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

Reformulating AI-based Multi-Object Relative State Estimation for Aleatoric Uncertainty-based Outlier Rejection of Partial Measurements

Thomas Jantos[1], Giulio Delama[1], Stephan Weiss[1] and Jan Steinbrener[1]

***Abstract*— Precise localization with respect to a set of objects of interest enables mobile robots to perform various tasks. With the rise of edge devices capable of deploying deep neural networks (DNNs) for real-time inference, it stands to reason to use artificial intelligence (AI) for the extraction of object-specific, semantic information from raw image data, such as the object class and the relative six degrees of freedom (6-DoF) pose. However, fusing such AI-based measurements in an Extended Kalman Filter (EKF) requires quantifying the DNNs' uncertainty and outlier rejection capabilities.**

This paper presents the benefits of reformulating the measurement equation in AI-based, object-relative state estimation. By deriving an EKF using the direct object-relative pose measurement, we can decouple the position and rotation measurements, thus limiting the influence of erroneous rotation measurements and allowing partial measurement rejection. Furthermore, we investigate the performance and consistency improvements for state estimators provided by replacing the fixed measurement covariance matrix of the 6-DoF object-relative pose measurements with the predicted aleatoric uncertainty of the DNN.

I. Introduction

Object-relative state estimation enables mobile robots to localize and navigate with respect to objects of interest, crucial for tasks such as object following and critical infrastructure inspection [1]. In previous work, Jantos et al. [2] introduced an extended Kalman filter (EKF)-based approach for object-relative state estimation, concurrently estimating the state of the mobile robot and the pose of the objects with respect to the navigation frame. An inertial measurement unit (IMU) is used as the propagation sensor and fused with artificial intelligence (AI)-based six degrees of freedom (6-DoF) object-relative pose measurements. The authors trained a deep learning (DL)-based object pose predictor to directly predict the 6-DoF pose of known objects from RGB images.

While AI-based methods excel in extracting semantic information from images, e.g., object class and 6-DoF object poses, they still can output erroneous predictions. Especially in the context of EKF-based sensor fusion, it is important to model the measurement covariance correctly, quantifying the sensor observations' uncertainty and ultimately the measurement's importance compared to the predicted state during the update step. In [3], Jantos et al. extended the DL-based object pose estimator for aleatoric uncertainty quantification, the uncertainty inherent in the input data, and replaced the fixed measurement covariance in an EKF with the dynamically predicted aleatoric uncertainty. They showed that the per-image and -object predicted aleatoric uncertainty captures the error characteristics of the full 6-DoF pose, even indicating ambiguous and difficult scenarios through increased uncertainty values. The latter gives ground for introducing aleatoric uncertainty-based outlier rejection (`AOR`) of measurements by determining suitable thresholds.

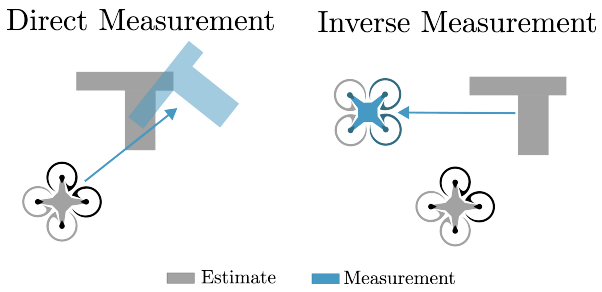

Fig. 1. Comparison of the 6-DoF object-relative pose measurements (`blue`) influence on the state estimation task (`gray`) for the proposed direct measurement and the inverse measurement [2] approaches. In both cases, the 6-DoF object pose is initially measured with respect to the mobile robot. By inverting the measurement, the relative pose measurement is expressed in the estimated object frame, thus rotation measurement errors lead to discrepancies between the measured and estimated mobile robot 6-DoF pose. In contrast, the proposed direct approach shifts the measurement error towards the object's estimate rather than the mobile robot's state. Moreover, our approach enables the decoupling of the translation and rotation measurement, allowing the rejection of either part.

