

# Confidence-Aware ESKF–MSCKF Fusion of Visual–Inertial and UWB/TDOA Measurements for Robust UAV Localization in GPS-Denied Environments

Md Shahanur Islam Shagor

`smshagor.ru@gmail.com`

Voronezh State University of Forestry and Technologies <https://orcid.org/0009-0003-7730-3202>

---

## Research Article

**Keywords:** error-state Kalman filter, GPS-denied navigation, MSCKF, sensor fusion, time difference of arrival, ultra-wideband, unmanned aerial vehicle, visual–inertial odometry

**Posted Date:** September 18th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11073266/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** The authors declare no competing interests.

---

# Confidence-Aware ESKF–MSCKF Fusion of Visual–Inertial and UWB/TDOA Measurements for Robust UAV Localization in GPS-Denied Environments

Md Shahanur Islam Shagor

Independent Researcher, Department of Computer Science and Microelectronics Engineering,  
Voronezh State University of Forestry and Technologies, Voronezh, Russia

Corresponding author: smshagor.dev@gmail.com

**Abstract**—Accurate localization remains a limiting requirement for small unmanned aerial vehicles operating where global navigation satellite system signals are unavailable, intermittent, or deliberately excluded. Visual–inertial odometry provides locally smooth motion estimates but can drift or lose reliability under weak texture, blur, or illumination changes, whereas ultra-wideband time-difference-of-arrival (UWB/TDOA) measurements provide drift-bounded geometric information but are vulnerable to non-line-of-sight bias, unfavorable anchor geometry, and timing uncertainty. This article presents a confidence-aware fusion architecture that combines error-state Kalman filtering (ESKF), multi-state constraint Kalman filtering (MSCKF), and asynchronous UWB/TDOA corrections without allowing either sensing modality to dominate when its measurement quality deteriorates. The estimator propagates an inertial nominal state and covariance, constructs multi-view visual constraints through cloned camera states and feature elimination, and applies TDOA range-difference observations after geometry and timing assessment. Visual and radio updates are assigned bounded confidence factors derived from innovation consistency, tracking quality, residual dispersion, anchor geometry, and measurement freshness; these factors inflate the corresponding covariance before chi-square gating and Joseph-form correction. A reproducible Monte Carlo study comprising 50 independent 90-s trials per condition evaluates nominal operation, visual degradation, UWB non-line-of-sight contamination, and combined visual/radio/timing stress. Under the combined condition, the proposed fusion achieved a mean three-dimensional position RMSE of 0.0856 m, compared with 0.1584 m for fixed-covariance fusion, a 45.9% reduction, while the mean 95th-percentile error decreased from 0.2975 m to 0.1688 m. Ablation results show that visual-only or TDOA-only confidence adaptation does not reproduce the full benefit, indicating that the two mechanisms address complementary failure modes. The reported numerical results are simulation-derived and do not constitute physical-flight validation.

**Index Terms**—error-state Kalman filter, GPS-denied navigation, MSCKF, sensor fusion, time difference of arrival, ultra-wideband, unmanned aerial vehicle, visual–inertial odometry.

## I. INTRODUCTION

Reliable state estimation is fundamental to autonomous flight because guidance, mapping, obstacle avoidance, and control depend on a sufficiently accurate and internally consistent estimate of vehicle motion. The problem becomes more difficult in indoor, underground, urban-canyon, and infrastructure-inspection environments where satellite navigation is unavailable or too unreliable to serve as a global reference. In these environments, no single lightweight sensing modality is uniformly dependable. An inertial measurement unit (IMU) is fast and self-contained but accumulates bias-driven drift; cameras provide rich geometric constraints but their information content changes with texture, illumination, motion blur, and occlusion; and UWB radios provide metric geometric observations but can incur positive and negative biases under non-line-of-sight (NLOS), multipath, anchor geometry, and clock uncertainty.

Visual–inertial estimation has therefore become a primary solution for GPS-denied motion estimation. The MSCKF formulation introduced a computationally attractive way to exploit multi-view feature constraints without augmenting long-lived feature positions in the filter state [1]. Subsequent work improved estimator consistency through first-estimates Jacobians (FEJ) [2], demonstrated robust stereo filtering for high-speed aerial flight [3], and developed optimization-based alternatives such as VINS-Mono [4], OKVIS [5], and on-manifold preintegration formulations [6]. These systems establish that visual and inertial sensing can provide accurate short- to medium-horizon motion estimation, but they cannot eliminate the fundamental possibility of drift or visual information loss.

UWB offers a complementary metric reference. Classical TDOA localization exploits differences in signal arrival time to define hyperbolic constraints [9], and modern UWB ranging literature has characterized both the precision advantages and the severe multipath sensitivity of wideband radio measurements [10]. Recent UAV systems have coupled UWB with visual–inertial estimators to reduce drift [11], used learned visual–inertial–UWB fusion [12], introduced UWB-assisted visual–inertial correction with uncertain anchor coordinates [15], and developed tightly coupled ESKF methods with explicit UWB outlier rejection [14]. A 2026 TDOA–IMU study further showed the relevance of adaptive noise treatment for dark and confined UAV environments [16].

