

EATF: Evidence-aware temporal filtering for recovering low-confidence UAV detections

Pham Xuan Cong*

Institute of Information Technology and Electronics, Academy of Military Science and Technology, 17 Hoang Sam, Nghia Do, Hanoi, Vietnam.

*Corresponding author: congpx@gmail.com

Received 13 Jun. 2026; Revised 4 Aug. 2026; Accepted 15 Sep. 2026; Published 15 Oct. 2026.

DOI: <https://doi.org/10.54939/1859-1043.j.mst.114.2026.159-166>

ABSTRACT

Small UAVs often produce weak and intermittent detections because of their small image size, background clutter, and rapid apparent motion. This paper presents Evidence-Aware Temporal Filtering (EATF), an interpretable offline post-processing method for a frozen detector. High-confidence detections are retained directly, whereas weaker detections are recovered only when their completed associated tracks satisfy fixed mean-confidence and cumulative-confidence thresholds. EATF does not retrain the detector, estimate motion states, interpolate boxes, or generate new detections. On DUT Anti-UAV Tracking, EATF increased frame-level recall from 0.4377 to 0.5319 and F1 from 0.5870 to 0.6416 relative to detector thresholding at 0.55. At matched recall, it improved precision from 0.7657 to 0.8084 and F1 from 0.6277 to 0.6416. Compared with default ByteTrack using post-tracking top-1 selection, EATF achieved higher recall and F1, with paired sequence-level bootstrap intervals supporting both gains. The current method remains retrospective, single-target, and vulnerable to persistent background false alarms under the evaluated DUT protocol.

Keywords: UAV detection; Temporal filtering; Low-confidence detections.

1. INTRODUCTION

Reliable visual detection of small unmanned aerial vehicles (UAVs) is important for airspace monitoring and counter-UAV surveillance. The task remains difficult because UAVs are often small, distant, visually similar to birds, and observed under substantial variations in scale, illumination, motion, and background complexity. Drone-vs-Bird studies identify small target size, motion blur, clutter, long observation distance, and inter-class similarity as major sources of error [1, 2]. Attention mechanisms and optimized YOLO-based architectures have been investigated to improve small- and long-range UAV detection [3, 4].

Most detectors nevertheless make confidence-based decisions independently for each frame. A conservative threshold suppresses false detections but removes valid weak UAV responses, whereas lowering the threshold improves recall while admitting more background detections. This trade-off is aggravated by imperfect confidence calibration [5]. Temporal methods based on optical flow, background modeling, temporal consistency, and integrated detection-tracking can improve robustness [6-10].

This paper presents Evidence-Aware Temporal Filtering (EATF), an interpretable post-processing rule for a frozen detector. EATF is not proposed as a new multi-object tracker. Greedy IoU association is used only to group spatially consistent detections and accumulate confidence evidence. High-confidence detections are retained directly, whereas weaker detections are recovered when their completed tracks satisfy fixed mean-confidence and cumulative-confidence thresholds. The method does not estimate motion states, maintain long-term identities, interpolate boxes, or generate new detections. Because decisions use completed-track statistics, the current implementation is retrospective and intended for offline video post-processing rather than zero-latency online tracking.

The contributions are threefold. First, EATF provides a transparent two-branch rule for direct retention and track-supported recovery of weak detections. Second, the rule is evaluated using local parameter sensitivity, temporal averaging, track-confirmation baselines, and ByteTrack [11]. Third, the frozen configuration is externally validated on DUT Anti-UAV Tracking through matched-recall comparison, paired sequence-level bootstrap analysis, top-1 protocol analysis, qualitative cases, and complete post-processing runtime measurements.

