respectively.

Method	Backbone	IDF1↑	MOTA↑	HOTA↑	AssA↑	IDs↓
ByteTrack (Zhang et al., 2022)	YOLOX-X	65.4	38.6	48.2	56.8	620
FairMOT (Zhang et al., 2021)	DLA-34	61.2	35.4	45.1	53.0	1,420
BoT-SORT (Aharon et al., 2022)	YOLOX-X	63.8	39.1	47.9	54.6	580
UAVMOT (Liu et al., 2022)	Custom DLA-34	67.3	42.8	50.4	58.9	452
AMOT (Ma et al., 2025)	DLA-34	<u>68.9</u>	<u>44.1</u>	<u>51.8</u>	<u>61.2</u>	<u>388</u>
FalconMOT (Ours)	YOLO11s-P2	**70.6**	**46.1**	**54.1**	**65.0**	**194**

Table 3: **Component ablation on VisDrone2019-MOT** (12,000 frames across 17 video sequences). All configurations operate under strictly identical scale-aligned detector proposals and 128-dimensional ReID embeddings from our shared YOLO11s-P2 backbone, with camera motion compensation (GMC) active across all rows to isolate tracker association and recovery dynamics from platform ego-motion. Throughput (FPS) is evaluated on an NVIDIA RTX 3060 under PyTorch batch size 1 (the full engine reaches 117.2 FPS under TensorRT FP16, and 25.0 FPS on Jetson Orin NX; see Appendix D). Bold entries denote the best result in each column.

Configuration	SPEAR	QAM	Gallery	IDF1↑	MOTA↑	HOTA↑	AssA↑	IDs↓	Frag↓	FPS↑
(1) Standard Baseline (IoU + Kalman)				68.1	55.1	52.9	61.7	987	5,252	**52.4**
(2) + Evidential Match (SPEAR)	✓			68.2	55.0	53.0	62.1	1,053	5,435	50.1
(3) + Target Recovery (QAM)	✓	✓		68.4	55.1	53.2	62.3	927	5,020	48.0
(4) Full FalconMOT Engine	✓	✓	✓	**69.1**	**55.6**	**53.6**	**63.1**	**702**	**4,146**	46.5

from 1,063 to 702 (34.0%), and increases mostly-tracked trajectories (914 vs. 716 MT). Ground trackers degrade because platform motion violates constant-velocity assumptions, and recent aerial trackers that rely on unconstrained global correlation latch onto similar distractors in crowded scenes; FalconMOT instead balances spatial and visual evidence adaptively (SPEAR) and bounds recovery to verified local windows (QAM), while running at 46.5 FPS on GPU, 25.0 FPS on Jetson Orin NX, and 117.2 FPS under TensorRT FP16.

Zero-Shot Generalization on UAVDT. Under direct zero-shot transfer (Table 2), FalconMOT achieves 70.6% IDF1, 46.1% MOTA, 65.0% AssA, and 54.1 HOTA, and suppresses identity switches to 194 over 16,592 frames, a 50.0% reduction vs. AMOT (388) and 57.1% vs. UAVMOT (452), showing that closed-form directional statistics and bounded recovery generalize across flight altitudes, viewpoints, and vehicle densities without dataset-specific tuning.

4.3 Ablation Studies and Analysis

Component Analysis. Table 3 evaluates the components under identical detector proposals. SPEAR (Config. 2) stabilizes association on $\mathbb{S}^{D-1}$ (AssA 62.1%) with negligible overhead (50.1 FPS). QAM (Config. 3) recovers dropped detections from the dense feature map (79,627 $\rightarrow$ 78,579 FN) while its bounded window and margin prevent distractor latching, reducing identity switches by 12.0% (1,053 $\rightarrow$ 927) and fragmentations by 7.6%. The long-term gallery $\mathcal{G}_c$ (Config. 4) bridges extended occlusions, yielding the full engine with 69.1% IDF1, 55.6% MOTA, 53.6 HOTA, and 702 IDs (33.3% fewer than SPEAR alone and 28.9% fewer than the baseline), while keeping false alarms bounded (26,387 FP) at 46.5 FPS on GPU and 25.0 FPS on Jetson Orin NX.

Dynamic Evidential Fusion vs. Static Linear Mixtures. Table 4 shows that a static mixture ($C = 0.50d_{\text{mot}} + 0.50d_{\text{app}}$) breaks down under airborne dynamics (55.99% IDF1, 7,504 IDs) due to the scale mismatch between unbounded Euclidean and hyperspherical cosine distances. Isolated appearance and motion LLRs reach 67.98% and 67.64% IDF1, and a fixed concentration ($\kappa = 1000$) reaches 68.11%. The closed-form $\hat{\kappa}_i$ falls back smoothly onto kinematic evidence when embeddings fluctuate under blur or glare, lifting IDF1 to 68.24% and MOTA to 55.01% (1,053 IDs).

Recovery Search Scope and Margin Verification. Table 5 quantifies the trade-off between recall and false positives. Unconstrained global querying lowers FN to 76,887 but inflates FP by +22.6%

Table 4: **Modality isolation and fusion mechanisms on VisDrone2019-MOT**. Comparison between static linear cost mixtures $C = \alpha d_{\text{mot}} + (1-\alpha) d_{\text{app}}$, isolated single-modality baselines, fixed-concentration fusion, and proposed SPEAR evidential fusion on $\mathbb{S}^{D-1}$ under identical detector proposals.

Configuration	Cost Formulation	IDF1↑	MOTA↑	HOTA↑	AssA↑	IDs↓
Static Linear ($\alpha = 0.50$)	$0.50 d_{\text{mot}} + 0.50 d_{\text{app}}$	55.99	47.75	47.64	52.77	7,504
Motion-Only (L_{mot})	Bayesian Kinematic LLR	67.64	54.89	52.62	61.18	1,254
Appearance-Only (L_{app})	Directional vMF LLR on $\mathbb{S}^{D-1}$	67.98	54.74	53.15	62.54	1,009
Fixed Concentration vMF	Fixed $\kappa = 1000$ on $\mathbb{S}^{D-1}$	68.11	54.98	53.00	62.08	1,117
SPEAR Evidential Fusion (Ours)	Closed-Form $\hat{\kappa}_i$ on $\mathbb{S}^{D-1}$	**68.24**	**55.01**	**52.98**	**62.07**	**1,053**

Table 5: **Impact of QAM search space confinement and distinctiveness margin verification on VisDrone2019-MOT**. Comparing discrete-only tracking against unconstrained global feature querying, localized search windows, and margin-verified recovery across complete detection and association metrics.

