

Statistical Uncertainty Learning for Robust Visual-Inertial State Estimation

Seungwon Choi, Dong-Gyu Park, Seo-Yeon Hwang, Tae-Wan Kim

Abstract—A fundamental challenge in robust visual-inertial odometry (VIO) is to dynamically assess the reliability of sensor measurements. This assessment is crucial for properly weighting the contribution of each measurement to the state estimate. Conventional methods often simplify this by assuming a static, uniform uncertainty for all measurements. This heuristic, however, may be limited in its ability to capture the dynamic error characteristics inherent in real-world data. To improve this limitation, we present a statistical framework that learns measurement reliability assessment online, directly from sensor data and optimization results. Our approach leverages multi-view geometric consistency as a form of self-supervision. This enables the system to infer landmark uncertainty and adaptively weight visual measurements during optimization. We evaluated our method on the public EuRoC dataset, demonstrating improvements in tracking accuracy with average reductions of approximately 24% in translation error and 42% in rotation error compared to baseline methods with fixed uncertainty parameters. The resulting framework operates in real time while showing enhanced accuracy and robustness. To facilitate reproducibility and encourage further research, the source code will be made publicly available.

I. INTRODUCTION

State estimation is a fundamental research topic in robotics, given its indispensable role in the navigation of autonomous systems[1]. Among the various sensor modalities, the fusion of visual and inertial measurements has become a popular approach for state estimation [2], [3], primarily due to the complementary nature of cameras and Inertial Measurement Units (IMUs). While cameras provide rich geometric information about the environment, IMUs offer high-frequency motion measurements that are robust to visual degradation. This combination allows for robust tracking in a wide range of environments, from well-lit indoor spaces to challenging outdoor scenarios [4].

The problem is typically formulated as a non-linear least squares (NLS) optimization within a sliding window framework [5], where camera poses, 3D landmarks, and sensor biases are jointly estimated. Modern visual-inertial odometry (VIO) systems often employ factor graph optimization [6] to fuse heterogeneous sensor measurements, with each observation weighted by its assumed uncertainty. However, one potential limitation in many approaches is the reliance on static noise models, which may not fully capture the dynamic nature of measurement uncertainty [7].

S. Choi, D. Park, and S.-Y. Hwang are with the Research Center, Clobot Inc., Seongnam, Republic of Korea. (e-mail: eugene@clobot.co.kr, derrick@clobot.co.kr, ian@clobot.co.kr)

T.-W. Kim is with the Department of Naval Architecture and Ocean Engineering, Seoul National University (SNU), Seoul, Republic of Korea (e-mail: taewan@snu.ac.kr)

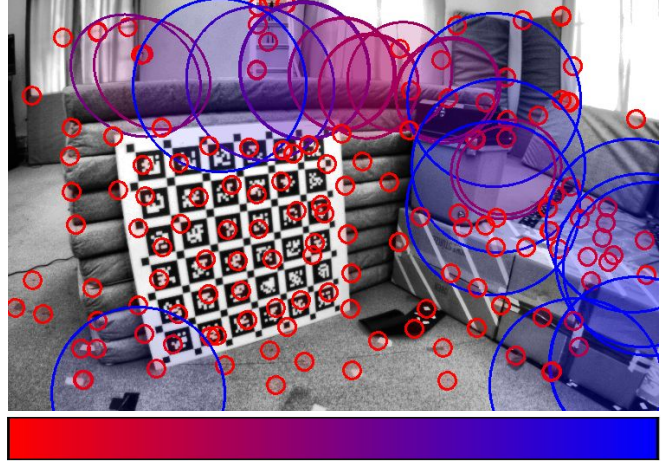


Fig. 1. Visualization of learned uncertainty estimates for visual landmarks. The ellipses represent the projected 2D covariance of 3D landmark uncertainties onto the image plane, illustrating the spatial distribution of measurement reliability. Color coding indicates uncertainty magnitude: red ellipses correspond to lower uncertainties (higher reliability), while blue ellipses indicate higher uncertainties (lower reliability). This adaptive uncertainty modeling enables observation-specific weighting in bundle adjustment optimization.

Consider the inherent variability in visual measurements: a feature tracked in well-textured, well-lit regions exhibits different error characteristics compared to one tracked near occlusion boundaries or in poorly illuminated areas. Similarly, the quality of stereo depth estimates can vary with baseline geometry and scene depth [8]. Traditional VIO frameworks often address this heterogeneity through uniform information matrices. This treatment of all visual observations as equally reliable is a simplifying assumption that may not always account for the rich statistical structure inherent in the measurements.

The effects of this limitation can become particularly pronounced in challenging scenarios. When a system encounters rapid motion, varying lighting conditions, or textureless environments, a mismatch between the assumed and actual measurement uncertainties can contribute to less accurate estimates, convergence issues, or tracking failures [9]. As autonomous systems operate in increasingly diverse environments, the ability to adapt measurement uncertainty models online could be crucial for maintaining consistent performance.

