AeroGraph: Window-Level Large-Graph Association and Uncertainty-Guided Fusion for Multi-Object Tracking in UAV Videos

1st Xianchang Wang
Shenyang Aerospace University
School of Computer Science
Shenyang, China
wxc@email.sau.edu.cn

2nd Qi Li*
Shenyang Aerospace University
School of Computer Science
Shenyang, China
liqi22@stu.sau.edu.cn

***Abstract*—Multiple object tracking (MOT) in Unmanned Aerial Vehicle (UAV) videos is crucial for a wide range of vision applications. However, UAV tracking remains challenging due to dense small targets, frequent occlusions, drastic viewpoint/scale variations, and non-stationary motion caused by platform jitter and detection noise. Most online tracking-by-detection pipelines mainly rely on frame-wise matching with limited temporal evidence, where unreliable motion priors and noisy observations can easily trigger identity drift and trajectory fragmentation. To tackle these issues, we propose *AeroGraph*, a window-level large-graph association framework for online UAV MOT. Specifically, we formulate data association as inference on a unified large-graph within a sliding window, which jointly models intra-frame structural context and inter-frame correspondences to aggregate multi-frame cues. Moreover, an uncertainty-guided motion fusion strategy is designed to adaptively balance motion consistency and graph affinity according to motion reliability, preventing error propagation under jitter and detection noise. In addition, a trajectory memory with window-level consistency learning is introduced to stabilize long-term identity representations and mitigate feature drift during online updates. Extensive experiments on representative UAV tracking benchmarks demonstrate that AeroGraph consistently improves association robustness and achieves competitive performance under challenging aerial conditions.**

***Index Terms*—UAV, Multi-Object Tracking, Data Association, large-graph, Uncertainty, Platform Motion**

I. Introduction

Multi-object tracking (MOT) aims to localize multiple targets in video streams while maintaining identity-consistent trajectories, and it is fundamental to spatiotemporal video understanding. In online settings, most systems adopt tracking-by-detection: detections are produced per frame and then linked to existing tracks under strict causality constraints. While practical, this formulation is vulnerable to compounding association errors when detection noise, occlusion, and motion ambiguity occur simultaneously.

Early online trackers establish strong efficiency–accuracy baselines. SORT [1] couples a motion model with one-to-one assignment, enabling a lightweight and fast pipeline. DeepSORT [2] improves identity preservation by introducing learned appearance representations and similarity metrics that complement motion cues, which is particularly beneficial under short-term occlusions. These foundations highlight a key trade-off in online MOT: association must remain efficient while being resilient to noisy observations and identity ambiguity.

Recent progress advances MOT through stronger representation learning and association design. FairMOT [3] unifies detection and ReID embedding learning to better align localization and identity discrimination, reducing the mismatch between detection features and association features. QDTrack [4] learns similarity with dense supervision and provides more discriminative matching signals under crowding and appearance confusion. Transformer-based trackers further tighten detection–association coupling; TrackFormer [5] uses track queries and temporal propagation to reduce reliance on manually designed association rules.

Once detectors and embeddings stabilize, association strategy remains a dominant factor. ByteTrack [6] improves trajectory continuity by associating all detections through a confidence-aware procedure, reducing fragmentation caused by discarding low-score boxes. BoT-SORT [7], OC-SORT [8], and StrongSORT [9] refine motion–appearance fusion and improve robustness under occlusion, non-linear motion, and drifting updates. Evaluation protocols have also evolved; HOTA [10] jointly measures detection, localization, and association quality, encouraging methods that improve association without sacrificing detection consistency.

UAV videos introduce substantially stronger non-stationarity than ground-based benchmarks. Targets are often small and densely distributed, viewpoint and scale vary rapidly, occlusions and close interactions are frequent, and platform jitter coupled with detection noise makes motion cues unstable. The VisDrone benchmark provides standardized evaluation and highlights these challenges [11]–[13], while UAVDT emphasizes the impact of camera motion, small objects, and background complexity on detection and tracking [14]. UAV-oriented online trackers adapt modeling to aerial dynamics; UAVMOT [15], SGT [16], and UGT [17] improve robustness in UAV scenarios, yet association still suffers from limited