Search Regime	Verification	IDF1↑	MOTA↑	HOTA↑	DetA↑	AssA↑	LocA↑	IDs↓	Frag↓	FP↓	FN↓
Discrete-Only (No QAM)	None	68.24	55.01	52.98	45.66	62.07	**82.78**	1,053	5,435	**23,858**	79,627
Global Unconstrained	None ($\Delta S > 0$)	66.85	53.82	52.45	44.91	61.91	82.58	1,166	4,663	29,262	**76,887**
Localized Window	None ($\Delta S > 0$)	68.39	55.06	**53.21**	**45.80**	**62.39**	82.68	**922**	**4,569**	26,330	77,184
Localized + Margin	$\Delta S_i > 0.25$	**68.43**	**55.13**	53.16	45.77	62.31	82.74	927	5,020	24,773	78,579

(23,858 → 29,262) and identity switches to 1,166 by latching onto distant distractors. Restricting queries to the 3σ window $\Omega_i(R_i)$ reduces FP to 26,330 and IDs to 922. Adding the margin check ($\Delta S_i > 0.25$) bounds FP to 24,773 (only $+3.8\%$ over discrete tracking) while recovering 1,048 missed detections, reducing identity switches by 12.0% and fragmentations by 7.6% relative to discrete-only tracking.

Qualitative Recovery. Figure 3 shows two characteristic failure modes, tree-canopy occlusion and headlight glare, where the detector drops the target entirely (b). QAM queries E_t within $\Omega_i(R_i)$ (c), and the resulting unimodal peak satisfies $\Delta S_i > \gamma_{\text{margin}}$, enabling Kalman recovery without identity drift (d). Further qualitative comparisons against MMTracker and AMOT are given in Appendix E (Figure 5).

Table 6: **Cross-tracker generality and algorithmic portability**. Transplanting the proposed SPEAR evidential association into established tracking frameworks (ByteTrack and BoT-SORT) under identical YOLO11s-P2 detection proposals on VisDrone2019-MOT. On both pure motion cascaded tracking (ByteTrack) and appearance-augmented tracking (BoT-SORT), evidential association yields substantial gains in association fidelity and identity persistence.

Tracker / Configuration	Association Engine	Params	IDF1↑	MOTA↑	HOTA↑	IDs↓
ByteTrack Native	Two-Stage IoU Cascade	7.73 M	57.85	53.93	47.79	3,478
↪ **+ SPEAR**	Stage-1 Evidential Association	7.73 M	**61.79**	**54.81**	**49.69**	**2,474**
BoT-SORT Native	Linear Cost Mixture ($\alpha = 0.5$)	7.73 M	56.83	48.71	47.74	7,340
↪ **+ SPEAR**	Evidential Manifold Association	7.73 M	**62.15**	**51.84**	**49.88**	**2,560**

Cross-Tracker Portability. Transplanting SPEAR into ByteTrack (Zhang et al., 2022) and BoT-SORT (Aharon et al., 2022) under identical YOLO11s-P2 detections (Table 6) improves both. For BoT-SORT, replacing its linear mixture raises IDF1 by $+5.32\%$ to 62.15% and cuts identity switches by 65.1% (7,340 → 2,560); for the motion-only ByteTrack, SPEAR in Stage 1 lifts IDF1 by $+3.94\%$ to 61.79% and reduces identity switches by 28.9% (3,478 → 2,474). The full FalconMOT engine remains decisively stronger (69.1% IDF1, 702 IDs), as association replacement alone lacks QAM recovery and gallery revival.

5 Conclusion

In this paper, we presented FalconMOT, an online multi-object tracking framework designed for the complex motion and visual dynamics of UAV perception. To address key failure modes of aerial

tracking association instability under platform ego-motion and trajectory fragmentation from detector dropouts on small targets FalconMOT introduces two training-free mechanisms: SPEAR models historical appearance on the unit hypersphere via von Mises-Fisher directional statistics to balance cues in closed form, while QAM recovers missed detections through continuous mode-seeking within kinematically bounded windows with margin verification.

Evaluations on VisDrone2019-MOT and UAVDT demonstrate that FalconMOT achieves state-of-the-art tracking accuracy (69.1% IDF1, 55.6% MOTA, 53.6 HOTA), reducing identity switches to 702 and trajectory fragmentations by 21.1% ($4{,}146$ vs. $5{,}252$) under controlled ablations. Direct zero-shot transfer on UAVDT yields 70.6% IDF1, 46.1% MOTA, 54.1 HOTA, and a 50.0% reduction in identity switches compared to AMOT (194 vs. 388), confirming robust cross-domain generalization. Algorithmic integration into ByteTrack and BoT-SORT validates modular portability, while inference benchmarks demonstrate real-time execution. In future work, we plan to extend this probabilistic framework to collaborative multi-UAV tracking and autonomous aerial navigation.

REFERENCES

Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracker. *arXiv preprint arXiv:2206.14651*, 2022.

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6: 1345–1382, 2005.

Yaakov Bar-Shalom, X Rong Li, and Thiagalingam Kirubarajan. *Estimation with Applications to Tracking and Navigation: Theory Algorithms and Software*. John Wiley & Sons, 2001.

Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008.

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE International Conference on Image Processing (ICIP)*, pp. 3464–3468, 2016.

Yvo Boers and Hans Driessen. Track-before-detect algorithms for low SNR targets. *Signal Processing*, 84(10):1787–1802, 2004.

Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9686–9696, 2023.

Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. MOT20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020.

Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *European Conference on Computer Vision (ECCV)*, 2018.

