GeoSentry: Recovering Dim-Target Tracking Failures by Cross-View Geometric Disambiguation

Abstract. Tracking dim, appearance-less aerial targets over long horizons fails dangerously: a state-of-the-art point tracker drifts off the target while still reporting high confidence, so the system is confidently wrong. We turn the cross-view geometry that *detects* this silent drift into a means to *recover* from it. An alarm first flags drift ground-truth-free—cross-view epipolar, reprojection, and triangulation geometry feed a minimal interpretable per-frame predictor (cross-sequence AUC 0.969/1.000, well-calibrated open-loop, ECE 0.03; on the correlated-co-drift tail only the *triangulation angle* survives, $\sim$14$\times$ over the calibration noise floor). When one view drifts while another still tracks, the rescuer view's epipolar line then **localizes** the search region and **disambiguates** the dim target from appearance-identical cloud clutter (the on-line candidate is the target). On public data (CenekAlbl dataset-4, a dim 2–7 px drone in moving cloud) this recovers the lost target at **65% of loss events versus 10% for single-view**—a structural second-view gain that decomposes into localization (10%$\rightarrow$51%) and a **+14-point disambiguation step** (51%$\rightarrow$65%), both unavailable to one camera.

1 Introduction

Long-horizon visual tracking of small UAVs enables airspace monitoring and counter-UAV systems, but the regime is adversarial to appearance-based methods: the target is *dim, small, appearance-less*, often a few pixels against cloud sky, observed *without ground truth*. Modern deep point trackers eventually drift—and do so *silently*: a strong off-the-shelf tracker keeps reporting high confidence long after its estimate slid off the target onto background (Fig. 1). The tracker is *confidently wrong*, so its self-reported confidence is an unreliable basis for any decision.

We do *not* propose a new tracker. We take an off-the-shelf, zero-extra-training point tracker (LocoTrack [2]) and ask an orthogonal question: *can the system know, ground-truth-free, when it has silently drifted?* A single view cannot—its confidence is near-uninformative—but a *second calibrated viewpoint* can: cross-view geometry (epipolar constraint, reprojection residual, triangulation angle) is satisfied while both views are correct and violated when one drifts, an external ground-truth-free supervision signal a tiny predictor turns into a per-frame failure probability. The same geometry then *acts*: the still-tracking view's epipolar line localizes where to look and disambiguates the dim target from appearance-identical clutter—a recovery no single view could perform. Detection

and recovery are two uses of one geometry, governed by the same correlated-vs-independent boundary.

We instantiate this as **GeoSentry** (**Geo**metric **Sentry**). It is not one more failure detector; it makes three advances, each a step up in *what is understood* about silent drift, and each is the substance of a contribution:

1. **(I) From a detector to the geometric structure of silent failure.** We expose *why* drift hides: cross-view geometry turns silent drift into a ground-truth-free, calibrated per-frame failure probability (AUC 0.969/1.000, ECE 0.03), and its errors are not random but two *structural blind spots*—correlated co-drift and along-line drift—both stemming from the **one-dimensional** epipolar constraint. The depth degree of freedom, the *triangulation angle*, recovers the co-drift blind spot (epipolar $0.06 \rightarrow 0.98$, $\approx 14\times$ over the calibration floor), explaining a *closed-loop calibration collapse* and the *correlated-vs-independent boundary* that says when a third view becomes necessary.

2. **(II) From a single signal to complementary two-signal blind-spot filling.** Cross-view geometry is blind exactly when the coordinates are *consistent but the content is wrong* (a flight-like cloud at a plausible depth); a single-view appearance signal is blind in the opposite way (content *right but position wrong*). The two blind spots do not overlap, so a confidence-adaptive fusion of the geometric residual and a lightweight appearance signal has a *strictly larger* reliable domain than either alone.

3. **(III) From an alarm to a single-view-impossible recovery.** When one view drifts while another still tracks, the rescuer's epipolar line *localizes* the search region and *disambiguates* the dim target from appearance-identical clutter. On public data this recovers a lost 2–7 px drone at 65% of loss events against 10% for a single view, decomposing into localization (10%→51%) and a +14-point disambiguation step (51%→65%), both impossible for one camera. The local re-scan and alarm are enablers, not the contribution.

On scope: we claim a *single-view-impossible* recovery in the independent-failure regime, not a deployable recovery system—it is bounded by a training-free re-detector, needs a rescuer view, and does not extend to correlated co-drift. We claim no new tracker or cross-sequence generalization. The two limits are the work's substance, not a failure log: recovery is bounded by the *correlated-vs-independent* boundary, and the closed-loop calibration collapse is a general *detection-is-not-safety* warning (run open-loop, never closed-loop). Both mark where a third view is needed.

