2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

# MUSE: Multimodal Uncertainty Quantification of State Estimation

Minkyung Kim*[1], Henry Che*[2], Bhargav Chandaka[2], Bhumsitt Pramuanpornsatid[2], Chengyu Yang[1], Sheng Cheng[1], Xiaofeng Wang[3], Naira Hovakimyan[1], Shenlong Wang[2]

***Abstract*— Accurate visual state estimation has been a central topic in robotics with a wide range of applications in robot navigation, autonomous driving, and autonomous flight. Recent advances in robot perception have led to significant improvements in the accuracy and robustness of state estimation, yet a fundamental challenge remains in how to quantify and calibrate its precision, i.e., how confident we are in an estimate and whether failures can be detected. This issue is particularly pronounced in visual–inertial odometry (VIO), where the heteroscedastic and multimodal nature of the problem makes uncertainty quantification especially difficult. This paper introduces MUSE (Multimodal Uncertainty Quantification of State Estimation), a novel real-time learning-based framework that leverages the strong and efficient sequential modeling capacity of Mamba to estimate localization uncertainty from multiple asynchronous sensor streams. Experiments on both public and in-house datasets demonstrate that MUSE achieves superior reliability and robustness compared to existing uncertainty quantification methods, and ablation studies justify the benefits of its key design choices. We release our source code and dataset at `https://github.com/hungdche/MUSE`.**

## I. INTRODUCTION

State estimation is fundamental in robotics, enabling applications ranging from autonomous driving and drone navigation to robot operation in GPS-denied environments. Research over the past few decades in perception-based state estimation, such as visual odometry (VO) and visual–inertial odometry (VIO), has yielded significant advances in estimating a robot's pose (position and orientation), making these methods indispensable for safety-critical tasks. However, accuracy alone is not sufficient: robotic systems must also reason about the reliability of their state estimates. In practice, odometry estimates remain prone to errors from sensor noise, environmental ambiguity, and algorithmic limitations, making uncertainty quantification essential for safe and reliable decision-making.

VO and VIO approaches can be broadly categorized into three paradigms. Filter-based approaches [1], [2] propagate covariance recursively but suffer from inconsistency due to linearization errors, model mismatch, drift, and limited belief representation. Optimization-based methods [3]–[8] improve accuracy and robustness, yet estimate uncertainty post-hoc from the final solution, yielding local and poorly calibrated measures. Learning-based approaches [9]–[11] achieve state-of-the-art (SOTA) accuracy but remain overconfident, providing unreliable uncertainty estimates.

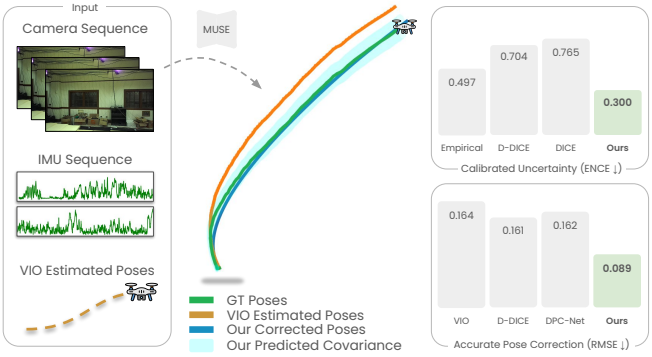


Fig. 1: We present MUSE, a framework to jointly correct poses and predict heteroscedastic uncertainty across time. Our method is able to achieve improvement over various baselines in both pose correction accuracy and uncertainty.

Recent works attempt to learn pose uncertainty directly from sensory cues [12]–[14]. While effective to some extent, these methods have key limitations: they typically rely on instantaneous inputs without exploiting temporal information and can use only visual information as their primary input. Intuitively, as depicted in Fig. 2, pose drifts and uncertainty in VIO are usually more predictable via temporal and multimodal cues (e.g., sudden drop in IMU data, motion blur detected in images, or sudden sharp changes in raw odometry that disagree with IMU) rather than a single image cue. As a result, these methods' uncertainty predictions remain limited in both reliability and generality.

In this paper, we present **MUSE**, a novel, real-time, and versatile framework for quantifying VO/VIO uncertainty from temporal multi-sensor streams. MUSE takes raw odometry and multi-sensor streams as input and predicts a *non-zero-mean* Gaussian distribution over pose errors in SE(3), enabling simultaneous pose correction and calibrated uncertainty estimation. At its core, MUSE leverages Mamba [15], a powerful and efficient structured state-space model (SSM), to capture uncertainty over time windows and account for the heteroscedasticity of pose errors. *To the best of our knowledge, this is the first framework that jointly addresses the multimodal, temporal, and heteroscedastic nature of odometry uncertainty.* Moreover, MUSE is directly deployable as a real-time plugin for any VO/VIO system.

*These authors contributed equally to this work.

[1]Authors are with the Department of Mechanical Science and Engineering, University of Illinois Urbana-Champaign, Champaign, IL 61801, USA. (email: {mk58,cy45,chengs,nhovakim}@illinois.edu)

[2]Authors are with Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign, Champaign, IL 61801, USA. (email: {hungdc2, bhargav9, bp17, shenlong}@illinois.edu)

[3]Author is with the Department of Electrical Engineering, University of South Carolina, Columbia, SC 29208, USA. (email: wangxi@cec.sc.edu)

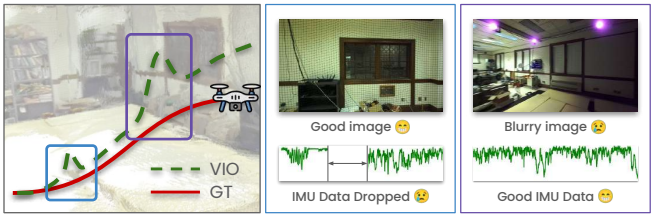


Fig. 2: Motivation. Uncertainty and failures in visual–inertial odometry (VIO) arise from diverse and multimodal factors (e.g., IMU anomalies, sensor contamination). Our model is designed to learn to capture these multimodal cues for better uncertainty calibration.

We validate MUSE on EuRoC [16] and our challenging in-house UnCal-Flight dataset across diverse VO/VIO frameworks. Results show improved pose correction and well-calibrated uncertainty estimates, outperforming various baselines (Fig. 1). A comprehensive ablation study further demonstrates the benefits of multimodal integration for robust uncertainty prediction.