In order to realize their object-relative EKF [2], it is necessary to invert the DL-based 6-DoF object pose measurements, i.e., instead of using the direct 6-DoF object pose expressed in the camera frame, the EKF measures the camera in the estimated object frame. While this allows for the straightforward derivation of the update Jacobians, it introduces a dependence on the estimated and measured object orientation, as visualized in Fig. 1 for a 2D example. Predicting the wrong object orientation also negatively impacts the measured position due to inverting the full 6-DoF pose measurement based on the estimated object frame.

In this work, we propose reformulating the EKF to use the direct 6-DoF object pose measurement, thus disentangling the object-relative position and orientation measurement. In addition to improved robustness against incorrect object orientation measurements, often caused by symmetric objects, this gives rise to the possibility of rejecting partial measurements, i.e., only the translation or rotation part of the full 6-DoF pose measurement. In the course of this paper, we will compare classical outlier rejection methods, i.e., χ^2-test, to the `AOR` approach introduced in [3], further underlining the benefits of AI-based uncertainty prediction for the state estimation task.

[1]The authors are with the Control of Networked Systems Group, University of Klagenfurt, 9020 Klagenfurt am Wörthersee, Austria `firstname.lastname@ieee.org`

This work was supported by the Christian Doppler Forschungsgesellschaft within the project AIONIC.

Our contributions are the following:

- Introducing a reformulated filter-based approach for object-relative state estimation that removes the need for measurement inversion.
- Decoupling object-relative position and rotation measurements allows for partial measurement rejection and increases robustness towards noisy rotation measurements.
- Extending aleatoric uncertainty-based outlier rejection to partial measurements for improved state estimation performance and consistency.
- Validating the proposed approach with several experiments and comparisons.

The remainder of the paper is organized as follows. We summarize the related work in Section II. In Section III, the reformulated EKF for object-relative state estimation, aleatoric uncertainty as dynamic measurement noise covariance, and partial measurement rejection are presented. The experiments and the corresponding results are discussed in Section IV. Finally, the paper is concluded in Section V.

II. Related Work

Classical approaches for mobile robot state estimation typically combine IMU measurements with sensor modalities such as GNSS [4], radars [5], or cameras [6] for visual inertial odometry (VIO). However, these approaches are usually unsuitable for object-relative state estimation and navigation as their measurements do not include the necessary semantic information about the objects of interest. Attaching cooperative markers to the objects allows for directly incorporating the 6-DoF object pose into the estimation process [7], [8], [9]. While straightforward and easy to use, it is often not feasible to equip objects of interest in the wild with markers.

Alternatively, object-relative state estimation can be realized by inferring the 6-DoF object pose from its geometric features. Thomas et al. [10] proposed a method for localization with respect to cylindrical objects, assuming a known radius. Loianno et al. [11] employed a parametric ellipse representation to detect objects in images and estimate their pose from visual attributes such as size, combined with camera parameters. Máthé et al. [12] applied classical techniques to recover full 6-DoF object poses for relative localization, while also exploring the use of machine learning to detect object presence in images.

Uncertainty quantification enables a better assessment of a deep neural network's (DNN) behavior and gives confidence to its predictions [13]. There are two main sources of uncertainty in DL: aleatoric uncertainty, inherent in the data, and epistemic uncertainty, which is the lack of knowledge in the model. Different approaches exist to model and utilize the uncertainty of 6-DoF pose predictors. Zorina et al. [14] empirically determine a 6-DoF pose uncertainty to be subsequently used in an object-relative SLAM approach. The covariance matrix assumes a single variance value for the rotational components and a shared variance value for the x and y components. In NVINS [15], a DL-based 6-DoF camera pose predictor is combined with IMU measurements in factor graph optimization. Besides the camera pose, the DNN also predicts its aleatoric and epistemic uncertainty in the form of a 3D Gaussian for the translation and a one-dimensional Langevin distribution of the rotation.