To explore this area, this work proposes a **statistical uncertainty learning** approach for visual-inertial odometry. Our framework is designed as an alternative to static noise

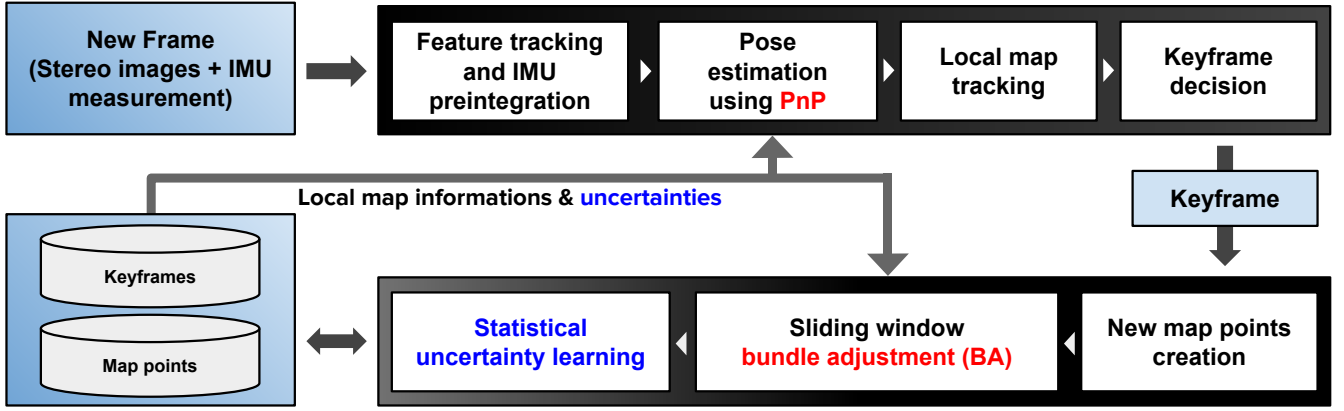


Fig. 2. Overview of the proposed visual-inertial system architecture featuring a dual-pipeline design. The upper pipeline handles real-time pose tracking via Perspective-n-Point (PnP) optimization, while the lower pipeline performs map refinement through sliding window bundle adjustment (BA) integrated with our statistical uncertainty learning module. The learned uncertainty estimates are propagated back to the tracking pipeline, establishing a feedback mechanism that enhances system performance.

models, aiming to statistically learn and propagate measurement uncertainties throughout the estimation pipeline. By leveraging multi-view geometric consistency and covariance analysis, we investigate whether a system can better estimate the error characteristics of its sensor measurements. Figure 1 illustrates our approach to modeling measurement variability through adaptive covariance representations.

Our system employs a dual-pipeline architecture that integrates statistical uncertainty learning with real-time VIO estimation. The upper pipeline handles real-time pose tracking via Perspective-n-Point (PnP) optimization, while the lower pipeline performs map refinement through sliding window bundle adjustment integrated with our statistical uncertainty learning module. The learned uncertainty estimates are propagated back to the tracking pipeline, establishing a feedback mechanism that enhances system performance. Figure 2 provides an overview of this dual-pipeline architecture.

Specifically, the key contributions of our proposed framework include:

- **Statistical multi-view uncertainty learning:** We propose a method to estimate 3D landmark uncertainty by analyzing the distribution of triangulated positions across multiple viewpoints. This approach aims to provide a data-driven measure of geometric consistency and measurement reliability.
- **Adaptive statistical weighting:** We investigate transforming learned 3D world uncertainties to 2D pixel-space information matrices through uncertainty propagation. This enables observation-specific weighting that may better reflect the statistical reliability of each measurement.
- **Open-source statistical VIO framework:** We present a complete, open-source implementation of our approach, which integrates with existing optimization frameworks while maintaining computational efficiency suitable for real-time applications.

Through evaluation on the EuRoC dataset [10], we demonstrate that our statistical learning approach achieves improved

accuracy and robustness compared to baseline methods using static noise models across various conditions. The proposed method represents a step toward more adaptive and statistically-grounded visual-inertial state estimation.

II. RELATED WORK

Visual-inertial odometry (VIO) has become a cornerstone of autonomous robot navigation and state estimation [11], [12], with state-of-the-art systems broadly categorized into filtering-based and optimization-based methods [7]. A critical component within these systems is the modeling of measurement uncertainty, which directly influences the weighting of visual observations in bundle adjustment and ultimately determines the accuracy and robustness of the state estimate. This section reviews how prominent VIO frameworks handle visual measurement uncertainty and highlights the limitations of current static approaches.

Early influential filtering-based methods such as MSCKF [13] established the foundation for tightly-coupled visual-inertial fusion. In MSCKF, the covariance of visual feature measurements is typically assumed to be a fixed identity matrix, implying that all feature observations are treated with equal confidence regardless of their actual reliability. This simplification, while computationally efficient, overlooks significant variations in feature quality, lighting conditions, and tracking ambiguity that naturally occur in real-world scenarios. Subsequent filtering-based systems like ROVIO [14] and OpenVINS [15] perpetuate this limitation by employing constant and uniform noise models for image feature re-projection errors. These models are typically determined through empirical tuning on specific datasets, making them poorly suited for diverse operating conditions encountered in practical applications.

