Journal of AI-Driven Engineering Innovation and Applications Volume 1 Issue 1
https://doi.org/10.67014/jadeia.v1i1.0015

Communication

Ultra-lightweight ByteTrack-YOLOX-S for UAV-Based Vehicle Tracking on UAVDT

Mohammad Fatin Fatihur Rahman[1,*], Mohammed Juhaer Hamim[2], Rafid Rafsani[2]

[1] *Department of Electrical and Computer Engineering, University of Windsor, Windsor, N9B 3P4, Canada*

[2] *Department of Electrical and Electronic Engineering, Ahsanullah University of Science and Technology, Dhaka-1208, Bangladesh*

*** Corresponding author:** Mohammad Fatin Fatihur Rahman, rahman7q@uwindsor.ca

ABSTRACT

Unmanned aerial vehicle (UAV) traffic videos contain small targets, frequent occlusion, moving-camera effects, and large confidence variations, which make lightweight multi-object tracking difficult. This short communication evaluates the authors' existing YOLOX-S detector and ByteTrack association pipeline on UAVDT under a 10 GB VRAM constraint; it does not introduce a new detector or a new association algorithm. Across 20 evaluated files, 16,583 frames, and 340,906 ground-truth object instances, the training protocol produced 74.8% IDF1, 62.7% recall, 97.4% precision, and 61.0% MOTA, whereas the test protocol produced 48.0% IDF1, 37.2% recall, 78.9% precision, and 27.2% MOTA. On the test protocol, false negatives account for 86.23% of the errors counted by MOTA, while identity switches account for only 0.052%; consequently, the low switch count mainly reflects limited detector coverage rather than consistently strong identity preservation. The revised analysis further explains the two-confidence threshold mechanism, the train-to-test degradation, the algorithmic complexity, and the limitations of inferring edge-device latency from desktop-GPU experiments. The results position ByteTrack–YOLOX-S as a compact and transparent baseline whose most important improvement direction is small-object candidate recall under altitude, illumination, and occlusion changes.

Keywords: UAVDT; ByteTrack; YOLOX-S; multi-object tracking; vehicle tracking; intelligent transportation systems; UAV surveillance

ARTICLE INFO

Received: 14 July 2026
Accepted: 26 August 2026
Available online: 16 September 2026

1. Introduction

UAV-based traffic monitoring can be used to monitor road segments that are not easily accessible by roadside sensors, and provide support to the transportation planning, road safety, congestion analysis, incident awareness and infrastructure management processes. Aerial sensing is especially beneficial for highway and intersection monitoring, temporary work zones and emergency response situations in intelligent transportation systems (ITS). The problem is, however, that UAV traffic video also presents challenging computer-vision scenarios: Road users are small in size, the camera is not fixed, they are frequently occluded, and the same vehicle can appear significantly different over short time spans as a result of a change in viewpoint or altitude.

The UAVDT benchmark is popular for the UAV-view vehicle detection and tracking task due to its urban road scenes recorded under different weather, altitude, camera view, occlusion, illumination attributes[1,2]. UAVDT is relevant to ITS research, but it poses a challenging tracking environment. A practical implementation of a

UAV tracker should therefore be able to achieve a compromise between accuracy, identity consistency, and computational simplicity.

ByteTrack is a tracking-by-detection approach that allows for the retention of low-confidence detections, thereby enhancing association[3,4]. YOLOX-S is a small, anchor-free detector with a decoupled detection head, which is suitable for the lightweight UAV tracking as a front end[5,6]. In recent years, UAV tracking has been enhanced by small object features, temporal context, appearance embedding, and transformers, optical flow, occlusion reasoning, or camera-motion compensation[5,7–17]. These additions help improve the accuracy of the benchmarks, but make it more challenging to determine whether a simple resource-constrained pipeline fails primarily at detection or at association.

The problem we study in this work is thus the diagnostic assessment of an unchanged YOLOX-S–ByteTrack pipeline under 10 GB VRAM. The scientific contribution is not the compact no-ReID pipeline itself, but the careful investigation of the behaviour of the pipeline on the UAVDT training and test protocol, the reasons for the degradation of the test results, and the nature of the errors that should be targeted first before deployment at the edge. The contributions are fourfold: (i) the original ultra-lightweight ByteTrack– YOLOX-S experiment and all reported metrics are preserved; (ii) the tracking formulation is explicitly notated in terms of detection, association, lifecycle, and complexity; (iii) train to test degradation, MOTA error composition are quantified; and (iv) confidence-threshold behavior, visual failure-prone conditions, computational efficiency, edge limitations, and mobile-network applications are discussed without inventing unmeasured results.