III. Method

In this section, we present our approach for 6-DoF object-relative state estimation. After introducing the notation used throughout this paper, the EKF formulation, the update step, and outlier rejection are presented.

A. Notation

Given two coordinate frames A and B, the homogeneous transformation $\mathbf{T}_{\mathsf{AB}}$ defines frame B with respect to frame A. The transformation consists of the translation $\mathbf{p}_{\mathsf{AB}} \in \mathbb{R}^3$ and the rotation $\mathbf{R}_{\mathsf{AB}} \in SO(3)$. The rotation can also be represented by the quaternion $\mathbf{q}_{\mathsf{AB}} = [\mathbf{q}_\mathbf{v}\ q_w]^T = [q_x\ q_y\ q_z\ q_w]^T$. The quaternion multiplication is represented by $\otimes$. An alternative is taking the matrix logarithm of the rotation to get an element of its Lie algebra and the axis-angle representation given by

$$Log(\mathbf{R}) = \theta[\mathbf{u}]_\times \tag{1}$$

$$\boldsymbol{\vartheta} = \theta\,\mathbf{u} \in \mathbb{R}^3 \quad , \tag{2}$$

with θ and $\mathbf{u}$ being the angle and the axis of rotation, and $[\cdot]_\times$ is the skew-symmetric operator as defined in [16]. The inverse rotation is expressed by the transposed rotation $\mathbf{R}_{\mathsf{AB}} = \mathbf{R}_{\mathsf{BA}}^T$ or the conjugate quaternion $\mathbf{q}_{\mathsf{AB}} = \mathbf{q}_{\mathsf{BA}}^{-1}$. The inverse position is then given by

$$\mathbf{p}_{\mathsf{BA}} = -\mathbf{R}_{\mathsf{BA}}\,\mathbf{p}_{\mathsf{AB}} \quad . \tag{3}$$

$\mathbf{I}_N$ and $\mathbf{0}_N$ refer to the identity and the null matrix in $\mathbb{R}^{N\times N}$.

B. EKF Formulation

In object-relative state estimation, the goal is to estimate the state of a mobile robot (I) in an arbitrary but fixed navigation frame (W) with respect to a set of objects of interest ($\mathsf{O_i}$). The DL-based object pose predictor provides the relative pose measurements between a camera (C) and $\mathsf{O_i}$. The different frames are visualized in Fig. 2. Depending on the total number of objects N in a scene, the full state vector $\mathbf{X}$ is then defined as:

$$\begin{aligned}\mathbf{X} = [&\mathbf{p}_{\mathsf{WI}}^T, \mathbf{v}_{\mathsf{WI}}^T, \mathbf{q}_{\mathsf{WI}}^T, \mathbf{b}_\omega^T, \mathbf{b}_\mathsf{a}^T,\\ &\mathbf{p}_{\mathsf{IC}}^T, \mathbf{q}_{\mathsf{IC}}^T, \mathbf{p}_{\mathsf{WO_0}}^T, \mathbf{q}_{\mathsf{WO_0}}^T, \ldots, \mathbf{p}_{\mathsf{WO_N}}^T, \mathbf{q}_{\mathsf{WO_N}}^T].\end{aligned} \tag{4}$$

The core states necessary for state propagation are the position $\mathbf{p}_{\mathsf{WI}}$ of the IMU, its velocity $\mathbf{v}_{\mathsf{WI}}$ and its orientation $\mathbf{q}_{\mathsf{WI}}$ as well as the gyroscopic bias $\mathbf{b}_\omega$ and the accelerometer bias $\mathbf{b}_\mathsf{a}$. The pose and velocity dynamics are given as [17]:

$$\dot{\mathbf{p}}_{\mathsf{WI}} = \mathbf{v}_{\mathsf{WI}} \tag{5}$$

$$\dot{\mathbf{v}}_{\mathsf{WI}} = \mathbf{R}_{\mathsf{WI}}(\mathbf{a}_\mathsf{m} - \mathbf{b}_\mathsf{a} - \mathbf{n}_\mathsf{a}) - \mathbf{g} \tag{6}$$