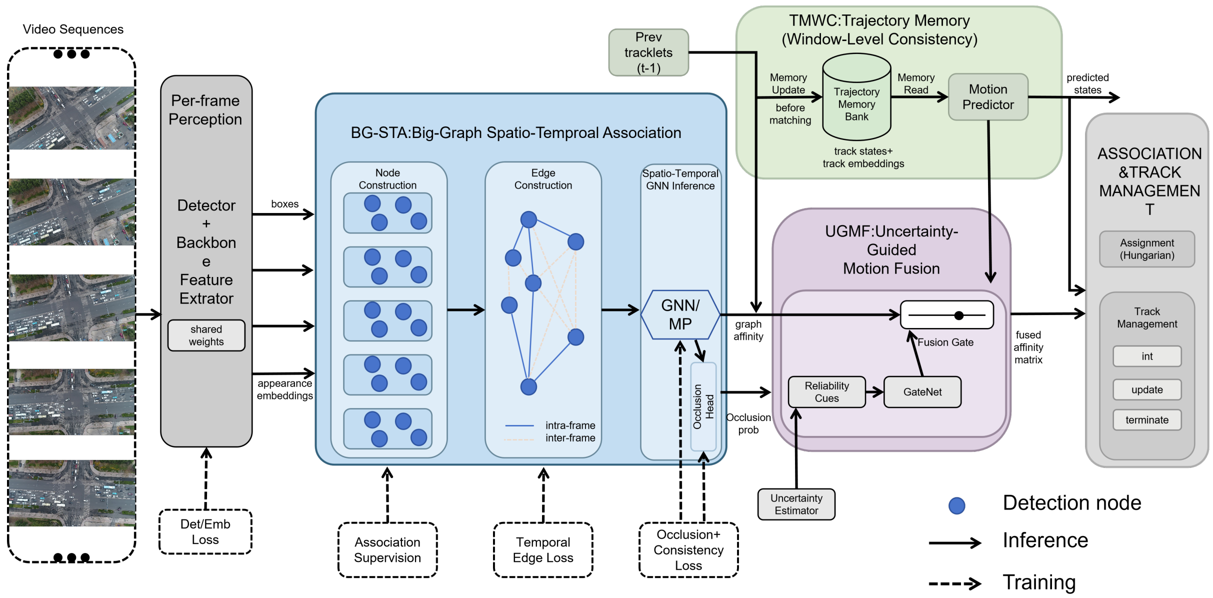

Fig. 1: Architecture of AeroGraph. It takes detections from a 5-frame sliding window and extracts ReID embeddings with a separate ReID network. BG-STA infers cross-frame affinities on a spatiotemporal graph, TMWC stabilizes track representations with trajectory memory and motion prediction, and UGMF fuses graph and motion affinities for online assignment.

multi-frame evidence aggregation, unreliable motion priors under jitter, and long-term feature drift.

To address these issues, we propose AeroGraph, an online UAV MOT framework built on three ideas: window-level large-graph inference, uncertainty-guided motion fusion, and trajectory memory with window-level consistency. The main contributions are:

- We perform window-level spatiotemporal association on a unified large-graph within a 5-frame window via BG-STA to aggregate multi-frame evidence and reduce ambiguities in crowded and occluded scenes.
- We estimate motion reliability and adaptively fuse motion consistency with graph affinity via UGMF to prevent noisy motion priors from dominating matching under platform jitter and detection noise.
- We stabilize long-term identity representations with trajectory memory and window-level consistency learning via TMWC, mitigating feature drift and trajectory fragmentation in long UAV sequences.

II. Methodology

A. Overall Framework

As illustrated in Fig. 1, AeroGraph performs online tracking with a 5-frame sliding window. At time step t, the window is defined as

$$\mathcal{W}_t = \{I_{t-4},\, I_{t-3},\, I_{t-2},\, I_{t-1},\, I_t\}. \tag{1}$$

For each frame in $\mathcal{W}_t$, we run the detector to obtain bounding boxes and extract ReID embeddings for each detection using a separate ReID network.

All detections in $\mathcal{W}_t$ are treated as nodes in a unified spatiotemporal graph. BG-STA builds intra-frame structural edges to encode local context and inter-frame edges to represent candidate correspondences, and outputs window-level graph affinity for cross-frame association. TMWC maintains a trajectory memory bank to provide temporally smoothed track embeddings and stabilizes online updates, while a motion predictor estimates track states for motion-based affinity computation. UGMF then adaptively fuses the graph affinity and motion affinity into a fused affinity matrix between active tracks and current-frame detections.