2. METHODOLOGY

2.1. Temporal association and track-level confidence evidence

Let frame t contain detections $D_t = \{d_{t,i}\}_{i=1}^{N_t}$, where each detection $d_{t,i}$ consists of a bounding box $b_{t,i}$, a class label, and a confidence score $s_{t,i}$. Predictions below the candidate threshold τ_{cand} are discarded before association. Detections are processed independently for each video in temporal order. For each remaining detection and active same-class track, EATF computes the IoU between the candidate box and the most recent track box. A pair is eligible when $IoU(b_{t,i}, b_{r,last}) \geq \tau_{IoU}$ and the number of intervening unmatched frames does not exceed G_{max} . Eligible pairs are sorted by decreasing IoU and greedily accepted under one-to-one assignment. Unmatched detections initialize new tracks, and tracks exceeding the gap limit are removed. No motion prediction, appearance descriptor, or camera-motion model is used.

After processing the video, each completed track r contains L_r detections with confidence scores $\{s_{r,k}\}_{k=1}^{L_r}$. EATF computes:

$$\bar{s}_r = \frac{1}{L_r} \sum_{k=1}^{L_r} s_{r,k}, \quad S_r = \sum_{k=1}^{L_r} s_{r,k} \quad (1)$$

where $\bar{s}_r$ is the mean confidence, and S_r is the cumulative confidence. Track length is used only to define these statistics and is not an independent acceptance condition.

2.2. Simplified EATF decision and output reconstruction

For a candidate detection $d_{t,i}$ assigned to completed track r , EATF applies:

$$\text{Keep}(d_{t,i}) = I[s_{t,i} \geq \tau_{high}] \vee (I[\bar{s}_r \geq \tau_{mean}] \wedge I[S_r \geq \tau_{sum}]) \quad (2)$$

where $I[\cdot]$ denotes the indicator function. The first branch retains a detection according to its own confidence. The second branch recovers weaker detections only when both track-level confidence conditions are satisfied. Because the statistics are computed from completed tracks, earlier detections may use evidence from later frames. Retained detections preserve their original frame indices, boxes, class labels, and confidence scores. EATF does not interpolate missing boxes, predict unobserved locations, or generate new detections.

An earlier formulation also required $L_r \geq 2$. This condition is redundant under the fixed thresholds. If $L_r = 1$, then $S_r = s_{t,i}$, so $S_r \geq 0.80$ already implies $s_{t,i} \geq 0.80$, and the detection is retained by the high-confidence branch. If $L_r \geq 2$, the condition is automatically satisfied. It is therefore omitted from the final rule.

3. EXPERIMENTAL SETUP

3.1. Datasets and evaluation protocol

The detector was developed on processed Drone-vs-Bird 2023 data [2], while DUT Anti-UAV Tracking was reserved for external validation [12]. Table 1 summarizes the evaluated splits. Drone-vs-Bird includes open-sky, cloud, mountainous, urban, and backlit scenes, whereas DUT additionally contains vegetation, fields, buildings, haze, and low-contrast backgrounds. These

environment descriptions are qualitative rather than exhaustive scene labels.

Table 1. Summary of the datasets used in this study.

Dataset/split	Sequences	Frames	Positive/Negative frames	Resolution	Median target size
Drone-vs-Bird train	53	56,622	55,615 / 1,007	720×576 - 3840×2160	29×18 px
Drone-vs-Bird validation	11	10,469	10,176 / 293	1280×720 - 3840×2160	48×28 px
Drone-vs-Bird test	12	19,279	19,133 / 146	720×576 - 3840×2160	38×28 px
DUT external test	20	24,804	22,218 / 2,586	1280×720; 1920×1080	75×42 px

The internal diagnostic set contained 1,092 frozen tracks, including 980 positive and 112 negative tracks. A track was positive when it contained at least one detection matched to ground truth at $\text{IoU} \geq 0.30$ and at least 30% of its detections were matched. Because this subset was positive-dominant, it was used only for rule analysis. External evaluation used all 20 DUT sequences and 24,804 frames. Predictions were correct at $\text{IoU} \geq 0.50$. As DUT contains at most one annotated target per frame, the principal protocol retained one final hypothesis per frame. Associating all detector candidates before post-filter top-1 selection yielded $F1 = 0.6407$, compared with 0.6416 for the principal pre-association top-1 protocol. The evaluation therefore represents a single-target rather than general multi-object setting.