$$\dot{\mathbf{q}}_{\mathsf{WI}} = \frac{1}{2}\Omega(\boldsymbol{\omega} - \mathbf{b}_\omega - \mathbf{n}_\omega)\,\mathbf{q}_{\mathsf{WI}}, \tag{7}$$

where $\mathbf{a}_m$ is the measured acceleration in I, $\mathbf{n}_a$ is the accelerometer noise parameter, $\mathbf{g}$ is the gravity vector in W, $\boldsymbol{\omega}_b$

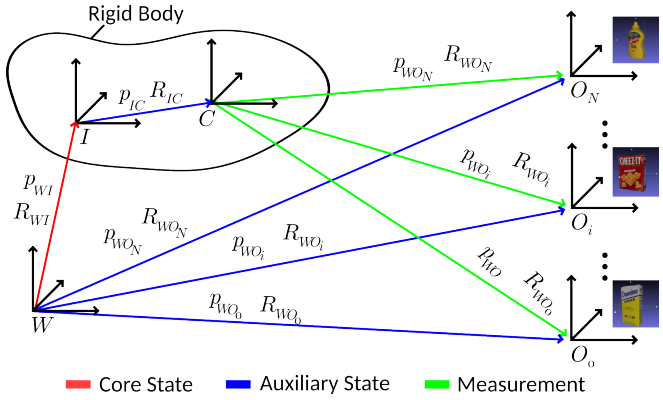

Fig. 2. Visualization of the different reference frames. The goal is to estimate the position and orientation of a rigid body (red) consisting of an IMU (I) and camera (C) in a fixed but arbitrary navigation frame (W) with respect to a set of objects of interest (O_i, blue). A DL-based pose predictor provides the object-relative 6-DoF pose measurements to the state estimator (green). The extrinsic calibration between the IMU and the camera is an additional auxiliary state in our formulation.

is the measured angular velocity in I, $\mathbf{n}_\omega$ is the gyroscopic noise parameter, and $\Omega(\boldsymbol{\omega})$ is the quaternion multiplication matrix of $\boldsymbol{\omega}$. The IMU biases are modeled as random walks.

Similar to [2], we estimate the pose of the objects in the navigation frame ($\mathbf{p}_{\mathsf{WO_i}}, \mathbf{q}_{\mathsf{WO_i}}$), modeled to stay consistent over time. In contrast and in tone with the direct measurement definition, we estimate the pose of the objects in the navigation frame, rather than the navigation frame with respect to the object frame. The extrinsic calibration between the IMU and the camera is also part of the state and can be estimated. In this work, we assume the calibration to be fixed and known.

The DL-based object pose predictor outputs a 6-DoF pose measurement for each image for each detected and known object. Given the current estimated mobile robot pose ($\mathbf{p}_{\mathsf{WI}}, \mathbf{q}_{\mathsf{WI}}$) and the measured relative object poses ($\hat{\mathbf{p}}_{\mathsf{CO_i}}, \hat{\mathbf{q}}_{\mathsf{CO_i}}$), the projected object frames are calculated with

$$\hat{\mathbf{p}}_{\mathsf{WO_i}} = \mathbf{p}_{\mathsf{WI}} + \mathbf{R}_{\mathsf{WI}}(\mathbf{p}_{\mathsf{IC}} + \mathbf{R}_{\mathsf{IC}}\hat{\mathbf{p}}_{\mathsf{CO_i}}) \tag{8}$$

$$\hat{\mathbf{R}}_{\mathsf{WO_i}} = \mathbf{R}_{\mathsf{WI}}\mathbf{R}_{\mathsf{IC}}\hat{\mathbf{R}}_{\mathsf{CO_i}} \quad . \tag{9}$$