A remaining difficulty is that fusion quality depends not only on nominal sensor covariance but also on whether the measurement is trustworthy at the current instant. Fixed-covariance fusion can become overconfident when visual tracks degrade or when TDOA residuals are biased. Hard rejection alone is also insufficient: a measurement can be partially informative without being consistent enough to deserve its nominal weight. This article addresses that interval between unconditional acceptance and binary rejection by introducing a confidence-aware ESKF–MSCKF–TDOA fusion architecture. The proposed method retains filter-level interpretability and computational boundedness while adapting visual and radio measurement covariances according to measurable indicators of quality.

The contribution is deliberately restricted to localization and estimation so that the technical claims remain separable from higher-level coordination, system-level fault-injection methodology, and platform-security mechanisms. Within this boundary, the article makes three contributions. First, it formulates a unified error-state estimator in which IMU propagation, MSCKF multi-view constraints, and asynchronous UWB/TDOA range-difference updates operate on a common state while preserving covariance consistency through FEJ-informed visual linearization and Joseph-form updates. Second, it introduces modality-specific confidence functions that combine visual innovation and tracking quality with TDOA residual dispersion, anchor geometry, and measurement age; the resulting bounded covariance inflation enables gradual down-weighting before a gross-outlier gate is reached. Third, it evaluates the method using paired Monte Carlo experiments under nominal, visual-degradation, NLOS, and combined timing-stress scenarios, including ablations and statistical comparisons against fixed-covariance, VIO-only, and TDOA-only baselines.

The remainder of this article is organized as follows. Section II reviews visual–inertial and UWB localization literature. Section III defines the system and measurement models. Section IV presents the proposed confidence-aware estimator. Section V specifies the simulation and statistical protocol. Section VI reports the results and ablation study. Section VII positions the method relative to representative literature without conflating results obtained on different datasets or hardware. Section VIII states limitations and threats to validity, and Section IX concludes the article.

## II. RELATED WORK

### A. Filter- and Optimization-Based Visual-Inertial Estimation

The original MSCKF exploits the fact that a static feature observed from multiple camera poses induces constraints among those poses even after the feature itself is algebraically eliminated from the measurement equation [1]. This property makes MSCKF attractive for resource-constrained aerial platforms because the filter does not maintain every feature as a persistent state. FEJ techniques were subsequently developed to mitigate inconsistency caused by relinearizing unobservable directions at changing state estimates [2]. Stereo MSCKF variants demonstrated that filter-based VIO can support fast autonomous flight while remaining computationally bounded [3], and more recent work has revisited state pruning and feature marginalization to improve MSCKF efficiency [13].

Optimization-based estimators provide a different accuracy–complexity trade-off. OKVIS formulates keyframe-based nonlinear visual-inertial optimization [5], while VINS-Mono combines sliding-window optimization with initialization, marginalization, and loop-related capabilities [4]. On-manifold IMU preintegration provides a principled way to summarize high-rate inertial data between optimization states [6]. Public benchmarks such as EuRoC [7] and TUM VI [8] have been central to comparing these methods under motion blur, illumination changes, and dynamic trajectories. The present work does not claim to replace these visual front ends. Instead, it uses an MSCKF-compatible filtering structure and focuses on how radio measurements should be weighted when visual and UWB reliability vary over time.

### B. UWB/TDOA Localization and Visual–Radio Fusion

TDOA localization converts inter-anchor arrival-time differences into range-difference constraints. Chan and Ho derived a computationally efficient hyperbolic estimator with favorable small-error behavior [9]. In practice, however, UWB errors are not stationary Gaussian disturbances: multipath and NLOS propagation can create biased measurements whose statistics depend on geometry and environment [10]. For UAV navigation, Lin and Zhan combined UWB positioning with VIO through nonlinear optimization to reduce long-term drift [11]. VIUNet used a deep fusion architecture for visual, inertial, and UWB sensing [12], demonstrating another route to multi-modal indoor UAV localization.

More recent studies have concentrated on robustness to radio imperfections. Zhao et al. developed an ESKF-based indoor UAV framework with multi-epoch UWB outlier rejection and heterogeneous sensing [14]. Xing et al. considered UWB-assisted correction when anchor positions are initially uncertain [15]. Shi et al. used TDOA with IMU measurements and adaptive noise weighting for confined UAV environments [16]. These studies support the value of adaptive UWB treatment but leave room for a filter in which visual and TDOA confidence are adapted jointly, using distinct quality indicators, while retaining a clear separation between soft covariance inflation and hard innovation rejection.

### C. Research Gap

The gap addressed here is therefore not the existence of VIO–UWB fusion, nor the isolated use of adaptive covariance. The specific problem is simultaneous, interpretable confidence management for two failure-prone complementary modalities inside an ESKF–MSCKF architecture. Visual degradation and radio NLOS have different signatures: the former appears through track support, reprojection consistency, and relative-motion innovation, whereas the latter appears through range-difference residuals, geometric dilution, and timing freshness. A common fixed covariance cannot represent both processes. The proposed design assigns each modality its own confidence path, then lets both paths influence a shared gated correction step.