3.2. Detector, EATF configuration, and baselines

A frozen YOLO11n checkpoint was fine-tuned for 100 epochs on Drone-vs-Bird at 1280-pixel input size and evaluated at 960 pixels using pretrained Ultralytics weights [13]. No DUT data were used for detector training or EATF parameter selection. EATF uses $\tau_{cand}=0.01$, $\tau_{IoU}=0.15$, $G_{max}=1$, $\tau_{high}=0.55$, $\tau_{mean}=0.36$ and $\tau_{sum}=0.80$. The configuration was frozen before DUT evaluation. Because the historical records do not establish a single global optimization procedure, we report local sensitivity and make no claim of parameter optimality. Baselines included detector thresholds 0.55 and 0.242, length-only, mean-confidence, cumulative-confidence, trailing means with windows 3 and 5, and ByteTrack [11]. ByteTrack used Ultralytics 8.4.23 defaults: high/low/new-track thresholds of 0.25/0.10/0.25, $track_buffer=30$, $match_thresh=0.80$, score fusion, and 30 fps. It processed all cached candidates, followed by post-tracking top-1 selection. No DUT-specific ByteTrack tuning was performed.

3.3. Metrics, bootstrap analysis, and runtime measurement

We report precision, recall, F1, and the target-absent false-positive rate (AFPR). AFPR is the number of target-absent frames containing at least one retained prediction divided by the total number of target-absent frames. Each target-absent frame contributes at most one false-alarm event, regardless of the number of retained predictions in that frame; lower AFPR is preferable.

Uncertainty was assessed using 10,000 paired sequence-level bootstrap replicates with seed 42. In each replicate, 20 DUT sequences were sampled with replacement using identical indices for all compared methods. TP, FP, FN, target-absent frames, and false-positive absent frames were pooled before metric computation. Undefined metrics were excluded pairwise; all reported comparisons retained 10,000 valid replicates. The favorable proportion uses $\Delta > 0$ for precision, recall, and F1, and $\Delta < 0$ for AFPR, where $\Delta = \text{EATF} - \text{baseline}$.

Detector inference required 29.99 ms/frame on an RTX 5060 Ti. Complete EATF post-processing required a median 0.595 ms/frame on an Intel Xeon Silver 4310: 0.004 ms for top-1 reduction, 0.484 ms for association, 0.106 ms for feature extraction, and 0.001 ms for rule evaluation. The measurements represent amortized offline processing. Post-processing used

Python 3.13.5 and Ultralytics 8.4.23.

4. RESULTS AND DISCUSSION

4.1. Rule simplification and local sensitivity

The original rule containing the minimum-track-length predicate and the simplified rule in (2) produced identical decisions. No mismatch was found among the 1,092 frozen internal tracks or among the 24,318 associated DUT detections. On the internal diagnostic subset, both rules produced 662 true-positive and 13 false-positive tracks, corresponding to a precision of 0.9807, a recall of 0.6755, and an F1 score of 0.8000. The DUT frame-level metrics were also unchanged. These results confirm the logical analysis in Section 2.2 and show that removing the length predicate simplifies the method without altering any reported decision. Local threshold sensitivity was evaluated on tracks reconstructed from the internal frame-level data. At the frozen association configuration $\tau_{IoU}=0.15$ and $G_{max}=1$, this analysis contained 2,014 tracks, including 1,480 positive and 534 negative tracks. Table 2 varies one decision threshold at a time while fixing all other parameters at $\tau_{high}=0.55$, $\tau_{mean}=0.36$, and $\tau_{sum}=0.80$.

Table 2. One-factor threshold sensitivity at the frozen internal association setting.

