# Uncertainty-Aware Spatial Perception for Generalizable State Estimation

Author Names Omitted for Anonymous Review.

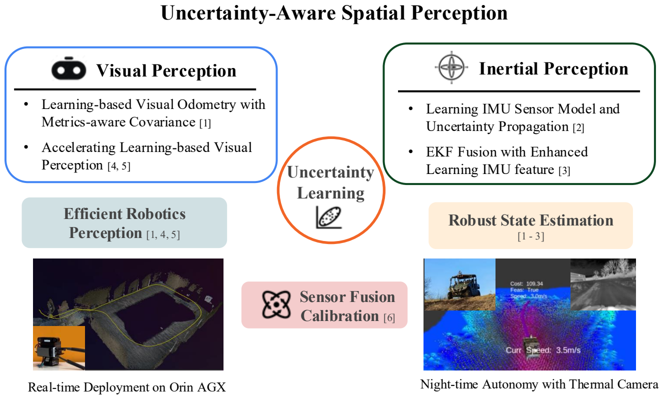
Uncertainty-Aware Spatial Perception
Visual Perception
• Learning-based Visual Odometry with Metrics-aware Covariance [1]
• Accelerating Learning-based Visual Perception [4, 5]
Inertial Perception
• Learning IMU Sensor Model and Uncertainty Propagation [2]
• EKF Fusion with Enhanced Learning IMU feature [3]
Uncertainty Learning
Efficient Robotics Perception [1, 4, 5]
Robust State Estimation [1 - 3]
Sensor Fusion Calibration [6]
Cost: 109.34
Feas: True
Speed: 3.0m/s
Curr Speed: 3.5m/s
Real-time Deployment on Orin AGX
Night-time Autonomy with Thermal Camera

Fig. 1. Overview of my research on uncertainty-aware spatial perception for generalizable state estimation. Learned uncertainty from vision and inertial sensing supports robust state estimation, efficient deployment, and principled sensor fusion and calibration.

## I. Introduction

Robots need spatial perception that remains reliable across platforms, sensors, and operating conditions, yet achieving this reliability is difficult because both the environment and the robot itself are sources of variability. On the environment side, lighting shifts, scenes contain dynamic objects and textureless regions, and fast motion introduces blur and occlusions [13, 12]. On the robot side, sensor characteristics vary across cameras and IMUs, vibration and temperature change noise statistics, and calibration parameters can drift over time [11, 10]. As a consequence, most visual SLAM and inertial navigation systems break when they leave the regimes they were calibrated for [8, 13], blocking the goal of spatial perception that works for **any robot, anywhere, anytime without manual tuning**.

I argue that the limiting factor is not only models or data, but also how systems learn and calibrate uncertainty. Classical pipelines rely on hand-tuned, fixed covariance matrices that assume stationary Gaussian noise [8, 13, 12], and because these covariances do not adapt to context, they become miscalibrated: overconfident under lighting changes, texture dropouts, or platform-dependent IMU vibration, and underconfident in benign conditions. When two sensors are fused with miscalibrated weights, the estimator either overtrusts a degraded modality or ignores a reliable one, producing state estimates that are worse than either sensor alone. Deployment on a new robot therefore requires human retuning; sensor fusion inherits brittle weights that degrade estimation; and downstream tasks such as planning and control receive unreliable state estimates.

Learning-based odometry has improved accuracy [14], but it typically outputs scale-agnostic confidence weights rather than calibrated, metric covariances, which prevents principled fusion and weakens failure detection. Generic efficiency methods such as post-training quantization [9] likewise rarely exploit the geometry and uncertainty structure of perception pipelines.

My research develops *learned uncertainty* as a unifying principle for generalizable, robust, and efficient spatial perception from vision and inertial sensing. Rather than treating covariance as a fixed parameter, I learn it from data as a metrics-aware, calibrated belief that transfers across cameras, IMUs, platforms, and conditions. This learned uncertainty enables three coupled capabilities: principled multi-modal fusion via calibrated weighting, robustness via explicit reliability estimation, and efficient deployment via uncertainty-guided compute allocation. Figure 1 illustrates how these capabilities connect across the vision and inertial sensing pipeline.