Our main contributions are summarized as follows:

(i) A novel architecture that learns multimodal, heteroscedastic uncertainty in odometry estimation.
(ii) An open-source, challenging drone navigation UnCal-Flight dataset for evaluating VO/VIO robustness.
(iii) Extensive experiments on EuRoC and our dataset across multiple VIO modules, showing MUSE's effectiveness as a real-time plugin for existing systems.

## II. RELATED WORK

### A. Visual(-Inertial) Odometry

Traditional VO/VIO pipelines include (i) filter-based estimators [1], [2] that fuse inertial and visual constraints online with recursive updates, and (ii) optimization-based estimators, such as VINS-Mono [4], VINS-Fusion [5], and ORB-SLAM [6]–[8], that refine a sliding window or global pose graph with geometric or photometric objectives. Kalman filter offers real-time efficiency but can suffer from inconsistency stemming from linearization, leading to overconfident or conservative covariances and drift. Optimization-based methods generally achieve higher accuracy, yet their uncertainty is typically local and not calibrated for predictive use, making it challenging to quantify heteroscedastic or multimodal error under changing conditions.

Recently, deep-learning-based approaches (e.g., DeepVO [9], TartanVO [10], DPVO [11]) achieve strong accuracy by directly estimating poses from visual inputs, but they frequently exhibit limited reliability in their uncertainty, especially under distribution shift or degraded sensing. To address this issue, unsupervised approaches like GANVO [17] and SelfVIO [18] remove the need for labeled trajectories by leveraging view-synthesis and adversarial training techniques. Meanwhile, Adaptive VIO [19] combines continual learning with traditional optimization to enable the system to adapt to the new environment.

Despite these advancements, challenges remain in ensuring the reliability of VO/VIO under out-of-distribution or adverse conditions, including dynamic environments, sensor degradation, and low-light scenarios. Our work seeks to close these gaps by incorporating uncertainty quantification for a VO/VIO framework, which can be utilized alongside existing VO/VIO systems in safety-critical applications.

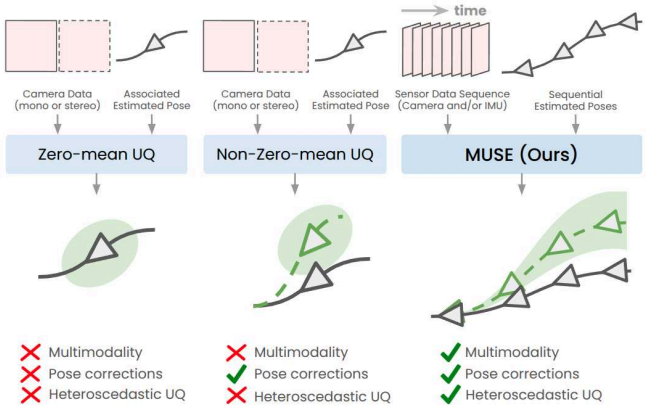


Fig. 3: Comparison between prior uncertainty quantification (UQ) methods and MUSE. Our approach leverages multimodal sensor streams to capture the heteroscedastic nature of pose uncertainty, which prior methods cannot.

### B. Uncertainty Quantification in SLAM

Uncertainty quantification has been widely studied in machine learning [20] and is increasingly being addressed in VO/VIO through learning-based approaches. DICE [12] and D-DICE [13] utilize convolutional neural networks (CNNs) to learn a pose error model from raw images in the form of Gaussian distributions. DPC-Net [14] predicts SE(3) pose error directly from images and performs corrections by employing pose graph relaxation techniques. More recent works [21], [22] have explored statistically principled calibration, including conformal prediction frameworks that wrap pre-trained VO/SLAM to guarantee coverage. There are also research directions that seek to quantify heteroscedastic uncertainties when estimating poses. For instance, Uncertainty-Aware VO (UA-VO) [23] provides confidence measures by accounting for epistemic and aleatoric uncertainties in VO. Similarly, D3VO [24], a self-supervised method, incorporates predicted pose, depth, and photometric uncertainties. However, these visual-focused approaches typically do not consider uncertainties arising from other sensor modalities.

Our work extends this line by explicitly targeting multimodal, sequential uncertainty quantification. Unlike prior visual-only-based or per-frame uncertainty predictors, we introduce a framework that fuses asynchronous visual, inertial, and odometry streams over long horizons. Our approach enables simultaneous pose correction and calibrated predictive uncertainty. This positions our method as the first to unify multimodal inputs, long-sequence modeling, and rigorous covariance prediction into a deployable plugin for existing VO/VIO systems, improving both accuracy and reliability in dynamic, resource-constrained environments.

### C. State-Space Models for Robotics

While standard sequence models, such as Transformers [25], excel at capturing complex relationships through self-attention, they suffer from quadratic computational complexity as sequence length increases. State Space Models (SSMs) have recently emerged as a powerful architecture inspired by classical dynamic systems. Early variants of SSM [26] first utilize specialized HiPPO parameterizations [27], enabling effective management of long-context information. Recently, Mamba [15] introduced a selective mechanism, enabling the model to selectively emphasize or forget information, thereby significantly improving its expressive power.

These properties make SSMs particularly appealing for robotics, which relies on high-throughput sensory streams under real-time constraints. Applications in perception include multimodal temporal fusion for 3D perception [28] and adaptive parameter tuning in SLAM [29]. Research in policy learning [30] shows that SSMs can replace heavier sequence models while preserving latency and robustness across long horizons. Frameworks like RoboMamba [31] further integrate Mamba into vision-language-action models, allowing robots to follow natural-language instructions grounded in visual context. Together, these advances highlight SSMs as a scalable backbone for real-time robotic systems that demand long-horizon and multimodal reasoning.

In this paper, we adopt Mamba [15] as the backbone for quantifying uncertainty in localization. Its ability to selectively process information from long, complex sequences is ideal for efficiently quantifying uncertainty from demanding multi-sensor streams.