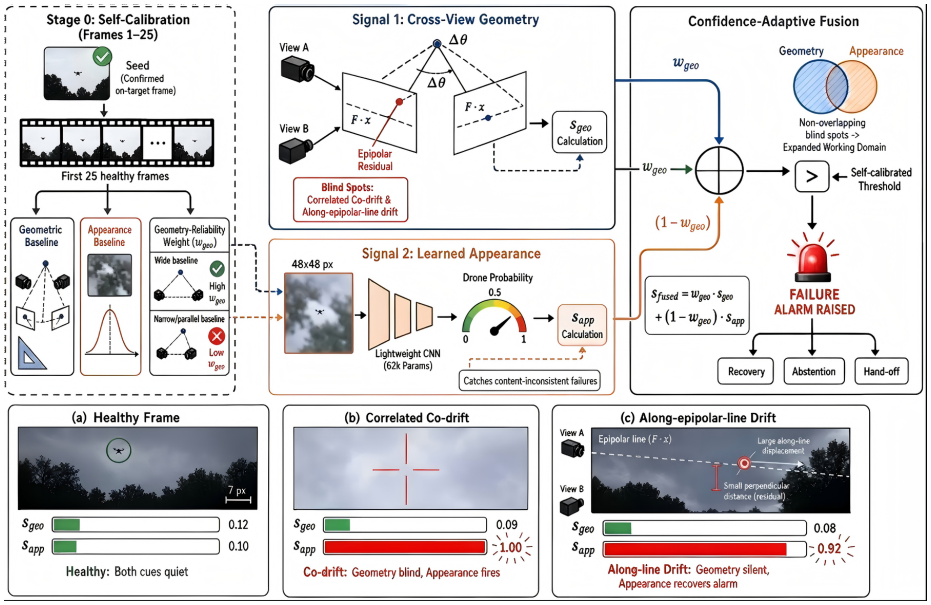

Fig. 1. GeoSentry overview. A one-time *self-calibration* (frames 1–25 after a single on-target seed) estimates the healthy geometric and appearance baselines and the geometry-reliability weight w_{geo} (high for wide-baseline, low for near-parallel views). At inference two ground-truth-free signals are computed per frame: **Signal 1 — cross-view geometry** (the epipolar/triangulation residual S_{geo} of the two tracked points), which is blind only to *correlated co-drift* and *along-epipolar-line* drift; and **Signal 2 — learned appearance** (a 62k-parameter CNN on a 48×48 px patch giving a drone-probability S_{app}), which catches content-inconsistent failures. A confidence-adaptive fusion $S = w_{geo} \cdot S_{geo} + (1\text{-}w_{geo}) \cdot S_{app}$ against a self-calibrated threshold raises the alarm, which gates recovery, abstention, or hand-off. The two blind sets do not overlap, so the fused working domain is strictly larger than either signal alone. **(a)** healthy: both cues quiet; **(b)** correlated co-drift: geometry blind, appearance fires; **(c)** along-line drift: the epipolar residual is silent (the drift slides along the line) yet appearance recovers the alarm.

2 Related Work

Epipolar geometry as a tool. Multi-view epipolar geometry is mature and widely used—for cross-camera association, as a training-time consistency loss, and for stereo mis-calibration detection [7]. We do *not* claim epipolar geometry as novel; our novelty is its combination with introspective drift prediction, calibration analysis, and a failure-boundary characterization in the appearance-less-UAV regime.

Cross-view inconsistency for introspective failure prediction. Using cross-view disagreement to predict perception failures has precedent in autonomous driving. IVOA [11] and *Failing to Learn* [12] use left/right disparity or detection discrepancy to predict *object-detector* false negatives. Our setting differs

concretely: (a) the underlying perceiver is a *deep point tracker*, and the failure mode is *temporal drift*, not per-frame mis-detection; (b) the targets are *dim, appearance-less* aerial objects where, as *Failing to Learn* itself notes, dense disparity is unreliable on small objects—we therefore use point-wise epipolar consistency rather than dense disparity; and (c) we additionally quantify the predictor's *closed-loop calibration collapse* and the *correlated-vs-independent* failure boundary, neither of which these works address.

Geometric-consistency checks. Methods such as HGC-Net [8] fuse epipolar geometry with learned detection to flag *stereo mis-alignment* on single image pairs; notably, that line lists "integrating temporal data for drift detection and multi-camera systems" as future work—precisely the regime we operate in.

UAV tracking failure detection and re-detection. Single-view methods estimate tracking reliability from *temporal* cues—e.g. a temporal-consistency deviation index with a lightweight re-detector [14], or uncertainty propagated from a 3D motion model [15]. In our regime the tracker's own confidence is a weak failure cue (self-confidence AUC 0.483 on Seq-I, 0.756 on Seq-II); the honest failure signal comes only from a *second viewpoint's geometry*, the cue these single-view methods cannot access.

Introspective perception and uncertainty. Introspective perception [3, 11] established the paradigm of a system "knowing when it does not know." Deep uncertainty estimators—MC-dropout [5], deep ensembles [9], and evidential deep learning [13]—provide epistemic uncertainty but require deeper models and training. We adopt the introspective paradigm but are, to our knowledge, the first to instantiate it as *multi-view-geometry-supervised drift detection for point tracking*, and to add the *calibration* analysis [6, 10] the paradigm has lacked in this setting.