Finally, one-to-one assignment is solved by the Hungarian algorithm [18], [19]. Track management updates matched tracks, initializes new tracks from unmatched high-confidence detections, and keeps unmatched tracks as tentative/lost until termination.

B. Big-Graph Spatio-Temporal Association

Online MOT typically relies on frame-to-frame matching, which becomes brittle in UAV videos where small targets and frequent occlusions amplify local ambiguities. BG-STA addresses this issue by performing window-level inference on a unified spatiotemporal large-graph within $\mathcal{W}_t$ (Eq. (1)), aggregating both intra-frame context and inter-frame correspondence evidence.

Let the detections in frame τ be $\mathcal{D}_\tau = \{d_i^\tau\}_{i=1}^{N_\tau}$. Each detection is represented as a node v_i^τ with appearance $\mathbf{a}_i^\tau$, geometric encoding $\mathbf{g}(\mathbf{b}_i^\tau)$, and a local topology descriptor $\mathbf{t}_i^\tau$ from K-NN neighborhoods (size K_s). Intra-frame structural edges $\mathcal{E}_s$ are built by K-NN connections.

Temporal edges $\mathcal{E}_t$ connect candidate correspondences across frames using IoU gating and a maximum temporal span G:

$$\begin{aligned} (u \to v) \in \mathcal{E}_t &\iff \mathrm{IoU}(\mathbf{b}_u, \mathbf{b}_v) \geq \theta, \\ \tau(v) = \tau(u) + \Delta t, &\quad \Delta t \in \{1, \ldots, G\}. \end{aligned} \tag{2}$$

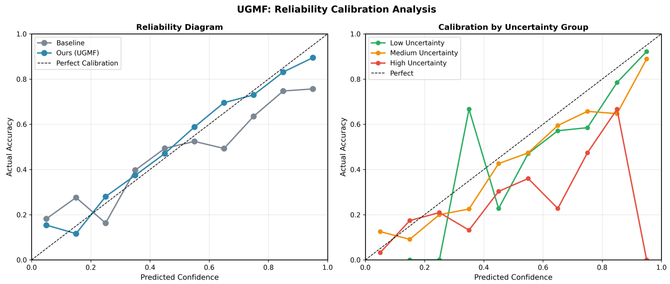

Fig. 2: Reliability diagrams of association-confidence scores with and without UGMF. UGMF improves confidence calibration: the curve is closer to the diagonal, with fewer over-confident errors under high uncertainty.

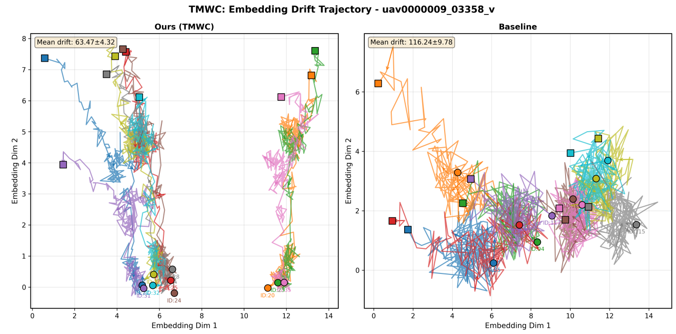

Fig. 3: Temporal drift of track embeddings with and without TMWC. TMWC reduces embedding drift over time (e.g., $116.24 \pm 9.78 \rightarrow 63.47 \pm 4.32$), indicating more stable track representations.

Each temporal edge carries attributes $\mathbf{r}_{uv}$ (IoU and normalized Δt), yielding the spatiotemporal graph

$$\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_s \cup \mathcal{E}_t). \tag{3}$$

We perform graph inference via message passing:

$$\mathbf{h}_u^{(\ell+1)} = \phi^{(\ell)}\Big(\mathbf{h}_u^{(\ell)}, \mathrm{AGG}\big(\{\psi^{(\ell)}(\mathbf{h}_u^{(\ell)}, \mathbf{h}_v^{(\ell)}, \mathbf{r}_{uv})\}_{v\in\mathcal{N}(u)}\big)\Big), \tag{4}$$

which updates node representations by aggregating neighborhood evidence [20].

Finally, BG-STA predicts association confidence on temporal edges as graph affinity:

$$s_{uv} = \sigma\big(g_{\text{edge}}\big([\mathbf{h}_u; \mathbf{h}_v; \mathbf{r}_{uv}]\big)\big), \quad (u \to v) \in \mathcal{E}_t, \tag{5}$$

where g_{edge} is an edge classifier and $\sigma(\cdot)$ is sigmoid. The scores form a window-level weighted association graph for subsequent fusion and decoding.