## III. Method

### A. Preliminaries

Let the ground-truth pose at time $i$ be $\mathbf{T}_{\mathrm{gt},i} \in \mathrm{SE}(3)$ and the estimated pose by an arbitrary VO/VIO module be $\hat{\mathbf{T}}_i$. Assume both trajectories share the same initial pose $\mathbf{T}_{\mathrm{gt},0} = \hat{\mathbf{T}}_0$. We define the rigid-body pose estimator error at time $i$ to be

$$\mathbf{T}_{\mathrm{err},i} = \mathbf{T}_{\mathrm{gt},i}\,\hat{\mathbf{T}}_i^{-1}. \tag{1}$$

Learning a parametric function to predict $\mathbf{T}_{\mathrm{err},i}$ is challenging and unstable due to the orthonormal constraints. Thus, we map the error $\mathbf{T}_{\mathrm{err},i}$ on SE(3) to the Lie algebra $\mathfrak{se}(3)$ via the logarithm map to obtain a six-value vector $\boldsymbol{\xi}_i = \log(\mathbf{T}_{\mathrm{err},i}) \in \mathbb{R}^6$. Throughout this paper, we adopt the "hat" operator $(\cdot)^\wedge : \mathbb{R}^6 \to \mathfrak{se}(3)$ so that $\exp(\boldsymbol{\xi}_i^\wedge) = \mathbf{T}_{\mathrm{err},i}$. Following the prior work [13], we model the error distribution as a non-zero-mean Gaussian distribution

$$\boldsymbol{\xi}_i \sim \mathcal{N}(\boldsymbol{\mu}_i,\, \boldsymbol{\Sigma}_i) \tag{2}$$

with mean $\boldsymbol{\mu}_i \in \mathbb{R}^6$ and full covariance $\boldsymbol{\Sigma}_i \in \mathbb{R}^{6\times 6}$.

As $\mathcal{N}(\boldsymbol{\mu}_i,\, \boldsymbol{\Sigma}_i)$ is assumed to model the true underlying pose error distribution, we can use $\boldsymbol{\mu}_i$ to "correct" estimated error. We define the corrected pose $\hat{\mathbf{T}}'_i$ as follows:

$$\hat{\mathbf{T}}'_i = \exp(\boldsymbol{\mu}_i^\wedge)\,\hat{\mathbf{T}}_i. \tag{3}$$

Then, the error of the corrected pose $\mathbf{T}'_{\mathrm{err},i}$ is given by

$$\mathbf{T}'_{\mathrm{err},i} = \mathbf{T}_{\mathrm{gt},i}\,\hat{\mathbf{T}}'^{-1}_i = \mathbf{T}_{\mathrm{err},i}\,\exp(-\boldsymbol{\mu}_i^\wedge). \tag{4}$$

Let $\boldsymbol{\xi}'_i = \log(\mathbf{T}'_{\mathrm{err},i}) \in \mathbb{R}^6$. Under the first-order Baker–Campbell–Hausdorff approximation, the error distribution of the corrected poses is approximately zero-mean with unchanged covariance:

$$\boldsymbol{\xi}'_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i). \tag{5}$$

### B. Problem Definition

Given a stream of multimodal sensor observations $\mathcal{Z}_t = \{z_0, \ldots, z_t\}$ where $z_i = \{z_{\text{image}}, z_{\text{imu}}\}_i$ and the sequence of estimated poses $\hat{\mathbf{T}}_{0:T}$ from a given odometry module over $T$ period, we seek to learn the distribution of the error between $\hat{\mathbf{T}}_{0:T}$ and the ground-truth poses $\mathbf{T}_{gt,0:T}$ through a parametric function $f_\theta$, such that:

$$\boldsymbol{\mu}_{0:T}, \boldsymbol{\Sigma}_{0:T} = f_\theta(\mathcal{Z}_T, \hat{\mathbf{T}}_{0:T}), \tag{6}$$

We expect our predicted $\boldsymbol{\mu}_{0:T}$ and $\boldsymbol{\Sigma}_{0:T}$ to accurately capture the true error distribution, where the mean provides pose correction and covariance offers posterior confidence. In this sense, our output is greatly desired for uncertainty-aware downstream planner or utilized as introspective model to improve perception accuracy.

### C. Input Feature Extractions

In light of our motivation to utilize multimodal data (Fig. 2), our pipeline takes as input multi-sensor data and raw odometry, which we process as follows.

*1) Sensor Inputs:* We consider stereo images and IMU data (linear accelerations and angular velocities) as the primary observational inputs of our model. We further remark that, owing to its modular design, MUSE can also support different sensor modalities and configurations. For visual inputs, we extract features using a pretrained *SuperPoint* [32] encoder $\mathcal{E}_{\text{image}}$ (operating on grayscale images converted from RGB) and project them to $d_{\text{image}}$ via a small MLP. For inertial measurement data, we sample them at 200 Hz, truncate them into chunks based on timestamps of every successive estimated poses, and encode them with a pretrained *RONIN* [33] encoder $\mathcal{E}_{\text{imu}}$. Similarly, an MLP is employed to project the output features to $d_{\text{imu}}$.

*2) Odometry Input:* As our model is designed to correct the VIO estimates, learning the high-dimensional underlying distribution of the estimates is valuable. The upstream VO/VIO engine supplies each pose and, when available, the associated $6 \times 6$ covariance (e.g. MSCKF [2]). We encode these quantities with a single-layer fully-connected network (FCN) $\mathcal{E}_{\text{odom}}$, where the resulting feature dimension is $d_{\text{odom}}$.

After extracting the features from each input modality, we concatenate them along the sequence length dimension. Thus, the resulting feature is obtained via:

$$f_{\text{input}} = \mathcal{E}_{\text{image}}(\mathcal{Z}_{\text{image}}) \oplus \mathcal{E}_{\text{imu}}(\mathcal{Z}_{\text{imu}}) \oplus \mathcal{E}_{\text{odom}}(\hat{T}_{0:T}), \tag{7}$$

where $\oplus$ denotes concatenation on the sequence dimension. During training, the sensor encoders $\mathcal{E}_{\text{image}}$ and $\mathcal{E}_{\text{imu}}$ are frozen for efficiency, while the projection layers and $\mathcal{E}_{\text{odom}}$ are updated by backpropagation.