## III. SYSTEM MODEL AND PROBLEM FORMULATION

### A. Frames, Sensors, and State

Let  $\{W\}$  denote the local world frame,  $\{I\}$  the IMU/body frame, and  $\{C\}$  the camera frame. The UWB anchor coordinates  $a_i \in R^3$  are expressed in  $\{W\}$ . The nominal inertial state at time  $k$  is  $x_k = [p_k^T, v_k^T, q_k^T, b_{a,k}^T, b_{g,k}^T]^T$ , where  $p$  and  $v$  are position and velocity in  $\{W\}$ ,  $q$  is the unit quaternion rotating vectors from  $\{I\}$  to  $\{W\}$ , and  $b_a$  and  $b_g$  are accelerometer and gyroscope biases. The corresponding 15-dimensional local error state is  $\delta x = [\delta p^T, \delta v^T, \delta \theta^T, \delta b_a^T, \delta b_g^T]^T$ . MSCKF camera clones are appended transiently as

pose error states so that multi-view feature constraints can be formed without retaining the feature coordinates after marginalization.

### B. Inertial Propagation

After subtracting estimated biases from accelerometer measurement  $a_m$  and angular-rate measurement  $\omega_m$ , the nominal dynamics are integrated between correction events. With gravity  $g$  expressed in the world frame and  $R(q)$  the body-to-world rotation matrix, the continuous-time model is

$$\dot{p} = v, \quad \dot{v} = R(q)(a_m - b_a - n_a) + g, \quad \dot{q} = \frac{1}{2} \Omega(\omega_m - b_g - n_g) q, \quad (1)$$

with bias random walks  $b_a = n_a$  and  $b_g = n_g$ . Linearizing the error dynamics gives  $\delta \dot{x} = F \delta x + G n$ . For a propagation interval  $\Delta t$ , the transition matrix  $\Phi$  is obtained from a second-order discretization  $\Phi \approx I + F \Delta t + (F \Delta t)^2/2$ , and covariance is propagated as

$$P_k^- = \Phi_k P_{k-1}^+ \Phi_k^T + Q_k. \quad (2)$$

The quaternion correction is injected through a small-angle rotation and the attitude-error coordinates are reset after each measurement update. When a stationary interval is detected, an optional zero-velocity constraint may be applied as an ordinary measurement update; it is not relied upon in the reported moving-trajectory experiments.

### C. MSCKF Multi-View Visual Constraint

At selected camera times, a clone of the current camera pose is appended to the augmented state and covariance. A feature  $f$  observed by  $m$  cloned camera poses generates stacked image residual  $r_f$ . Linearization about the first estimate used for the relevant pose produces

$$r_f \approx H_x \delta x_A + H_f \delta p_f + n_f, \quad (3)$$

where  $\delta x_A$  denotes the augmented state error and  $\delta p_f$  is the feature-position perturbation. To avoid introducing the feature into the persistent state, an orthonormal basis  $A$  of the left null space of  $H_f$  is used,  $A^T H_f = 0$ , giving the projected constraint

$$r_o = A^T r_f \approx H_o \delta x_A + n_o, \quad H_o = A^T H_x. \quad (4)$$

Triangulation requires positive depth and an acceptable reprojection residual. The projected residual is subjected to a chi-square consistency test before correction. FEJ-style linearization is retained for cloned poses involved in the constraint so that changes in the current estimate do not spuriously inject information into unobservable directions [2]. After their measurement information has been consumed, old camera clones are marginalized to bound state dimension.

### D. UWB/TDOA Measurement Model

For a reference anchor  $a_0$  and another anchor  $a_i$ , the ideal range-difference observation is the difference between the Euclidean distances from the vehicle to the two anchors. Expressed in meters after multiplying the measured arrival-time difference by propagation velocity, the observation model is

$$z_i = \|p - a_i\| - \|p - a_0\| + n_i. \quad (5)$$

Linearization with respect to position yields

$$H^p_i = (p - a_i)^T / \|p - a_i\| - (p - a_0)^T / \|p - a_0\|. \quad (6)$$

Stacking valid anchor pairs produces  $z = h(p) + n$ . A damped least-squares position can be computed for initialization or geometric diagnostics, but the proposed filter uses the raw range-difference residual in the ESKF correction. This avoids treating an already estimated TDOA position as if it were an independent raw measurement. Measurements older than a configured freshness horizon are rejected; smaller time offsets are translated into covariance inflation as described in Section IV.

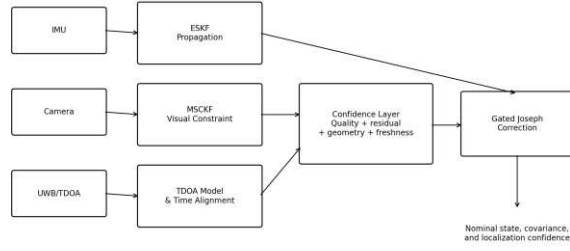


FIGURE 1. Proposed localization architecture. IMU propagation, MSCKF visual constraints, and UWB/TDOA observations are evaluated by modality-specific confidence mechanisms before gated Joseph-form correction.

## IV. PROPOSED CONFIDENCE-AWARE ESKF-MSCKF-TDOA FUSION

### A. Measurement Confidence as Covariance Inflation

The central design choice is to keep the state estimator probabilistically interpretable while allowing measurement quality to vary. Each measurement family begins with a nominal covariance  $R_0$ . A scalar confidence  $c \in [c_{\min}, 1]$  is computed from observable quality indicators and converted to an effective covariance

$$R_{\text{eff}} = R_0 / \max(c, c_{\min}). \quad (7)$$

Thus, a measurement with  $c = 1$  retains its nominal weight, whereas lower confidence increases uncertainty without changing the residual itself. Covariance inflation is bounded by  $c_{\min}$  to avoid numerical singularity or an uncontrolled jump in effective variance. Grossly inconsistent observations are still rejected by a separate innovation gate. This two-stage design distinguishes “uncertain but useful” from “incompatible with the predicted state.”