## II. Past and Current Research

My research asks: how can robots learn metrics-aware uncertainty that transfers across sensors and platforms, and then use it to make state estimation robust, fuseable, and efficient? To answer this question, I first investigate how to learn visual covariance from features so that a visual odometry system can select correspondences and weight residuals according to predicted reliability. I then extend the same idea to inertial sensing, learning uncertainty propagation and observability-aware representations that generalize across unseen IMUs and trajectories. With calibrated uncertainties from both modalities verified to be metrics-aware, I design tuning-free visual-inertial initialization and calibration that fuse vision and inertial constraints robustly across changing environments and sensor configurations. Finally, I turn uncertainty into an efficiency signal for deployment by using it to guide quantization of visual odometry models and token merging for geometry transformers on edge hardware.

### Metrics-Aware Covariance for Stereo Visual Odometry.

I developed a stereo visual odometry system [4] that learns metrics-aware covariance from visual features, replacing hand-tuned geometric noise models. We train the uncertainty model on a large-scale synthetic dataset [1] so that the learned covariance captures metric scale, and the model demonstrates sim-to-real generalizability to unseen real cameras. The key idea is to predict 2D matching uncertainty from appearance, context, and occlusion cues, then lift it through stereo geometry into a full 3D covariance that includes inter-axis correlations. This covariance serves two roles: filtering unreliable correspondences and weighting residuals in pose optimization, enabling strong cross-camera generalization without manual tuning.

**Data-Driven Inertial Odometry and Uncertainty Propagation.** IMU preintegration typically uses fixed white-noise parameters that fail to capture platform vibration, bias drift, and motion-dependent stochasticity. To address this, I developed a data-driven framework [2] that learns residual IMU correction and uncertainty propagation within a differentiable preintegration pipeline, preserving the classical kinematic structure while learning the parts that break under real hardware and dynamics. The learned uncertainty is verified to be metrics-aware, meaning the predicted covariance matches the empirical error distribution in metric units, which is essential for principled downstream fusion. Building on this uncertainty-aware preintegration, I developed an inertial odometry system [3] that targets a complementary bottleneck: limited observability of attitude under agile motion. Through feature analysis (PCA and t-SNE), I show that preserving IMU data in the body frame yields a more expressive and compact representation than the conventional global-frame transformation, because attitude couples linearly with gravitational acceleration in the body frame. Based on this finding, AirIO explicitly encodes attitude information alongside body-frame IMU features, and fuses the resulting velocity predictions with AirIMU-based preintegration in an EKF where both modules carry learned uncertainty. The system outperforms prior inertial odometry methods on UAV datasets without requiring thrust commands or additional sensors, and generalizes to unseen trajectories.

**Robust Visual-Inertial Initialization and Online Calibration.** Traditional visual-inertial initializers often fail under exposure changes, low texture, or dynamic objects because fixed noise parameters cannot reflect measurement quality. I developed a robust initialization and online calibration system [6] that uses metrics-aware uncertainty from both vision and inertial modules to overcome this limitation. The system derives visual pose covariance from learned feature-matching uncertainty and inertial covariance from learned preintegration, then weights the two sets of constraints according to their predicted reliability to produce a principled, stable initialization. Because each constraint carries a calibrated covariance, the optimizer naturally down-weights degraded measurements without requiring per-scenario heuristics or outlier thresholds. This removes per-platform parameter adjustment and enables tuning-free initialization across diverse environments and sensor configurations.