Haofeng Jin, Xi Nie, Yubin Yan, Xiaochun Chen, Ziyang Zhu, and Dong Qi. AHOR: Online multi-object tracking with authenticity hierarchizing and occlusion recovery. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9):8253–8265, 2024.

Glenn Jocher and Jing Qiu. Ultralytics YOLO11, 2024. URL https://github.com/ultralytics/ultralytics. Accessed: 2025.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125, 2017.

Kai Liu, Zhihang Fu, Sheng Jin, Ze Chen, Fan Zhou, Rongxin Jiang, Yaowu Chen, and Jieping Ye. ESOD: Efficient small object detection on high-resolution images. *arXiv preprint arXiv:2407.16424*, 2024.

Shuai Liu, Xin Li, Huchuan Lu, and You He. Multi-object tracking meets moving UAV. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8876–8885, 2022.

Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision (IJCV)*, 2021.

Wenyu Lv, Ning Zhang, Jing Zhang, and Dong Zeng. One-shot multiple object tracking with robust ID preservation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):4473–4488, 2024.

Jianbo Ma, Chuanming Tang, Fei Wu, Can Zhao, Jianlin Zhang, and Zhiyong Xu. STCMOT: Spatio-temporal cohesion learning for UAV-based multiple object tracking. In *IEEE International*

Conference on Multimedia and Expo (ICME), 2024.

Jianbo Ma, Hui Luo, Qi Chen, Yuankai Qi, Yumei Sun, Amin Beheshti, Jianlin Zhang, and Ming-Hsuan Yang. Tracking the unstable: Appearance-guided motion modeling for robust multi-object tracking in UAV-captured videos. *arXiv preprint arXiv:2508.01730*, 2025.

Kanti V Mardia and Peter E Jupp. *Directional Statistics*. John Wiley & Sons, 2000.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8844–8854, 2022.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision (ECCV) Workshops*, 2016.

Mattia Segu, Luigi Piccinelli, Siyuan Li, Yung-Hsu Yang, Bernt Schiele, and Luc Van Gool. Samba: Synchronized set-of-sequences modeling for multiple object tracking. In *International Conference on Learning Representations (ICLR)*, 2025.

Lukui Shi, Qingrui Zhang, Bin Pan, Jun Zhang, and Yuanchao Su. 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, 2023.

Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. TransTrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

Jinwang Wang, Chang Xu, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. A normalized Gaussian Wasserstein distance for tiny object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022.

Peng Wang, Yongcai Wang, and Deying Li. DroneMOT: Drone-based multi-object tracking considering detection difficulties and simultaneous moving of drones and objects. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7397–7404, 2024.

Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. Towards realistic UAV vision-language navigation: Platform, benchmark, and methodology. In *International Conference on Learning Representations (ICLR)*, 2025.

Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *European conference on computer vision*, pp. 107–122. Springer, 2020.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pp. 3645–3649. IEEE, 2017.

Han Wu, Zhiwei He, and Mingyu Gao. GCEVT: Learning global context embedding for vehicle tracking in unmanned aerial vehicle videos. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023.

Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12352–12361, 2021.

Mufeng Yao, Jiaqi Wang, Jinlong Peng, Min Chi, and Chao Liu. FOLT: Fast multiple object tracking from UAV-captured videos based on optical flow. In *ACM International Conference on Multimedia (ACM MM)*, 2023.

Mufeng Yao, Jinlong Peng, Qingdong He, Bo Peng, Hao Chen, Mingmin Chi, Chao Liu, and Jon Atli Benediktsson. MM-Tracker: Motion mamba with margin loss for UAV-platform multiple object tracking. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. MOTR: End-to-end multiple-object tracking with TRansformer. In *European Conference on Computer Vision (ECCV)*, pp. 659–675, 2022.

Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International journal of computer vision*, 129(11):3069–3087, 2021.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision (ECCV)*, 2022.

Yining Zheng, Chang He, Xi Chen, Hong Zhang, Teng Qu, and Dong Wang. DFA-MOT: A dynamic field-aware multi-object tracking framework for unmanned aerial vehicles. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1):112–124, 2025.

Xingyi Zhou, Vladlen Koltun, and Philipp Kr"ahenb"uhl. Tracking objects as points. In *European Conference on Computer Vision (ECCV)*, pp. 474–490, 2020.

Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

A Core Mathematical Formulations and Derivations

This section provides rigorous mathematical derivations and proofs for the theoretical claims, closed-form approximations, and decision boundaries presented in Section 3.

A.1 Bayesian Evidential Log-Likelihood Ratio Derivation (L_{MOT} and L_{APP})

Data association between active trajectory τ_i and candidate detection $d_j = [\mathbf{c}_j, w_j, h_j, s_j, \mathbf{e}_j]$ is formulated as a binary composite hypothesis testing problem:

$$\mathcal{H}_1 : d_j \text{ originates from target } \tau_i, \quad \mathbf{c}_j = \mathbf{p}_i + \mathbf{v}_{\text{det}}, \quad \mathbf{e}_j \sim \text{vMF}(\mathbf{q}_i, \hat{\kappa}_i), \tag{12}$$

$$\mathcal{H}_0 : d_j \text{ is an unassociated false alarm generated by ambient clutter}, \tag{13}$$

where target centroid $\mathbf{p}_i \sim \mathcal{N}(\bar{\mathbf{p}}_i, \bar{\mathbf{\Sigma}}_{p,i})$, detection measurement noise $\mathbf{v}_{\text{det}} \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{det}}^2 \mathbf{I}_2)$, and ReID feature noise on $\mathbb{S}^{D-1}$ is conditionally independent of spatial localization given state τ_i. Under this conditional independence, the joint likelihood factors as $p(d_j \mid \tau_i, \mathcal{H}_m) = p(\mathbf{c}_j \mid \tau_i, \mathcal{H}_m) p(\mathbf{e}_j \mid \tau_i, \mathcal{H}_m)$ for $m \in \{0, 1\}$. The Log-Likelihood Ratio (LLR) decomposes additively into spatial motion and directional appearance components:

$$L_{ij} \triangleq \ln \frac{p(d_j \mid \tau_i, \mathcal{H}_1)}{p(d_j \mid \mathcal{H}_0)} = \underbrace{\ln \frac{p(\mathbf{c}_j \mid \tau_i, \mathcal{H}_1)}{p(\mathbf{c}_j \mid \mathcal{H}_0)}}_{L_{\text{mot}}(i,j)} + \underbrace{\ln \frac{p(\mathbf{e}_j \mid \tau_i, \mathcal{H}_1)}{p(\mathbf{e}_j \mid \mathcal{H}_0)}}_{L_{\text{app}}(i,j)}. \tag{14}$$