### B. Visual Confidence

The visual confidence  $c_V$  combines front-end support and filter consistency. Let  $q_{\text{trk}} \in [0, 1]$  represent normalized tracking support, such as the fraction of successfully tracked and geometrically valid features, and let  $v_V = r_o^T S_V^{-1} r_o / d_V$  be the normalized innovation statistic for a projected visual constraint with  $d_V$  degrees of freedom. A bounded confidence can be written as

$$c_V = \text{clip}(q_{\text{trk}} \cdot \exp[-\alpha_V \max(0, v_V - 1)], c_{\min}, 1). \quad (8)$$

The exponential term leaves innovations near their expected normalized value almost unchanged, while progressively reducing the weight of constraints whose residuals grow beyond the expected covariance. A hard chi-square threshold is applied after this soft adaptation. In an image-based implementation,  $q_{\text{trk}}$  can additionally include texture, track length, disparity, and reprojection quality; in the simulation of Section V, the same mechanism is exercised through a controlled MSCKF-derived local-motion surrogate whose quality variable and innovation are degraded during the designated visual-failure interval.

### C. TDOA Confidence: Residual, Geometry, and Freshness

TDOA reliability is represented by three factors. First, residual quality  $c_r$  decreases as the median absolute deviation or normalized innovation of range-difference residuals increases. Second, geometry factor  $c_g$  decreases when the stacked TDOA Jacobian is ill-conditioned; a GDOP-like diagnostic derived from  $(H_p^T H_p)^{-1}$  is sufficient for this purpose when the matrix is well-conditioned. Third, freshness factor  $c_t$  decreases with absolute measurement age  $|\Delta t|$ . The composite confidence is

$$c_T = \text{clip}(c_r \cdot c_g \cdot c_t, c_{\min}, 1). \quad (9)$$

Because NLOS biases can be much larger than nominal radio noise, residual-based adaptation is supplemented with a hard innovation gate. Measurement age is handled differently from NLOS: a moderately delayed observation is down-weighted, whereas an observation beyond the freshness horizon is discarded. This prevents asynchronous sensing from being treated as perfectly synchronized while avoiding unnecessary rejection of small timing offsets.

### D. Gated Joseph-Form Correction

For either measurement family, let  $r = z - h(\hat{x})$  and  $S = H P^- H^T + R_{\text{eff}}$ . The innovation statistic  $d^2 = r^T S^{-1} r$  is compared with the chi-square threshold for the residual dimension. If accepted, the Kalman gain is  $K = P^- H^T S^{-1}$ , the error state is injected into the nominal state, and covariance is corrected using the Joseph form

$$P^+ = (I - KH)P^-(I - KH)^T + K R_{\text{eff}} K^T. \quad (10)$$

The Joseph expression is used because it preserves symmetry and positive-semidefinite structure more robustly under finite-precision arithmetic than the simplified covariance update. Camera-clone marginalization follows visual correction. TDOA updates are processed at their own timestamps and do not require camera synchronization, although freshness affects confidence.

### E. Processing Sequence and Bounded Complexity

IMU samples drive propagation. Camera events augment the state, construct multi-view constraints from tracks that have terminated or reached a configured length, apply visual confidence and gating, then marginalize stale clones. UWB events construct TDOA range differences, evaluate residual/geometry/freshness confidence, and apply the same generic correction operator. The number of camera clones and accepted visual tracks is bounded, so filter complexity does not grow with mission duration. The proposed confidence layer adds scalar and small-matrix diagnostics but does not introduce a learned model or an additional long-lived state.

## V. EVALUATION METHODOLOGY

### A. Reproducible Monte Carlo Design

The numerical study was designed to isolate estimator behavior under controlled sensing failures rather than to imitate a specific flight platform. Each configuration was evaluated in 50 paired Monte Carlo trials of 90 s, using the same random seed for competing methods within a scenario. The vehicle followed a smooth three-dimensional trajectory contained within an approximately  $18 \text{ m} \times 16 \text{ m} \times 6 \text{ m}$  anchor volume. Eight anchors were distributed around the lower and upper perimeter to provide vertical as well as horizontal range-difference geometry. Figure 2 shows the top-down trajectory and anchor layout.

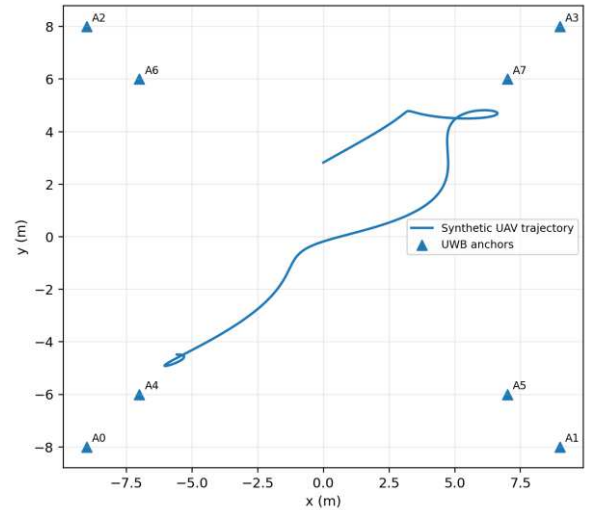


FIGURE 2. Top-down view of the simulated three-dimensional UAV trajectory and the eight-anchor UWB geometry used for the Monte Carlo evaluation.