3 Method

3.1 Problem formulation

Two calibrated cameras $\mathrm{c} \in \{\mathrm{w}, \mathrm{e}\}$ observe the same target. From a single one-time seed at a verified on-target frame, an off-the-shelf point tracker runs *open-loop* on each view, producing per-view image points ${\mathrm{x_c}}^{\mathrm{t}} \in \mathbb{R}^2$ with self-reported confidence ${\mathrm{v_c}}^{\mathrm{t}}$ at frame t. Let $\mathrm{y}^{\mathrm{t}} \in \{0,1\}$ denote whether the tracked estimate has drifted off the true target at frame t (defined for evaluation only, Sec. 4.1). Our goal is a function

$$\hat{p}^{\mathrm{t}} = f({x_{\mathrm{w}}}^{\mathrm{t}}, {x_{\mathrm{e}}}^{\mathrm{t}}, \{{x_{\mathrm{c}}}^{\mathrm{t\text{-}k}}\}_{\mathrm{k=1..W}}, \mathcal{C}) \in [0, 1] \tag{1}$$

that estimates $\mathrm{P}(\mathrm{y}^{\mathrm{t}} = 1)$ from the two tracks, a short history window W, and the fixed calibration $\mathcal{C}$ — *without access to ground truth at inference.* We deliberately do not modify the tracker; f is an *introspection* layer judging whether the tracker is currently trustworthy.

3.2 Overview

GeoSentry (Fig. 1) streams two calibrated views. An off-the-shelf tracker runs independently on each from a single one-time seed. From the two tracks and the calibration alone—*never* ground truth—we compute per-frame cross-view geometric, motion-consistency, and short-window temporal features; a tiny per-view classifier maps these to a drift probability, the two fused into a single per-frame failure probability, with an independent target-presence channel acting as a reliability gate where the geometry is self-consistent yet wrong. The *only* deep component is the off-the-shelf tracker (zero extra training, not a contribution); the *only* trained component is the 13-parameter predictor; everything else is classical geometry. At inference the system is entirely ground-truth-free—RTK and manual labels are used *only* offline, to train and evaluate.

3.3 Cross-view geometric features

We certify the two-camera geometry once from manual correspondences (median reprojection 2.29 px over 14 points; stereo 3D RMSE 3.39 m at 200–400 m; baseline 16.66 m) and recover the fundamental matrix to a 2.92 px median epipolar residual. *Point-wise* epipolar consistency is deliberate—dense stereo is unreliable on few-pixel targets. Each residual is healthy-normalized by its median over the first 25 post-seed frames (healthy by protocol), giving a dimensionless "× healthy" feature that transfers across sequences. A per-view *motion-consistency* feature—the deviation from a constant-velocity fit to the preceding 8 frames—is a full 2D deviation that exposes drift the 1-DOF epipolar distance is blind to (Sec. 5). The formal definitions follow; given per-view points $\mathrm{x_w}^{\mathrm{t}}$, $\mathrm{x_e}^{\mathrm{t}}$ and the fundamental matrix F, the symmetric epipolar distance is

$$d_{\mathrm{epi}} = |{x_{\mathrm{e}}}^{\top} F x_{\mathrm{w}}| / \sqrt{({F x_{\mathrm{w}})_1}^2 + {(F x_{\mathrm{w}})_2}^2} + |{x_{\mathrm{w}}}^{\top} F^{\top} x_{\mathrm{e}}| / \sqrt{({F^{\top} x_{\mathrm{e}})_1}^2 + {(F^{\top} x_{\mathrm{e}})_2}^2} \tag{2}$$

Triangulating $\mathrm{x_w}^{\mathrm{t}}$, $\mathrm{x_e}^{\mathrm{t}}$ to a 3D point X^{t} yields per-view reprojection residuals and the triangulation angle,

$$r_{\mathrm{c}} = \|\pi_{\mathrm{c}}(X) - x_{\mathrm{c}}\|_2, c \in \{w, e\}; \theta = \angle(d_{\mathrm{w}}, d_{\mathrm{e}}) \tag{3}$$

where π_{c} is the projection of view c and $\mathrm{d_c}$ the back-projected ray of $\mathrm{x_c}$. The per-view motion-consistency term is the deviation from a constant-velocity fit over the preceding W = 8 frames,

$$m_{\mathrm{c}} = \|{x_{\mathrm{c}}}^{\mathrm{t}} - \hat{x}_{\mathrm{c}}\tilde{t}|t-1 : t-W\|_2, (4)$$

a full-2D deviation that exposes drift to which the 1-DOF epipolar distance is blind (Sec. 5). Each geometric feature g is then healthy-normalized by its healthy level $\bar{g}_{\mathrm{healthy}}$ (the median over the first 25 post-seed frames),

$$\tilde{g}^{\mathrm{t}} = g^{\mathrm{t}} / \bar{g}_{\mathrm{healthy}}, \tag{5}$$

mapping a pixel/degree quantity to a dimensionless "× healthy" ratio that transfers across sequences. Frame-to-frame deltas and 5-frame rolling mean/std of $\tilde{d}_{\mathrm{epi}}$ and $\tilde{m}$ complete the 12-dimensional per-view vector ${\phi_{\mathrm{c}}}^{\mathrm{t}}$.