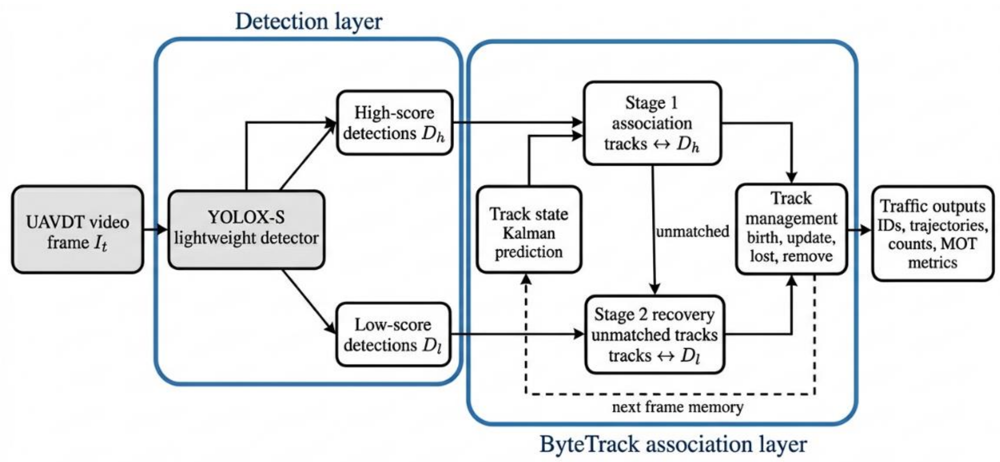

Figure 1. Algorithm paradigm of the ultra-lightweight ByteTrack–YOLOX-S tracking pipeline for UAVDT based traffic monitoring. UAV frames are processed by YOLOX-S, split into high- and low-confidence detections, associated with predicted tracks through ByteTrack's two-stage matching process, and converted into traffic outputs such as identities, trajectories, counts, and MOT metrics.

2. Related work

The growth of UAV-based intelligent transportation systems has drawn increasing attention to multi-object tracking (MOT) in aerial traffic monitoring. Prior research has developed deep-learning detectors and tracking-by-detection pipelines to address small object scale, occlusion, and moving-camera viewpoints. Modern lightweight approaches can maintain useful tracking accuracy on benchmarks such as UAVDT while emphasizing computational efficiency and real-time deployment.

Wang et al.[7] developed an ultra-lightweight aerial detection-and-tracking design for intelligent transportation, while You et al.[8] combined an improved YOLOv8 detector with ByteTrack for vehicle tracking. Xiao et al.[9] enhanced online UAV tracking with temporal context and spatial topology, Ma et al.[10] introduced asymmetric feature enhancement for one-shot UAV tracking, and Dang et al.[11] integrated occlusion perception and motion compensation. These works directly target difficult UAV conditions, but they use additional temporal, appearance, or motion modules beyond the plain YOLOX-S–ByteTrack pipeline evaluated here.

More recent systems continue this specialization. GAO-Tracker uses a holistic transformer and multiple-feature trajectory matching, Liu et al.[13] combine FB-YOLOv8 with multi-scale OSNet features for scale-variant vehicle tracking, and TASONet combines spatial enhancement with temporal modeling around a ByteTrack-based tracker[12,14]. Their published UAVDT results provide useful context, although direct ranking is limited by differences in detector backbone, data preparation, evaluation protocol, and hardware. **Table 1** summarizes the studies related to UAV tracking designs and evaluates their connection to this study.

Table 1. Focused summary of related UAV tracking designs and their connection to this study.

Method	Primary UAV challenge	Additional mechanism	Relation to this study
You et al.[8]	Vehicle detection and association	Improved YOLOv8 + ByteTrack	Same association family, different detector
Xiao et al.[9]	Temporal ambiguity and spatial relations	Temporal context + topology	Adds learned temporal/spatial cues
AsyUAV[11]	Detection–ReID task conflict	Asymmetric feature enhancemen	Uses appearance/ReID information
OMCTrack[10]	Occlusion and UAV motion	Occlusion perception + motion compensation	Adds explicit recovery and compensation
GAO-Tracker[12]	Scale, occlusion, and rapid motion	Holistic transformer + feature matching	Uses richer feature and trajectory modeling
TASONet[14]	Small-object representation and temporal coherence	Spatial enhancement + temporal modeling	ByteTrack-based but architecturaly heavier
This study	Resource-bounded baseline diagnosis	YOLOX-S + two-stage IoU association	No external ReID, flow, or compensation branch

In summary, the literature shows that specialized modules can substantially improve UAV tracking. The narrower purpose of this paper is to retain a simple baseline and identify, from its complete error profile, whether future effort should first be placed on detector recall, identity association, or deployment optimization.

3. Methodology

3.1. Dataset and evaluation workload

UAVDT contains UAV-view urban traffic sequences with vehicle annotations and scene attributes relevant to detection and tracking. The evaluated workload in this study contains 20 files, 16,583 analyzed frames, and 340,906 ground-truth object instances. **Table 2** summarizes the dataset and evaluation setting.

Table 2. Dataset and evaluation information.