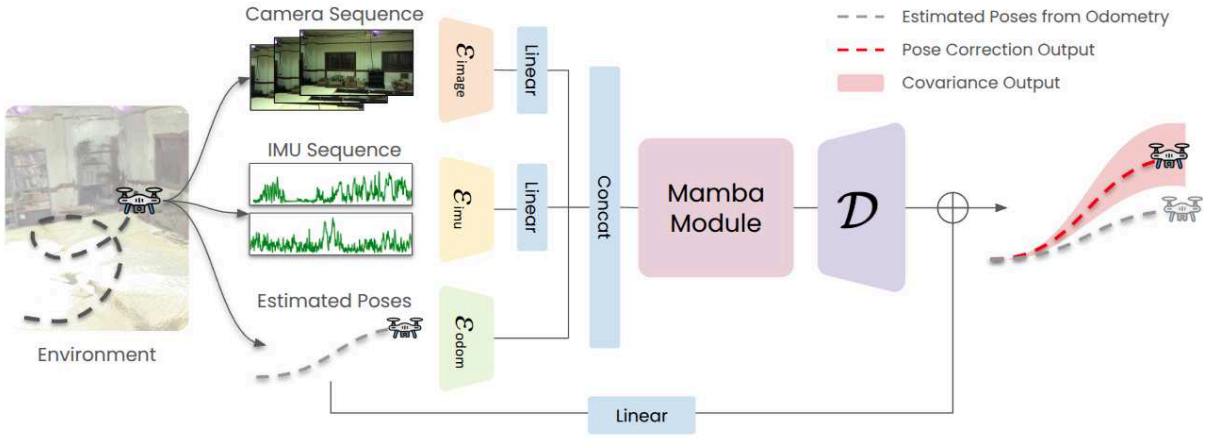


Fig. 4: MUSE Architecture. MUSE takes multiple asynchronous sensor streams and the estimated pose of a given VO/VIO as the input. Each input stream (camera, IMU, and Odometry) is first encoded by its corresponding encoder before being concatenated together on the sequence length dimension. The output of concatenation is then passed through a Mamba module to capture the sequential relationship before being decoded into pose corrections and covariances.

### D. Learning Error Distribution

To efficiently learn the sequential nature of pose estimation, we employ the recently published state-space model Mamba [15] at the heart of our pipeline. The features obtained from Eq. (7) are passed through a stack of Mamba blocks to learn the temporal correlations of the input modalities. For our Mamba module, we use $N$ blocks of Mamba, each consisting of a small linear layer followed by a Conv1D, SiLU activation, SSM [15], and a small linear layer. A residual connection, which consists of another linear layer and a SiLU activation, is then multiplied element-wise with the output of SSM, following [34]. In all experiments, we fix $N = 4$ as a balance between efficiency and performance.

### E. Predicting Pose Corrections and Uncertainty

The desired output of our model is the mean $\boldsymbol{\mu}_i \in \mathbb{R}^6$ and covariance $\boldsymbol{\Sigma}_i \in \mathbb{R}^{6\times6}$ for $i \in [0, 1, \cdots, T]$. However, directly predicting 36 parameters representing $\boldsymbol{\Sigma}_i$ would lead to training instability due to the positive semi-definite constraints. Thus, following [12], [13], we opt to reconstruct $\boldsymbol{\Sigma}_i$ through LDL Decomposition $\boldsymbol{\Sigma}_i = L_i D_i L_i^T$ by predicting 6 parameters $d_i$ and 15 parameters $l_i$ such that

$$\boldsymbol{\Sigma}_i = L_i D_i L_i^T = \mathrm{trilu}(l_i)\mathrm{diag}(\exp(d_i))\mathrm{trilu}(l_i)^T. \quad (8)$$

Here, $\mathrm{diag}(\exp(d_i))$ denotes the $6\times6$ diagonal matrix, where $\exp(d_i) \in \mathbb{R}^6$ contains its diagonal entries, and $\mathrm{trilu}(l_i)$ denotes the unit lower triangular $6 \times 6$ matrix with $l_i \in \mathbb{R}^{15}$ being the vector of nonzero and off-diagonal parameters. We apply the exponential function $\exp(\cdot)$ on $d_i$ to enforce positiveness and guarantee the existence and uniqueness of the decomposition.

Implementation-wise, we employ two MLP decoders to predict (i) six elements of $\boldsymbol{\mu}_i$ and (ii) 21 elements of $[d_i, l_i]$ that construct $\Sigma_i$ through Eq. (8). Together, they form the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, providing both an accurate state estimate and a calibrated measure of its uncertainty.

### F. Objective Functions

Although mean and covariance can be jointly optimized as in [13], we empirically found that decoupled optimizations lead to more stable training and better convergence.

*Mean (bias) loss*: Following DPC-Net [14], we minimize the geodesic distance between the bias-corrected estimate and ground-truth:

$$L_{\text{geo}} = \frac{1}{2T}\sum_{i=0}^{T}\left\|\log\left(\exp(\hat{\boldsymbol{\xi}}_i)\,\mathbf{T}_{\text{gt},i}^{-1}\right)\right\|_{\mathbf{P}}^2 - \left\|\log(\mathbf{T}_{\text{gt},i}^{-1})\right\|_{\mathbf{P}}^2 + \lambda_s \mathcal{L}_s, \quad (9)$$

where $\hat{\boldsymbol{\xi}}_i$ is the network prediction of error, and $\|\mathbf{v}\|_{\mathbf{P}}^2 = \mathbf{v}^\top \mathbf{P}\,\mathbf{v}$ with an empirical weighting matrix $\mathbf{P}$. For training Mamba, we introduce an additional smoothness penalty, denoted as $\mathcal{L}_s = \frac{1}{T-1}\sum_{i=1}^{T}(\Delta\log(\mathbf{T}_{\text{gt},i}^{-1}) - \Delta\log\left(\exp(\hat{\boldsymbol{\xi}}_i)\,\mathbf{T}_{\text{gt},i}^{-1}\right))$, where $\Delta(\cdot)$ denotes the change in pose between the previous and current states, to capture the smoothness of the trajectory.

*Covariance loss*: To calibrate uncertainty, we minimize the negative log-likelihood (NLL):

$$L_{\text{nll}} = \sum_{i=0}^{T} -\log \mathbb{P}(\boldsymbol{\xi}_i \mid (\boldsymbol{\mu}, \boldsymbol{\Sigma})_i) \quad (10)$$

where $\mathbb{P}$ is probability measure, $\boldsymbol{\xi}_i$ is the observed error, and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})_i$ is the network prediction of error distribution. The total objective is $\mathcal{L} = L_{\text{geo}} + L_{\text{nll}}$ and is optimised end-to-end with AdamW.

