Vision-Based Relative State Estimation for Non-Cooperative Spacecraft Using Deep Learning and Adaptive Kalman Filtering

Safa Mesut Bostancı[1] Halil Ersin Söken[1]

[1]Department of Aerospace Engineering, Middle East Technical University, Ankara, Türkiye

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

ICRA IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION 2026 VIENNA

Motivation

Autonomous rendezvous and docking with an uncooperative spacecraft underpins on-orbit servicing (OOS) and active debris removal (ADR). Unlike cooperative targets, an uncooperative object may be non-responsive, poorly characterized, and tumbling, so proximity operations become a tightly coupled estimation-and-control problem. Growing congestion in Low Earth Orbit and the rising population of defunct satellites raise the risk of cascading collisions (Kessler syndrome) [1]. Terminal approach imposes stringent requirements [2]: relative attitude accuracy $\sim 0.3°$, position < 0.2 m at 10 m, and translational velocity ~ 0.01 m/s — all while measurement quality degrades under self-occlusion, harsh shadows, and changing illumination against an Earth-limb background.

The Gap: Pose ≠ Motion

Deep-learning pipelines estimate single-frame pose (position and attitude) accurately, but typically provide no velocity or angular rate, precisely the dynamic states required for trajectory prediction and control. Naive finite differencing of noisy DL pose amplifies error by orders of magnitude (see results). Bridging single-frame pose and full dynamic state is the core problem addressed here.

Contributions

Recent methods pair DL pose estimators with recursive filters [3, 4]; surveys [5] categorize monocular pipelines as direct (end-to-end regression) or indirect (CNN features + PnP). We close the pose-to-motion gap with:

1. A framework fusing a keypoint-based CNN with a Multiplicative Extended Kalman Filter (MEKF) on $SO(3)$ to recover full 6-Degree of Freedom (DoF) motion (position, velocity, attitude, angular rate) while avoiding the quaternion-norm and singularity issues of a standard EKF.
2. Adaptive, tuning-free measurement covariance $\mathbf{R}_k$ derived directly from Perspective-n-Point (PnP) reprojection geometry — trust is set per-frame by keypoint quality, range, and viewing angle, not by a hand-tuned constant.
3. Bayesian-optimized process-noise tuning and validation on SPEED-UE-Cube, recovering motion from *pose-only* measurements while simultaneously improving pose accuracy by over an order of magnitude.

State & Measurement Model

The relative state is propagated in the chief's local-vertical local-horizontal (LVLH) frame using linearized Clohessy–Wiltshire (CW) dynamics, valid for near-circular chief orbits and short relative ranges. The 13-state vector is

$$\mathbf{x} = \left[\, \mathbf{p}^\top,\ \mathbf{v}^\top,\ \mathbf{q}^\top,\ \boldsymbol{\omega}^\top \,\right]^\top \in \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{H} \times \mathbb{R}^3,$$

with position $\mathbf{p}$, velocity $\mathbf{v}$, attitude quaternion $\mathbf{q}$, and body angular rate $\boldsymbol{\omega}$. The vision front-end supplies pose-only measurements

$$\mathbf{z}_k = \left[\, \mathbf{p}_k^{m\,\top},\ \mathbf{q}_k^{m\,\top} \,\right]^\top = h(\mathbf{x}_k) + \mathbf{v}_k, \quad \mathbf{v}_k \sim \mathcal{N}(\mathbf{0},\, \mathbf{R}_k),$$

so velocity and angular rate are observed only *indirectly* through the dynamics — making the choice of $\mathbf{R}_k$ critical.

Vision Pipeline & Dataset

SPEED-UE-Cube [6] is a photorealistic synthetic dataset of a non-cooperative 3U CubeSat rendered in Unreal Engine 5, with a full-orbit rendezvous trajectory sampled at 5 s intervals, ground-truth 6-DoF pose, and varied lighting and Earth backgrounds.