Item	Description
Dataset	UAVDT
Scene type	Urban roads, highways, intersections, and traffic-surveillance scenes
Scale	100 sequences and approximately 80,000 annotated frames in the benchmark
Object focus	Vehicle categories, including cars, trucks, buses, and related vehicle labels depending on the annotation format
Frame properties	30 FPS; benchmark frames reported at 1080 × 540 pixels
Challenges	Small objects, occlusion, moving camera, viewpoint shift, altitude change, fog, night, and daylight variation
Evaluated splits	ByteTrack–YOLOX-S training and test protocol splits

Two ground-truth quantities are distinguished in the revised notation. NGT denotes the number of annotated object instances across frames and is the denominator used by MOTA, whereas Ntraj denotes the number of ground-truth trajectories used in the mostly tracked, partially tracked, and mostly lost diagnostics. This distinction avoids interpreting the trajectory counts as frame-level object totals.

3.2. Tracking pipeline

Figure 1 shows the algorithm paradigm. For each UAV frame *It*, YOLOX-S produces a detection set *Dt* = *{d t i} n I* = 1 *t*, where *d t i* = (*b t i*, *ct i*, *st i*) contains bounding box *b t i*, class label *c t i*, and confidence score *s t i*. Detections are divided into high and low-confidence sets:

$$D_h^t = \{d_i^t \mid s_i^t \geq \tau_h\}, \qquad D_l^t = \{d_i^t \mid \tau_l \leq s_i^t < \tau_h\} \tag{1}$$

where $0 \leq \tau l < \tau h \leq 1$. Detections below τl are removed before association. Let $Tt-1 = \{Tj\}^{m\,t}_{j=1}$be the active tracks. A constant-velocity Kalman model predicts the current box ˆ*b t j* for every track. The geometric association cost between predicted track *j* and detection *i* is

$$C_{ji}^t = 1 - IoU(\hat{b}_j^t, b_i^t) \tag{2}$$

with pairs outside the matching gate treated as infeasible. The Hungarian algorithm first matches predicted tracks to *Dh t*. Unmatched tracks are then matched to *Dl t*, allowing low-score candidates caused by blur, partial occlusion, small scale, or weak illumination to recover an existing trajectory. Unmatched reliable detections initialize new tracks, matched tracks update their Kalman states, and tracks that remain unmatched beyond the buffer are removed.

Algorithm 1 Two-stage ByteTrack association used in the evaluated pipeline

Require: UAV frame *It*, active tracks *Tt*−1, thresholds τh, τl

Ensure: Updated tracks *Tt*

1: Obtain *Dt* from YOLOX-S and partition it into *Dh t* and *Dl t* using Equation (1)

2: Predict each active track by the Kalman motion model

3: Compute IoU costs and match predicted tracks with *Dh t*

4: Match the remaining tracks with *Dl t*

5: Update matched tracks and initialize tracks from unmatched reliable detections

6: Retain or remove unmatched tracks according to the track buffer

7: Return active identities, boxes, and trajectories

The formulation relies on three practical assumptions. First, adjacent-frame motion is sufficiently smooth for Kalman prediction to place the track near the next detection. Second, a valid vehicle candidate survives detector scoring above τl. Third, sufficient box overlap remains for IoU matching. UAV altitude changes, abrupt camera motion, severe occlusion, and low illumination can violate these assumptions. ByteTrack's second stage helps only when a low-score box still exists; it cannot recover a vehicle that the detector fails to output. The design remains lightweight because it does not require a separate appearance gallery, re-identification network, optical-flow branch, or camera-motion-compensation module.

3.3. Experimental setup

The experiment was conducted on a single desktop GPU. The same ByteTrack–YOLOX-S configuration was evaluated on the training and test protocol splits. **Table 3** records the implementation setup.

Table 3. Experimental Setup for ByteTrack–YOLOX-S.

Parameter	Value
Operating system	Ubuntu 24.04.1 LTS
CPU	AMD Ryzen 5 3500X 6-Core Processor
GPU	NVIDIA RTX 3080, 10 GB VRAM

Table 3. (*Continued*).

Parameter	Value
RAM	46 GB
Framework	PyTorch/Torch 2.4.1
CUDA version	11.8
Detector	YOLOX-S
Tracker	ByteTrack
Batch size	6
Learning rate	0.0001
Input size	540 × 1024 pixels
Epochs	80
Dataset	UAVDT training/test protocol splits

The high- and low-score thresholds, IoU gates, and track-buffer settings were held fixed when generating the two aggregate result sets. Their exact numerical values were not preserved in the supplied aggregate record; therefore, the threshold analysis in Section 4.5 is expressed symbolically rather than assigning unsupported values. No embedded edge device was used, and end-to-end FPS, latency, power, and thermal measurements were not recorded.