**Compute-Efficient Perception Deployment.** Learned perception models often exceed the compute budget of edge devices carried by real robots, so I use uncertainty to decide where approximation is safe and where fidelity matters. First, I developed a full-stack quantization framework for learning-based visual odometry with metrics-aware covariance [7] that combines a sim-to-real calibration strategy for zero-shot generalization with a kurtosis-guided hybrid scheme that applies expensive operations only where outliers dominate, yielding a $1.83\times$ speedup on embedded GPU hardware with negligible accuracy loss. The framework quantizes both the feature extractor and the covariance head jointly, ensuring that the compressed model retains calibrated uncertainty. Second, I introduced uncertainty-guided token merging for visual geometry transformers [5], where a lightweight predictor ranks tokens by uncertainty and selectively merges uninformative ones, achieving up to $11.3\times$ speedup for dense 3D reconstruction and $7.2\times$ for semantic mapping on edge hardware while preserving quality in geometrically informative regions.

## III. Future Work

Building on learned uncertainty for odometry and fusion, my future research extends the core idea into three directions: efficient deployment, semantic memory, and resilience in extreme environments.

**Efficient Robotic Spatial Perception.** Spatial perception must run on the robots themselves, and our compression results show that learned models can be accelerated with minimal accuracy loss. A deeper challenge, however, is to co-design uncertainty estimation with hardware-aware architectures so that the compute budget adapts online to predicted reliability and scene complexity. I aim to build perception systems that allocate more computation to uncertain regions and less to confident ones, enabling always-on spatial perception on power-constrained platforms from handheld devices to aerial robots.

**Spatial Memory with Semantic Understanding.** Current SLAM systems build geometric maps that often lack the semantics needed for long-horizon tasks such as object retrieval, task planning, and human-robot collaboration. I plan to extend learned uncertainty into a spatial memory that jointly encodes geometry, semantics, and language-grounded embeddings, supporting queries such as “where did I last see the red toolbox?” and enabling robots to reason about scene changes over time. By integrating vision-language representations with uncertainty-aware mapping, robots can plan and interact more reliably across repeated missions in evolving environments, using uncertainty to identify stale or unreliable map regions that need re-observation.

**Robustness in Extreme Environments.** Spatial perception research often targets urban and indoor settings, but high-impact applications lie in extreme and resource-limited domains such as subterranean rescue, deep-sea exploration, and planetary operations, where degraded GPS, limited communication, harsh lighting, and strict power budgets co-occur. I aim to use learned uncertainty to make systems degrade gracefully by adapting fusion weights and compute allocation online: when a sensor modality becomes unreliable, the system should automatically re-weight or disable it, maintaining safe operation even under severe distribution shift.

Together, these directions advance spatial perception that is accurate, deployable, and resilient for real-world robots.

## References

[1] Anonymous authors. Title omitted for anonymous review. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020.

[2] Anonymous authors. Title omitted for anonymous review, 2024. arXiv preprint (details omitted for anonymous review).

[3] Anonymous authors. Title omitted for anonymous review. *IEEE Robotics and Automation Letters*, 2025. Anonymized preprint information omitted for anonymous review.

[4] Anonymous authors. Title omitted for anonymous review. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025.

[5] Anonymous authors. Title omitted for anonymous review, 2025. arXiv preprint (details omitted for anonymous review).

[6] Anonymous authors. Title omitted for anonymous review, 2025. In submission.

[7] Anonymous authors. Title omitted for anonymous review, 2025. In submission.

[8] Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation. In *Robotics: Science and Systems (RSS)*, 2015.

[9] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

[10] Paul Furgale, Joern Rehder, and Roland Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1280–1286. IEEE, 2013.

[11] Patrick Geneva, Kevin Eckenhoff, Woosik Lee, Yulin Yang, and Guoquan Huang. Openvins: A research platform for visual-inertial estimation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4666–4672. IEEE, 2020.

[12] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[13] Tong Qin, Peiliang Li, and Shaojie Shen. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. doi: 10.1109/TRO.2018.2853729.

[14] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. *Advances in Neural Information Processing Systems*, 34: 16558–16569, 2021.