Optimization-based approaches, which have gained prominence due to their superior handling of nonlinear constraints [16], often retain similarly simplified uncertainty models. VINS-Mono [11], one of the most widely adopted VIO systems, models the re-projection error of visual features

using unit covariance matrices in its bundle adjustment formulation. Specifically, the information matrix for each feature observation is set to $\Omega = \mathbf{I}$, effectively assigning uniform weight to every measurement regardless of factors such as illumination quality, feature distinctiveness, or geometric configuration. Other notable systems such as direct-methods [17], [18] and semi-direct methods [19] also typically rely on simplified photometric error models.

Some systems have made modest progress toward adaptive uncertainty modeling. The ORB-SLAM family [20], [21], [12] introduces scale-dependent weighting based on the pyramid level at which features are detected. The information matrix is scaled by σ^{-2} , where σ represents the scale factor of the detection level. While this approach acknowledges that features detected at different scales have varying precision, it remains a predefined heuristic rather than a learned model. OKVIS [5] provides infrastructure for more sophisticated uncertainty modeling, with provisions for incorporating feature tracking scores and detection quality metrics. However, in practice, many implementations often rely on simple isotropic uncertainty models, partly due to the challenges in developing robust methods for learning appropriate weightings online.

The predominant reliance on static, hand-tuned noise models represents a fundamental limitation in current VIO systems. These approaches fail to capture the inherent variability in measurement quality that arises from environmental factors such as lighting conditions, weather, and scene texture; geometric factors including viewing angle, distance, and baseline geometry; and dynamic factors like camera motion, exposure settings, and focus changes. The mismatch between assumed and actual measurement uncertainties can lead to overconfident estimates, poor convergence in challenging scenarios, and suboptimal information weighting during optimization.

Recent work in uncertainty quantification has explored learning-based approaches in related domains. In structure from motion, some methods have investigated uncertainty propagation through multi-view geometry [22], while robust estimation techniques have focused on outlier detection and covariance inflation [23]. However, these approaches either assume known uncertainty models or focus on outlier rejection rather than systematic uncertainty learning. The integration of uncertainty learning directly into the VIO optimization loop—where bundle adjustment results provide statistical evidence for refining uncertainty models—remains relatively underexplored, although some works have begun to explore this direction [24]. Deep learning methods have also been proposed to estimate uncertainty directly from images [25], but often require large amounts of training data.

With the goal of improving upon the limitations of static noise models, this work introduces a statistical uncertainty learning framework. Our method learns measurement uncertainties from multi-view geometric consistency by using bundle adjustment results as statistical evidence. This approach allows uncertainty estimates to adapt to observed data characteristics, aiming for more effective information

weighting in visual-inertial optimization. The core of our contribution is the integration of this uncertainty learning within the bundle adjustment loop, where optimized poses and landmarks provide the empirical evidence to refine uncertainty models for subsequent optimizations. This represents a step toward more adaptive, data-driven uncertainty estimation.

III. VISUAL-INERTIAL SYSTEM OVERVIEW

Our approach builds upon established visual-inertial odometry principles, employing a sliding window-based optimization strategy that combines real-time pose tracking with batch refinement.

A. System Architecture

The visual-inertial system follows a two-stage estimation pipeline: (1) real-time pose estimation using Perspective-n-Point (PnP) optimization for immediate tracking [26], and (2) sliding window bundle adjustment for refined state estimation [5]. This architecture balances computational efficiency with estimation accuracy, making it suitable for real-time applications while maintaining high precision through batch optimization. The final optimization is performed using a modern nonlinear solver like Ceres solver.

B. State Representation

The system state at time k consists of the camera pose, velocity, and IMU biases:

$$\mathbf{x}_k = [\mathbf{T}_{wb,k}, \mathbf{v}_{w,k}, \mathbf{b}_{a,k}, \mathbf{b}_{g,k}]^T \quad (1)$$

where $\mathbf{T}_{wb,k} \in SE(3)$ represents the transformation from body frame to world frame, $\mathbf{v}_{w,k} \in \mathbb{R}^3$ is the velocity in world coordinates, and $\mathbf{b}_{a,k}, \mathbf{b}_{g,k} \in \mathbb{R}^3$ are the accelerometer and gyroscope biases, respectively.

For optimization purposes, the $SE(3)$ pose is parameterized using the Lie algebra representation [27]:

$$\mathbf{T}_{wb} = \exp(\boldsymbol{\xi}^\wedge), \quad \boldsymbol{\xi} = [\boldsymbol{\rho}^T, \boldsymbol{\phi}^T]^T \in \mathbb{R}^6 \quad (2)$$

where $\boldsymbol{\rho} \in \mathbb{R}^3$ and $\boldsymbol{\phi} \in \mathbb{R}^3$ represent the translation and rotation components in the tangent space.

C. Real-time Pose Tracking via PnP

For each incoming frame, we perform real-time pose estimation using the Perspective-n-Point algorithm [26]. Given a set of 3D landmarks $\{\mathbf{p}_j\}$ and their corresponding 2D observations $\{\mathbf{z}_j\}$, the PnP optimization minimizes:

$$\boldsymbol{\xi}^* = \arg \min_{\boldsymbol{\xi}} \sum_j e_j^T \Omega_j e_j \quad (3)$$

where the re-projection error is defined as:

$$e_j = \mathbf{z}_j - \pi(\mathbf{T}_{cb} \exp(\boldsymbol{\xi}^\wedge) \mathbf{p}_j) \quad (4)$$

and $\pi(\cdot)$ represents the camera projection function, $\mathbf{T}_{cb}$ is the body-to-camera transformation, and Ω_j is the information matrix for observation j .