3.4. Evaluation metrics

Let IDTP, IDFP, and IDFN denote identity true positives, identity false positives, and identity false negatives, respectively. Let TP, FP, FN, and IDs denote true positives, false positives, false negatives, and identity switches. Let *N*GT denote the total number of ground-truth object instances across frames.

The primary metrics are

$$IDP = \frac{IDTP}{IDTP + IDFP} \tag{3}$$

$$IDR = \frac{IDTP}{IDTP + IDFN} \tag{4}$$

$$IDF1 = \frac{2IDTP}{2IDTP + IDFP + IDFN} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$MOTA = 1 - \frac{FP + FN + IDs}{N_{GT}} \tag{8}$$

For matched target j, let dj be the localization distance or overlap error and let M be the number of matches. Then

$$MOTP = \frac{\sum_{j=1}^{M} d_j}{M} \tag{9}$$

IDF1 summarizes identity consistency, recall measures coverage, precision measures reliability, MOTA penalizes missed detections, false alarms, and identity switches, and MOTP describes localization quality. MT, PT, and ML denote mostly tracked, partially tracked, and mostly lost trajectories. FM denotes track fragmentation, while IDt, IDa, and IDm denote identity transfer, identity ascension, and identity migration errors.

3.5. Algorithmic complexity and efficiency interpretation

Let m be the number of predicted tracks, $nh = |Dh^t|$, $nl = |Dl^t|$, and mu the unmatched tracks after the first association. Kalman prediction is $O(m)$; constructing the two IoU cost matrices is $O(mn_h + m_u n_l)$ in time and memory; and the two Hungarian assignments have worst-case complexity $O(k_h^3 + k_l^3)$, where $k_h = \max(m, n_h)$ and $k_l = \max(m_u, n_l)$. In ordinary traffic frames these candidate sets are much smaller than the image tensor, so detector inference is expected to dominate total processing cost.

The original YOLOX report lists YOLOX-S at approximately 9.0 million parameters, 26.8 GFLOPs, and 9.8 ms detector latency at 640 × 640 with FP16 and batch size 1 on a Tesla V100[3]. These published values describe the reference detector, not the present 540 × 1024 RTX 3080 experiment, and they must not be interpreted as measured end-to-end tracking latency for this study. Compared with recent UAV trackers in **Table 1**, the evaluated pipeline avoids the additional learned ReID, optical-flow, transformer, temporal-modeling, and camera-motion branches. Actual efficiency on an edge platform still requires device-specific profiling.

4. Results

4.1. Aggregate workload

The aggregate scale of the evaluation is shown in **Table 4**. The workload contains a substantial number of vehicle instances, so false-negative behavior is particularly important for interpreting the results.

Table 4. Aggregate statistics across evaluated files.

Statistic	Value
Total files processed	20
Total frames analyzed	16,583
Total ground-truth object instances	340,906

4.2. Tracking performance

Table 5 reports the core metrics. The training split achieved 74.80% IDF1 and 61.00% MOTA, while the test split achieved 48.00% IDF1 and 27.20% MOTA. Precision remained relatively high on both splits, but recall decreased sharply on the test split.

Table 5. Core ByteTrack–YOLOX-S metrics on UAVDT training and test protocol splits.

Experiment	IDF1	IDP	IDR	Recall	Precision	MOTA	MOTP
ByteTrack–YOLOX-S training set	74.80%	61.50%	95.50%	62.70%	97.40%	61.00%	0.196
ByteTrack–YOLOX-S test set	48.00%	74.80%	35.30%	37.20%	78.90%	27.20%	0.212

Table 6 evaluates the trajectory and error diagnostics for ByteTrack-YOLOX-S. The test set achieved more False Positives and False Negatives than the training set while achieving lower MT, PT etc.

Table 6. Trajectory and Error Diagnostics for ByteTrack–YOLOX-S.

Experiment	*N* traj	MT	PT	ML	FP	FN	IDs	FM	IDt	IDa	IDm
ByteTrack–YOLOX-S training set	1366	544	448	374	7190	157,686	121	834	41	70	19
ByteTrack–YOLOX-S test set	1227	363	337	527	34,040	213,952	128	617	27	94	14