C. Uncertainty-Guided Motion Fusion

BG-STA provides window-level association evidence, while motion offers a complementary geometric cue. In UAV videos, platform jitter and detection noise make motion priors unstable, so UGMF models motion uncertainty and adaptively fuses graph affinity (Eq. (5)) with motion consistency.

For each track i, we predict the next-step center $\hat{\mathbf{p}}_i$ and an uncertainty scale to form a covariance $\mathbf{\Sigma}_i$ (diagonal in practice). Given detection center $\mathbf{p}_j$, we measure prediction–observation consistency by

$$d_{ij}^2 = (\mathbf{p}_j - \hat{\mathbf{p}}_i)^\top \mathbf{\Sigma}_i^{-1}(\mathbf{p}_j - \hat{\mathbf{p}}_i). \tag{6}$$

We map d_{ij}^2 to motion affinity s_{ij}^{mot} via an exponential decay.

For each candidate pair (i, j), we build a gating feature $\mathbf{g}_{ij}$ from track reliability p_i^{occ}, detection confidence c_j, uncertainty scale, and geometric deviations. A lightweight gating network outputs $\omega_{ij} \in (0, 1)$, and the fused affinity is

$$s_{ij} = (1 - \omega_{ij})\, s_{ij}^{\text{graph}} + \omega_{ij}\, s_{ij}^{\text{mot}}, \tag{7}$$

where s_{ij}^{graph} comes from BG-STA. ω_{ij} increases when motion is reliable and decreases under high uncertainty to avoid error propagation.

D. Trajectory Memory with Window-Level Consistency

Occlusions and noisy detections can make online appearance updates unstable, which degrades association and motion modeling. TMWC improves long-term identity stability via a trajectory memory for temporally smoothed embeddings and a window-level consistency supervision during training.

For each confirmed track i, we maintain a memory vector $\mathbf{m}_i \in \mathbb{R}^d$ and update it with an exponential moving average:

$$\mathbf{m}_i \leftarrow (1 - \alpha)\mathbf{m}_i + \alpha\, \mathbf{x}_i^{t-1}, \tag{8}$$

where $\mathbf{x}_i^{t-1}$ is the previous track representation and α is a momentum coefficient. During matching, $\mathbf{m}_i$ serves as the track-side embedding to reduce short-term fluctuations. Moving-average stabilization is widely used in momentum-based representation learning [21], and memory-based designs have also been explored for online MOT [22].

We further leverage node-level reliability predicted by BG-STA to suppress low-quality observations by down-weighting candidate edges incident to low-confidence or low-reliability detections.

During training, we construct temporal candidate edges $\mathcal{E}_{\text{temp}}$ within the window (as in Eq. (2)) and supervise them with

$$y_{pq} = \mathbb{I}[\mathrm{id}_p = \mathrm{id}_q]\,, \tag{9}$$

yielding a temporal consistency loss $\mathcal{L}_{\text{temp}}$. This encourages short-term identity coherence while keeping standard one-to-one online matching at inference time.

E. Training & Inference Strategy

During online inference, we compute the fused affinities between active tracks and current-frame detections using Eq. (7) and solve one-to-one assignment with the Hungarian algorithm [18], [19]:

$$\pi^* = \arg\max_{\pi} \sum_i s_{i,\pi(i)}. \tag{10}$$

Matched tracks are updated, unmatched detections may initialize new tracks, and unmatched tracks are retained for a limited tolerance before termination. We update trajectory memory

(a) VisDrone 2021 dataset

(b) UAVDT dataset

Fig. 4: Qualitative comparison on representative challenging clips from VisDrone2021-MOT and UAVDT. Red arrows indicate identity drift, erroneous reconnection, or track interruption. Yellow dashed boxes highlight missed tracking for small or heavily occluded targets.

before matching (Eq. (8)) and apply reliability-aware gating to suppress error propagation under occlusion.

During training, we jointly optimize graph inference (Eq. (5)), uncertainty and occlusion cues, and online decoding (Eqs. (7) and (10)). The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{det/emb}} + \mathcal{L}_{\text{assoc}} + \lambda_t \, \mathcal{L}_{\text{temp}} + \lambda_o \, \mathcal{L}_{\text{occ}} + \lambda_c \, \mathcal{L}_{\text{occ-cons}}. \quad (11)$$

Here $\mathcal{L}_{\text{assoc}}$ uses Focal Loss to handle class imbalance [23] and a Dice-style term to improve stability [24]. λ_t, λ_o, and λ_c are weighting factors.