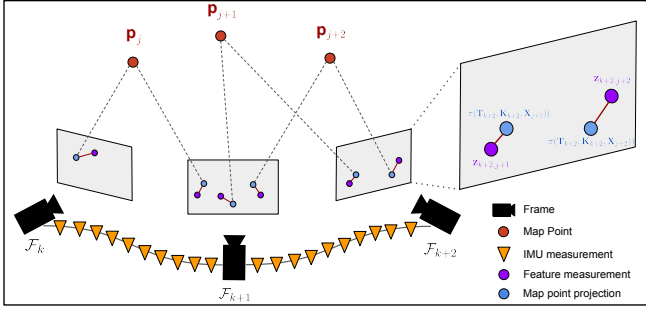


Fig. 3. Sliding window bundle adjustment framework. The system maintains a fixed-size window of keyframes while marginalizing older frames. Visual observations connect 3D landmarks to multiple keyframes, while IMU measurements provide constraints between consecutive poses. The optimization jointly estimates camera poses, landmark positions, and IMU biases within the sliding window.

The projection function for a pinhole camera model is:

$$\pi(\mathbf{p}_c) = \begin{bmatrix} f_x \frac{X_c}{Z_c} + c_x \\ f_y \frac{Y_c}{Z_c} + c_y \end{bmatrix} \quad (5)$$

where $\mathbf{p}_c = [X_c, Y_c, Z_c]^T$ is the 3D point in camera coordinates, and (f_x, f_y, c_x, c_y) are the camera intrinsic parameters.

D. IMU Preintegration

Between consecutive keyframes, IMU measurements are preintegrated to provide motion constraints, following the on-manifold preintegration theory [28]. Given IMU measurements $\{\omega_t, \mathbf{a}_t\}$ over the time interval $[t_i, t_{i+1}]$, we compute the preintegrated quantities:

$$\alpha_{i,i+1} = \iint_{t_i}^{t_{i+1}} \mathbf{R}_t (\mathbf{a}_t - \mathbf{b}_{a,t} - \mathbf{n}_a) dt^2 \quad (6)$$

$$\beta_{i,i+1} = \int_{t_i}^{t_{i+1}} \mathbf{R}_t (\mathbf{a}_t - \mathbf{b}_{a,t} - \mathbf{n}_a) dt \quad (7)$$

$$\gamma_{i,i+1} = \int_{t_i}^{t_{i+1}} \frac{1}{2} (\omega_t - \mathbf{b}_{g,t} - \mathbf{n}_g)^\wedge dt \quad (8)$$

where $\mathbf{R}_t$ is the rotation at time t , and $\mathbf{n}_a, \mathbf{n}_g$ represent IMU noise terms.

The IMU constraint between two consecutive states is formulated as:

$$\mathbf{r}_{IMU} = \begin{bmatrix} \mathbf{R}_i^T (\mathbf{p}_{i+1} - \mathbf{p}_i - \mathbf{v}_i \Delta t - \frac{1}{2} \mathbf{g} \Delta t^2) - \alpha_{i,i+1} \\ \mathbf{R}_i^T (\mathbf{v}_{i+1} - \mathbf{v}_i - \mathbf{g} \Delta t) - \beta_{i,i+1} \\ \log((\gamma_{i,i+1})^{-1} \mathbf{R}_i^T \mathbf{R}_{i+1})^\vee \end{bmatrix} \quad (9)$$

E. Sliding Window Bundle Adjustment

The sliding window optimization jointly refines camera poses, 3D landmarks, velocities, and IMU biases over a fixed-size window of keyframes [5]. The optimization objective combines visual re-projection errors and IMU constraints:

$$\min_{\mathcal{X}, \mathcal{L}} \sum_{(i,j) \in \mathcal{O}} e_{ij}^T \Omega_{ij} e_{ij} + \sum_{k=0}^{N-1} \mathbf{r}_{IMU,k}^T \Sigma_{IMU,k}^{-1} \mathbf{r}_{IMU,k} \quad (10)$$

where $\mathcal{X} = \{\mathbf{x}_0, \dots, \mathbf{x}_N\}$ represents the sliding window states, $\mathcal{L} = \{\mathbf{p}_1, \dots, \mathbf{p}_M\}$ are the 3D landmarks, $\mathcal{O}$ denotes the set of visual observations, and $\Sigma_{IMU,k}$ is the IMU measurement covariance. This optimization framework is illustrated in Figure 3.

IV. STATISTICAL UNCERTAINTY LEARNING

Our statistical uncertainty learning framework extends the traditional VIO pipeline by creating a closed-loop system where bundle adjustment optimization results continuously refine uncertainty models, which in turn improve subsequent optimizations. This section details our adaptive approach that learns from optimization outcomes to enhance measurement reliability estimation.

A. Learning-based Uncertainty Framework Overview

Unlike traditional VIO systems that rely on fixed noise models, our approach implements a statistical learning cycle where bundle adjustment optimization provides empirical evidence for uncertainty refinement, which then improves adaptive weighting in subsequent optimizations, creating a self-improving system. This iterative process ensures that uncertainty estimates continuously adapt to the observed statistical properties of the sensor measurements and environmental conditions.