A Hungarian algorithm, based on the Euclidean and geodesic distance and the object class, matches the predicted object frames to the currently estimated object frames ($\mathbf{p}_{\mathsf{WO_i}}, \mathbf{q}_{\mathsf{WO_i}}$). When viewed for the first time, Eqs. (8) and (9) are used to initialize a new object frame in the state $\mathbf{X}$. During the EKF update step, the position and orientation residual for each matched object $\mathsf{O_i}$ is calculated independently:

$$\begin{aligned}\tilde{\mathbf{z}}_{\mathbf{p}_{\mathsf{O_i}}} &= \hat{\mathbf{z}}_{\mathbf{p}_{\mathsf{O_i}}} - \mathbf{z}_{\mathbf{p}_{\mathsf{O_i}}} \\ &= \hat{\mathbf{p}}_{\mathsf{CO_i}} - (\mathbf{p}_{\mathsf{CI}} + \mathbf{R}_{\mathsf{CI}}(\mathbf{p}_{\mathsf{IW}} + \mathbf{R}_{\mathsf{IW}}\mathbf{p}_{\mathsf{WO_i}})) && (10)\\ &= \hat{\mathbf{p}}_{\mathsf{CO_i}} - (\mathbf{R}_{\mathsf{IC}}^T(-\mathbf{p}_{\mathsf{IC}} + \mathbf{R}_{\mathsf{WI}}^T(\mathbf{p}_{\mathsf{WO_i}} - \mathbf{p}_{\mathsf{WI}})) && (11)\end{aligned}$$

$$\tilde{\mathbf{z}}_{\mathbf{R}_{\mathsf{O_i}}} = 2\frac{\tilde{\mathbf{z}}_{\mathbf{q}_{\mathsf{v,O_i}}}}{\tilde{\mathbf{z}}_{q_{\mathsf{w,O_i}}}} \tag{12}$$

$$\begin{aligned}\tilde{\mathbf{z}}_{\mathbf{q}_{\mathsf{O_i}}} &= \mathbf{z}_{\mathbf{q}_{\mathsf{O_i}}}^{-1} \otimes \hat{\mathbf{z}}_{\mathbf{q}_{\mathsf{O_i}}} \\ &= (\mathbf{q}_{\mathsf{CI}} \otimes \mathbf{q}_{\mathsf{IW}} \otimes \mathbf{q}_{\mathsf{WO}})^{-1} \otimes \hat{\mathbf{q}}_{\mathsf{CO_i}} && (13)\\ &= (\mathbf{q}_{\mathsf{IC}}^{-1} \otimes \mathbf{q}_{\mathsf{WI}}^{-1} \otimes \mathbf{q}_{\mathsf{WO}})^{-1} \otimes \hat{\mathbf{q}}_{\mathsf{CO_i}} && (14)\end{aligned}$$

$$\tilde{\mathbf{z}}_{\mathsf{O_i}} = \begin{bmatrix}\tilde{\mathbf{z}}_{\mathbf{p}_{\mathsf{O_i}}} \\ \tilde{\mathbf{z}}_{\mathbf{q}_{\mathsf{O_i}}}\end{bmatrix} \quad . \tag{15}$$

The filter formulation using the inverse 6-DoF object pose measurement [2] suffers from the inverse position measurement being dependent on the currently measured object orientation, see Eq. (3). Hence, rotation measurement errors are also expressed in the position measurement, as visualized in Fig. 1.

In view of these residuals, it is necessary to determine derivation rules for expressions containing the transpose rotation, see Appendix:

$$\frac{\partial \mathbf{Q}\mathbf{R}^T\mathbf{S}}{\partial \mathbf{R}} = -\mathbf{S}^T\mathbf{R} \qquad \frac{\partial \mathbf{Q}\mathbf{R}^T\mathbf{v}}{\partial \mathbf{R}} = \mathbf{Q}[\mathbf{R}^T\mathbf{v}]_\times \quad , \tag{16}$$