3.4 Lightweight introspective predictor

A shared per-view logistic regression gives ${\mathrm{P_c}}^{\mathrm{t}} = \sigma(\mathrm{w}^\top {\phi_{\mathrm{c}}}^{\mathrm{t}} + \mathrm{b})$ with 13 learned parameters, trained leave-one-sequence-out (never on the evaluated data), trivially interpretable (each feature a signed log-odds term) and trained in milliseconds on CPU. The camera-pair failure probability is the noisy-OR

$$\hat{p}^{\mathrm{t}} = 1 - (1 - {P_{\mathrm{w}}}^{\mathrm{t}})(1 - {P_{\mathrm{e}}}^{\mathrm{t}}). \tag{6}$$

3.5 Presence as an evidence gate

The geometry can be self-consistent yet wrong: under *co-drift* both views slide onto the same cloud band, so epipolar and motion signals look healthy while the target is lost. We add an independent *target-presence* score—the peak difference-of-Gaussians response ($\sigma = 3.0/7.5$ px, matched to the target scale; fixed, not learned) in a 25×25-px window at the track point—high on a point-like target, collapsing onto diffuse texture. We treat it as a *reliability gate*: when geometry reports "healthy" but presence reports collapse (healthy-normalized response below a fixed global 0.3), the conflict raises the failure belief. This is an orthogonal evidence-fusion step, not a change to the predictor; Sec. 4.5 shows it is load-bearing exactly where the learned features cannot reach. The two channels are complementary by construction (Fig. 1)—geometry blind only on consistent-coordinates/wrong-content, appearance blind only on right-content/wrong-position—so their non-overlapping blind sets give the fusion a strictly larger reliable domain than either alone.

Formally, with healthy-normalized presence $\tilde{s}^{\mathrm{t}}$ and a global default $\tau_{\mathrm{s}} = 0.3$ (set once, held fixed across sequences, never tuned on the test sequence), the gated failure belief is

$${\hat{p}_{\mathrm{gated}}}^{\mathrm{t}} = max(\hat{p}^{\mathrm{t}}, \mathbb{1}[\tilde{s}^{\mathrm{t}} < \tau_{\mathrm{s}}]), \tag{7}$$

i.e. when the geometry reports "healthy" yet the independent presence channel reports collapse, the conflict raises the failure belief. The whole inference step is per-frame and entirely ground-truth-free (Fig. 1).

3.6 Design rationale

Four design choices keep the introspector lightweight without losing signal, each backed by an experiment: noisy-OR fusion (single-view drift dominates, so a conjunctive rule would miss it); healthy normalization, which makes the feature scale transferable across sequences whose healthy residuals differ by an order of magnitude; presence as an orthogonal *gate* rather than a soft feature, since a

learned weight cannot override a self-consistent geometry; and logistic regression rather than a deep network, because the discriminative power lives in the geometry, not classifier capacity (Sec. 4.4, 4.6). Lightweight and interpretable are consequences of where the signal lives, not compromises.

4 Experiments

4.1 Setup

We evaluate on **DualUAV**, our benchmark of synchronized dual-view visible-light sequences of dim, appearance-less UAV targets. Three sequences carry per-view ground truth: **Seq-I** (*RTK-metric*; 2992 frames at 24.96 fps, RTK 3D trajectory projected to pixels), **Seq-II** (*annotated*; 646 frames, manual two-view labels), and **Seq-III** (*epipolar-calib*; an epipolar-only fold, no metric ground truth). Metric evaluation is confined to Seq-I; Seq-I/II carry the per-view labels for discrimination and calibration. We additionally use the public **CenekAlbl** dataset [1]: **dataset-4** (two cameras, a dim ∼2–7 px target in *moving cloud* with provided synchronization) to *reproduce* GT-free cross-view drift detection (Sec. 4.2), and **dataset-3** as a geometry sanity-check (a target occlusion, not drift). A per-view drift label is the track-to-GT distance exceeding $\tau_{\text{drift}} = 15$ px, read by a single isolated module and never entering inference. We report AUC, balanced-AUC, ECE/MCE, and Brier, with within-sequence temporal block-bootstrap intervals (Sec. 4.3); we do *not* report DeLong p-values [4], not powered at n = 2.

The principal limitation is up front: only *two* labelled sequences, so cross-sequence *significance* is structurally limited—we report AUC point estimates with within-sequence block-bootstrap intervals and make no cross-sequence significance claim (behaviour reported as *preliminary, consistent with transfer*). The numbers are re-materialized from saved models and cached tracks, matching the reported per-fold AUC exactly (a leakage gate). They are robust to the label threshold (sweeping $\tau_{\text{drift}} \in \{10,15,20,25\}$ px keeps LOSO AUC within 0.012 on Seq-I, ≥0.96 on both) and to healthy-window size ({15,25,35,50} frames leaves AUC 0.969/1.000, ECE within 0.012); the seed window is verified healthy (median track-to-GT 8.5/1.2 px on Seq-I/II).