Three-stage monocular front-end:

- ODN (object detection network) localizes the spacecraft and crops the region of interest.
- KRN (keypoint regression network) predicts 11 2D semantic keypoints on the CubeSat body via heatmap regression.
- PnP solves the absolute pose $(\mathbf{p}_k^m, \mathbf{q}_k^m)$ from 2D–3D correspondences, with reprojection residuals reused to build $\mathbf{R}_k$.

PyTorch reimplementation; YOLOv4-tiny / CSPDarknet-tiny backbone with CIoU loss [7].

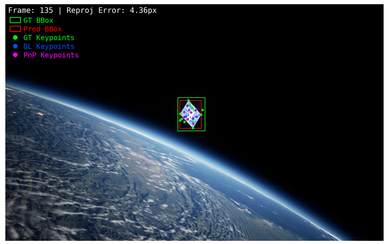

Figure 1. Detected bounding box and 11 keypoint predictions under cluttered Earth background.

Navigation Filter: 13-State MEKF

A Multiplicative EKF maintains the quaternion unit-norm constraint through multiplicative error-state updates [8]. The nominal state $\mathbf{x} = [\mathbf{p}^\top, \mathbf{v}^\top, \mathbf{q}^\top, \boldsymbol{\omega}^\top]^\top$ (position, velocity, attitude quaternion, angular velocity) propagates under Clohessy–Wiltshire (CW) dynamics, while the 12-dimensional error state replaces $\mathbf{q}$ with a minimal rotation vector $\delta\boldsymbol{\theta}$; the error-state covariance evolves as $\mathbf{P}_k^- = \mathbf{F}_k \mathbf{P}_{k-1} \mathbf{F}_k^\top + \mathbf{Q}_k$. Innovation gating rejects outliers, and bounded update clamping preserves stability when measurement quality degrades.

PnP-Derived Measurement Covariance

Rather than a fixed measurement noise, $\mathbf{R}_k$ is derived from the PnP solution geometry. For N correspondences with reprojection residuals $\mathbf{e}_i$ [9], the maximum-likelihood estimate gives

$$\hat{\sigma}^2_{\text{pix}} = \frac{\|\mathbf{e}\|^2}{2N - 6}, \qquad \boldsymbol{\Sigma}_{\boldsymbol{\xi}} = \hat{\sigma}^2_{\text{pix}} \, (\mathbf{J}^\top \mathbf{J})^{-1},$$

so $\mathbf{R}_k = \text{blkdiag}(\boldsymbol{\Sigma}_{t,k}, \boldsymbol{\Sigma}_{\theta,k})$ adapts per frame to each pose's geometric quality, down-weighting ambiguous configurations such as the monocular depth-scale direction.

Filter Tuning via Bayesian Optimization

The MEKF tuning parameters (process-noise scales, update-clamp limits, and reprojection threshold) are optimized with the Optuna framework using the Tree-structured Parzen Estimator [10] (20 random + 1000 optimization trials), minimizing a weighted multi-metric cost:

$$J = w_p \,\text{RMSE}_p + w_\phi \,\text{RMSE}_\phi + w_\omega \,\text{RMSE}_\omega + w_v \,\text{RMSE}_v.$$

The weights $(w_p, w_\phi, w_\omega, w_v) = (1.0,\ 0.21,\ 5.0,\ 333.33)$ normalize each term to a comparable contribution given the differing units and magnitudes of position (m), attitude (deg), angular rate (deg/s), and velocity (m/s).

Table 1. MEKF parameters and Bayesian-optimization search ranges.

Parameter	Search range	Used in
σ_v (vel. proc. noise)	$[10^{-6},\ 10^{-3}]$	Both
σ_θ (att. proc. noise)	$[10^{-4},\ 0.05]$	Both
σ_ω (rate proc. noise)	$[10^{-7},\ 5\times10^{-3}]$	Both
$\delta\theta_{\max}$	$[1,\ 30]°$	Both
$\delta r_{\max}$	$[0.005,\ 5]$ m	Both
$\delta v_{\max}$	$[5\times10^{-4},\ 0.5]$ m/s	Both
$\delta\omega_{\max}$	$[5\times10^{-4},\ 0.5]$ rad/s	Both
$\epsilon_{\text{reproj,max}}$	$[5,\ 100]$ px	Both
$\sigma_{\theta,\text{fix}}$	$[0.1,\ 10]°$	Fixed $\mathbf{R}$
$\sigma_{r,\text{fix}}$	$[0.001,\ 1]$ m	Fixed $\mathbf{R}$
k_Σ (cov. scale)	$[0.5,\ 20]$	PnP Cov.
$\sigma_{t,\min}/\sigma_{t,\max}$	$[0.005, 0.1]/[0.2, 2]$ m	PnP Cov.
$\sigma_{\theta,\min}/\sigma_{\theta,\max}$	$[0.1, 2]/[2, 30]°$	PnP Cov.