B. Uncertainty Propagation for Adaptive Weighting

For adaptive information matrix computation in bundle adjustment, we propagate learned 3D world uncertainties to 2D pixel space through a series of transformations. Given a 3D point $\mathbf{p}_{world}$ with learned covariance Σ_{world} , the geometric transformation chain follows:

$$\mathbf{p}_{world} \xrightarrow{\mathbf{T}_{cw}} \mathbf{p}_{cam} \xrightarrow{\pi} \mathbf{z} \quad (11)$$

To propagate uncertainties through this transformation, we first compute the projection Jacobian with respect to the camera coordinates:

$$\mathbf{J}_\pi = \frac{\partial \pi(\mathbf{p}_{cam})}{\partial \mathbf{p}_{cam}} = \begin{bmatrix} \frac{f_x}{Z_c} & 0 & -\frac{f_x X_c}{Z_c^2} \\ 0 & \frac{f_y}{Z_c} & -\frac{f_y Y_c}{Z_c^2} \end{bmatrix} \quad (12)$$

Applying the uncertainty propagation formula through both the rotation and projection transformations yields the pixel-space covariance:

$$\Sigma_{pixel} = \mathbf{J}_\pi \mathbf{R}_{cw} \Sigma_{world} \mathbf{R}_{cw}^T \mathbf{J}_\pi^T \quad (13)$$

To ensure numerical stability during optimization, we construct the adaptive information matrix by inverting the pixel covariance and adding a small regularization term:

$$\Omega_{adaptive} = \Sigma_{pixel}^{-1} + \lambda \mathbf{I} \quad (14)$$

where λ is a small regularization parameter that prevents singular matrices during inversion.

This adaptively weighted information matrix is then incorporated into the bundle adjustment optimization objective,

allowing each observation to contribute according to its learned reliability:

$$\min_{\mathcal{X}, \mathcal{L}} \sum_{(i,j) \in \mathcal{O}} e_{ij}^T \Omega_{adaptive,ij} e_{ij} \quad (15)$$

C. Post-Optimization Statistical Learning

After each bundle adjustment iteration, we extract statistical evidence from the optimized results to refine our uncertainty estimates. The core idea is to leverage the multi-view geometric consistency of optimized landmarks to assess measurement reliability. For each map point, we collect triangulated positions using the newly optimized camera poses and landmark positions, as detailed in Algorithm 1.

Algorithm 1 Statistical Learning from Bundle Adjustment Results

- 1: **Input:** Optimized poses $\{\mathbf{T}_{wb,i}^{opt}\}$, optimized points $\{\mathbf{p}_j^{opt}\}$
 - 2: **for** each MapPoint $\mathbf{p}_j^{opt}$ in optimization window **do**
 - 3: **Initialize:** Position collection $\mathcal{P}_j = \emptyset$
 - 4: **for** each observation $(frame_i, feature_k)$ of $\mathbf{p}_j$ **do**
 - 5: **Update depth:** $d_{ik} = (\mathbf{T}_{cw,i}^{opt} \mathbf{p}_j^{opt}).z()$
 - 6: **Triangulate:** $\mathbf{p}_{world}^{(ik)} = \mathbf{T}_{wc,i}^{opt} [d_{ik} \mathbf{p}_{norm}, d_{ik}]^T$
 - 7: **Collect:** $\mathcal{P}_j \leftarrow \mathcal{P}_j \cup \{\mathbf{p}_{world}^{(ik)}\}$
 - 8: **end for**
 - 9: **Compute:** Empirical covariance using optimized position as mean
 - 10: **Update:** MapPoint uncertainty Σ_j^{new}
 - 11: **end for**
-

D. Iterative Statistical Refinement

The complete learning cycle integrates seamlessly into sliding window optimization through an iterative refinement process. Algorithm 2 outlines the detailed procedure for statistical learning-enhanced bundle adjustment, which alternates between forward uncertainty propagation and backward statistical learning phases. This iterative process is designed to exhibit several desirable properties that enhance the overall system performance. First, it promotes self-consistency, where improved uncertainty estimates may contribute to more accurate optimization, potentially providing higher quality statistical evidence for further refinement. Second, the framework demonstrates adaptive convergence, as uncertainty models automatically adjust to specific sensor characteristics and prevailing environmental conditions. Finally, the process exhibits inherent robustness; even when initialized with suboptimal uncertainty estimates, the continuous accumulation of statistical evidence from new observations gradually improves the uncertainty models over time.

Through this iterative refinement process, the learned uncertainties $\Sigma_j^{(t)}$ progressively converge toward representations that better reflect true measurement reliability, creating observation-specific weighting that captures actual geometric consistency across multiple viewpoints.