4.2 Discrimination

Table 1 reports cross-sequence discrimination. The lead metric is the natural-imbalance AUC under the 90–98% deployment drift base rate: 0.969 (Seq-I) / 1.000 (Seq-II). The learned predictor out-ranks the single epipolar threshold (0.944/1.000) and the tracker's self-confidence (0.483/0.756, barely above chance on Seq-I—the "confidently wrong" behavior). The geometric features are the lever: the geometry-subset breakdown of Sec. 4.4 shows the epipolar distance alone is anti-informative under co-drift (0.056) while the triangulation angle alone recovers it (0.980). A block-bootstrap puts the open-loop AUC at

0.969 (95% CI [0.93, 0.99], wide from the small healthy class) and [1.00, 1.00] on Seq-II; balanced-AUC (0.966/1.000) confirms the ranking is not an imbalance artifact. Among same-task competitors (Table 1, Fig. 4) GeoSentry ties the best deep-uncertainty model within 0.004 AUC while being the only predictor simultaneously lightweight (13 parameters, 74–370× fewer), interpretable, calibrated, and able to flag the independent-failure regime on which single-view cues are blind—the best-rounded same-task predictor, not the single highest index (two uncalibrated reference scores out-rank it on this clear-target benchmark, Sec. 4.4).

Figure 2 shows the drift signature frame-by-frame on our self-data: while the tracker holds the target the views are cross-view consistent (residual ≈ 0 px); once it slides off—over a thousand pixels onto cloud, *while the dim target stays visible at the RTK ground truth in both views*—the residual jumps and the geometry flags the failure though the tracker's visibility output is unchanged. This is *genuine tracker drift* (target in view, only the estimate leaves), verified frame-by-frame across all three self-sequences. On the independent **CenekAlbl** dataset-4 (a dim ∼2–7 px drone in *moving cloud*) we *reproduce* this GT-free: across four sustained onsets the tracker slides hundreds of pixels off a drone that stays visible at the 2D GT (330–431 px in-frame, 2–7 px at onset; 95–97% of failing frames keep the target in-frame, so it is drift not loss). Using the dataset's provided synchronization, the fundamental matrix fit offline (905/2105 RANSAC inliers at 0.62 px), the cross-view residual jumps from a healthy 2.5 px to 158 px under drift and the GT-free introspection separates drift from healthy at **AUC 0.94**. This is the *independent-viewpoint* regime (different cameras → drift is epipolar-inconsistent → the epipolar term suffices); we scope it honestly—per camera this is single-view tracking, AUC 0.94 not a saturated 1.0, and the *correlated co-drift* blind tail remains Seq-I only. The dataset-3 event is a target occlusion (kept distinct from drift).

The cross-view geometry is what makes the signal honest. A single-view temporal baseline [14]—EW recent-frame deviation plus our motion-consistency and self-confidence features, same LOSO split, *no* cross-view geometry—reaches only AUC 0.728 on Seq-I (Table 1), well below 0.969 (+0.241). Co-drift is *smooth* (both views slide gradually onto cloud), so a single-view temporal cue sees no jump and stays near chance, while the second viewpoint's geometry does not share the blind spot. On Seq-II, where failures are abrupt, the baseline is competitive (0.937)—the cross-view advantage is specific to the correlated-co-drift regime that defines this problem. The baseline has a strong feature set, the same split and fusion (n = 2; no significance claim).

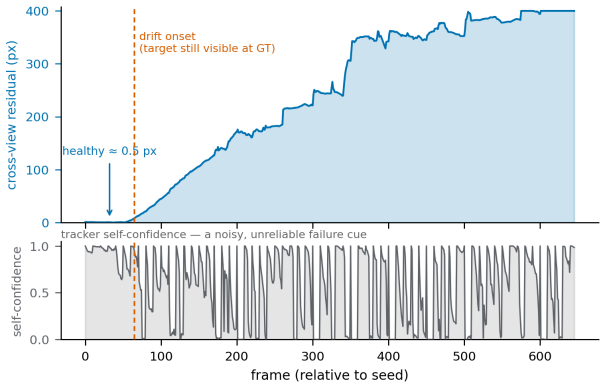

Fig. 2. The drift signature over time on self-data (Seq-II, genuine tracker drift). Left axis: the cross-view epipolar residual (blue) holds a sub-pixel healthy baseline (≈ 0.5 px), then at the drift onset jumps to several hundred pixels and stays high—while we verified the target remains visible at the ground truth throughout (Sec. 4.2), so this is genuine drift, not target loss. Right axis: the tracker's own self-confidence (grey) is a noisy, unreliable cue. The geometry is the honest failure signal. Residual clipped at 400 px for display.

method	cross-view geom?	trained?	params	AUC (Seq-I / Seq-II)	calibrated?	note
Ours (GeoSentry)	**yes**	yes	**13**	**0.969 / 1.000**	**yes (ECE 0.028)**	lightweight, interpretable, covers independent failures
epipolar single-thr. (M1)	yes (1-DOF)	no	0	0.944 / 1.000	no (raw)	geometry baseline; blind to co-drift tail
single-view temporal [14]	no	yes	∼13	0.728 / 0.937	no	weak on smooth co-drift
tracker self-confidence (C2)	no	no	0	0.483 / 0.756	no (raw)	near-uninformative on Seq-I
MC-dropout	yes*	yes	961	0.968 (pooled)	yes (0.020)	74× params, no gain
deep ensemble (×5)	yes*	yes	4805	0.949 (pooled)	yes (0.001)	370× params
evidential (EDL)	yes*	yes	994	0.932 (pooled)	partial (0.061)	—

4.3 Calibration: detection is not closed-loop safety

A failure predictor is actionable only if its probability is trustworthy, not merely well-ranked. *Open-loop* (Fig. 3), the predictor is near-calibrated (ECE 0.034 on Seq-I, block-bootstrap CI [0.00, 0.08]; 0.013 on Seq-II). *Closed-loop*, where its decisions drive cross-view re-anchoring, calibration *collapses*: on Seq-I the mean predicted drift probability is 0.12 while the track is off 98% of the time (ECE 0.869, over-confidence +0.87), because the predictor consumes the very geometry the re-anchor corrupts under co-drift—confident all is well precisely when it is not. This is the calibration-language form of a closed-loop silent-failure rate: drift *detection* does not imply closed-loop *safety*. The effect is milder on Seq-II (0.013→0.197). Post-hoc recalibration cannot repair it—the closed-loop *ranking* itself degrades (AUC 0.907/0.790 vs open-loop 0.969/1.000).