The inertial propagation interval was 0.02 s (50 Hz). An MSCKF-derived local-motion observation was supplied at 10 Hz and a TDOA update at 5 Hz. The acceleration measurement noise standard deviation was  $0.11 \text{ m/s}^2$ , the accelerometer-bias random-walk parameter was  $0.0015 \text{ m/s}^2/\sqrt{\text{s}}$  in the numerical model, and nominal TDOA standard deviation was 0.08 m. The local-motion surrogate was intentionally used so that the reported study tests the estimator and confidence logic without falsely attributing performance to a particular feature detector, camera calibration, or image sequence. Public visual-inertial benchmarks such as EuRoC [7] and TUM VI [8] motivate the use of asynchronous camera/IMU sensing

and challenging visual conditions, but their image data were not used to generate the numerical values reported here.

TABLE I SIMULATION AND SENSOR PARAMETERS

| Parameter                              | Value                       |
|----------------------------------------|-----------------------------|
| Monte Carlo trials per method/scenario | 50                          |
| Trial duration                         | 90 s                        |
| Inertial propagation                   | 50 Hz                       |
| MSCKF-derived local-motion update      | 10 Hz                       |
| UWB/TDOA update                        | 5 Hz                        |
| UWB anchors                            | 8                           |
| Acceleration noise $\sigma$            | 0.11 m/s <sup>2</sup>       |
| Acceleration-bias random walk          | 0.0015 m/s <sup>2</sup> /√s |
| Nominal TDOA noise $\sigma$            | 0.08 m                      |
| Visual-degradation interval            | 60–80 s                     |
| Nominal timing jitter                  | ≈ 6 ms                      |
| Additional combined-stress jitter      | ≈ 25 ms standard deviation  |

### B. Stress Scenarios

Four scenarios were evaluated. The nominal scenario used stable local-motion and TDOA observations with nominal noise. The visual-degradation scenario reduced visual quality during 60–80 s and introduced increased local-motion noise, missed updates, and occasional outlying constraints. The NLOS scenario contaminated TDOA observations with a baseline outlier probability of approximately 16%, with stronger contamination later in the run and signed biases on the order of 0.9 m. The combined scenario applied both visual degradation and NLOS contamination and added approximately 25 ms standard-deviation timing jitter on top of the nominal synchronization jitter. The precise purpose of the combined case was to test whether confidence mechanisms that are individually sensible remain stable when both modalities deteriorate together.

### C. Baselines and Metrics

Four estimator variants were compared: VIO-only, consisting of inertial propagation and the local visual-motion correction; TDOA-only, consisting of inertial propagation and radio correction; fixed fusion, which used both observation families with constant nominal covariance; and the proposed confidence-aware fusion. Performance metrics were three-dimensional position RMSE over each run, the per-run 95th-percentile position error, maximum position error, and final position error. The proposed method also recorded rejected visual and TDOA updates. For the main proposed-versus-fixed comparison, paired two-sided t-tests were applied to run-level RMSE differences because identical random seeds generated paired sensing disturbances. Mean paired differences are reported with 95% confidence intervals; statistical significance is interpreted together with effect magnitude rather than as a substitute for engineering relevance.

## VI. RESULTS AND DISCUSSION

### A. Localization Accuracy Across Failure Conditions

Table II summarizes run-level RMSE. Under nominal sensing, fixed and adaptive fusion were almost indistinguishable in absolute terms: 0.02984 m and 0.02972 m, respectively. This is expected because confidence adaptation should not materially perturb well-behaved measurements. Under degraded sensing, however, the difference was substantial. Visual degradation increased fixed-fusion RMSE to 0.09206 m, whereas the proposed method retained 0.03995 m. Under TDOA NLOS, the corresponding values were 0.12648 m and 0.06364 m. In the combined condition, fixed fusion reached 0.15841 m while the proposed method achieved 0.08565 m.

TABLE II THREE-DIMENSIONAL POSITION RMSE ACROSS 50 MONTE CARLO TRIALS (m)

| Method   | Nominal         | Visual degr.    | UWB NLOS        | Combined        |
|----------|-----------------|-----------------|-----------------|-----------------|
| VIO only | 0.2190 ± 0.0695 | 1.0722 ± 0.3907 | 0.2190 ± 0.0695 | 1.0722 ± 0.3907 |

| Method       | Nominal         | Visual degr.    | UWB NLOS        | Combined        |
|--------------|-----------------|-----------------|-----------------|-----------------|
| TDOA only    | 0.0530 ± 0.0034 | 0.0530 ± 0.0034 | 0.2377 ± 0.0153 | 0.2408 ± 0.0178 |
| Fixed fusion | 0.0298 ± 0.0027 | 0.0921 ± 0.0260 | 0.1265 ± 0.0125 | 0.1584 ± 0.0180 |
| Proposed     | 0.0297 ± 0.0027 | 0.0400 ± 0.0049 | 0.0636 ± 0.0107 | 0.0856 ± 0.0137 |

Relative to fixed fusion, the proposed method reduced mean RMSE by 56.6% in the visual-degradation scenario, 49.7% in the NLOS scenario, and 45.9% under combined stress. Figure 3 shows the means and across-run standard deviations. The VIO-only estimator was especially sensitive to the imposed visual degradation because no global metric correction was available; conversely, TDOA-only accuracy remained bounded but was less precise than fused estimation and deteriorated under NLOS. These trends illustrate complementarity rather than a claim that either single modality is universally inferior.