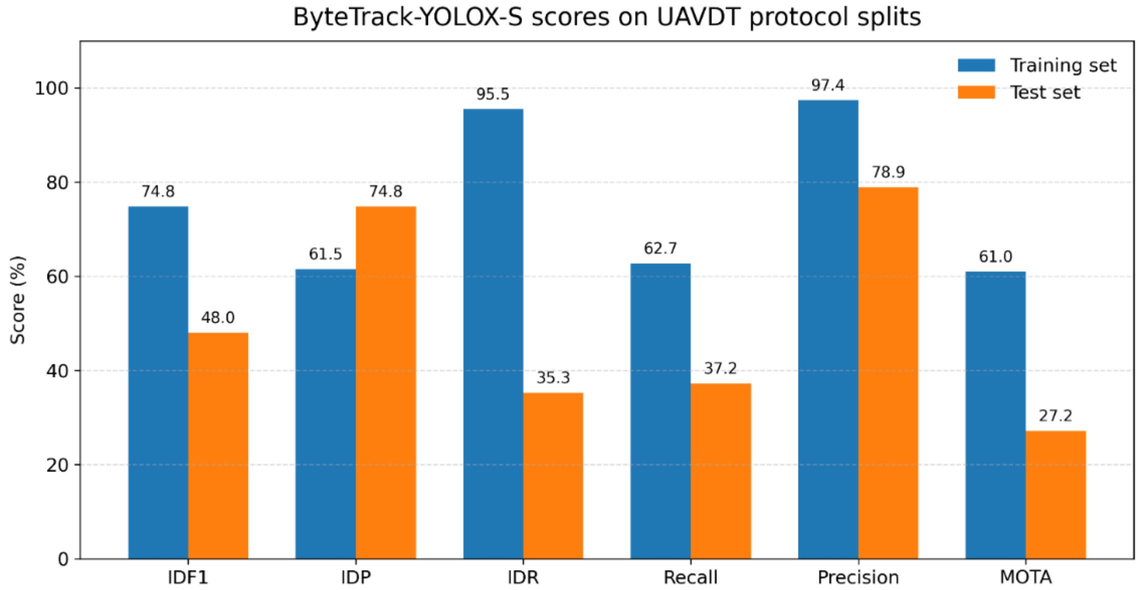

Figure 2. Core metric comparison for ByteTrack–YOLOX-S on the UAVDT training and test protocol splits.

Its evident from **Figure 2** that the test split maintains moderate precision but shows lower recall, IDR, IDF1, and MOTA.

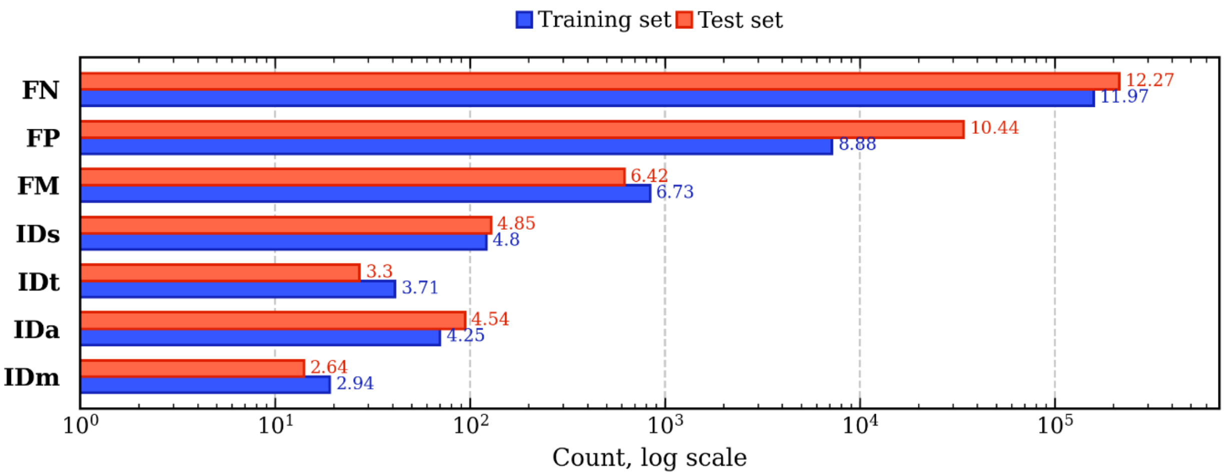

Figure 3. Training and test split error profiles for ByteTrack–YOLOX-S on UAVDT, shown on a logarithmic count scale. False negatives dominate both splits, while identity-related errors remain comparatively small.

Figure 3 depicts that both splits have a high number of false negatives, while identity-related errors are relatively small.

Figure 4. Qualitative UAVDT tracking examples using ByteTrack–YOLOX-S. The panels illustrate aerial traffic scenes with bounding boxes and identity labels over vehicles under different road geometries and traffic densities.

Figure 4 shows the aerial traffic scenes with the vehicles marked with a bounding box and identity label for various traffic situations and road layouts.

4.3. Training-to-test degradation

Table 7 quantifies the change rather than only listing the two result rows. The largest decrease is in IDR (-60.2 percentage points), followed by MOTA(-33.8), IDF1 (-26.8), recall (-25.5), and precision (-18.5). IDP increases by 13.3 points even as IDR collapses, indicating that the identities that are retained are comparatively selective, but a much smaller fraction of the ground-truth identity evidence is covered.

Table 7. Explicit performance change from the training to the test protocol.