III. Experiments

A. Implementation Details

We use YOLOX-M as the detector [25] and train it separately on VisDrone2021-MOT and UAVDT. YOLOX-M is trained for 80 epochs with SGD (Nesterov momentum) using a base per-image learning rate of $0.001/64$ with batch-size scaling (default: 64).

At inference, the input resolution is 1280 with a confidence threshold of 0.01 and an NMS threshold of 0.45. ReID features are extracted by OSNet (osnet_x1_0) [26] pretrained on MSMT17 [27]; each detection is cropped to 256×128 and encoded as a 512-d embedding.

We use a 5-frame window with $\Delta t \leq 2$ and IoU gating at 0.2 for association. Track management sets $T_{\text{lost}} = 60$. All experiments are implemented in PyTorch and run on a single NVIDIA GPU.

B. Online Inference

We follow a strictly online pipeline and adopt the high-confidence-first, low-confidence-second association strategy in ByteTrack [6]. For each frame, we run YOLOX-M with a confidence threshold of 0.01 and an NMS threshold of 0.45 [25], and split detections by $t_{\text{high}} = 0.5$ and $t_{\text{low}} = 0.15$.

We first match active tracks to high-confidence detections. BG-STA provides cross-frame scores within the 5-frame sliding window; candidates satisfy $\Delta t \leq 2$ and IoU ≥ 0.2. We keep edges with graph confidence > 0.5 and solve one-to-one assignment with the Hungarian algorithm [18], [19].

We then match remaining tracks to low-confidence detections to recover missed targets using a Wasserstein-based geometric cost (threshold 0.45) [28], [29]. Appearance features are extracted by OSNet [26]; when reliability is available, UGMF gates appearance updates, otherwise we update with the matched detection feature.

We set $T_{\text{lost}} = 60$ and initialize new tracks from unmatched high-confidence detections with score > 0.2, while discarding unmatched low-confidence detections. We remove tiny boxes (area < 50 pixels), disable aspect-ratio filtering, and do not apply Sinkhorn normalization [29].

C. Comparison with State-of-the-Art

We evaluate AeroGraph on VisDrone2021-MOT and UAVDT against representative online MOT methods. As shown in Table I, AeroGraph achieves state-of-the-art performance on both benchmarks, with strong improvements in HOTA [10] and IDF1 , indicating more reliable association under occlusions, scale variations, and platform-induced motion.

On VisDrone2021-MOT, AeroGraph ranks first on the key metrics (HOTA/MOTA/IDF1) in Table I. The higher MT and lower ML further suggest improved trajectory continuity in crowded scenes.

On UAVDT, AeroGraph attains the top HOTA and IDF1 in Table I, demonstrating robust association across diverse aerial sequences. The gains on HOTA and IDF1 indicate that the improvements mainly come from stronger association quality under noisy motion and crowded interactions.

D. Ablation Study and Analysis

We conduct ablation studies on VisDrone2021-MOT using an online baseline that removes BG-STA, UGMF, and TMWC. As shown in Table II, the baseline achieves HOTA=45.423, MOTA=46.123, and IDF1=56.546. Adding TMWC improves the results to HOTA=46.163, MOTA=46.500, and IDF1=57.574, indicating that trajectory memory and window-level consistency stabilize track representations. Incorporating UGMF further increases IDF1 to 58.262 and HOTA to 46.339,