System Architecture & Angular-Rate Tracking

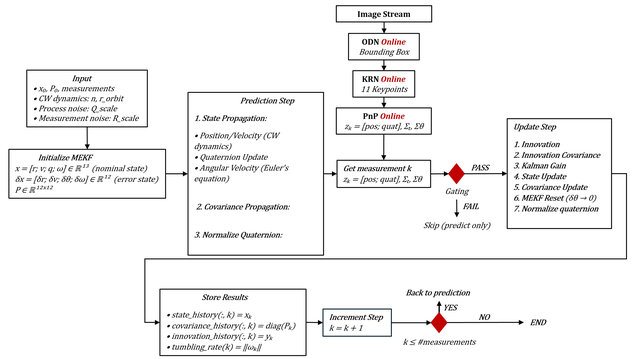

Figure 2. ODN+KRN+PnP pose measurements with geometry-derived covariance are fused with CW dynamics; innovation gating rejects outliers (failed frames fall back to a predict-only step).

Each frame is processed by a three-stage vision front-end: ODN localizes the target, KRN regresses semantic keypoints, and PnP solves for the relative pose. Crucially, the PnP residual geometry yields a per-frame measurement covariance $\mathbf{R}_k$ that reflects keypoint quality, target range, and viewing angle. The MEKF predict–update loop (Fig. 1) propagates state through CW dynamics and corrects with the DL pose, gated by a Mahalanobis innovation test to suppress catastrophic outliers. The result is a recovered 6-DoF state position, attitude, linear velocity, and the angular rates from a purely pose-based measurement stream, with no IMU or rate gyro in the loop.

Estimation Accuracy: Pose & Motion

The MEKF with PnP-derived covariance and clamping reduces position RMSE $\sim 26\times$ and attitude RMSE $\sim 4.4\times$ versus raw DL pose (Table 2); velocity and angular rate—unavailable via finite differencing—are recovered to 0.002 m/s and $0.27°$/s. The contribution is clearest *without* clamping (Fig. 3): fixed $\mathbf{R}$ suffers excursions up to 19 cm and $> 100°$, while PnP-derived $\mathbf{R}_k$ suppresses them automatically—intrinsic outlier robustness, no manual safety mechanism. Angular-rate tracking remains accurate throughout (Fig. 4).

Table 2. Pose and motion estimation performance. Baseline row: raw DL pose (translation/attitude) and finite differencing (velocity/angular rate).

		Trans. (m)		Att. (deg)		Vel. (cm/s)		ω (deg/s)	
Method	Clamp	Mean	RMSE	Mean	RMSE	$\bar{e}$	RMSE	$\bar{e}$	RMSE
Raw DL / Finite Diff.	–	1.095	6.963	16.37	34.35	28.8	167.5	3.275	7.358
MEKF (Fixed $\mathbf{R}$)	Yes	**0.205**	0.272	7.01	8.53	0.158	0.189	0.256	0.286
MEKF (Fixed $\mathbf{R}$)	No	1.085	2.258	15.21	22.10	0.497	0.838	0.440	0.517
MEKF (PnP Cov.)	Yes	0.208	**0.271**	**6.64**	**7.84**	**0.151**	**0.183**	**0.245**	**0.274**
MEKF (PnP Cov.)	No	0.210	0.280	7.30	9.42	0.219	0.400	0.270	0.318

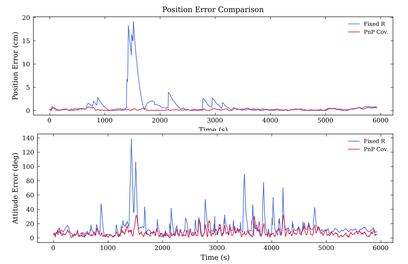

Figure 3. Position (top) and attitude (bottom) error, MEKF *without* update clamping. Geometry-derived $\mathbf{R}_k$ suppresses the excursions that fixed $\mathbf{R}$ leaves untreated.