Metric	**Training**	**Test**	**Change**
IDF1	74.80%	48.00%	-26:8 points
IDP	61.50%	74.80%	+13:3 points
IDR	95.50%	35.30%	-60:2 points
Recall	62.70%	37.20%	-25:5 points
Precision	97.40%	78.90%	-18:5 points
MOTA	61.00%	27.20%	-33:8 points
MOTP	0.196	0.212	+0:016

The degradation curve is similar to that of a detector coverage problem, in which recall and IDR drop much more quickly than the number of ID switches changes. The “held out” data include challenging scales, illumination conditions, Vehicle scores can be distorted by occlusion states, camera viewpoints, and any of these can decrease valid vehicle scores or diminish box overlap. This mechanism is an evidence-informed interpretation and not a condition-wise causal experiment as the results retained are aggregated without per-attribute logs.

4.4. Contextual comparison with recent UAV trackers

Table 8 compares the test result with the latest published UAVDT results. The methods were not reimplemented in exactly the same protocol so the values are not a controlled leaderboard; they are contextual. Specifically, the variety of detector backbones, training data, image resolution, split handling and hardware can differ between the studies.

Table 8. Contextual UAVDT performance comparison reported by the respective studies.

Method	MOTA	IDF1	Additional design elements
GAO-Tracker[12]	61.7%	–	Holistic transformer and multiple-feature trajectorymatching
Scale-variant BoT-SORT[13]	46.1%	65.3%	FB-YOLOv8 and multi-scale OSNet appearance features
OMCTrack[10]	85.5%	88.5%	Occlusion perception and adaptive motion compensation
TASONet[14]	75.97%	88.51%	Spatial enhancement and temporal modeling
ByteTrack–YOLOX-S (this work, test)	**27.2%**	**48.0%**	No external ReID, optical flow. transformer, or motion-compensation branch

The comparison does not support a claim of state-of-the-art superiority. Instead, it shows the accuracy cost of retaining a plain compact pipeline and supports the paper's revised contribution: a transparent baseline diagnosis that identifies which missing capabilities account for the gap to specialized methods.

4.5. Error interpretation and threshold sensitivity

As illustrated in the training split, the ByteTrack–YOLOX-S achieves identities preservation when there are enough candidate detections from the detector. It has an IDR of 95.50% and a precision of 97.40%, demonstrating a good rate of coverage of identities and positive-prediction quality in that context. The test split is more difficult: IDR decreases to 35.30%, recall decreases to 37.20%, and MOTA decreases to 27.20%. The main drawback of the test-set precision is not false alarms - it is missed vehicles - 78.90% precision.

This is confirmed by **Table 6**. The number of identity switches is nearly identical for both splits (121 for the training split and 128 for the test split). In comparison, there were 213,952 test-set false negatives. The test-set is used for counting the three error terms. The total number of MOTA is 248,120, of which the false negative rate is 86.23%, the false positive rate is 13.72%, and only 0.052% are identity switches. The shares corresponding to the training protocol are 95.57%, 4.36%, and 0.073%.

Low identity switch count is not an indicator of high false negative count. The identity switch can only take place following the detection and association of a target. If the vehicle is not in the output of the detector, then it is a false negative and there is no association event and thus no chance of an identity switch. If the trajectory is either shorter or broken, the number of possible switch events may also be lessened. It is not advisable, therefore, to conclude from the low switch count alone that there is a strong identity tracking.

The score thresholds provide a direct mechanism for this behavior. Let S_c be the confidence score of a valid vehicle under condition c, such as high altitude or low illumination, and let F_c be its cumulative distribution. The score-induced probability that a valid candidate is discarded before ByteTrack is

$$P_{miss}^{score}(c;\tau_l) = P(S_c < \tau_l) = F_c(\tau_l) \tag{10}$$

Increasing τ_l cannot reduce this component of the false-negative rate. High altitude, small target size, blur, and poor illumination can shift Sc toward lower values, placing more valid boxes near or below τ_l. Lowering τ_l may recover candidates, but it also increases false positives, the second-stage candidate set, and association cost. Changing τ_h mainly reallocates candidates between the first and second stages and can affect which detections are considered reliable for track initialization. A condition-stratified threshold sweep is required to identify the best trade-off; it cannot be reconstructed from the available aggregate tables.

4.6. Visual failure-prone conditions

Figure 5 enlarges difficult regions from the original qualitative examples. The bottom-left daytime scene contains many very small vehicles; the nighttime intersection contains overlapping vehicles and reflections; and the elevated nighttime view contains distant low-contrast targets. These are precisely the conditions in which confidence can fall below τ_l or IoU overlap can become unreliable.