TABLE I: Comparison Results on UAVDT and VisDrone2021 Datasets

Dataset	Method	HOTA↑	MOTA↑	IDF1↑	FN↓	FP↓	IDS↓	MT↑	ML↓	f_{ps} ↑
UAVDT	SORT [1]	60.4	66.414	77.119	19034	6505	160	201	36	20.02
	DeepSORT [2]	61.97	68.498	78.615	20035	4008	**61**	179	43	30.34
	ByteTrack [6]	62.179	68.754	78.76	20010	**3796**	102	182	42	**32.07**
	ByteTrack+ReID [6]	61.629	68.765	77.205	18067	6021	118	196	44	22.82
	BoT-SORT [7]	61.361	67.341	78.508	20296	4323	64	181	43	25
	UAVMOT [15]	61.22	67.901	78.084	19371	5121	69	190	42	28
	OC-SORT [8]	59.784	69.484	74.518	18246	6480	249	205	21	13.32
	StrongSORT [9]	58.296	49.25	72.61	**10668**	28032	134	**248**	**17**	10.58
	QDTrack [4]	61.244	70.259	76.146	17304	6420	92	198	45	13.42
	FairMOT [3]	49.126	51.185	66.471	33103	4136	110	107	87	25.5
	TrackFormer [5]	43.195	37.924	53.343	45197	5585	680	67	75	7.94
	SGT [16]	58.163	57.975	79.948	11658	20598	107	211	**17**	18
	UGT [17]	63.587	**71.628**	79.798	17285	4357	62	207	33	30.34
	Ours(AeroGraph)	**64.439**	70.412	**83.031**	14131	8450	**58**	238	19	21.5
VisDrone2021	SORT [1]	35.021	33.148	42.844	112483	20392	3525	329	523	28.65
	DeepSORT [2]	36.98	34.397	46.714	110899	21077	1784	386	517	18.53
	ByteTrack [6]	40.661	39.541	51.398	105518	16257	1581	507	538	**31.15**
	ByteTrack+ReID [6]	41.222	40.988	50.264	103245	15254	1512	510	525	18.63
	BoT-SORT [7]	42.42	41.652	56.843	103505	14114	1430	543	537	25.33
	UAVMOT [15]	38.133	38.685	45.15	108134	13610	3357	463	557	10.02
	OC-SORT [8]	42.917	41.573	52.049	132209	22019	2859	270	631	15.36
	StrongSORT [9]	36.594	33.282	42.525	185573	12214	1890	**836**	**368**	12.1
	QDTrack [4]	44.737	39.053	55.316	104570	24242	2627	669	504	10.25
	FairMOT [3]	31.102	12.806	37.745	114834	59997	3072	397	581	21.5
	TrackFormer [5]	35.544	25	51.11	141526	25856	1534	515	946	6.97
	SGT [16]	43.381	39.207	54.80	116052	**7707**	951	456	372	27.88
	UGT [17]	45.491	41.799	57.76	101074	15174	**618**	560	524	16.21
	Ours(AeroGraph)	**47.289**	**46.767**	**59.677**	**84167**	22758	2227	695	457	10.78

TABLE II: Ablation results on VisDrone2021-MOT.

Setting	HOTA↑	MOTA↑	IDF1↑
baseline	45.423	46.123	56.546
baseline + TMWC	46.163(+0.740)	46.500(+0.377)	57.574(+1.028)
baseline + TMWC + UGMF	46.339(+0.916)	46.210(+0.087)	58.262(+1.716)
baseline + TMWC + UGMF + BG-STA	**47.289**(+1.866)	**46.767**(+0.644)	**59.677**(+3.131)

showing that uncertainty-guided fusion suppresses mismatches and harmful updates when motion cues become unreliable. With BG-STA enabled, AeroGraph achieves the best performance (HOTA=47.289, MOTA=46.767, IDF1=59.677), corresponding to gains of +1.866, +0.644, and +3.131 over the baseline, respectively.

To better interpret these gains, we analyze behaviors related to association stability. For UGMF, Fig. 2 shows reliability diagrams of association-confidence scores. Compared with the baseline, UGMF produces a curve closer to the diagonal, indicating better calibration for fusion and update control. When binned by uncertainty, UGMF separates confident matches from risky ones: under high uncertainty it reduces over-confident false associations, while under low uncertainty it maintains high correctness, which is consistent with the IDF1 improvements in Table II.

For TMWC, Fig. 3 visualizes temporal drift of track embeddings. TMWC reduces both the mean drift and its variance; in an example sequence, the drift decreases from 116.24 ± 9.78 to 63.47 ± 4.32. The smaller magnitude and tighter distribution indicate smoother identity representations over time, making appearance matching less sensitive to short-term occlusions and noisy observations. This helps mitigate long-term feature drift and reduces identity confusion, consistent with the gains in Table II.

E. Qualitative Visualization

To qualitatively assess robustness in challenging UAV scenes, we compare AeroGraph with UGT on representative clips from VisDrone2021-MOT and UAVDT, as shown in Fig. 4. UGT is a strong online baseline in Table I, which enables a direct examination of association stability under dense occlusions, rapid scale changes, and platform-induced disturbances.