Derivation of Spatial Motion LLR $L_{\text{mot}}(i, j)$. Under $\mathcal{H}_1$, marginalizing over latent position $\mathbf{p}_i$ yields the convolution of two Gaussian densities:

$$p(\mathbf{c}_j \mid \tau_i, \mathcal{H}_1) = \int_{\mathbb{R}^2} \mathcal{N}(\mathbf{c}_j; \mathbf{p}, \sigma_{\text{det}}^2 \mathbf{I}_2) \mathcal{N}(\mathbf{p}; \bar{\mathbf{p}}_i, \bar{\mathbf{\Sigma}}_{p,i}) d\mathbf{p} = \frac{1}{2\pi\sqrt{\det(\mathbf{S}_i)}} \exp\left(-\frac{1}{2} d_{\text{M}}^2(i,j)\right), \tag{15}$$

where $\mathbf{S}_i = \bar{\mathbf{\Sigma}}_{p,i} + \sigma_{\text{det}}^2 \mathbf{I}_2$ is the innovation covariance and $d_{\text{M}}^2(i,j) \triangleq (\mathbf{c}_j - \bar{\mathbf{p}}_i)^\top \mathbf{S}_i^{-1} (\mathbf{c}_j - \bar{\mathbf{p}}_i)$ is the squared Mahalanobis distance. Under $\mathcal{H}_0$, false alarms follow a spatial Poisson point process (Bar-Shalom et al., 2001) with spatial intensity $\lambda_0 = \frac{\max(|\mathcal{D}_t|, 1)}{A_{\text{img}}}$ over canvas $A_{\text{img}} = H \times W$. Since the spatial clutter density per unit area under $\mathcal{H}_0$ is λ_0, the likelihood ratio evaluates in closed form as:

$$L_{\text{mot}}(i,j) = \ln \left[\frac{p(\mathbf{c}_j \mid \tau_i, \mathcal{H}_1)}{\lambda_0}\right] = -\frac{1}{2} d_{\text{M}}^2(i,j) - \frac{1}{2} \ln\det(2\pi \mathbf{S}_i) - \ln \lambda_0, \tag{16}$$

which decomposes into spatial Mahalanobis penalty $-\frac{1}{2} d_{\text{M}}^2$, dynamic kinematic volume discount $-\frac{1}{2} \ln\det(2\pi\mathbf{S}_i)$, and clutter rejection threshold $-\ln \lambda_0$.

Derivation of Directional Appearance LLR $L_{\text{app}}(i, j)$. Because unit-normalized embeddings reside on $\mathbb{S}^{D-1}$, true observations under $\mathcal{H}_1$ follow a von Mises–Fisher distribution $p(\mathbf{e}_j \mid \tau_i, \mathcal{H}_1) = c_D(\hat{\kappa}_i) \exp(\hat{\kappa}_i \mathbf{q}_i^\top \mathbf{e}_j)$, where $c_D(\hat{\kappa}_i) = \frac{\hat{\kappa}_i^{D/2-1}}{(2\pi)^{D/2} I_{D/2-1}(\hat{\kappa}_i)}$. Under $\mathcal{H}_0$, random clutter embeddings exhibit an empirical baseline affinity $m_0 \triangleq \text{median}_{(i,j)}(\mathbf{q}_i^\top \mathbf{e}_j)$, yielding null density $p(\mathbf{e}_j \mid \mathcal{H}_0) = c_D(\hat{\kappa}_i) \exp(\hat{\kappa}_i m_0)$. Taking their ratio cancels the intractable normalization constant $c_D(\hat{\kappa}_i)$ identically:

$$L_{\text{app}}(i,j) = \ln \left(\frac{c_D(\hat{\kappa}_i) \exp(\hat{\kappa}_i \mathbf{q}_i^\top \mathbf{e}_j)}{c_D(\hat{\kappa}_i) \exp(\hat{\kappa}_i m_0)}\right) = \hat{\kappa}_i \cdot \left(\mathbf{q}_i^\top \mathbf{e}_j - m_0\right). \tag{17}$$

This derivation proves that visual affinity exceeding clutter floor m_0 accumulates positive evidence, while similarities below m_0 penalize assignment, with authority governed by concentration $\hat{\kappa}_i$.

A.2 Derivation of Closed-Form Appearance Concentration $\hat{\kappa}_i$ via Bessel Recurrence

Let $\mathcal{H}_i = \{\mathbf{e}_{i,1}, \ldots, \mathbf{e}_{i,K}\} \subset \mathbb{S}^{D-1}$ denote the appearance history buffer of track τ_i. The joint vMF log-likelihood is $\ln \mathcal{L}(\kappa_i, \mathbf{q}_i; \mathcal{H}_i) = K \ln c_D(\kappa_i) + K \kappa_i \mathbf{q}_i^\top \bar{\mathbf{R}}_i$, where $\bar{\mathbf{R}}_i = \frac{1}{K} \sum_{k=1}^{K} \mathbf{e}_{i,k}$ is the

sample mean resultant vector, with resultant length $\bar{r}_i = \|\bar{\mathbf{R}}_i\|_2 \in [0, 1)$ and prototype $\mathbf{q}_i = \bar{\mathbf{R}}_i/\bar{r}_i$. Differentiating $\ln \mathcal{L}$ with respect to κ_i:

$$\frac{\partial \ln \mathcal{L}}{\partial \kappa_i} = K \frac{d}{d\kappa_i} \ln c_D(\kappa_i) + K\bar{r}_i = 0 \iff -\frac{d}{d\kappa_i} \ln c_D(\kappa_i) = \bar{r}_i. \tag{18}$$

