Learning Uncertainty for Generalizable Vision–Inertial Spatial Perception

Author Names Omitted for Anonymous Review.

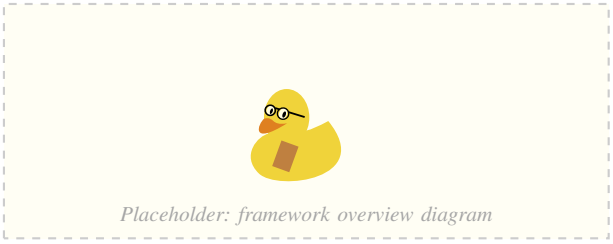

Fig. 1. Learned uncertainty as the unifying principle for vision–inertial spatial perception. Metrics-aware covariance learned from visual and inertial data drives principled fusion, cross-platform generalization, and compute-efficient deployment, advancing toward spatial perception for any robot, anywhere.

I. Introduction

Autonomous robots must perceive and localize reliably across diverse platforms, sensors, and operating conditions—from handheld devices to aerial vehicles, indoors to outdoors, day to night. The ultimate goal of spatial perception is a system that works for *any robot, anywhere*, without manual tuning. Yet despite decades of progress in simultaneous localization and mapping (SLAM) and visual-inertial odometry (VIO), current systems remain brittle when deployed beyond their calibrated settings [7, 9].

A central cause of this brittleness is the treatment of measurement uncertainty. Classical SLAM pipelines model sensor noise with hand-tuned, fixed covariance matrices that assume stationary Gaussian distributions [7, 9]. These parameters cannot capture environment-dependent sensor reliability—a stereo camera calibrated indoors produces overconfident estimates in outdoor lighting, and an IMU noise model tuned on a ground vehicle fails on an aerial platform with different vibration characteristics. Consequently, deploying a perception system on a new robot or in a new environment requires extensive per-platform recalibration by domain experts, a process that does not scale to the diversity of robots and conditions encountered in real-world autonomy.

Learning-based odometry methods have demonstrated strong accuracy in recent years [10], yet they typically produce scale-agnostic confidence weights that do not reflect actual estimation error in metric space. This lack of calibrated, metrics-aware uncertainty limits their integration into multi-sensor fusion, prevents reliable failure detection, and makes downstream planning and control fragile. Meanwhile, efficiency-oriented quantization methods [8] target general network acceleration but do not account for the geometry-aware structure unique to perception pipelines.

My research addresses these limitations by developing **learned uncertainty** as the unifying principle for generalizable, robust, and efficient vision–inertial spatial perception. Rather than treating uncertainty 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 tightly coupled capabilities: (i) principled multi-modal fusion through calibrated measurement weighting, (ii) robustness via explicit knowledge of when and where perception is reliable, and (iii) system acceleration through uncertainty-guided, compute-efficient inference that allocates resources where they matter most.

II. Past and Current Research

My research develops learned uncertainty for each sensing modality, accelerates perception for edge deployment, and fuses the components into a robust visual-inertial system. Concretely, I first learn metrics-aware visual uncertainty to select correspondences and weight optimization without hand tuning. I then learn inertial uncertainty propagation and observability-aware inertial representations that transfer across IMUs and platforms. To make these models usable on real robots, I develop uncertainty-guided compression and acceleration for edge hardware. Finally, I bring the learned components together in tuning-free visual-inertial initialization and calibration, using uncertainty as the common currency that makes fusion principled and reliable.

A. Metrics-Aware Stereo Visual Odometry

We developed a stereo visual odometry system [3] that learns to predict metrics-aware covariance directly from visual features, replacing hand-tuned geometric noise models. Unlike the scale-agnostic diagonal confidence weights used by prior learning-based methods [10], our system introduces two innovations: a learning-based model that quantifies 2D matching uncertainty by reasoning over appearance, context, and occlusion cues, and a novel covariance propagation scheme that lifts these 2D uncertainties into full 3D covariance (including inter-axis correlations) via the stereo geometry. The resulting metrics-aware covariance serves a dual role: it drives keypoint selection by filtering unreliable correspondences and weights residuals in the two-frame pose graph optimization proportional to each measurement's actual reliability. Without any manual tuning, this two-frame system outperforms existing VO algorithms and even multi-frame SLAM systems on public benchmarks across diverse cameras and conditions.

B. Data-Driven Inertial Odometry and Uncertainty Propagation

IMU preintegration traditionally relies on fixed white-noise parameters that cannot capture platform-specific vibration, bias drift, or motion-dependent stochastic errors. We developed a data-driven framework [1] that jointly learns to correct raw IMU measurements and propagate the resulting uncertainty through a fully differentiable preintegration pipeline. Rather than replacing the kinematic model with a black-box network, the approach preserves the classical preintegration structure while learning a residual noise correction and a covariance propagation function that captures platform- and motion-specific noise patterns. Building on this, we proposed an inertial odometry system [2] that addresses a complementary challenge: limited observability of certain motion states under constrained dynamics. The system enhances IMU feature representations to improve state observability and combines learned uncertainty with robust estimation, achieving significantly reduced drift in long-trajectory, GPS-denied environments such as subterranean and aerial operations.