In the crowded intersection clip from VisDrone2021 (Fig. 4(a)), small and visually similar targets frequently interact and undergo short-term occlusions. UGT presents unstable associations in crowded regions, with identity drift and track interruptions indicated by red arrows, while missed tracking of small targets under heavy occlusion is marked by yellow dashed boxes. AeroGraph maintains more continuous trajectories and recovers correct identities after short occlusions, resulting in fewer interruptions and more stable identity assignments.

In the nighttime traffic clip from UAVDT (Fig. 4(b)), illumination changes, motion blur, detection jitter, and platform motion jointly degrade motion cues and observation quality. UGT suffers from target loss and erroneous reconnections in distant small-target areas and during close vehicle interactions, which leads to frequent identity switches. AeroGraph preserves stable trajectories across consecutive frames and maintains consistent identities for key targets, producing more continuous tracks.

Overall, Fig. 4 is consistent with the quantitative results in Table I: AeroGraph reduces identity drift and trajectory fragmentation under dense occlusions and strong platform motion. This behavior also aligns with the ablation results in Table II, where BG-STA strengthens window-level association evidence, UGMF stabilizes fusion and updates under unreliable motion, and TMWC mitigates long-term representation drift.

IV. Conclusion

We presented AeroGraph, an online UAV MOT framework designed for stable association under dense small targets, frequent occlusions, and unreliable motion cues in aerial videos. AeroGraph integrates BG-STA for window-level spatiotemporal large-graph inference, UGMF for reliability-aware motion fusion, and TMWC for stabilizing long-term identity representations with trajectory memory and window-level consistency. Experiments on VisDrone2021-MOT and UAVDT show consistent gains on HOTA and IDF1 with fewer identity errors, demonstrating more robust association under challenging UAV conditions. Future work will investigate cross-window long-term memory and more efficient large-graph inference for real-time deployment.

References

[1] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. Ieee, 2016, pp. 3464–3468.

[2] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[3] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *International journal of computer vision*, vol. 129, no. 11, pp. 3069–3087, 2021.

[4] J. Pang, L. Qiu, X. Li, H. Chen, Q. Li, T. Darrell, and F. Yu, "Quasi-dense similarity learning for multiple object tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 164–173.

[5] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "Trackformer: Multi-object tracking with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8844–8854.

[6] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," in *European conference on computer vision*. Springer, 2022, pp. 1–21.

[7] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "Bot-sort: Robust associations multi-pedestrian tracking," *arXiv preprint arXiv:2206.14651*, 2022.

[8] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, "Observation-centric sort: Rethinking sort for robust multi-object tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 9686–9696.

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

[10] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe, "Hota: A higher order metric for evaluating multi-object tracking," *International journal of computer vision*, vol. 129, no. 2, pp. 548–578, 2021.

[11] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," *arXiv preprint arXiv:1804.07437*, 2018.

[12] P. Zhu, L. Wen, D. Du, X. Bian, H. Ling, Q. Hu, H. Wu, Q. Nie, H. Cheng, C. Liu *et al.*, "Visdrone-vdt2018: The vision meets drone video detection and tracking challenge results," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[13] L. Wen, P. Zhu, D. Du, X. Bian, H. Ling, Q. Hu, J. Zheng, T. Peng, X. Wang, Y. Zhang *et al.*, "Visdrone-mot2019: The vision meets drone multiple object tracking challenge results," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[14] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 370–386.

[15] S. Liu, X. Li, H. Lu, and Y. He, "Multi-object tracking meets moving uav," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8876–8885.

[16] J. Hyun, M. Kang, D. Wee, and D.-Y. Yeung, "Detection recovery in online multi-object tracking with sparse graph tracker," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 4850–4859.

[17] C. Deng, J. Wu, Y. Han, W. Wang, and J. Chanussot, "Learning a robust topological relationship for online multiobject tracking in uav scenarios," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[18] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[19] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the society for industrial and applied mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[20] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International conference on machine learning*. Pmlr, 2017, pp. 1263–1272.

[21] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[22] J. Cai, M. Xu, W. Li, Y. Xiong, W. Xia, Z. Tu, and S. Soatto, "Memot: Multi-object tracking with memory," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8090–8100.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[24] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[25] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[26] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3702–3712.

[27] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 79–88.

[28] G. Peyré and M. Cuturi, *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

[29] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, vol. 26, 2013.