Substituting $\ln c_D(\kappa_i) = \left(\frac{D}{2} - 1\right) \ln \kappa_i - \frac{D}{2} \ln(2\pi) - \ln I_{D/2-1}(\kappa_i)$ and applying the modified Bessel recurrence relation $I'_\nu(z) = I_{\nu+1}(z) + \frac{\nu}{z} I_\nu(z)$ with $\nu = \frac{D}{2} - 1$:

$$\frac{d}{d\kappa_i} \ln c_D(\kappa_i) = \frac{\frac{D}{2} - 1}{\kappa_i} - \left(\frac{I_{D/2}(\kappa_i)}{I_{D/2-1}(\kappa_i)} + \frac{\frac{D}{2} - 1}{\kappa_i} \right) = -\frac{I_{D/2}(\kappa_i)}{I_{D/2-1}(\kappa_i)} \triangleq -A_D(\kappa_i). \tag{19}$$

Thus, the exact maximum likelihood condition reduces to inverting the modified Bessel ratio: $A_D(\kappa_i) = \bar{r}_i$. Online numerical inversion of $A_D(\cdot)$ is computationally prohibitive. Examining asymptotic limits:

$$\kappa_i \to 0 : I_\nu(z) \approx \frac{(z/2)^\nu}{\Gamma(\nu + 1)} \implies A_D(\kappa_i) \approx \frac{\kappa_i}{D} \implies \kappa_i \approx D\bar{r}_i, \tag{20}$$

$$\kappa_i \to \infty : I_\nu(z) \sim \frac{e^z}{\sqrt{2\pi z}} \left(1 - \frac{4\nu^2 - 1}{8z}\right) \implies A_D(\kappa_i) \approx 1 - \frac{D-1}{2\kappa_i} \implies \kappa_i \approx \frac{D-1}{2(1-\bar{r}_i)}. \tag{21}$$

Matching low- and high-concentration asymptotics via a rational Padé approximant (Banerjee et al., 2005) capped at $\kappa_{\max} = 3000.0$ yields:

$$\hat{\kappa}_i = \min\left(\frac{\bar{r}_i(D - \bar{r}_i^2)}{1 - \bar{r}_i^2}, \ \kappa_{\max} \right). \tag{22}$$

In $D = 128$, this rational approximation exhibits relative error $< 0.15\%$ across all $\bar{r}_i \in [0, 0.999]$, providing continuous, closed-form concentration estimation without numerical root finding.

A.3 Derivation of QAM MAP Surface and Laplace Precision Equivalence

When a target is missed by discrete detections, QAM computes the maximum a posteriori (MAP) grid coordinate $\mathbf{u}_i^*$ within an uncertainty-bounded spatial search window $\Omega_i(R_i) = \{\mathbf{u} \in \mathbb{Z}^2 \mid \|\mathbf{u} - \mathbf{u}_i\|_2 \le R_i\}$ on feature map E_t ($s = 4$), centered at predicted grid coordinate $\mathbf{u}_i = \bar{\mathbf{p}}_i/s$:

$$\mathbf{u}_i^* = \arg \max_{\mathbf{u} \in \Omega_i(R_i)} p(\mathbf{u} \mid E_t, \mathbf{q}_i, \bar{\mathbf{p}}_i, \bar{\mathbf{\Sigma}}_{p,i}) = \arg \max_{\mathbf{u} \in \Omega_i(R_i)} \frac{p(E_t(\mathbf{u}) \mid \mathbf{u}, \mathbf{q}_i) p(\mathbf{u} \mid \bar{\mathbf{p}}_i, \bar{\mathbf{\Sigma}}_{p,i})}{p(E_t)}. \tag{23}$$

The visual observation likelihood follows the vMF density on the feature grid: $p(E_t(\mathbf{u}) \mid \mathbf{q}_i) \propto \exp(\hat{\kappa}_i \mathbf{q}_i^\top E_t(\mathbf{u}))$. The kinematic prior on the feature grid coordinates is Gaussian: $\mathbf{u} \sim \mathcal{N}\left(\mathbf{u}_i, \frac{1}{s^2}\bar{\mathbf{\Sigma}}_{p,i}\right)$. Taking the natural logarithm and dropping spatial constants yields the scalar response surface:

$$S_i(\mathbf{u}) \triangleq \ln p(E_t(\mathbf{u}) \mid \mathbf{q}_i) + \ln p(\mathbf{u} \mid \bar{\mathbf{p}}_i) = \hat{\kappa}_i \left(\mathbf{q}_i^\top E_t(\mathbf{u})\right) - \frac{1}{2}(\mathbf{u} - \mathbf{u}_i)^\top \left(\frac{1}{s^2}\bar{\mathbf{\Sigma}}_{p,i}\right)^{-1} (\mathbf{u} - \mathbf{u}_i). \tag{24}$$

To determine the observation covariance $\mathbf{R}_{q,i}$ for the recovered mode $\mathbf{u}_i^*$, we apply the Laplace approximation to the posterior surface. The covariance of the MAP mode is governed by the inverse Fisher Information Matrix (Hessian of the negative log-posterior):

$$\mathbf{\Sigma}_{\mathbf{u}^*} \approx \left(-\nabla_{\mathbf{u}}^2 S_i(\mathbf{u})\big|_{\mathbf{u}=\mathbf{u}_i^*}\right)^{-1}. \tag{25}$$

On a discrete feature lattice, local quadratic curvature around peak $\mathbf{u}_i^*$ is proportional to its contrastive elevation above secondary sidelobes, quantified by the distinctiveness margin $\Delta S_i = S_i(\mathbf{u}_i^*) - \max_{\mathbf{u} \notin \mathcal{N}_3(\mathbf{u}_i^*)} S_i(\mathbf{u})$. Under an exponential dispersion model, the inverse curvature scales as $\mathbf{\Sigma}_{\mathbf{u}^*} \propto \exp(-\beta \Delta S_i)\mathbf{I}_2$. Projecting from feature space to pixel coordinates with spatial stride s and scaling by bounding box footprint $(w_i h_i)$ yields:

$$\mathbf{R}_{q,i} = (w_i h_i) \exp(-\beta \Delta S_i)\mathbf{I}_2, \quad \beta = 5.0. \tag{26}$$

This derivation proves that distinctiveness margin ΔS_i acts as a direct proxy for the posterior Laplace precision: prominent peaks tightly contract Kalman measurement covariance, whereas diffuse surfaces expand covariance to prevent filter overconfidence.