Varied parameter	Precision	Recall	F1	$\Delta F1$
Frozen configuration ($\tau_{high}=0.55$, $\tau_{mean}=0.36$, $\tau_{sum}=0.80$)	0.9355	0.7061	0.8048	0
$\tau_{high}=0.50$	0.9200	0.7615	0.8333	+ 0.0285
$\tau_{high}=0.60$	0.9423	0.6074	0.7387	− 0.0661
$\tau_{mean}=0.30$	0.9310	0.7203	0.8122	+ 0.0074
$\tau_{mean}=0.42$	0.9421	0.6932	0.7988	− 0.0060
$\tau_{sum}=0.60$	0.9307	0.7074	0.8038	− 0.0010
$\tau_{sum}=1.20$	0.9407	0.6865	0.7938	− 0.0110

The high-confidence threshold had the largest effect. Lowering it from 0.55 to 0.50 increased recall and F1 but reduced precision and increased the number of false-positive tracks from 72 to 98. Increasing it to 0.60 produced the opposite trade-off and reduced F1 to 0.7387. The mean- and cumulative-confidence thresholds were less sensitive around the frozen point: F1 varied from 0.7988 to 0.8122 over $\tau_{mean} \in [0.30, 0.42]$ and from 0.7938 to 0.8048 over $\tau_{sum} \in [0.60, 1.20]$. The frozen configuration was therefore not the maximum-F1 setting on this diagnostic data; it represents a more conservative precision–recall operating point. This analysis assesses local robustness and does not retune the configuration used on DUT.

On the original 1,092-track diagnostic subset, applying the mean-confidence condition without the cumulative constraint increased internal F1 but retained 41 of 112 negative tracks, compared with 13 for simplified EATF. This motivated retaining cumulative confidence as a conservative false-track rejection condition rather than as a source of large F1 improvement. Changing τ_{IoU} or G_{max} repartitions the detections into different track sets, so track-level F1 values are not directly comparable across association settings. These parameters were therefore retained as part of the frozen configuration and are not claimed to be optimal.

4.2. External validation and baseline comparison

Table 3 compares simplified EATF with detector thresholding, temporal rules, trailing averaging, and ByteTrack. EATF achieved the highest F1, although the conservative detector retained the highest precision and lowest AFPR.

At matched recall, EATF improved precision and F1 while reducing the AFPR point estimate relative to threshold reduction. Length-only and cumulative-only filtering produced higher recall but excessive AFPR, whereas trailing averaging and ByteTrack were more conservative but recovered fewer targets.

Mean-confidence filtering nearly matched EATF in F1 (0.6413 versus 0.6416). The cumulative-confidence condition mainly increased precision and reduced AFPR, with no statistically conclusive F1 gain over mean-only filtering. It therefore acts primarily as a conservative false-alarm rejection constraint.

Table 3. External validation and temporal-baseline comparison on DUT Anti-UAV Tracking.

Method	Precision	Recall	F1	AFPR
Detector, confidence 0.55	0.8908	0.4377	0.5870	0.0197
Detector, matched recall, $\tau=0.242$	0.7657	0.5319	0.6277	0.1524
Length-only track confirmation	0.5488	0.5718	0.5601	0.7765
Mean-confidence filtering	0.8049	0.5329	0.6413	0.1168
Cumulative-confidence filtering	0.5970	0.5647	0.5804	0.5104
Trailing mean, window 5	0.8022	0.5085	0.6225	0.0762
ByteTrack, default post-top-1	0.7925	0.5082	0.6193	0.0750
Simplified EATF	0.8084	0.5319	0.6416	0.1071

ByteTrack was applied to the full cached detector output, followed by highest-confidence post-tracking selection in each frame. The other temporal methods used the principal single-target input protocol.

Table 4. Paired sequence-level bootstrap comparisons on DUT Anti-UAV Tracking.