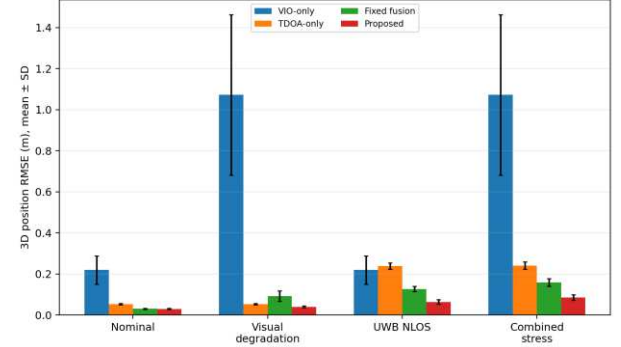


FIGURE 3. Mean three-dimensional position RMSE ± one standard deviation across 50 trials for each scenario and estimator variant.

### B. Paired Statistical Analysis and Tail Error

Because runs were paired by random seed, the proposed-minus-fixed comparison can be evaluated on within-run RMSE differences. The mean improvement (fixed minus proposed) was 0.00013 m in the nominal case (95% CI 0.00003–0.00022 m; paired  $t = 2.74$ ,  $p = 0.0085$ , Cohen's  $d_z = 0.39$ ). Although statistically detectable, this nominal difference is operationally small. Under visual degradation, the mean paired improvement was 0.05211 m (95% CI 0.04521–0.05901 m;  $p < 10^{-19}$ ,  $d_z = 2.15$ ); under NLOS it was 0.06284 m (95% CI 0.05992–0.06576 m;  $p < 10^{-39}$ ,  $d_z = 6.11$ ); and under combined stress it was 0.07276 m (95% CI 0.06843–0.07709 m;  $p < 10^{-34}$ ,  $d_z = 4.78$ ).

Tail behavior follows the same pattern. In the combined scenario, mean per-run 95th-percentile error decreased from 0.2975 m for fixed fusion to 0.1688 m for the proposed method, while mean maximum error decreased from 0.5997 m to 0.3985 m. Figure 4 shows the distribution of run-level RMSE under combined stress and demonstrates that the improvement is not produced by only a few favorable trials.

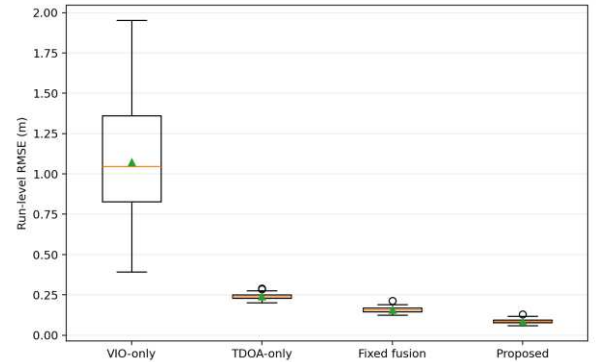


FIGURE 4. Distribution of run-level position RMSE for the combined visual/NLOS/timing stress condition (50 trials per method).

### C. Ablation of Confidence Mechanisms

The combined-stress case was used for ablation because it activates both failure modes. Visual-confidence-only fusion reduced mean RMSE from 0.15841 m to 0.14685 m, while TDOA-confidence-only fusion reduced it

to 0.14891 m. The full method reached 0.08565 m. The result is important because it shows that the two adaptations are not interchangeable: visual confidence protects the estimator when relative-motion observations become unreliable, whereas TDOA confidence protects it from biased or stale global constraints. Applying only one mechanism leaves the other failure channel able to dominate the correction.

**TABLE III CONFIDENCE ABLATION UNDER COMBINED STRESS**

| Variant                | RMSE (m)            | 95th pct. (m) | Max. (m) |
|------------------------|---------------------|---------------|----------|
| Fixed covariance       | $0.1584 \pm 0.0180$ | 0.2975        | 0.5997   |
| Visual confidence only | $0.1468 \pm 0.0130$ | 0.2617        | 0.5994   |
| TDOA confidence only   | $0.1489 \pm 0.0411$ | 0.3458        | 0.5497   |
| Full proposed          | $0.0856 \pm 0.0137$ | 0.1688        | 0.3985   |

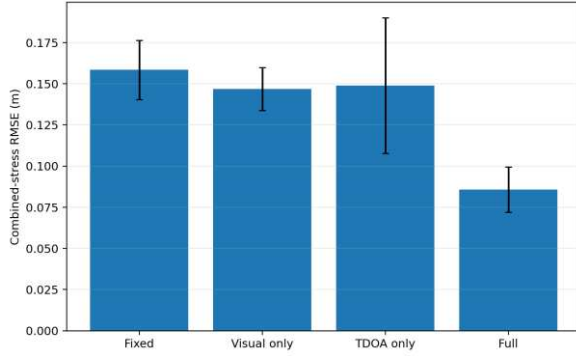


**FIGURE 5. Combined-stress ablation.** Joint visual and TDOA confidence adaptation provides a larger reduction in RMSE than either mechanism alone.