where $\mathbf{Q}, \mathbf{R}, \mathbf{S} \in SO(3)$ and $\mathbf{v} \in \mathbb{R}^3$. Thus, the Jacobians for position $\mathbf{H}_\mathbf{p}$ and orientation $\mathbf{H}_\mathbf{R}$ for a single relative pose measurement matched to object $\mathsf{O_i}$ with respect to the states are [16]:

$$\begin{aligned}\mathbf{H}_{\mathbf{p},\mathbf{p}_{\mathsf{WI}}} &= -\mathbf{R}_{\mathsf{IC}}^T\mathbf{R}_{\mathsf{WI}}^T && (17)\\ \mathbf{H}_{\mathbf{p},\mathbf{R}_{\mathsf{WI}}} &= \mathbf{R}_{\mathsf{IC}}^T\left([\mathbf{R}_{\mathsf{WI}}^T\mathbf{p}_{\mathsf{WO}}]_\times - [\mathbf{R}_{\mathsf{WI}}^T\mathbf{p}_{\mathsf{WI}}]_\times\right) && (18)\\ \mathbf{H}_{\mathbf{p},\mathbf{p}_{\mathsf{IC}}} &= -\mathbf{R}_{\mathsf{IC}}^T && (19)\\ \mathbf{H}_{\mathbf{p},\mathbf{R}_{\mathsf{IC}}} &= -[\mathbf{R}_{\mathsf{IC}}^T\mathbf{p}_{\mathsf{IC}}]_\times \\ &\quad + [\mathbf{R}_{\mathsf{IC}}^T\mathbf{R}_{\mathsf{WI}}^T\mathbf{p}_{\mathsf{WO}}]_\times - [\mathbf{R}_{\mathsf{IC}}^T\mathbf{R}_{\mathsf{WI}}^T\mathbf{p}_{\mathsf{WI}}]_\times && (20)\\ \mathbf{H}_{\mathbf{p},\mathbf{p}_{\mathsf{WO_i}}} &= \mathbf{R}_{\mathsf{IC}}^T\mathbf{R}_{\mathsf{WI}}^T && (21)\\ \mathbf{H}_{\mathbf{p},\mathbf{R}_{\mathsf{WO_i}}} &= \mathbf{0}_3 && (22)\\ \mathbf{H}_{\mathbf{R},\mathbf{p}_{\mathsf{WI}}} &= \mathbf{0}_3 && (23)\\ \mathbf{H}_{\mathbf{R},\mathbf{R}_{\mathsf{WI}}} &= -\mathbf{R}_{\mathsf{WO}}^T\mathbf{R}_{\mathsf{WI}} && (24)\\ \mathbf{H}_{\mathbf{R},\mathbf{p}_{\mathsf{IC}}} &= \mathbf{0}_3 && (25)\\ \mathbf{H}_{\mathbf{R},\mathbf{R}_{\mathsf{IC}}} &= -\mathbf{R}_{\mathsf{WO}}^T\mathbf{R}_{\mathsf{WI}}\mathbf{R}_{\mathsf{IC}} && (26)\\ \mathbf{H}_{\mathbf{R},\mathbf{p}_{\mathsf{WO_i}}} &= \mathbf{0}_3 && (27)\\ \mathbf{H}_{\mathbf{R},\mathbf{R}_{\mathsf{WO_i}}} &= \mathbf{I}_3, && (28)\end{aligned}$$

where, e.g. $H_{\mathbf{p},\mathbf{p}_{WI}}$ only considers the part of the residual $\tilde{\mathbf{z}}_{\mathbf{p}_{\mathsf{O_i}}}$ that depends on the state $\mathbf{p}_{WI}$. The rest of the Jacobians are equal to $\mathbf{0}_3$. As relative pose measurements for different objects are independent of each other, the Jacobians for the other $(i \neq n)$ object-world states, i.e., $\mathbf{H}_{\mathbf{p},\mathbf{p}_{\mathsf{WO_n}}}$, $\mathbf{H}_{\mathbf{p},\mathbf{R}_{\mathsf{WO_n}}}$, $\mathbf{H}_{\mathbf{R},\mathbf{p}_{\mathsf{WO_n}}}$, $\mathbf{H}_{\mathbf{R},\mathbf{R}_{\mathsf{WO_n}}}$ are all equal to $\mathbf{0}_3$. For a single object $\mathsf{O_i}$,