Metric	$\Delta = \text{EATF} - \text{baseline}$	95% CI	Favorable proportion
EATF vs. detector, confidence 0.55			
Precision $\uparrow$	-0.0824	[-0.2423, 0.0039]	0.0500
Recall $\uparrow$	+0.0942	[0.0555, 0.1301]	1.0000
F1 $\uparrow$	+0.0547	[0.0007, 0.1001]	0.9762
AFPR $\downarrow$	+0.0874	[0.0112, 0.1377]	0.0000
EATF vs. matched-recall detector			
Precision $\uparrow$	+0.0427	[0.0058, 0.1039]	0.9945
Recall $\uparrow$	-0.00005	[-0.0076, 0.0062]	0.4986
F1 $\uparrow$	+0.0139	[0.0017, 0.0265]	0.9901
AFPR $\downarrow$	-0.0452	[-0.1593, 0.0450]	0.8151
EATF vs. ByteTrack post-top-1			
Precision $\uparrow$	+0.0159	[-0.0049, 0.0477]	0.9184
Recall $\uparrow$	+0.0237	[0.0087, 0.0395]	0.9999
F1 $\uparrow$	+0.0223	[0.0084, 0.0364]	0.9999
AFPR $\downarrow$	+0.0321	[-0.0410, 0.0688]	0.1490

Table 4 reports paired bootstrap differences, defined as $\Delta = \text{EATF} - \text{baseline}$. Positive differences favor EATF for precision, recall, and F1; negative differences favor EATF for AFPR.

Relative to detector thresholding at 0.55, EATF showed statistically supported gains in recall and F1, but lower precision and higher AFPR. Against the matched-recall detector, the precision and F1 improvements were supported, whereas the recall difference was negligible and the AFPR difference remained inconclusive. Compared with ByteTrack, EATF achieved statistically supported gains in recall and F1; however, the confidence intervals for precision and AFPR

included zero. These results indicate that EATF primarily improves target recovery and overall F1 rather than uniformly improving all operating characteristics.

4.3. Sequence-level and qualitative analysis

Figure 1 shows that EATF increased recall in 16 of 20 DUT sequences, with the largest gains in videos 07, 11, 08, and 10. The AFPR effect was less uniform: video 09 combined substantial recovery with the largest false-alarm increase, confirming that persistent background responses remain sequence-dependent.

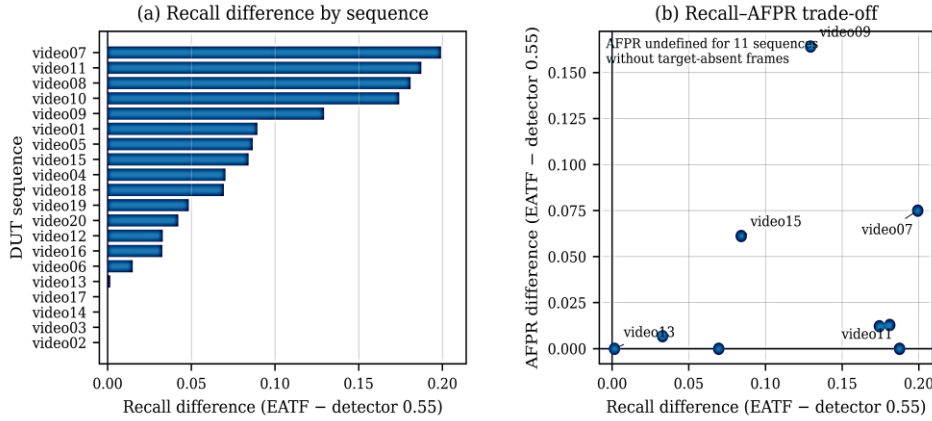


Figure 1. Sequence-level changes of simplified EATF relative to detector thresholding at 0.55 on DUT: (a) Recall difference by sequence; (b) Recall-AFPR trade-off.

Sequence-level AFPR was defined only for the nine sequences containing at least one target-absent frame. It was undefined for the remaining eleven sequences and was therefore omitted from Figure 1(b), rather than assigned a value of zero. The scatter plot consequently characterizes false-alarm trade-offs only for sequences in which AFPR can be evaluated.