### G. Implementation Details

We fix the sequence length $T = 100$ and truncate the input data accordingly. For the dimensions of the input features, we set $d_{\text{image}} = 256$, $d_{\text{imu}} = 128$, and $d_{\text{odom}} = 128$. Additionally, we employ a skip connection from the raw VIO output to the decoder, which is added element-wise to the output mean through a shallow MLP of dimension 6. Both the final

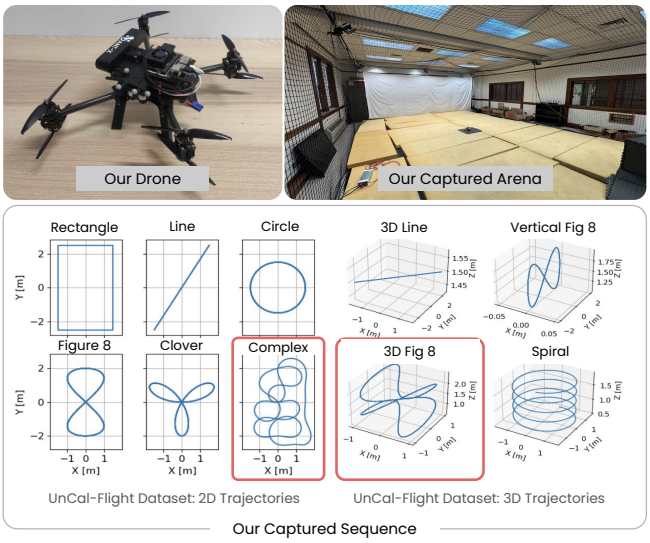


Fig. 5: Our drone, captured arena, and trajectories included in UnCal-Flight dataset. Red denotes evaluation trajectories.

decoder and the skip connection MLP are zero-initialized for training stability. For training parameters, we set $\lambda_s = 100$. We train all our models on 4xNVIDIA RTX 6000 GPUs with an effective batch size of 128 (32 per GPU) and learning rate set to $1e-6$ for mean learning and $1e-4$ for covariance learning. The training time of our model is around 1 hour.

## IV. EXPERIMENTS & RESULTS

We evaluate MUSE against other SOTA uncertainty estimation methods for pose estimation on various odometry methods. Moreover, we demonstrate our method's ability to estimate well-calibrated uncertainty in both zero-mean and non-zero-mean settings when compared with different baselines. Finally, we showcase the robustness of MUSE on the challenging environments in UnCal-Flight dataset.

### A. Datasets

*1) EuRoC Dataset:* The EuRoC dataset [16] is a widely used benchmark for evaluating VO/VIO and SLAM algorithms. It consists of multiple sequences recorded in indoor environments using a Micro Aerial Vehicle (MAV) equipped with a synchronized stereo camera and an inertial measurement unit (IMU). The dataset provides precise ground-truth trajectories obtained through a motion capture system, making it suitable for assessing pose estimation accuracy and uncertainty quantification methods.

We use the Machine Hall trajectories from the EuRoC dataset, with `MH_01` to `MH_04` for training and `MH_05` for evaluation. We run MSCKF [2], VINS-Fusion [5], ORB-SLAM 3 [8], and DPVO [11] to obtain their respective estimations.

*2) UnCal-Flight Dataset:* While the EuRoC dataset is valuable, its number and diversity of trajectories are limited. To address this limitation, we collected an uncalibrated flight (UnCal-Flight) dataset, featuring more complex motion patterns in varied environments. We used ROG-X [35] to gather UAV-based visual-inertial data. The drone is equipped with a ZED 2i stereo camera that captures RGB images at 15 Hz and an onboard IMU recording at 200 Hz. The flight arena, measuring $3 \times 5 \times 7\,m^3$, is equipped with a Vicon motion capture system providing accurate ground-truth measurements at 100 Hz.

To increase dataset diversity, we conducted drone flights along various trajectories at different speeds and with dynamic yaw motions inspired by [36], under varying lighting conditions and with a human moving in and out of the sensor's frame. The shapes of the collected trajectories are illustrated in Fig. 5, with minor deviations due to safety margins and tracking performance. In total, 145 trajectories were collected.

### B. Evaluation Setup

*1) Evaluation Settings:* We evaluate our methods on two settings: (1) zero-mean and (2) non-zero-mean uncertainty estimation. In the first setting, we assume the mean pose error to be zero and modify our decoder to output only the covariance. This setting is valuable for well-tuned and highly accurate pose estimation methods such as ORB-SLAM 3 [8] or DPVO [11]. The second setting is to evaluate the performance of our pose correction, which is valuable for underperforming odometry methods or challenging scenarios. To better evaluate the heteroscedasticity of the covariance estimation, instead of computing the metrics on a frame-to-frame basis, we split the poses into 100-frame overlapping chunks and evaluate the metrics on a chunk-to-chunk basis.

*2) Baselines:* We compare our method against state-of-the-art uncertainty quantification approaches, including DICE [12], D-DICE [13], and DPC-Net [14]. We implement DICE and D-DICE with parameters that best follow their respective papers. The images are resized to $48 \times 64$ for DICE training, and to $80 \times 128$ for D-DICE and DPC-Net. In the zero-mean setting, we include the empirical uncertainty of each VO/VIO model. For the MSCKF, we utilize the covariance obtained from the filter, while for others, we fit a Gaussian distribution at each timestep throughout the sample of chunks. Though simple, we observe that this baseline is quite competitive in capturing the heteroscedastic nature of the covariance. Additionally, to evaluate the impact of multimodal input on uncertainty estimation, we train our network using different input configurations: (i) image-only, (ii) image with IMU, (iii) image with estimations from VO/VIO models, and (iv) full multimodal input incorporating all of the above. This allows us to analyze the contribution of each modality to the overall uncertainty quantification.