Algorithm 2 Statistical Learning-Enhanced Bundle Adjustment

- 1: **Initialize:** Use prior uncertainty estimates $\{\Sigma_j^{(0)}\}$
 - 2: **for** each bundle adjustment iteration t **do**
 - 3: **Forward Pass:** Create adaptive information matrices
 - 4: **for** each observation (i, j) **do**
 - 5: $\Omega_{ij}^{(t)} = (\mathbf{J}_\pi \mathbf{R}_{cw} \Sigma_j^{(t)} \mathbf{R}_{cw}^T \mathbf{J}_\pi^T)^{-1}$
 - 6: **end for**
 - 7: **Optimization:** Solve bundle adjustment with adaptive weighting
 - 8: $\{\mathbf{T}_{wb,i}^{opt}, \mathbf{p}_j^{opt}\} = \arg \min \sum_{(i,j)} e_{ij}^T \Omega_{ij}^{(t)} e_{ij}$
 - 9: **Backward Pass:** Learn from optimization results
 - 10: **for** each MapPoint j **do**
 - 11: **Collect:** Multi-view positions using $\{\mathbf{T}_{wb,i}^{opt}\}$
 - 12: **Compute:** Empirical covariance around $\mathbf{p}_j^{opt}$
 - 13: **Update:** $\Sigma_j^{(t+1)} \leftarrow \Sigma_{empirical,j}$
 - 14: **end for**
 - 15: **end for**
-

E. Integration with Visual-Inertial Pipeline

The statistical learning framework operates as a background process during sliding window optimization, with minimal computational overhead:

$$\begin{aligned} \text{Total Cost} = & \sum_{(i,j) \in \mathcal{O}} e_{visual,ij}^T \Omega_{learned,ij} e_{visual,ij} \\ & + \sum_k \mathbf{r}_{IMU,k}^T \Sigma_{IMU,k}^{-1} \mathbf{r}_{IMU,k} \end{aligned} \quad (16)$$

where $\Omega_{learned,ij}$ represents the statistically learned information matrices that continuously improve through the optimization feedback loop.

V. EXPERIMENTAL RESULTS

A. Experimental Setup

We evaluated our framework on the EuRoC MAV dataset [10]. For the quantitative analysis, we compared three configurations in a progressive manner. We started with a **Baseline** method, representing a standard VIO approach that assumes uniform uncertainty for all visual observations. The first step, **Phase 1**, introduced an enhanced model incorporating triangulation-based uncertainty derived from geometric constraints. This allowed us to investigate our first key hypothesis: whether explicitly modeling uncertainty, even with a simple geometric approach, is fundamentally more effective than treating all measurements as equally reliable. The final step, **Phase 2**, implemented our full proposed framework, which adds a learning-based module to adaptively calibrate uncertainty from statistical feedback. The goal of this phase was to test our second hypothesis: that a statistical learning approach can yield additional improvements by capturing complex uncertainty patterns that a fixed geometric model may not represent.

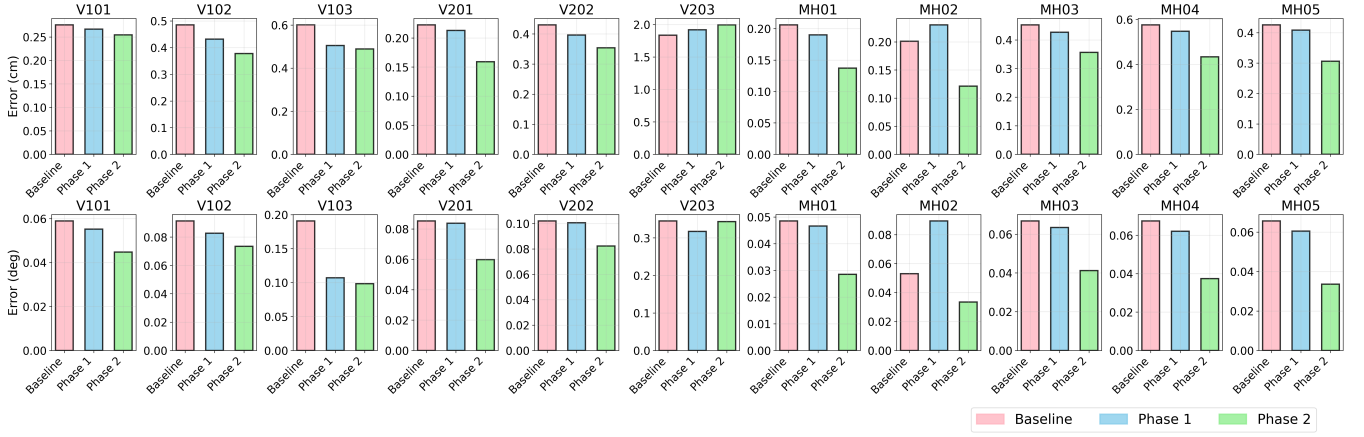


Fig. 4. Quantitative evaluation of frame-to-frame relative pose accuracy across EuRoC sequences. Upper panels display translation RMSE (cm), lower panels show rotation RMSE (degrees). Our statistical uncertainty learning approach (Phase 2) demonstrates consistent improvements over baseline and geometric uncertainty methods (Phase 1) across most sequences. The progressive improvement trend (Baseline $\rightarrow$ Phase 1 $\rightarrow$ Phase 2) validates the effectiveness of adaptive uncertainty learning for visual-inertial odometry.

TABLE I
FRAME-TO-FRAME RELATIVE POSE RMSE COMPARISON ON EUROC DATASET.