The full proposed configuration rejected, on average, 65.0 TDOA updates and 67.98 visual updates per combined-stress trial after confidence assessment and hard gating. These counts should not be interpreted as a target rejection rate; they are a consequence of the deliberately aggressive disturbance model. Their value is diagnostic: the estimator remains bounded when both sensing channels periodically become inconsistent, while nominal-case accuracy remains essentially unchanged.

## VII. COMPARATIVE ANALYSIS AND POSITIONING RELATIVE TO PRIOR WORK

Direct numerical ranking against published localization systems would be misleading because prior studies use different cameras, radio hardware, trajectories, anchor geometry, datasets, and ground-truth systems. The comparison in Table IV is therefore architectural rather than a cross-paper accuracy leaderboard. It identifies the estimation structure, radio modality, and robustness mechanism most relevant to the present contribution.

**TABLE IV ARCHITECTURAL COMPARISON WITH REPRESENTATIVE PRIOR WORK**

| Method            | Core estimator         | Radio integration    | Primary robustness emphasis                           |
|-------------------|------------------------|----------------------|-------------------------------------------------------|
| MSCKF [1]         | EKF/MSCKF              | None                 | Feature elimination; bounded filter state             |
| S-MSCKF [3]       | Stereo MSCKF           | None                 | Stereo visual redundancy for fast flight              |
| Lin and Zhan [11] | Nonlinear VIO fusion   | UWB position         | Global drift correction                               |
| VIUNet [12]       | Deep multimodal fusion | UWB                  | Learned visual-inertial-UWB fusion                    |
| Zhao et al. [14]  | Tightly coupled ESKF   | UWB ranging          | Multi-epoch UWB outlier rejection                     |
| Shi et al. [16]   | Adaptive KF            | UWB/TDOA             | Adaptive noise and anchor weighting                   |
| Proposed          | ESKF-MSCKF             | Raw TDOA differences | Joint visual + residual/geometry/freshness confidence |

The proposed architecture differs from radio-only adaptive filtering [16] by explicitly coupling a multi-state visual constraint model to the same confidence-managed correction path. It differs from systems that primarily reject UWB outliers [14] by allowing soft down-weighting of both visual and radio observations before a hard gate. It also remains model-based rather than learned, which makes the source of each covariance change

inspectable. This interpretability is useful when the goal is not only low error but also understanding why an observation was trusted.

The simulation results should consequently be read as evidence for the proposed confidence mechanism under a controlled state-estimation model, not as evidence that it outperforms the physical systems in Table IV. A valid experimental ranking would require common raw sensor sequences, identical calibration, identical anchor geometry, and a shared reference trajectory. Section VIII identifies this as a priority for subsequent validation.

## VIII. LIMITATIONS AND THREATS TO VALIDITY

The principal limitation is that all numerical results in this article are simulation-derived. No claim is made about physical-flight accuracy, UWB radio qualification, real NLOS channel statistics, camera exposure behavior, thermal effects, or embedded timing latency. The simulation deliberately exposes the estimator to structured degradations, but a real environment can contain correlated multipath, anchor-specific clock errors, feature-tracking failures, rolling-shutter effects, calibration drift, and dynamic occluders that are not fully represented by the disturbance model.

A second limitation is the abstraction of the image front end. The state estimator is formulated with MSCKF camera clones, multi-view geometry, null-space projection, FEJ treatment, and chi-square gating; however, the Monte Carlo study drives the visual correction through an MSCKF-derived local-motion surrogate rather than raw images. The reported RMSE therefore evaluates confidence-aware fusion and estimator response, not feature detection, optical flow, descriptor quality, or end-to-end camera processing. Image-sequence evaluation on EuRoC, TUM VI, or an equivalent synchronized UAV dataset is required before drawing conclusions about complete visual-front-end performance.

Third, the TDOA anchors are known exactly in the current model. Deployment error in anchor coordinates, antenna delay calibration, clock drift, and environment-dependent channel bias can couple with vehicle state and change observability. Recent work on uncertain anchor coordinates [15] indicates that joint or online calibration is a meaningful extension. Finally, the paired t-test assumes that the distribution of paired RMSE differences is sufficiently regular for mean-based inference. The large stress-condition effect sizes make the engineering conclusion robust to modest deviations, but a future hardware study should additionally report nonparametric intervals, trajectory-wise ATE/RPE, estimator consistency metrics such as NEES/NIS, and computational latency on the target processor.

## IX. CONCLUSION

This article presented a confidence-aware localization architecture that fuses inertial propagation, MSCKF-compatible visual constraints, and asynchronous UWB/TDOA observations within an ESKF. The key mechanism is a separation between soft confidence adaptation and hard innovation gating. Visual confidence responds to tracking support and innovation consistency, while TDOA confidence responds to residual dispersion, anchor geometry, and measurement freshness. Both modalities modify effective measurement covariance before a Joseph-form error-state correction.

Across 50 paired Monte Carlo trials per scenario, confidence adaptation had negligible practical cost under nominal sensing and provided substantial benefit under controlled failures. Relative to fixed-covariance fusion, mean three-dimensional RMSE was reduced by 56.6% under visual degradation, 49.7% under UWB NLOS contamination, and 45.9% under combined visual, radio, and timing stress. The combined-stress ablation showed that adapting only one modality did not reproduce the full improvement, supporting the hypothesis that visual and UWB confidence address complementary error processes.