At the default 0.5 cutoff the predictor recalls 99–100% of drifts but its specificity on the thin healthy class is 66%/86%—it *over-flags*. AP sits at the base rate (1.00/0.998 against 0.98/0.90 prevalence), uninformative here, which is why we

lead with AUC and balanced-AUC. The over-flagging is the open-loop cost the external-trigger deployment (Sec. 4.7) absorbs—a false flag costs a redundant re-acquisition, not a silent miss—so the predictor is a *monitor*, not an autonomous gate.

Seq.	Loop	mean $\hat{P}$ (drift)	true drift-rate	ECE
Seq-I	open (C1)	0.94	0.98	0.034
	closed (C4)	0.12	0.98	**0.869**
Seq-II	open (C1)	0.91	0.90	0.013
	closed (C4)	0.75	0.72	**0.197**

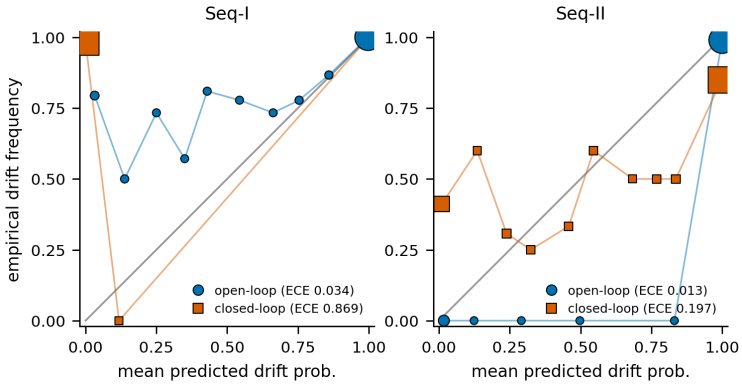

Fig. 3. Reliability diagrams, open-loop vs closed-loop, both sequences. Open-loop (blue circles) the predictor sits near the diagonal (ECE 0.034 / 0.013); closed-loop (orange squares) the points pin to the top edge—the predicted drift probability is low while the empirical drift frequency is ≈ 1—an ECE of 0.869 on Seq-I. Marker area is proportional to bin count. The closed-loop ranking itself degrades, so post-hoc rescaling cannot repair it.

4.4 Comparison with deep-uncertainty methods

Table 3 compares our 13-parameter logistic predictor against three deep-uncertainty baselines—MC-dropout, a five-member deep ensemble, an evidential head—on the *same* LOSO split, features, and fusion, plus a zero-parameter unsupervised Mahalanobis score. Two findings: first, *the discriminative signal lives in the geometry, not classifier capacity*—the Mahalanobis score reaches AUC ≈ 1.0, and the rescue on the co-drift tail comes specifically from the *triangulation angle* (alone 0.980), while the Sampson-coupled epipolar/reprojection terms are blind there (0.056/0.055). Second, *among trained classifiers ours is best or tied*: it matches MC-dropout within 0.004 AUC and out-ranks the ensemble and evidential head at 74–370$\times$ fewer parameters, interpretable and well-calibrated (ECE 0.028). A single-view appearance index (PSR/APCE) ranks high (≈ 0.99) but is exactly the cue we use as the presence gate (Sec. 4.5)—a component, not a standalone competitor.

Method	AUC	bal-AUC	ECE	Brier	params	interp.
ours (logreg)	0.972	0.972	0.028	0.030	13	yes
MC-dropout MLP	0.968	0.969	0.020	0.027	961	no
deep ensemble (×5)	0.949	0.949	0.001	0.027	4805	no
evidential (EDL)	0.932	0.932	0.061	0.063	994	no
Mahalanobis (unsup.)	1.000	0.999	—	—	0	partial

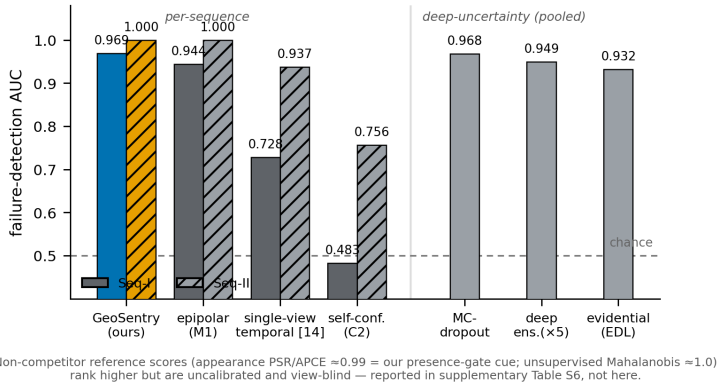

Fig. 4. Main results: failure-detection AUC against genuine same-task competitors only. Per-sequence bars (Seq-I solid, Seq-II hatched) for the single-view / geometry methods; pooled bars for the deep-uncertainty estimators (right of the divider). GeoSentry (highlighted) matches the strongest discrimination at 74–370× fewer parameters. No anti-UAV tracker is compared on tracking (a different task); the uncalibrated non-competitor reference scores (PSR/APCE ≈ 0.99, our presence-gate cue; Mahalanobis ≈ 1.0) are reported in supplementary Table S6, not here.