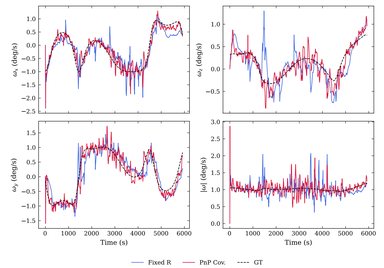

Figure 4. Angular-rate estimates (Fixed $\mathbf{R}$, PnP Cov., GT) recovered from *pose-only* data without visible drift across the full 5930 s trajectory.

Conclusion

- Pose-only DL measurements are elevated to full 6-DoF motion (velocity + angular rate) via multiplicative Kalman filtering.
- The PnP-reprojection covariance adapts the trust per frame, giving outlier robustness with less reliance on manual safety tuning.
- This measurement-adaptive framework is distinct from innovation-based adaptive filtering: $\mathbf{R}_k$ is set by the geometric conditioning of the PnP problem, not by filter residuals.
- *Future work:* learned covariance from keypoint detection heatmaps.

References

[1] D. J. Kessler and B. G. Cour-Palais, "Collision frequency of artificial satellites: The creation of a debris belt," *Journal of Geophysical Research: Space Physics*, vol. 83, no. A6, pp. 2637–2646, 6 1978.

[2] L. P. Cassinis, "Monocular Vision-Based Pose Estimation of Uncooperative Spacecraft," Ph.D. dissertation, Delft University of Technology, 2022. [Online]. Available: https://doi.org/10.4233/uuid:27dcbbc2-7d9e-4f67-925a-5e676ca4e43chttps://repository.tudelft.nl/islandora/object/uuid%3A27dcbbc2-7d9e-4f67-925a-5e676ca4e43c

[3] T. H. Park and S. D'Amico, "Adaptive Neural-Network-Based Unscented Kalman Filter for Robust Pose Tracking of Noncooperative Spacecraft," *Journal of Guidance, Control, and Dynamics*, vol. 46, no. 9, pp. 1671–1688, 6 2023.

[4] B. Candan and S. Servadio, "Relative Pose Estimation of an Uncooperative Target with Camera Marker Detection," *Aerospace*, vol. 12, no. 5, 5 2025.

[5] J. Song, D. Rondao, and N. Aouf, "Deep learning-based spacecraft relative navigation methods: A survey," *Acta Astronautica*, vol. 191, pp. 22–40, 8 2022. [Online]. Available: http://arxiv.org/abs/2108.08876http://dx.doi.org/10.1016/j.actaastro.2021.10.025

[6] Z. Ahmed, H. Park, A. Bhattacharjee, R. Fazel-Rezai, R. Graves, O. Saarela, R. T. ||, K. Vemulapalli, and S. D'amico, "SPEED-UE-CUBE: A MACHINE LEARNING DATASET FOR AUTONOMOUS, VISION-BASED SPACECRAFT NAVIGATION," in *46th Rocky Mountain AAS Guidance, Navigation and Control Conference*, 2024, pp. 24–027. [Online]. Available: https://github.com/tpark94/speed-ue-cube-baseline

[7] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," 4 2020. [Online]. Available: http://arxiv.org/abs/2004.10934

[8] F. Landis Markley and J. L. Crassidis, *Fundamentals of Spacecraft Attitude Determination and Control*. Springer, 2014, vol. 33. [Online]. Available: http://www.springer.com/series/6575

[9] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision* . Cambridge University Press, 2004, vol. 2.

[10] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 7 2019, pp. 2623–2631.

mesut.bostanci@metu.edu.tr