The study establishes a reproducible estimator-level basis for subsequent experimental work but deliberately stops short of physical-flight claims. The next validation step is end-to-end evaluation with raw synchronized camera/IMU data and measured UWB/TDOA observations, followed by hardware experiments that quantify calibration sensitivity, NLOS behavior, timing uncertainty, estimator consistency, and embedded computation. Maintaining this distinction between simulation evidence and physical evidence is essential for a defensible assessment of GPS-denied UAV localization.

## ACKNOWLEDGMENT

OpenAI ChatGPT (GPT-5.6 Sol) was used to assist with manuscript drafting, language organization, presentation of the Introduction through Conclusion, and preparation of figure/table descriptions. The author defined the research scope, estimator architecture, evaluation constraints, and non-overlap boundaries; reviewed the mathematical formulation, simulation assumptions, numerical outputs, statistical analysis, references, and conclusions; and remains responsible for the accuracy and integrity of the submitted work. The numerical results reported in this article were generated from the explicitly described simulation model rather than generated as textual estimates by the AI system.

## REFERENCES

- [1] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint Kalman filter for vision-aided inertial navigation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Rome, Italy, 2007, pp. 3565–3572, doi: 10.1109/ROBOT.2007.364024.
- [2] G. P. Huang, A. I. Mourikis, and S. I. Roumeliotis, "A first-estimates Jacobian EKF for improving SLAM consistency," in *Experimental Robotics, Springer Tracts in Advanced Robotics*, vol. 54, 2009, pp. 373–382, doi: 10.1007/978-3-642-00196-3\_43.
- [3] K. Sun et al., "Robust stereo visual inertial odometry for fast autonomous flight," *IEEE Robot. Autom. Lett.*, vol. 3, no. 2, pp. 965–972, Apr. 2018, doi: 10.1109/LRA.2018.2793349.
- [4] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, Aug. 2018, doi: 10.1109/TRO.2018.2853729.
- [5] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual-inertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, 2015, doi: 10.1177/0278364914554813.
- [6] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual-inertial odometry," *IEEE Trans. Robot.*, vol. 33, no. 1, pp. 1–21, Feb. 2017, doi: 10.1109/TRO.2016.2597321.
- [7] M. Burri et al., "The EuRoC micro aerial vehicle datasets," *Int. J. Robot. Res.*, vol. 35, no. 10, pp. 1157–1163, 2016, doi: 10.1177/0278364915620033.
- [8] D. Schubert, T. Goll, N. Demmel, V. Usenko, J. Stückler, and D. Cremers, "The TUM VI benchmark for evaluating visual-inertial odometry," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2018, pp. 1680–1687, doi: 10.1109/IROS.2018.8593419.
- [9] Y. T. Chan and K. C. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Trans. Signal Process.*, vol. 42, no. 8, pp. 1905–1915, Aug. 1994, doi: 10.1109/78.301830.
- [10] D. Dardari, A. Conti, U. Ferner, A. Giorgetti, and M. Z. Win, "Ranging with ultrawide bandwidth signals in multipath environments," *Proc. IEEE*, vol. 97, no. 2, pp. 404–426, Feb. 2009, doi: 10.1109/JPROC.2008.2008846.
- [11] H.-Y. Lin and J.-R. Zhan, "GNSS-denied UAV indoor navigation with UWB incorporated visual inertial odometry," *Measurement*, vol. 206, Art. no. 112256, 2023, doi: 10.1016/j.measurement.2022.112256.
- [12] P.-Y. Kao, H.-J. Chang, K.-W. Tseng, T. Chen, H.-L. Luo, and Y.-P. Hung, "VIUNet: Deep visual-inertial-UWB fusion for indoor UAV localization," *IEEE Access*, vol. 11, pp. 61525–61534, 2023, doi: 10.1109/ACCESS.2023.3279292.
- [13] M. R. Abdollahi, S. H. Pourtakdoust, M. H. Yoosefian Nooshabadi, and H. N. Pishkenari, "An improved Multi-State Constraint Kalman Filter for visual-inertial odometry," *J. Franklin Inst.*, vol. 361, no. 15, Art. no. 107130, 2024, doi: 10.1016/j.jfranklin.2024.107130.
- [14] J. Zhao, Z. Deng, E. Hu, W. Su, B. Lou, and Y. Liu, "An indoor UAV localization framework with ESKF tightly-coupled fusion and multi-epoch UWB outlier rejection," *Sensors*, vol. 25, no. 24, Art. no. 7673, 2025, doi: 10.3390/s25247673.
- [15] Y. Xing, W. Li, F. Pan, and X. Feng, "Ultra-wideband assisted visual-inertial localization correction system with position-unknown UWB anchors," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2025, pp. 12884–12891, doi: 10.1109/IROS60139.2025.11246170.
- [16] H. Shi, Q. Wang, Z. Wang, J. Zhan, H. Liang, and B. Yang, "TDOA based localization mechanism for the UAV positioning in dark and confined environments," *Displays*, vol. 92, Art. no. 103346, 2026, doi: 10.1016/j.displa.2026.103346.

## AUTHOR BIOGRAPHY

**MD SHAHANUR ISLAM SHAGOR** is an independent researcher and software engineer working on autonomous systems, GPS-denied navigation, multi-sensor state estimation, trustworthy distributed systems, and applied artificial intelligence. His research interests include UAV autonomy, visual-inertial estimation, UWB localization, federated learning, and security mechanisms for cyber-physical systems.