### C. Metrics

*a) Uncertainty Calibration:* We evaluate calibration using Log-Likelihood (LL) and Expected Normalized Calibration Error (ENCE) [37]. Similar to Expected Calibration Error (ECE) in classification, ENCE measures the mismatch between observed errors and predicted uncertainty in regression tasks. Low ENCE indicates well-calibrated confidence, while high ENCE suggests over- or under-confidence.

| Methods | VINS-Fusion | | MSCKF | | ORB-SLAM 3 | | DPVO | |
|---|---|---|---|---|---|---|---|---|
| | RMSE [m]↓ | GEO [rad]↓ | RMSE [m]↓ | GEO [rad]↓ | RMSE [m]↓ | GEO [rad]↓ | RMSE [m]↓ | GEO [rad]↓ |
| Raw Odom | 0.164 | 0.014 | 0.063 | **0.008** | **0.027** | **0.002** | **0.048** | **0.005** |
| D-DICE [13] | 0.161 | 0.020 | 0.066 | 0.010 | 0.032 | 0.010 | 0.070 | 0.020 |
| DPC-Net [14] | 0.162 | 0.012 | 0.063 | **0.008** | **0.027** | **0.002** | **0.048** | **0.005** |
| MUSE (Ours) | **0.089** | **0.009** | **0.062** | **0.008** | 0.028 | **0.002** | **0.048** | **0.005** |

TABLE I: Pose correction on EuRoC.

| Methods | VINS-Fusion | | MSCKF | | ORB-SLAM 3 | | DPVO | |
|---|---|---|---|---|---|---|---|---|
| | LL↑ | ENCE↓ | LL↑ | ENCE↓ | LL↑ | ENCE↓ | LL↑ | ENCE↓ |
| *Zero-Mean Uncertainty Quantification* | | | | | | | | |
| Empirical | 36.26 | 0.497 | 25.08 | 0.904 | **57.03** | **0.080** | 48.92 | 0.135 |
| DICE [12] | 26.11 | 0.765 | 33.54 | 0.439 | 44.97 | 0.432 | 37.00 | 0.473 |
| D-DICE [13] | 21.74 | 0.704 | 28.34 | 0.611 | 38.08 | 0.561 | 31.14 | 0.601 |
| MUSE (Zero-Mean) | **37.95** | **0.300** | **44.33** | **0.138** | **57.03** | 0.097 | **49.09** | **0.123** |
| *Non-Zero-Mean Uncertainty Quantification* | | | | | | | | |
| D-DICE [13] | 21.33 | 0.331 | 28.69 | 0.500 | 31.46 | 0.639 | 24.06 | 0.646 |
| MUSE (Ours) | **40.46** | **0.278** | **43.51** | **0.145** | **54.83** | **0.087** | **49.06** | **0.239** |

TABLE II: Uncertainty quantification on EuRoC.

| Methods | Yaw Constant | | | | Yaw Forward | | | |
|---|---|---|---|---|---|---|---|---|
| | RMSE [m]↓ | GEO [rad]↓ | LL↑ | ENCE↓ | RMSE [m]↓ | GEO [rad]↓ | LL↑ | ENCE↓ |
| Raw Odom | 0.045 | **0.022** | - | - | 0.065 | 0.042 | - | - |
| D-DICE [13] | 0.044 | **0.022** | 24.36 | 0.631 | 0.062 | 0.042 | 22.13 | 0.522 |
| DPC-Net [14] | 0.045 | **0.022** | - | - | 0.059 | 0.041 | - | - |
| MUSE (Ours) | **0.035** | **0.022** | **32.00** | **0.421** | **0.040** | **0.033** | **29.46** | **0.379** |

TABLE III: Pose correction and uncertainty quantification on UnCal-Flight Dataset. Raw Odom is obtained from ZED-VO.

| Input | | | Yaw Forward | | | |
|---|---|---|---|---|---|---|
| Cam | IMU | Odom | RMSE [m]↓ | GEO [rad]↓ | LL ↑ | ENCE ↓ |
| ✓ | | | 0.061 | 0.043 | 26.13 | 0.677 |
| ✓ | ✓ | | 0.058 | 0.043 | 25.62 | 0.628 |
| ✓ | | ✓ | 0.050 | 0.038 | 26.25 | **0.359** |
| ✓ | ✓ | ✓ | **0.040** | **0.033** | **29.46** | 0.379 |

TABLE IV: Ablation of multimodal inputs.

For uncertainty quantification, we define the uncertainty score over 6D pose vector as $u_t = \sqrt{\mathrm{tr}(\mathbf{\Sigma}_t)}$. Samples are grouped into equal-sized M bins $\{B_j\}_{j=1}^{M}$, and we compute the root mean variance $\mathrm{RMV}(j) = \sqrt{\frac{1}{|B_j|}\sum_{t\in B_j} u_t^2}$ and the empirical root mean square error (RMSE) as $\mathrm{RMSE}(j) = \sqrt{\frac{1}{|B_j|}\sum_{t\in B_j} \|\boldsymbol{\xi}_t - \boldsymbol{\mu}_t\|^2}$. Then, ENCE is computed as: $\mathrm{ENCE} = \frac{1}{M}\sum_{j=1}^{M} \frac{|\mathrm{RMSE}(j) - \mathrm{RMV}(j)|}{\mathrm{RMV}(j)}$.

*b) Pose Correction:* For the mean values, which correspond to the corrected pose estimation, we use the Root Mean Square Error (RMSE) for the translational error and Geodesic Distance (GEO) for the rotational error $\mathrm{GEO}(R_{est}, R_{gt}) = \arccos\left(\frac{\mathrm{tr}(R_{est}R_{gt}^T)-1}{2}\right)$.

### D. Pose Corrections

In Table I, we present our results on pose correction compared with baseline methods on the same set of odometry. For odometry methods with relatively higher raw odometry errors, such as VINS-Fusion, our method is able to decrease its RMSE by 45.7%, whereas other baselines show minor improvement or regression. For other methods (e.g., MSCKF, ORB-SLAM, and DPVO), their raw odometry has already obtained relatively low pose errors thanks to their well-tuned parameters on EuRoC. Our pose correction shows no regression in performance, whereas D-DICE degrades the raw odometry's performance.

Moreover, we present qualitative results in Fig. 6. Here, we zoom in on the portion of the trajectory where VIO starts to drift from the GT, and provide the according sensor input. Thanks to our utilization of multimodal inputs, our method is able to detect and correct VIO poses, whereas single-model methods such as DPC-Net and D-DICE fail to. This further validates our intuition outlined in Fig. 2 about the importance of multimodality in capturing pose drifts and uncertainty.

### E. Uncertainty Quantification

We present our uncertainty quantification performance for both zero-mean and non-zero-mean settings in Table II.