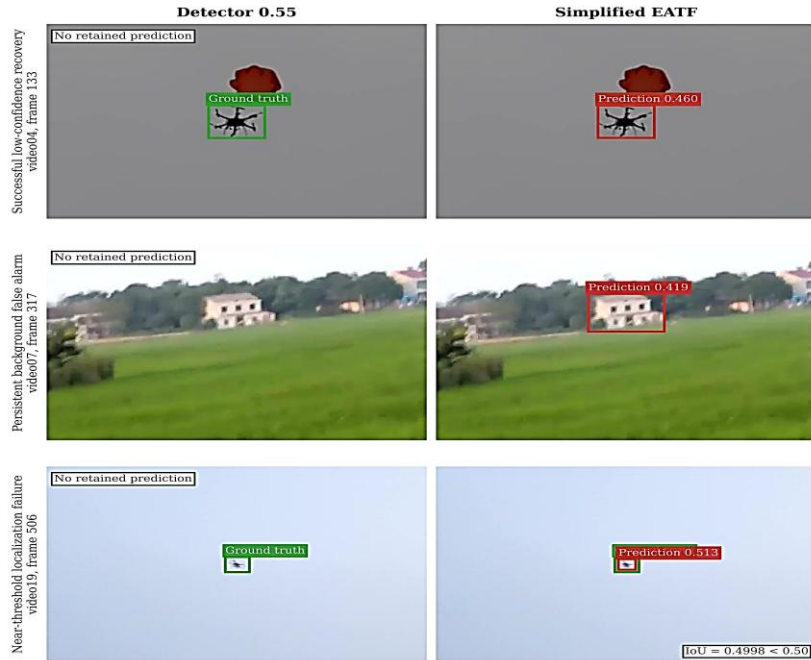


Figure 2. Representative DUT cases selected from the frozen evaluation outputs: low-confidence recovery, persistent background false alarm, and near-threshold localization failure.

Figure 2 illustrates successful recovery of a weak UAV response, a persistent background false alarm, and a localization failure just below the IoU threshold. EATF retained the original boxes and therefore could neither correct the near-threshold localization error nor recover frames without detector candidates. The method consequently addresses confidence-based rejection, but not localization or semantic background confusion. All three cases were selected from the same frozen DUT evaluation outputs used for the quantitative analysis.

Overall, the sequence-level and qualitative analyses indicate that EATF is most effective when weak UAV responses remain spatially consistent, but it remains vulnerable to persistent background tracks and detector localization errors.

5. CONCLUSIONS

This paper presented simplified EATF, an interpretable offline post-processing method for recovering low-confidence UAV detections using mean and cumulative track confidence. On DUT Anti-UAV Tracking, EATF increased recall from 0.4377 to 0.5319 and F1 from 0.5870 to 0.6416 relative to detector thresholding at 0.55. At matched recall, it achieved higher precision and F1 than threshold reduction, and it also obtained higher recall and F1 than default ByteTrack under the evaluated single-target protocol.

The current method remains limited by retrospective completed-track processing, persistent background false alarms, and its inability to correct inaccurate or missing detector boxes. Future work will investigate online running statistics, motion and appearance consistency, background-track rejection, and extension to multi-target video.