C. Compute-Efficient Perception Deployment

Learned perception models impose heavy computational demands that limit deployment on edge devices carried by real robots. We addressed this challenge on two fronts. First, we developed a full-stack quantization framework for learning-based VO with metrics-aware covariance [6]: a sim-to-real calibration strategy on synthetic data enables zero-shot generalization, and a kurtosis-based hybrid quantization scheme selectively applies expensive rotations only to outlier-dominated layers, achieving $1.83\times$ speedup on NVIDIA Jetson Thor with negligible accuracy loss. Second, we introduced confidence-guided token merging [4] for visual geometry transformers, where a lightweight predictor ranks tokens by uncertainty and selectively merges uninformative ones, yielding up to $11.3\times$ speedup for dense 3D reconstruction and $7.2\times$ for semantic mapping on edge hardware.

D. Tuning-Free Visual-Inertial Initialization

We designed a robust VI initialization and online calibration system [5] that leverages metrics-aware uncertainty from both the visual and inertial modules. Traditional VI initialization fails under extreme exposure, textureless scenes, or dynamic objects because it relies on manually set noise parameters that cannot reflect actual measurement quality. By deriving visual pose covariance from learned feature-matching uncertainty and inertial covariance from the learned preintegration model, our system formulates a principled initialization that weights visual and inertial constraints according to their true reliability. This removes the manual parameter adjustment that typically accompanies each new platform, achieving tuning-free initialization across diverse environments and sensor configurations.

III. Future Work

Building on learned uncertainty for odometry and fusion, my future research pursues three directions that broaden the scope from estimation to deployment, memory, and resilience.

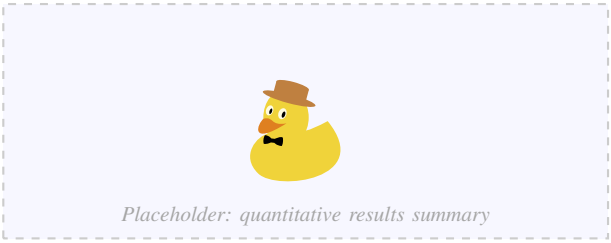

Fig. 2. Learned uncertainty improves spatial perception across modalities and platforms. (a) Metrics-aware visual covariance yields robust odometry on diverse stereo sequences. (b) Learned IMU uncertainty reduces drift in GPS-denied environments. (c) Uncertainty-guided quantization and token merging enable real-time deployment on edge hardware.

Efficient Embodied Deployment. Spatial perception must ultimately run on the robots themselves. While our quantization and token merging work demonstrates that learned models can be compressed with minimal accuracy loss, a deeper challenge remains: co-designing uncertainty estimation with hardware-aware architectures so that perception adapts its compute budget in real time based on scene complexity. I aim to develop uncertainty-aware perception systems that automatically trade off fidelity and latency, enabling always-on spatial AI on power-constrained embodied platforms, from handheld devices to aerial robots.

Spatial Memory with Semantic Understanding. Current SLAM systems build geometric maps that lack the semantic richness required for long-horizon tasks. I plan to extend learned uncertainty into a spatial memory that jointly encodes geometry, semantics, and language-grounded embeddings, enabling robots to answer queries such as "where did I last see the red toolbox?" and to reason about scene changes over time. By integrating vision-language representations with uncertainty-aware mapping, this spatial memory will support task planning, object retrieval, and human-robot interaction across repeated missions in evolving environments.

Robustness in Extreme Environments. Spatial perception research has largely focused on urban and indoor settings, yet some of the most impactful applications lie in extreme and resource-limited domains: subterranean rescue, deep-sea exploration, and planetary surface operations. These environments impose severe constraints, including degraded or absent GPS, limited communication, harsh lighting, and strict power budgets. I aim to leverage learned uncertainty as the mechanism that allows perception systems to degrade gracefully rather than fail catastrophically, adapting sensor fusion strategies and compute allocation to maintain safe operation when conditions push beyond the training distribution.

Together, these directions advance toward spatial perception that is not only accurate but also deployable, semantically rich, and resilient—equipping robots to operate reliably in the environments that matter most.

References

[1] Anonymous authors. AirIMU: Learning uncertainty propagation for inertial odometry, 2024. URL https:

//arxiv.org/abs/2310.04874.

[2] Anonymous authors. AirIO: Learning inertial odometry with enhanced IMU feature observability. *IEEE Robotics and Automation Letters*, 2025. URL https://arxiv.org/abs/2501.15659.

[3] Anonymous authors. MAC-VO: Metrics-aware covariance for learning-based stereo visual odometry. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3803–3814. IEEE, 2025. URL https://arxiv.org/abs/2409.09479.

[4] Anonymous authors. Confidence-guided token merging for visual geometric transformers, 2025.

[5] Anonymous authors. Learned metrics-aware covariance for robust visual-inertial initialization and calibration, 2025. In submission.

[6] Anonymous authors. Quantization for learning-based visual odometry with metrics-aware covariance, 2025. In submission.

[7] 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.

[8] 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.

[9] 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.

[10] 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.