*1) Zero-Mean Uncertainty Quantification:* Under zero-mean settings, we compare MUSE with DICE [12], D-DICE [13], and empirical covariance. Our method outperforms all the baselines in estimating the uncertainty of VINS-Fusion, MSCKF, and DPVO poses, and achieves competitive results on ORB-SLAM 3 with the empirical covariance. Notably, DICE and D-DICE fall short of empirical uncertainty in this setting, which may follow from the lack of sequential and

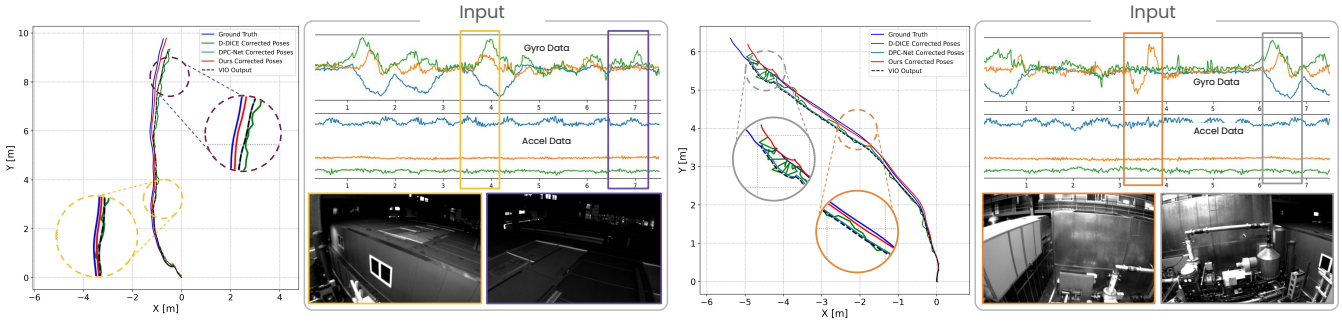


Fig. 6: Qualitative results of pose correction performance on EuRoC Dataset (`MH_05_difficult`). MUSE (red), which utilizes visual, inertial, and odometry signals as inputs, is able to accurately detect and correct pose drifts (zoomed in regions), whereas DPC-Net (light blue) and D-DICE (green) can only support visual input and thus fail.

multimodal modeling in their methods.

*2) Non-zero-mean Uncertainty Quantification:* Under non-zero-mean settings, since DPC-Net does not predict uncertainty, we only compare MUSE with D-DICE. Our uncertainty metrics are better across the board from D-DICE.

### F. Challenging Scenarios

Under this setting, we want to evaluate our performance on the challenging UnCal-Flight Dataset. For the odometry, we use the proprietary odometry algorithm provided by the onboard ZED 2i camera, which we call ZED-VO. We report both pose corrections and non-zero-mean uncertainty quantification in Table III. Our method outperforms all baselines in pose correction for both yaw-constant and yaw-forward subsets. Notably, in terms of RMSE metrics, our method achieves 22.2% and 38.4% improvement over the raw odometry on the above two subsets, respectively. For uncertainty quantification, our method outperforms D-DICE in both LL and ENCE metrics in both subsets. Moreover, we showcase the qualitative results in this setting in Fig. 7, with uncertainty plotted. Our method provides better correction with reasonable uncertainty compared to D-DICE. The average inference time for a single window with all modalities as input is under 3 ms on a single NVIDIA RTX 6000 GPU.

### G. Ablative Studies

To better understand the importance of multimodal input when learning the error distribution, we conduct an ablative study comparing our design with different combinations of input modalities and report the numbers in Table IV. We observe that using all modalities (camera, IMU, and estimated odometry) usually leads to the best performance in both pose correction and uncertainty estimation.

## V. CONCLUSIONS

In this paper, we introduce MUSE, a novel framework for multimodal uncertainty quantification for any black-box pose estimation method. Using a highly efficient and powerful state-space model, MUSE is able to estimate the error distribution of a given odometry with well-calibrated uncertainty, where the predicted mean can be used to accurately correct the pose estimation. Experiments on both EuRoC and the challenging UnCal-Flight dataset demonstrate the effectiveness of our approach across diverse odometry methods and sensing conditions. Beyond improved calibration and correction, our ablation results highlight the importance of multimodal integration, showing consistent gains when fusing visual, inertial, and odometry streams. In future work, we plan to extend MUSE to additional sensing modalities and integrate into downstream planning and control frameworks, moving toward a fully uncertainty-aware autonomy stack.

## ACKNOWLEDGMENT

This work is supported by the Air Force Office of Scientific Research Grant (AFOSR) Grant AF FA9550-25-1-0274, the National Aeronautics and Space Administration (NASA) under Grant 80NSSC22M0070, and by the National Science Foundation (NSF) under Grants CMMI-2135925, CPS-2311085, IIS-2331878, IIS-2331879, IIS-2340254, IIS-2312102, CNS-2414227, IIS-2404385, and CCF-2525287. Henry Che is supported by the NSF GRFP fellowship.

## REFERENCES

[1] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint kalman filter for vision-aided inertial navigation," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*, 2007, pp. 3565–3572.

[2] 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," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 965–972, 2018.

[3] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2018.

[4] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[5] T. Qin, J. Pan, S. Cao, and S. Shen, "A general optimization-based framework for local odometry estimation with multiple sensors," 2019.

[6] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[7] 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.

[8] 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*, vol. 37, no. 6, pp. 1874–1890, 2021.

[9] S. Wang, R. Clark, H. Wen, and N. Trigoni, "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 2043–2050.