the Jacobian is given by stacking the individual components:

$$\mathbf{H}_{\mathbf{p},\mathsf{O_i}} = [\mathbf{H}_{\mathbf{p},\,\mathbf{p}_{\mathsf{WI}}}, \mathbf{H}_{\mathbf{p},\,\mathbf{v}_{\mathsf{WI}}}, \mathbf{H}_{\mathbf{p},\mathbf{R}_{\mathsf{WI}}}, \mathbf{H}_{\mathbf{p},\,\mathbf{b}_{\boldsymbol{\omega}}}, \mathbf{H}_{\mathbf{p},\,\mathbf{b}_{\mathsf{a}}}, \mathbf{H}_{\mathbf{p},\,\mathbf{p}_{\mathsf{IC}}}, \mathbf{H}_{\mathbf{p},\mathbf{R}_{\mathsf{IC}}}, \mathbf{H}_{\mathbf{p},\,\mathbf{p}_{\mathsf{WO}_0}}, \mathbf{H}_{\mathbf{p},\mathbf{R}_{\mathsf{WO}_0}} \ldots, \mathbf{H}_{\mathbf{p},\,\mathbf{p}_{\mathsf{WO}_N}}, \mathbf{H}_{\mathbf{p},\mathbf{R}_{\mathsf{WO}_N}}] \tag{29}$$

$$\mathbf{H}_{\mathbf{R},\mathsf{O_i}} = [\mathbf{H}_{\mathbf{R},\,\mathbf{p}_{\mathsf{WI}}}, \mathbf{H}_{\mathbf{R},\,\mathbf{v}_{\mathsf{WI}}}, \mathbf{H}_{\mathbf{R},\mathbf{R}_{\mathsf{WI}}}, \mathbf{H}_{\mathbf{R},\,\mathbf{b}_{\boldsymbol{\omega}}}, \mathbf{H}_{\mathbf{R},\,\mathbf{b}_{\mathsf{a}}}, \mathbf{H}_{\mathbf{R},\,\mathbf{p}_{\mathsf{IC}}}, \mathbf{H}_{\mathbf{R},\mathbf{R}_{\mathsf{IC}}}, \mathbf{H}_{\mathbf{R},\,\mathbf{p}_{\mathsf{WO}_0}}, \mathbf{H}_{\mathbf{R},\mathbf{R}_{\mathsf{WO}_0}} \ldots, \mathbf{H}_{\mathbf{R},\,\mathbf{p}_{\mathsf{WO}_N}}, \mathbf{H}_{\mathbf{R},\mathbf{R}_{\mathsf{WO}_N}}] \tag{30}$$

$$\mathbf{H}_{\mathsf{O_i}} = \begin{bmatrix} \mathbf{H}_{\mathbf{p},\mathsf{O_i}} \\ \mathbf{H}_{\mathbf{R},\mathsf{O_i}} \end{bmatrix}. \tag{31}$$

Depending on the current image, as it can capture multiple objects of interest simultaneously, the update is conducted simultaneously. The final residual $\tilde{\mathbf{z}}$ and observation matrix $\mathbf{H}$ for the state update are determined by vertically stacking the residuals and Jacobians, see Eq. (15) and Eq. (31), for each matched object for the current image. As discussed in [2], one of the object frames ($\mathbf{p}_{\mathsf{WO_A}}$, $\mathbf{q}_{\mathsf{WO_A}}$) needs to be fixed, i.e., set its Jacobian $\mathbf{H}_{\mathsf{O_A}}$ to $\mathbf{0} \in \mathbb{R}^{6\times 21+3\cdot N}$, to prevent observability issues in pure object-relative state estimation.