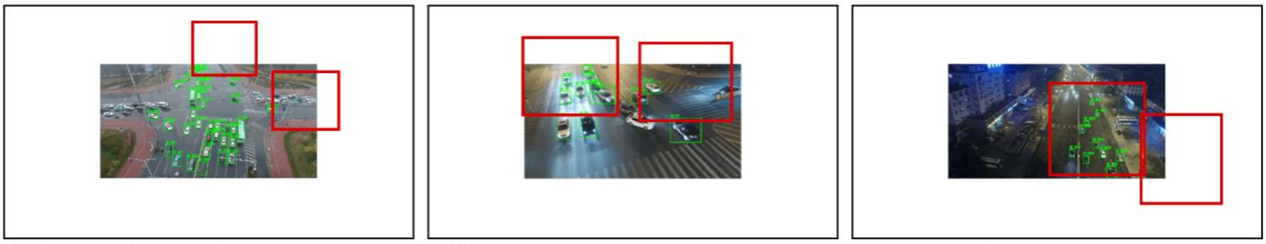

Figure 5. Failure-prone conditions extracted from the original qualitative UAVDT examples: tiny targets and scale variation, overlap or partial occlusion, and low illumination or contrast. Red boxes identify visually difficult regions for mechanism-level interpretation; they do not represent an additional frame-level ground-truth experiment.

5. Discussion

Ultra-lightweight UAV tracking is relevant to ITS because aerial platforms can rapidly observe traffic conditions without depending on fixed infrastructure. The ByteTrack–YOLOX-S pipeline has several practical advantages: the detector and tracker are modular, the association logic is transparent, and the model can be trained and evaluated on a 10 GB VRAM GPU. These properties make it an efficient baseline for traffic monitoring, vehicle counting, queue analysis, and incident-aware UAV surveillance.

The results also reveal that there is a limit to how much information can be gained from light weight association if detections go missing. Recall is the main constraint for the test split. This is significant for transportation applications as vehicle under-counts can result in inaccurate estimates of traffic density, queue length, and incident severity. Going forward, these improvements should focus on detection recall with varying scales at low angles, temporal recovery for forgotten vehicles, and confidence calibration for missed detections with high scores but insufficient data.

Efficiency-wise, the current evidence suggests feasibility only on the desktop-GPU in 10 GB VRAM. Does not support real-time edge deployment. Any deployment study should include end-to-end preprocessing, detector inference, association (both association stages), data transfer, power draw, memory peak, thermal throttling, and FPS, all with the actual platform that it will be deployed on. The analytical complexity in Section 3.5 determines variables that affect association overhead, but it cannot be used to replace such measurements.

The tracking outputs also have potential value in mobile and edge networks. A UAV can transmit compact identities, trajectories, counts, or confidence summaries to a mobile-edge-computing server instead of continuously uploading all raw video. Learning-aided sensor activation and mobile charging schedules can help maintain energy-efficient sensing coverage[18]; multi-UAV MEC schedulers can jointly coordinate trajectories, energy renewal, and application placement[19]; and dynamic clustering or stochastic-game scheduling can distribute computation among UAV swarms and edge servers [20]. Within such systems, the lightweight tracker acts as the perception front end, while communication and computation policies decide where and when the resulting traffic information should be processed.

Limitations

This study has six primary limitations. First, it evaluates one detector–tracker family, so the results should be considered a compact baseline rather than a universal conclusion about UAV MOT. Second, testing was carried out on a desktop RTX 3080 GPU; embedded-device latency, power consumption, memory peaks, and thermal behavior were not measured. Third, the analysis is aggregate-level and does not include a controlled per-weather, per-altitude, per-view, per-density, or per-illumination threshold sweep. Fourth, the exact numerical tracking thresholds and lifecycle gates were not preserved in the supplied aggregate record. Fifth, no external re-identification module, cameramotion-compensation module, optical-flow branch, or multi-sensor fusion input was used. Sixth, the literature comparison uses values reported by the respective papers and is not an identical-protocol reimplementation. These limitations are now stated explicitly to prevent overgeneralization.

6. Conclusion

This paper retained and re-examined the original ByteTrack–YOLOX-S experiment as an ultra-lightweight UAV-based vehicle-tracking baseline. The training protocol reached 74.80% IDF1 and 61.00% MOTA, while the test protocol reached 48.00% IDF1 and 27.20% MOTA. The 33.8-point MOTA decrease is accompanied by a 25.5-point recall decrease and a 60.2-point IDR decrease. False negatives account for 86.23% of the test errors counted by MOTA, whereas identity switches account for only 0.052%, confirming that the dominant limitation is candidate generation rather than excessive identity reassignment.

The revised formulation shows why the low-score association stage can recover weak detections but cannot recover boxes that fall below the low threshold or are never generated. It also clarifies that the compact design avoids external ReID, optical-flow, transformer, and camera-motion modules, but that actual edge latency and power remain unmeasured. Future work should therefore prioritize condition-aware small-object recall, threshold validation, low-cost temporal recovery, and device-level profiling while preserving the simplicity of the original pipeline.

Data availability

The UAVDT benchmark is available from its original release.

Ethical approval

Not applicable.

Informed consent statement

Not applicable.

Conflict of interest

The authors declare no conflict of interest.