REFERENCES

- [1]. A. Coluccia et al., “Drone vs. Bird Detection: Deep Learning Algorithms and Results from a Grand Challenge”, *Sensors*, Vol. 21, No. 8, p. 2824, (2021). DOI: 10.3390/s21082824.
- [2]. A. Coluccia et al., “Drone-vs-Bird Detection Grand Challenge at ICASSP2023”, *ICASSP 2023*, IEEE, pp. 1–2, (2023).
- [3]. F. N. M. Zamri et al., “Enhanced small drone detection using optimized YOLOv8 with attention mechanisms”, *IEEE Access*, Vol. 12, pp. 90629–90643, (2024).
- [4]. N. V. Ngon et al., “Optimizing long-range UAV detection on YOLOv8: Breaking-point distance analysis and combining adaptive tiling with AdamW optimizer”, *Journal of Military Science and Technology*, Vol. 109, pp. 154–163, (2026). DOI: 10.54939/1859-1043.j.mst.109.2026.154-163.
- [5]. F. Kuppens et al., “Multivariate Confidence Calibration for Object Detection”, *Proceedings of the IEEE/CVF Conference on CVPRW*, pp. 326–327, (2020). DOI: 10.1109/CVPRW50498.2020.00171.
- [6]. U. Seidaliyeva et al., “Real-time and accurate drone detection in a video with a static background”, *Sensors*, Vol. 20, No. 14, p. 3856, (2020).
- [7]. J. Zhu et al., “Bird and UAVs recognition detection and tracking based on improved YOLOv9-DeepSORT”, *IEEE Access*, Vol. 12, pp. 147942–147957, (2024). DOI: 10.1109/access.2024.3475629.
- [8]. X. Cheng et al., “IRWT-YOLO: A Background Subtraction-Based Method for Anti-Drone Detection”, *Drones*, Vol. 9, No. 4, p. 297, (2025). DOI: 10.3390/drones9040297.
- [9]. Y. Sun et al., “Enhancing UAV detection in surveillance camera videos through spatiotemporal information and optical flow”, *Sensors*, Vol. 23, No. 13, p. 6037, (2023).
- [10]. L. K. Thanh et al., “Real-time long-term target tracking on an ARM platform with NPU acceleration and integration into UAV line-of-sight stabilization”, *Journal of Military Science and Technology*, Vol. 109, pp. 25–34, (2026). DOI: 10.54939/1859-1043.j.mst.109.2026.25-34.
- [11]. Y. Zhang et al., “ByteTrack: Multi-object tracking by associating every detection box”, *European Conference on Computer Vision*, Springer, pp. 1–21, (2022).
- [12]. J. Zhao et al., “Vision-Based Anti-UAV Detection and Tracking”, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 23, No. 12, pp. 25323–25334, (2022). DOI: 10.1109/TITS.2022.3177627.
- [13]. Ultralytics, “Ultralytics YOLO11”, Ultralytics Docs, Available: <https://docs.ultralytics.com/models/yolo11>

TÓM TẮT

EATF: Lọc theo thời gian dựa trên bằng chứng để khôi phục các phát hiện UAV có độ tin cậy thấp

Các UAV cỡ nhỏ thường tạo ra phát hiện yếu và không liên tục do kích thước ảnh nhỏ, nhiễu nền và chuyển động nhanh. Bài báo đề xuất Evidence-Aware Temporal Filtering (EATF), một phương pháp hậu xử lý ngoại tuyến cho bộ phát hiện cố định. Phát hiện có độ tin cậy cao được giữ trực tiếp; phát hiện yếu hơn chỉ được khôi phục khi quỹ đạo hoàn thành thỏa ngưỡng độ tin cậy trung bình và tích lũy. EATF không yêu cầu huấn luyện lại, không ước lượng chuyển động, không nội suy hợp giới hạn và không sinh phát hiện mới. Trên DUT Anti-UAV Tracking, EATF tăng Recall từ 0.4377 lên 0.5319 và F1 từ 0.5870 lên 0.6416 so với ngưỡng 0.55. Ở mức Recall tương đương, Precision tăng từ 0.7657 lên 0.8084 và F1 tăng từ 0.6277 lên 0.6416. So với ByteTrack sau post-tracking top-1, EATF đạt Recall và F1 cao hơn. Phương pháp mang tính hồi cứu, giới hạn ở một mục tiêu và còn chịu ảnh hưởng của quỹ đạo nền bền vững.

Từ khoá: Phát hiện UAV; Lọc theo thời gian; Các phát hiện có độ tin cậy thấp.