Method	Error Type	EuRoC Sequence										
		V1_01	V1_02	V1_03	V2_01	V2_02	V2_03	MH_01	MH_02	MH_03	MH_04	MH_05
Baseline	Trans (cm)	0.276	0.485	0.601	0.222	0.430	1.84	0.206	0.201	0.452	0.576	0.426
	Rot (°)	0.059	0.091	0.191	0.085	0.102	0.345	0.049	0.053	0.067	0.067	0.066
Phase 1	Trans (cm)	0.267	0.432	0.505	0.213	0.396	1.919	0.190	0.230	0.427	0.548	0.409
	Rot (°)	0.055	0.083	0.107	0.084	0.101	0.317	0.047	0.089	0.064	0.062	0.061
Phase 2	Trans (cm)	0.255	0.378	0.489	0.159	0.354	1.993	0.137	0.121	0.356	0.434	0.306
	Rot (°)	0.045	0.074	0.098	0.060	0.082	0.344	0.029	0.033	0.041	0.037	0.034

B. Experimental Environment

All experiments were conducted on a desktop system running a Linux-based operating system. The hardware was equipped with an Intel(R) Core(TM) i7-14700KF processor, which features a hybrid architecture of 20 cores (8 Performance-cores and 12 Efficient-cores) and 28 threads, with a maximum clock speed of 5.6 GHz and 20 MiB of L2 cache.

C. Evaluation Metrics

To ensure a robust assessment of local tracking accuracy, we employ the frame-to-frame relative transform error, following standard evaluation methodologies [29]. This metric is less sensitive to long-term drift than absolute trajectory error. For each consecutive frame pair $(i, i+1)$, we compute the error between the estimated and ground truth relative poses. The estimated relative pose is defined as:

$$\mathbf{T}_{rel}^{est} = (\mathbf{T}_i^{est})^{-1} \mathbf{T}_{i+1}^{est} \quad (17)$$

while the ground truth relative pose is:

$$\mathbf{T}_{rel}^{gt} = (\mathbf{T}_i^{gt})^{-1} \mathbf{T}_{i+1}^{gt} \quad (18)$$

The relative pose error is then decomposed into two components:

a) *Translation Error* (e_{trans}): The Euclidean distance between the relative translation vectors:

$$e_{trans} = \|\mathbf{t}_{rel}^{est} - \mathbf{t}_{rel}^{gt}\|_2 \quad (19)$$

b) *Rotation Error* (e_{rot}): The angular difference between the relative rotation matrices:

$$e_{rot} = \arccos \left(\frac{\text{trace}((\mathbf{R}_{rel}^{est})^T \mathbf{R}_{rel}^{gt}) - 1}{2} \right) \quad (20)$$

The results presented in Table I are the Root Mean Square Error (RMSE) of these metrics over all frame pairs in each sequence.

D. Quantitative Performance Analysis

As summarized in Figure 4 and Table I, our experiments demonstrate a progressive improvement in tracking accuracy with each phase, where the final learning-based approach in **Phase 2** achieves the highest performance. Overall, Phase 2 attains the lowest error in the majority of test cases, specifically in 10 out of 11 sequences for both translation and rotation metrics. A notable exception was observed in the highly challenging V2_03 sequence, where rapid motion and challenging conditions seemed to highlight the complementary strengths of different approaches. Despite

this outlier, the results support the overall effectiveness of our proposed framework.

The performance gains align with the intended effects of our two-phase approach. The initial comparison of Phase 1 to the Baseline showed modest improvements of 3–11% in most sequences, suggesting that basic uncertainty modeling may provide benefits over uniform weighting assumptions. This lends support to the fundamental idea of considering measurement reliability in VIO systems. Building on this foundation, Phase 2 demonstrated more substantial additional improvements of 4–25% over Phase 1, with the larger gains typically observed in sequences with richer geometric structure. This outcome is consistent with the hypothesis that statistical learning can capture complex uncertainty patterns that simple geometric models may not fully represent.

Cumulatively, the complete framework in Phase 2 resulted in overall accuracy improvements ranging from 8–34% compared to the Baseline across different sequences, suggesting clear benefits from the statistical uncertainty learning process. The degree of improvement appeared to correlate with sequence characteristics and motion complexity. On sequences with stable motion and good visual features (e.g., V1_01, MH_01), Phase 2 achieved substantial error reductions of approximately 25–34%. This trend extended to sequences with more complex motion patterns (e.g., V1_02, V1_03, MH_03), where the adaptive nature of Phase 2 continued to yield meaningful gains in the range of 15–25%.

Interestingly, the most challenging sequence, V2_03, presented a mixed result that may highlight fundamental trade-offs under extreme conditions. In this aggressive scenario, no single method excelled consistently across both metrics; the Baseline achieved the most stable translation performance, while Phase 1 with geometric constraints produced the best rotation accuracy.

E. Computational Efficiency

These accuracy improvements were observed with what appears to be minimal computational overhead. The Phase 1 uncertainty propagation added approximately a 2.3% computational cost, while the learning-based calibration in Phase 2 added a mere 0.8% on top of Phase 1. On our experimental system, equipped with an Intel Core i7-14700KF CPU, the entire framework maintained real-time performance at 90–100 Hz, suggesting its potential suitability for practical visual-inertial odometry applications.