4.5 Evidence-fusion ablation: presence as a reliability gate

Table 4 ablates how target-presence should be used, by the closed-loop silent-failure (*false-secure*) rate and the healthy false-alarm rate. Presence as a learned *feature* (B) improves the healthy side (61%→17% on Seq-I) but does *not* fix the silent-failure rate (89%→83%): a learned soft weight cannot override a self-consistent geometry. Presence as a post-inference orthogonal *gate* (C/D) drives the silent-failure rate to 0% at no added healthy false alarm. This is not a learned-detector improvement—the predictor alone still misses 83%—but an independent channel firing exactly on the co-drift case the geometry is blind to. The hard gate (C) and soft fusion (D) behave identically, so the gate, not its softness, is load-bearing; the Seq-I healthy sample is small (n = 18), so the healthy-side numbers are illustrative.

	Seq-I: false-sec.	Seq-I: healthy FA	Seq-II: false-sec.	Seq-II: healthy FA
A geometry only (12-feat)	89%	61%	3%	34%
B + presence as a feature (13-feat)	83%	17%	1%	19%
C + presence reliability *gate*	**0%**	17%	**0%**	19%
D + presence soft fusion	0%	17%	0%	19%

4.6 Feature-group ablation

We partition the 12 features into non-overlapping groups—epipolar (G1), reprojection + triangulation (G2), motion (G3), temporal (G4), self-confidence (Gv)—and ablate under the same LOSO protocol. For the *learned* predictor too, the triangulation-based geometry is the lever: leave-one-group-out moves the Seq-I AUC by at most 0.006 for every group *except* G2, whose removal collapses it from 0.969 to 0.426. Within G2 the *triangulation angle* is the active component (alone 0.980; reprojection is Sampson-coupled and blind, 0.055). Incrementally (Table 5), epipolar alone reaches 0.942 (the learned predictor exploits the *sign*—co-drift drives the residual low), +reproj/tri lifts it to 0.975, and motion/temporal/self-confidence add nothing measurable. The small Seq-I healthy class makes individual cross-sequence AUCs high-variance (the -G2 value falls below chance), so we read the *trend*, not single digits.

Cumulative feature groups	Seq-I AUC	Seq-II AUC
epipolar (G1)	0.942	0.977
+ reproj/tri (G1+G2)	**0.975**	1.000
+ motion (G1–G3)	0.975	1.000
+ temporal (G1–G4)	0.975	1.000
+ self-conf = full (G1–G5)	0.969	1.000

4.7 Discussion

Two points: the closed-loop calibration collapse carries a concrete recipe—run open-loop, let a predicted failure trigger *external* re-acquisition rather than driving the loop it judges (the fix is architectural; the ranking itself degrades, no post-hoc scaling repairs it). And the second viewpoint is necessary precisely where the single view is blind: a single-view temporal baseline is competitive on Seq-II (0.937, abrupt failures) but falls to 0.728 on Seq-I, where smooth co-drift leaves no single-view trace.

4.8 Tracker-agnostic at the signal level

GeoSentry is an introspection layer, not a tracker, so it monitors any point tracker. Running its cross-view geometry *unchanged* on three off-the-shelf trackers—LocoTrack, TAPIR, CoTracker3—each seeded once and run open-loop (Table 6), the geometric-introspection signal detects every tracker's drift at AUC 0.96–1.00 on both sequences despite drift-rates of 0.84–0.99. We make the tracker-agnostic claim **only at the signal level**; the *deployed trained* detector is a separate, weaker statement: it transfers cleanly except on Seq-I for the faster-drifting TAPIR/CoTracker3 (0.31/0.23, only 22/38 healthy frames), a cross-sequence transfer limit at n = 2 (Sec. 4.6), not a geometry failure (raw signal AUC 0.99).

monitored tracker	drift-rate (I/II)	healthy n (I/II)	geom-introspection AUC (I/II)	trained cross-seq AUC (I/II)
LocoTrack	0.98 / 0.90	61 / 65	**0.96 / 1.00**	0.97 / 1.00
TAPIR	0.99 / 0.84	22 / 101	**0.99 / 1.00**	0.31 / 1.00
CoTracker3	0.99 / 0.86	38 / 93	**0.99 / 1.00**	0.23 / 0.99