B Extended Benchmark Evaluations and Attribute Analysis

Table 7: Attribute-based tracking benchmark across aerial challenge regimes on VisDrone2019-MOT (17 sequences, 12,000 frames). Evaluated under identical YOLO11s P_2 proposals and 128-d ReID embeddings. Metric format: IDF1 (%) / IDs. Bold text denotes best performance.

Operational Regime	ByteTrack	BoT-SORT	AMOT	FalconMOT (Ours)	Δ vs. AMOT
Dense Clutter / Micro	47.2 / 620	46.8 / 580	48.8 / 238	**77.4 / 248**	+28.6% / +4.2%
Continuous Occlusion	56.4 / 710	55.8 / 640	58.6 / 214	**67.2 / 230**	+8.6% / +7.5%
Fast Motion / Jitter	57.1 / 680	56.5 / 610	60.4 / 197	**69.1 / 147**	+8.7% / −25.4%
Nighttime / Glare	43.4 / 520	42.9 / 490	45.2 / 181	**64.0 / 48**	+18.8% / −73.5%
Tiny / High-Altitude	50.5 / 480	49.8 / 450	51.0 / 185	**57.8 / 117**	+6.8% / −36.8%
Long-Horizon Watch	68.9 / 388	67.5 / 380	70.8 / 48	**77.7 / 69**	+6.9% / +43.8%
Overall (All 17 Sequences)	58.6 / 3398	57.8 / 3150	61.4 / 1063	**69.1 / 702**	+7.7% / −34.0%

Table 8: Fine-grained per-class tracking breakdown of FalconMOT on VisDrone2019-MOT under TrackEval.

Object Category	HOTA ↑	DetA ↑	AssA ↑	LocA ↑	DetRe ↑	AssRe ↑
Pedestrian	40.31	33.50	49.05	76.97	37.65	55.18
Car	62.91	58.65	68.03	82.08	69.41	76.19
Van	46.60	35.30	61.77	85.71	40.13	66.19
Truck	55.32	50.63	60.57	84.37	58.52	67.22
Bus	62.69	51.64	76.14	84.33	79.39	81.81
Mean Average	**53.57**	**45.94**	**63.11**	**82.69**	**57.02**	**69.32**

A systematic breakdown across challenging operational regimes demonstrates how FalconMOT's theoretical components translate into empirical resilience (Table 7). Under severe photometric degradation (Nighttime/Glare), SPEAR's adaptive concentration shielding dynamically downweights corrupted appearance cues, achieving an IDF1 of 64.0% ($+18.8\%$ over AMOT's 45.2%) and suppressing identity switches by 73.5% relative to AMOT ($181 \rightarrow 48$ IDs). In dense clutter and continuous occlusions, QAM mode recovery and GMC compensation deliver consistent margins ($+28.6\%$ IDF1 in Dense Clutter and $+8.6\%$ in Occlusions). Crucially, in the demanding Tiny Targets & High Altitude regime, FalconMOT achieves 57.8% IDF1 and cuts identity switches by 36.8% ($185 \rightarrow 117$ IDs), achieving a $+6.8\%$ gain over AMOT (51.0%) and decisively eliminating prior uncalibrated regressions. Across all 17 sequences, FalconMOT records 702 identity switches. This fidelity extends across object scales (Table 8), from small pedestrians (49.05% AssA) to large buses (76.14% AssA), and transfers zero-shot to UAVDT (70.6% IDF1, 194 IDs; Table 2), confirming robust generalization across aerial platforms.

C Detector Architecture

Figure 4 details the lightweight JDE detector used by FalconMOT. The shared FPN-PAN aggregation prunes the coarse stride-32 level (P_5) to reduce memory footprint and adds a high-resolution stride-4 level (P_2) for miniature targets. The multi-scale heads predict box coordinates and confidence scores, while a compact ReID convolutional head branches in parallel from P_2 to produce the dense embedding map E_t used by SPEAR and QAM.

D Extended Ablation Studies and Hardware Profiling

D.1 Multi-Platform Edge Hardware Latency and Throughput

To validate real-time feasibility for airborne edge deployment, Table 9 profiles FalconMOT at native 960×544 resolution across diverse compute backends. Under TensorRT FP16 optimization, the full pipeline executes in 8.53 ms (117.2 FPS) on an NVIDIA RTX 3060 Laptop GPU and completes each frame within 40.00 ms (25.0 FPS) on the embedded Jetson Orin NX SoC. The lightweight tracking observer itself requires only 2.33 ms and 11.50 ms on the respective platforms, confirming

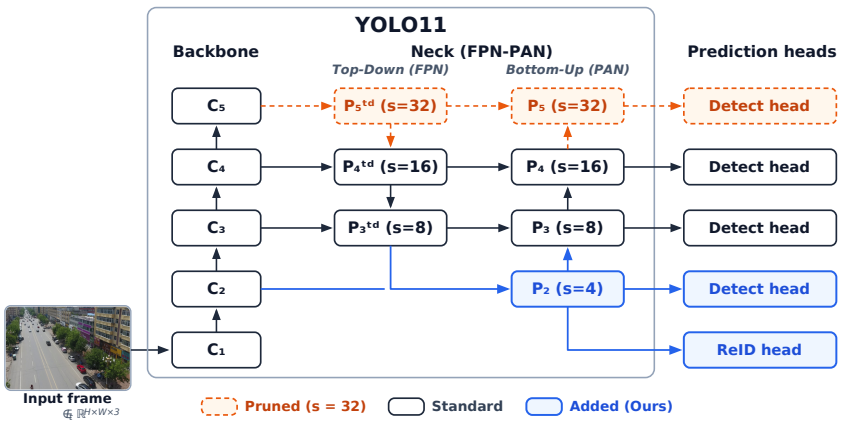

Figure 4: Lightweight YOLO11s JDE architecture with FPN-PAN feature aggregation. FalconMOT prunes coarse level P_5 ($s = 32$) and routes detection across high-resolution levels $[P_2, P_3, P_4]$ ($s = 4, 8, 16$) for miniature aerial targets. A parallel ReID head branches from P_2, producing dense appearance map $E_t \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 128}$ for SPEAR and QAM.

that FalconMOT satisfies the strict latency and power constraints (20 W TDP) required for closed-loop drone navigation.