References

1. Du D, Qi Y, Yu H, et al. The unmanned aerial vehicle benchmark: Object detection and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018; pp. 370–386. doi: 10.1007/978-3-030-01249-6_23.
2. Li H, Liu X, Li G. A benchmark for UAV-view natural language-guided tracking. *Electronics* 2024. 13(9): 1706, 2024. doi: 10.3390/electronics13091706.
3. Zhang Y, Sun P., 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 Computer Vision – ECCV 2022, Lecture Notes in Computer Science, 2022, pp. 1–21. https://doi.org/10.1007/978-3-031-20047-2_1.
4. Ma LV, Hussain MI, Park J, et al. Adaptive confidence threshold for ByteTrack in multi-object tracking. In: Proceeding of the 2023 International Conference on Control, Automation and Information Sciences (ICCAIS); 2023. doi: 10.1109/ICCAIS59597.2023.10382403.
5. Ge Z, Liu S, Wang F, et al. YOLOX: Exceeding YOLO series in 2021. arXiv preprint arXiv:2107.08430, 2021. doi: 10.48550/arXiv.2107.08430.
6. Jiao L, Abdullah MI. YOLO series algorithms in object detection of unmanned aerial vehicles: A survey. *Service Oriented Computing and Applications* 2024. 18(3): 269–298. doi: 10.1007/s11761-024-00388-w.
7. Wang Q, Ye G, Chen Q, et al. A UAV perspective based ultra-lightweight target detection and tracking algorithm for intelligent transportation. *Complex & Intelligent Systems* 2025. 11: 73. doi: 10.1007/s40747-024-01687-7.
8. You L, Chen Y, Xiao C, et al. Multi-object vehicle detection and tracking algorithm based on improved YOLOv8 and ByteTrack. *Electronics* 2024. 13(15): 3033. doi: 10.3390/electronics13153033.
9. Xiao C, Cao Q, Zhong Y, et al. Enhancing online UAV multi-object tracking with temporal context and spatial topological relationships. *Drones* 2023. 7(6): 389. doi: 10.3390/drones7060389.

10. Dang Z, Sun X, Sun B, et al. OMCTrack: Integrating occlusion perception and motion compensation for UAV multi-object tracking. *Drones* 2024. 8(9): 480. doi: 10.3390/drones8090480.
11. Ma J, Liu D, Qin S, et al. An asymmetric feature enhancement network for multiple object tracking of unmanned aerial vehicle. *Remote Sensing* 2024. 16(1): 70. doi: 10.3390/rs16010070.
12. Yuan Y, Wu Y, Zhao L, et al. Multiple object tracking in drone aerial videos by a holistic transformer and multiple feature trajectory matching pattern. *Drones* 2024. 8(8): 349. doi: 10.3390/drones8080349.
13. Liu S, Shen X, Xiao S, et al. A multi-scale feature-fusion multi-object tracking algorithm for scale-variant vehicle tracking in UAV videos. *Remote Sensing* 2025. 17(6): 1014. doi: 10.3390/rs17061014.
14. Ma R, Lai C, Sheng Q, et al. TASONet: A spatial enhancement and temporal modeling framework for UAV small object tracking. *Remote Sensing* 2026. 18(4): 561. doi: 10.3390/rs18040561.
15. Chen Z, Zhu E, Guo Z, et al. Predictive autonomy for UA1V remote sensing: A survey of video prediction. *Remote Sensing* 2025. 17(20): 3423. doi: 10.3390/rs17203423.
16. Xu T, Gu J, Zhu X, et al. A tri-modal dataset and a baseline system for tracking unmanned aerial vehicles. arXiv preprint arXiv:2511.18344. 2025. doi.org/10.48550/arXiv.2511.18344.
17. Mai L, Dai C, Ma H, et al. Challenges, progress, and future directions of visual tracking in UAVs. In: Proceeding of 2025 12th International Conference on Dependable Systems and Their Applications (DSA); 2025. doi: 10.1109/DSA66321.2025.
18. Chen J, Yi C, Wang R, et al. Learning aided joint sensor activation and mobile charging vehicle scheduling for energy-efficient WRSN-based industrial IoT. *IEEE Transactions on Vehicular Technology* 2023. 72(4): 5064–5078. doi: 10.1109/TVT.2022.3224443.
19. Chen J, Yi C, Li J, et al. A triple learner based energy efficient scheduling for multi-UAV assisted mobile edge computing. In: Proceedings of the IEEE International Conference on Communications (ICC); 2023; Rome, Italy. pp. 3240–3245. doi: 10.1109/ICC45041.2023.10279356.
20. Li J, Yi C, Chen J, et al. A reinforcement learning-based stochastic game for energy-efficient UAV swarm-assisted MEC with dynamic clustering and scheduling. *IEEE Transactions on Green Communications and Networking* 2025. 9(1): 255–270. doi: 10.1109/TGCN.2024.3424449.