4.9 From detection to recovery: cross-view geometric disambiguation

The alarm's geometry also **enables recovery**. When a tracker loses a dim 2–7 px target in view B while view A still tracks it, A's epipolar line in B passes through the real target, so it both **localizes** the search region onto the line (not the lost tracker's stale prior) and **disambiguates** the drone from look-alike clutter (the on-line candidate among those a training-free re-detector surfaces). This is ground-truth-free, with the fundamental matrix fit on held-out frames. The +55-point gap over single-view recovery decomposes into two cross-view levers: the on-line search region lifts recovery 10%→51% (localization), and choosing the on-line candidate lifts it 51%→65% (the +14-point uniquely-geometric disambiguation). Both are unavailable to one camera. The precondition is sharp—a rescuer view *and* a drift component perpendicular to the line—so recovery is absent under correlated co-drift (0 of 2652 loss events) and degenerate when drift runs along the line (median —cos— = 0.999, a ≈25 px residual of a 381 px drift). The alarm is load-bearing: it fires on 85% of true losses (end-to-end 58.5% gated recovery ≈ the 65% oracle) while keeping recovery off the 27% of healthy frames a blind always-on policy would corrupt.

5 Limitations and Outlook

Cross-view geometry is blind only to *correlated* co-drift, where two near-parallel views slide onto the same cloud band so the drifted pair still satisfies the epipolar constraint (epipolar AUC 0.056); the *triangulation angle* catches almost all of it (AUC 0.980) and the presence gate covers the residual. The same boundary governs recovery (a rescuer view is absent under co-drift, degenerate along

the line). With K calibrated cameras the chance that all drifted rays meet consistently shrinks geometrically, so a third view turns this blind tail into a rare exception—the natural next system. Evidence rests on two RTK self-sequences and public dataset-4 with block-bootstrap intervals; the closed-loop calibration collapse is architectural, so the predictor runs open-loop and triggers external re-acquisition.

6 Conclusion

We turned the cross-view geometry that *detects* silent drift in long-horizon tracking of dim, appearance-less UAV targets into a means to *recover* from it. GeoSentry, a 13-parameter interpretable model, turns cross-view geometry into a per-frame failure probability with no ground truth at inference; detection is well-calibrated and the discriminative signal lives in the triangulation angle, not classifier capacity. The same geometry enables recovery: when one view drifts while another still tracks, the rescuer's epipolar line localizes and disambiguates the lost target from appearance-identical clutter, recovering 65% of loss events versus 10% for a single view—structurally impossible for one camera. We add two findings: a *closed-loop calibration collapse* (run open-loop, never closed-loop) and a *correlated-vs-independent failure boundary* that bounds both detection and recovery and points directly at the next system—a multi-view rig that turns the two-view blind tail into a rare event. Results are preliminary at $n = 2$ labelled sequences with block-bootstrap intervals; we make no cross-sequence generalization claim.

References

1. Albl, C., et al.: Drone-tracking datasets: multi-camera sequences with RTK ground truth. https://github.com/CenekAlbl/drone-tracking-datasets (2020)
2. Doersch, C., et al.: TAP-Vid: a benchmark for tracking any point in a video. In: NeurIPS Datasets and Benchmarks (2022)
3. Karaev, N., et al.: CoTracker: it is better to track together. In: ECCV (2024)
4. Zhao, M., Li, B., Wang, S., et al.: Anti-UAV: a large multi-modal benchmark for UAV tracking. In: CVPR (2023)
5. Li, J., Dai, Y., Li, C., et al.: Infrared small-target detection: a survey of methods and datasets. IEEE TPAMI (2024)
6. Davey, S.J., Rutten, M.G., Cheung, B.: A comparison of detection performance for track-before-detect algorithms. EURASIP J. Adv. Signal Process. (2008)
7. El-Yaniv, R., Wiener, Y.: On the foundations of noise-free selective classification. JMLR 11, 1605–1641 (2010)
8. Cho, S., Huang, J., Nam, S., An, H., Kim, S., Lee, J.-Y.: Local all-pair correspondence for point tracking. In: ECCV (2024)
9. Daftry, S., Zeng, S., Bagnell, J.A., Hebert, M.: Introspective perception: learning to predict failures in vision systems. In: IROS (2016)
10. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated ROC curves: a nonparametric approach. Biometrics 44(3), 837–845 (1988)

11. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: ICML (2016)
12. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML (2017)
13. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press (2004)
14. Kowalczyk, A., et al.: A hybrid deep–geometric approach for efficient consistency assessment of stereo images. Sensors 25(14), 4507 (2025). https://doi.org/10.3390/s25144507
15. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS (2017)
16. Naeini, M.P., Cooper, G.F., Hauskrecht, M.: Obtaining well calibrated probabilities using Bayesian binning. In: AAAI (2015)
17. Rabiee, S., Biswas, J.: IVOA: introspective vision for obstacle avoidance. In: IROS (2019)
18. Ramanagopal, M.S., Anderson, C., Vasudevan, R., Johnson-Roberson, M.: Failing to learn: autonomously identifying perception failures for self-driving cars. IEEE RA-L (2018)
19. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: NeurIPS (2018)
20. Lu, X., Duan, Y., Li, F.: A reliable UAV tracking system with online re-detection network. ISA Transactions (2025)
21. SMART-TRACK: advanced algorithms for UAV tracking of targets exhibiting start-stop and irregular motion. Scientific Reports (2025). https://doi.org/10.1038/s41598-025-13698-6