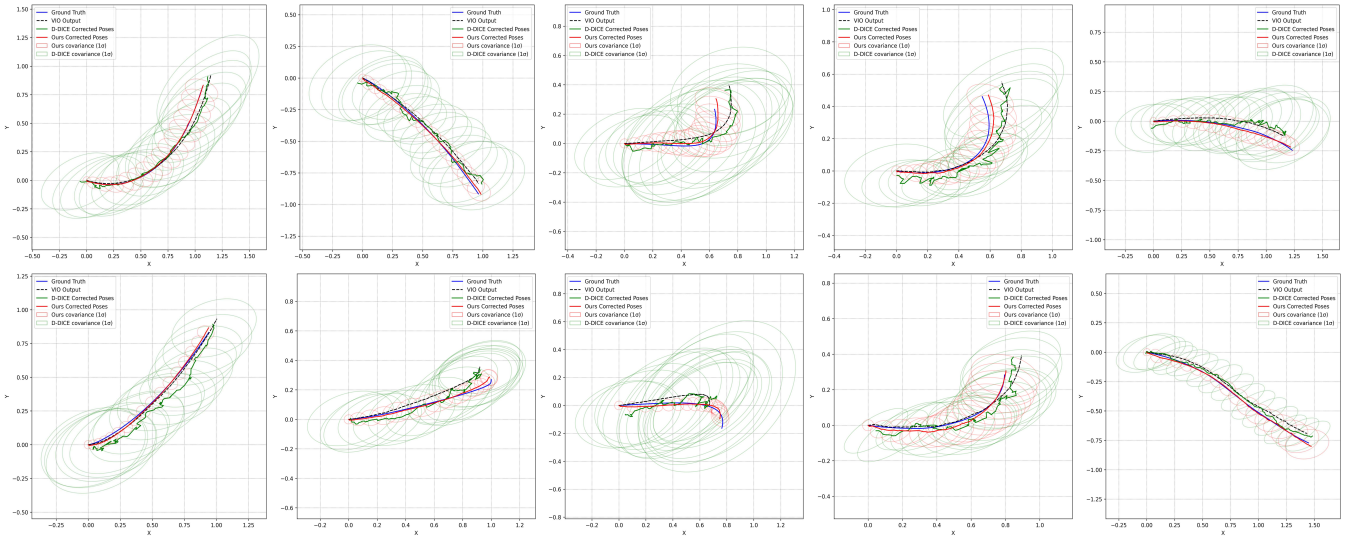


Fig. 7: Qualitative Results on UnCal-Flight Dataset. Our method can predict accurate pose correction (red curve) with reasonable uncertainty (red circles) compared to D-DICE (green curve and green circles). Moreover, our pose correction shows significant improvement over raw VIO odometry (black dotted lines).


[10] W. Wang, Y. Hu, and S. Scherer, "Tartanvo: A generalizable learning-based vo," in *Conference on Robot Learning*. PMLR, 2021, pp. 1761–1772.

[11] Z. Teed, L. Lipson, and J. Deng, "Deep patch visual odometry," *Advances in Neural Information Processing Systems*, vol. 36, pp. 39 033–39 051, 2023.

[12] K. Liu, K. Ok, W. Vega-Brown, and N. Roy, "Deep inference for covariance estimation: Learning gaussian noise models for state estimation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 1436–1443.

[13] A. De Maio and S. Lacroix, "Simultaneously learning corrections and error models for geometry-based visual odometry methods," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6536–6543, 2020.

[14] V. Peretroukhin and J. Kelly, "Dpc-net: Deep pose correction for visual localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2424–2431, 2018.

[15] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[16] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[17] Y. Almalioglu, M. R. U. Saputra, P. P. B. d. Gusmão, A. Markham, and N. Trigoni, "Ganvo: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 5474–5480.

[18] Y. Almalioglu, M. Turan, M. R. U. Saputra, P. P. De Gusmão, A. Markham, and N. Trigoni, "Selfvio: Self-supervised deep monocular visual–inertial odometry and depth estimation," *Neural Networks*, vol. 150, pp. 119–136, 2022.

[19] Y. Pan, W. Zhou, Y. Cao, and H. Zha, "Adaptive vio: Deep visual-inertial odometry with online continual learning," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 18 019–18 028.

[20] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.

[21] Z. Mei, A. Dixit, M. Booker, E. Zhou, M. Storey-Matsutani, A. Z. Ren, O. Shorinwa, and A. Majumdar, "Perceive with confidence: Statistical safety assurances for navigation with learning-based perception." SAGE Publications Sage UK: London, England, 2024, p. 02783649251378151.

[22] A. C. Stutts, D. Erricolo, T. Tulabandhula, and A. R. Trivedi, "Lightweight, uncertainty-aware conformalized visual odometry," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 7742–7749.

[23] G. Costante and M. Mancini, "Uncertainty estimation for data-driven visual odometry," *IEEE Transactions on Robotics*, vol. 36, no. 6, pp. 1738–1757, 2020.

[24] N. Yang, L. v. Stumberg, R. Wang, and D. Cremers, "D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1281–1292.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[26] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[27] A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Ré, "Hippo: Recurrent memory with optimal polynomial projections," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1474–1487, 2020.

[28] J. Luo, J. Cheng, Q. Xiang, J. Wu, R. Fan, X. Chen, and X. Tang, "Overlapmamba: A shift state space model for lidar-based place recognition," *IEEE Robotics and Automation Letters*, vol. 10, no. 8, pp. 8380–8387, 2025.

[29] X. Ma, C. Huang, X. Huang, and W. Wu, "Mamba-dqn: Adaptively tunes visual slam parameters based on historical observation dqn," *Applied Sciences*, vol. 15, no. 6, p. 2950, 2025.

[30] X. Jia, Q. Wang, A. Donat, B. Xing, G. Li, H. Zhou, O. Celik, D. Blessing, R. Lioutikov, and G. Neumann, "Mail: Improving imitation learning with mamba," *arXiv preprint arXiv:2406.08234*, 2024.

[31] J. Liu, M. Liu, Z. Wang, P. An, X. Li, K. Zhou, S. Yang, R. Zhang, Y. Guo, and S. Zhang, "Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 40 085–40 110, 2024.

[32] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 337–33 712.

[33] S. Herath, H. Yan, and Y. Furukawa, "Ronin: Robust neural inertial navigation in the wild: Benchmark, evaluations, new methods," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3146–3152.

[34] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to mlps," vol. 34, 2021, pp. 9204–9215.

[35] "ANT-X website." [Online]. Available: https://antx.it/

[36] A. Antonini, W. Guerra, V. Murali, T. Sayre-McCord, and S. Karaman, "The blackbird dataset: A large-scale dataset for uav perception in aggressive flight," in *International Symposium on Experimental Robotics*. Springer, 2018, pp. 130–139.

[37] D. Levi, L. Gispan, N. Giladi, and E. Fetaya, "Evaluating and calibrating uncertainty prediction in regression tasks," *Sensors*, vol. 22, no. 15, p. 5540, 2022.