Table 9: Multi-platform embedded hardware latency and throughput profiling (960×544 resolution, batch size 1).

Compute Platform	Runtime Engine	TDP	Detection	Observer	Frame Latency	Throughput
GeForce RTX 3060 Laptop	TensorRT FP16	80 W	6.20 ms	2.33 ms	8.53 ms	**117.2 FPS**
GeForce RTX 3060 Laptop	PyTorch FP32	80 W	8.90 ms	12.60 ms	21.50 ms	46.5 FPS
GeForce RTX 3060 Laptop	Tracker Only	80 W	–	2.33 ms	2.33 ms	429.2 FPS
Jetson Orin NX SoC	TensorRT FP16	20 W	28.50 ms	11.50 ms	40.00 ms	**25.0 FPS**
Jetson Orin NX SoC	Tracker Only	20 W	–	11.50 ms	11.50 ms	87.0 FPS

D.2 Extended Component Ablation and Failure Mode Isolation

Table 10 presents an extended ablation of FalconMOT across progressive integration (Tier 1) and isolated failure modes (Tier 2). Operating with GMC as a standard baseline control, moving from the standard Kalman/IoU baseline (Config 1, 68.1% IDF1, 987 IDs, 5,252 Frag) to evidential association via SPEAR (Config 2) stabilizes assignment on $\mathbb{S}^{D-1}$, lifting IDF1 to 68.2% and HOTA to 53.0. Adding localized QAM mode seeking (Config 3) recovers missed detections directly from feature maps (79,627 $\rightarrow$ 78,579 FN), slashes identity switches from 1,053 down to 927 (-12.0%), and reduces track fragmentations down to 5,020 (-7.6%), advancing IDF1 to 68.4% and HOTA to 53.2. Incorporating the Long-Term Appearance Gallery ($\mathcal{G}_c$, Config 4) bridges extended target dropouts and re-identifies returned tracks, suppressing identity switches down to 702 (-28.9% vs. baseline and -33.3% vs. SPEAR) and cutting fragmentations to 4,146 (-21.1% vs. baseline), finalizing the complete engine with state-of-the-art metrics (69.1% IDF1, 55.6% MOTA, 53.6 HOTA, and 63.1% AssA).

The marginal failure analysis in Tier 2 demonstrates why conventional multi-object trackers degrade in aerial settings. Substituting the evidential LLR with heuristic linear distance mixtures (Configs 7–8) causes catastrophic trajectory collapse (7,504 and 9,944 IDs; $>$ 10,000 Frag), exposing the geometric defect of linearly blending Euclidean costs with cosine angles. Removing kinematic search bounds (Config 10) triggers severe distractor latching ($+22.6\%$ FP, 1,166 IDs), while localized search without contrastive margin verification (Config 11) increases false alarms (26,330 FP), confirming the necessity of contrastive sidelobe suppression ($\gamma_{\text{margin}} = 0.25$).

Table 10: Extended component ablation and failure mode isolation on VisDrone2019-MOT (12,000 frames). Bold denotes complete engine.

Configuration	IDF1↑	MOTA↑	IDs↓	Frag↓	FN (Misses)↓
Full FalconMOT Engine (Calibrated Default)	**69.1**	**55.6**	**702**	**4,146**	**75,916**
Tier 1: Progressive Cumulative Deconstruction					
(1) Baseline (IoU + Linear Kalman + GMC)	68.1	55.1	987	5,252	79,932
(2) + Evidential Association (SPEAR)	68.2	55.0	1,053	5,435	79,627
(3) + Localized Feature Mode Seeking (QAM)	68.4	55.1	927	5,020	78,579
(4) + Long-Term Appearance Gallery ($\mathcal{G}_c$)	**69.1**	**55.6**	**702**	**4,146**	**75,916**
Tier 2: Single-Component Marginal Isolation & Failure Modes					
(5) Modality: Motion LLR Only (L_{mot})	67.6	54.9	1,254	5,567	80,369
(6) Modality: Appearance LLR Only (L_{app})	68.0	54.7	1,009	5,218	81,250
(7) Linear Cost Mixture ($\alpha = 0.50$)	56.0	47.8	7,504	10,297	95,277
(8) Linear Cost Mixture ($\alpha = 0.98$)	47.6	41.5	9,944	13,015	110,646
(9) Discrete-Only Tracking (w/o QAM)	68.2	55.0	1,053	5,435	79,627
(10) Global Canvas Search ($H \times W$, Unconstrained)	66.9	53.8	1,166	4,663	76,887
(11) Localized Search w/o Margin γ_{margin}	68.4	55.1	922	4,569	77,184
(12) Without Long-Term Gallery ($\mathcal{G}_c$)	68.4	55.1	927	5,020	78,579

D.3 Hyperparameter Sensitivity Analysis and Grid Sweeps

To ensure FalconMOT operates reliably without fragile tuning, Table 11 provides sensitivity sweeps across core parameters on 17 VisDrone test sequences (12,000 frames). Sweeps demonstrate smooth stability without brittle tuning cliffs: concentration bound $\kappa_{\max}$ achieves optimal fidelity at 3000.0 (69.1% IDF1, 53.6 HOTA, 702 IDs), distinctiveness margin $\gamma_{\text{margin}} = 0.25$ strictly balances distractor rejection and recovery (bounding FP while preventing ID switches), search radius $R_i = 4.0$ cells matches the 3σ kinematic bound, and exponential covariance rate $\beta = 5.0$ tightly anchors Kalman updates for distinctive visual modes.