VI. CONCLUSION

Our findings suggest that the explicit modeling of measurement uncertainty is a valuable component of visual-inertial odometry. The initial phase of our study indicated that incorporating even a basic geometric uncertainty model can yield modest performance gains of 3–11% over assuming uniform covariance. This observation highlights the potential importance of properly weighting observations rather than treating them as equally reliable. Building on this, the second phase showed that a statistical learning approach could offer further improvements of 4–25% over the geometric model.

This suggests that data-driven methods may be capable of capturing more complex uncertainty patterns that are not fully represented by simpler, pre-defined models.

Taken together, the progressive comparison from our baseline to Phase 1 and Phase 2 resulted in overall trajectory accuracy improvements ranging from 8–34% across different sequences. This progression lends support to the value of both foundational uncertainty modeling and more advanced statistical learning techniques. Furthermore, the proposed framework demonstrated practical viability, maintaining real-time performance at 90–100 Hz with a computational overhead of less than 3%, suggesting its suitability for autonomous navigation tasks.

In summary, the systematic comparison presented in Table I is consistent with our core hypothesis that principled uncertainty modeling is beneficial and that statistical learning represents a promising direction for enhancing traditional geometric approaches in visual-inertial state estimation. Future work will investigate extending this framework to dynamic environments, exploring uncertainty learning for semantic features, and analyzing theoretical convergence properties of the learning algorithm. Additionally, we plan to develop more robust approaches that can maintain superior stability and accuracy in highly challenging scenarios such as the V2_03 sequence, where rapid motion and extreme conditions currently present limitations for statistical learning methods.

REFERENCES

- [1] S. Thrun, W. Burgard, and D. Fox, *Probabilistic robotics*. MIT press, 2005.
- [2] T. Lupton and S. Sukkarieh, “Visual-inertial-aided navigation for high-dynamic motion in built environments without gps,” vol. 28, no. 1, 2012, pp. 61–76.
- [3] D. Scaramuzza and F. Fraundorfer, “Visual odometry,” *IEEE Robotics & Automation Magazine*, vol. 18, no. 4, pp. 80–92, 2011.
- [4] K. Sun, K. Mohta, B. Pfrommer, M. Watterson, S. Liu, Y. Mulgaonkar, C. J. Taylor, and V. Kumar, “Robust stereo visual inertial odometry for fast autonomous flight,” vol. 3, no. 2, 2018, pp. 967–974.
- [5] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, “Keyframe-based visual-inertial odometry using nonlinear optimization,” vol. 34, no. 3, 2015, pp. 314–334.
- [6] F. Dellaert and M. Kaess, “Factor graphs for robot perception,” in *Foundations and Trends in Robotics*, vol. 6, no. 1–2, 2017, pp. 1–139.
- [7] H. Strasdat, J. M. Montiel, and A. J. Davison, “Visual slam: why filter?” *Image and Vision Computing*, vol. 30, no. 2, pp. 65–77, 2012.
- [8] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.
- [9] T. Schneider, M. Dymczyk, M. Fehr, K. Egger, I. Gilitschenski, M. Hamer, R. Siegwart, and P. Furgale, “maplab: An open framework for research in visual-inertial mapping and localization,” *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1418–1425, 2018.
- [10] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, “The euroc micro aerial vehicle datasets,” vol. 35, no. 10, 2016, pp. 1157–1163.
- [11] T. Qin, P. Li, and S. Shen, “Vins-mono: A robust and versatile monocular visual-inertial state estimator,” in *IEEE Transactions on Robotics (T-RO)*, vol. 34, no. 4, 2018, pp. 1004–1020.
- [12] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, “ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM,” *IEEE Transactions on Robotics (T-RO)*, vol. 37, no. 6, pp. 1874–1890, 2021.
- [13] A. I. Mourikis and S. I. Roumeliotis, “A multi-state constraint kalman filter for vision-aided inertial navigation,” in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2007, pp. 3565–3572.

- [14] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct ekf-based approach," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 298–304.
- [15] P. Geneva, K. Eickenhoff, W. Lee, Y. Yang, and G. Huang, "Openvins: A research platform for visual-inertial estimation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2036–2043, 2020.
- [16] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g2o: A general framework for graph optimization," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 3607–3613.
- [17] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 834–849.
- [18] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," vol. 40, no. 3, 2018, pp. 611–625.
- [19] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 15–22.
- [20] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics (T-RO)*, vol. 31, no. 5, pp. 1147–1163, 2015.
- [21] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.
- [22] W. Förstner and B. P. Wrobel, *Photogrammetric computer vision: statistics, geometry, orientation and reconstruction*. Springer, 2016.
- [23] P. J. Huber and E. M. Ronchetti, *Robust statistics*. John Wiley & Sons, 2009.
- [24] V. Usenko, N. Demmel, D. Schubert, and D. Cremers, "Visual-inertial mapping with non-linear factor recovery," vol. 5, no. 2, 2019, pp. 422–429.
- [25] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in neural information processing systems (NeurIPS)*, 2017.
- [26] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An accurate o(n) solution to the pnp problem," vol. 81, no. 2, 2009, pp. 155–166.
- [27] T. D. Barfoot, *State estimation for robotics*. Cambridge University Press, 2017.
- [28] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual-inertial odometry," vol. 33, no. 1, 2017, pp. 1–21.
- [29] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012, pp. 573–580.