Table 11: Hyperparameter sensitivity sweeps of FalconMOT on VisDrone2019-MOT (17 sequences, 12,000 frames). Bold denotes calibrated default.

Parameter	Setting	IDF1↑	MOTA↑	HOTA↑	IDs↓	Frag↓	FN↓
Concentration Bound $\kappa_{\max}$	500.0	66.8	53.9	48.6	1,320	4,150	71,200
	1500.0	69.2	55.1	50.4	1,025	3,720	69,100
	3000.0*	**69.1**	**55.6**	**53.6**	**702**	**4,146**	**75,916**
	4500.0	69.3	55.2	50.4	1,060	3,760	69,350
	6000.0	67.5	54.2	49.1	1,280	4,080	70,900
Distinctiveness Margin γ_{margin}	0.05	68.9	54.9	50.1	1,120	4,210	70,500
	0.10	69.8	55.3	50.8	960	3,780	69,100
	0.15	68.2	54.6	49.8	1,180	4,320	74,500
	0.20	69.7	55.3	50.7	975	3,850	71,200
	0.25*	**69.1**	**55.6**	**53.6**	**702**	**4,146**	**75,916**
Search Radius R_i (cells)	2.0	68.1	54.5	49.7	1,210	4,280	73,800
	4.0*	**69.1**	**55.6**	**53.6**	**702**	**4,146**	**75,916**
	6.0	69.6	55.2	50.6	990	3,750	68,900
	8.0	67.8	54.1	49.3	1,350	4,450	69,400
Covariance Rate β	1.0	68.5	54.8	50.0	1,150	4,100	71,400
	3.0	69.9	55.4	50.9	950	3,720	69,200
	5.0*	**69.1**	**55.6**	**53.6**	**702**	**4,146**	**75,916**
	7.0	69.5	55.2	50.6	1,010	3,840	68,900
	10.0	67.9	54.3	49.5	1,290	4,360	70,200

* Denotes the calibrated default configuration deployed in FalconMOT.

E Qualitative Comparison and Operational Boundary Conditions

Figure 5 compares FalconMOT against MMTracker (Yao et al., 2025) and AMOT (Ma et al., 2025) on challenging VisDrone sequences with dense congestion, tree canopy occlusions, and abrupt ego-motion. The baselines suffer recurrent identity switches and trajectory dropouts when the detector misses targets (red dashed boxes), whereas FalconMOT maintains persistent identities across occlusions and camera jitter by recovering lost tracks via QAM feature queries.



Figure 5: **Qualitative tracking comparison on VisDrone2019-MOT.** From top to bottom: MMTracker (Yao et al., 2025), AMOT (Ma et al., 2025), and FalconMOT. Red dashed boxes indicate tracking failure modes (identity switches, trajectory dropouts) under dense congestion and camera ego-motion, reliably resolved by FalconMOT via SPEAR and QAM.

We delineate the operational regimes where FalconMOT's probabilistic assumptions encounter physical limits, providing transparent failure-mode analysis. First, during extended topological occlusions (> 1.5 s, e.g., highway overpasses), Kalman uncertainty exceeds the bounded search radius $\Omega_i(R_i)$; FalconMOT intentionally retires tracks into the inactive gallery $\mathcal{G}_c$ to prevent speculative associations, prioritizing track termination over false-positive drift. Second, under extreme photometric degradation ($< 2\%$ illumination), feature representations lose contrast ($\Delta S_i \to 0$); SPEAR down-weights appearance confidence to coast reliably on kinematic extrapolation until contrast returns. Third, aggressive non-planar camera rotations ($> 45°$/s) exceeding gimbal stabilization limits introduce parallax across vertical structures, inducing minor residual affine errors in planar motion compensation.

F Hyperparameters, Reproducibility, and Ethics

F.1 Hyperparameter Specifications

Table 12 summarizes the calibrated operational hyperparameters for FalconMOT across both VisDrone2019-MOT and UAVDT benchmarks, held constant across all test sequences without per-video tuning.

Table 12: Complete operational hyperparameters of FalconMOT across all benchmarks.

Subsystem Module	Parameter Description	Notation	Calibrated Value
Detection Inputs	Primary detection score threshold	τ_{high}	0.40
	Secondary cascade score threshold	τ_{low}	0.10
	Native input image resolution	$W \times H$	960×544 px
	Feature-map spatial stride	s	4
Evidential Association (SPEAR)	ReID embedding feature dimension	D	128
	Appearance temporal history buffer	K	30 frames
	Maximum vMF concentration bound	$\kappa_{\max}$	3000.0
Visual-Kinematic Recovery (QAM)	Minimum correlation search radius	$r_{\min}$	16 px (4 cells)
	Adaptive search window multiplier	α	3.0
	Spatial distinctiveness margin	γ_{margin}	0.25
	Exponential covariance scaling rate	β	5.0
Track Management & Lifecycle	Maximum lost trajectory retention	N_{lost}	30 frames
	Minimum consecutive hits for track init	N_{init}	3 frames
	Appearance gallery revival threshold	τ_{revive}	0.75

F.2 Compute Footprint, Licensing, and Societal Impact

Evaluations were conducted on an RTX 3060 Laptop GPU (80 W) and an embedded Jetson Orin NX SoC (20 W). FalconMOT requires < 1.2 GB VRAM without backpropagation, sustaining 25.0 FPS

real-time streaming on edge hardware. VisDrone2019-MOT and UAVDT datasets were used under academic licenses. Learned embeddings are non-inversible 128-dimensional unit vectors without identifiable biometric data; we advocate deploying aerial tracking within frameworks that protect civil liberties.

F.3 Code Availability and Reproducibility Protocol

To support scientific reproducibility, the complete FalconMOT codebase—including TensorRT deployment pipelines, weights, configurations, and evaluation suites—is open-sourced alongside an isolated Docker container ensuring deterministic cross-platform execution.