C. Outlier Rejection

Reliably rejecting outlier measurements is crucial for EKF-based state estimation. A commonly used outlier rejection strategy is the χ^2-test. For a single measurement, the χ^2-test checks the statistical plausibility of its innovation covariance $\mathbf{S}$ with

$$\mathbf{S} = \mathbf{H}\mathbf{P}\mathbf{H}^T + \boldsymbol{\Sigma} \tag{32}$$

$$d^2 = \tilde{\mathbf{z}}^T \mathbf{S}^{-1} \tilde{\mathbf{z}} \quad , \tag{33}$$

where $\mathbf{P}$ is the state covariance, $\boldsymbol{\Sigma}$ is the measurement noise covariance matrix, and d^2 follows a χ^2 distribution with m degrees of freedom. Comparing it to an upper critical value, the statistical consistency of the measurement can be verified. Otherwise, the measurement is rejected. Eq. (32) indicates that the choice of $\boldsymbol{\Sigma}$ directly influences the χ^2-test. If $\boldsymbol{\Sigma}$ is chosen too conservatively, the χ^2 will reject too many measurements. On the other hand, large measurement uncertainty values will lead to the inclusion of outlier measurements. Ideally, $\boldsymbol{\Sigma}$ should capture the uncertainty of the measurement correctly, a task often not straightforward and requiring time-consuming engineering and tuning efforts.

We assume no cross-correlations between the individual translation and rotation components, with the latter expressed in the vector space tangent to the measured rotation [16]. The full measurement noise covariance matrix of the 6-DoF pose measurement expressed in the camera frame is given by

$$\boldsymbol{\Sigma}_{\mathbf{p},\mathsf{CO_i}} = \begin{pmatrix} \hat{\sigma}_x^2 & 0 & 0 \\ 0 & \hat{\sigma}_y^2 & 0 \\ 0 & 0 & \hat{\sigma}_z^2 \end{pmatrix} \tag{34}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\vartheta},\mathsf{CO_i}} = \begin{pmatrix} \hat{\sigma}_{\vartheta_1}^2 & 0 & 0 \\ 0 & \hat{\sigma}_{\vartheta_2}^2 & 0 \\ 0 & 0 & \hat{\sigma}_{\vartheta_3}^2 \end{pmatrix} \tag{35}$$

$$\boldsymbol{\Sigma}_{\mathsf{CO_i}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{p},\mathsf{CO_i}} & \mathbf{0}_{3\times 3} \\ \mathbf{0}_{3\times 3} & \boldsymbol{\Sigma}_{\boldsymbol{\vartheta},\mathsf{CO_i}} \end{pmatrix} \quad . \tag{36}$$

Similar to other measurement modalities, DL-based measurements can suffer from inaccuracies and outliers, due to, e.g., out-of-distribution data or ambiguous viewpoints in the case of 6-DoF object pose estimation. Hence, determining $\boldsymbol{\Sigma}$ for a deep neural network requires sophisticated methods. However, a (pre-trained) 6-DoF object pose predictor extended for aleatoric uncertainty prediction can capture its error characteristics [3], thus serving as a dynamic measurement noise covariance matrix. Moreover, increased aleatoric uncertainty levels indicate ambiguous and challenging situations, ultimately introducing the concept of aleatoric uncertainty-based outlier rejection `AOR`.

We adopt the aleatoric uncertainty prediction principle to our direct measurement filter formulation. Given an input image, the 6-DoF pose predictor outputs the translation $\hat{\mathbf{p}}_{\mathsf{CO_i}}$, the rotation $\hat{\boldsymbol{\vartheta}}_{\mathsf{CO_i}}$ and corresponding aleatoric uncertainties ($\hat{\boldsymbol{\Sigma}}_{\mathbf{p},\mathsf{CO_i}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\vartheta},\mathsf{CO_i}}$), modeled to represent Gaussian distributions with $\mathcal{N}(\hat{\mathbf{p}}_{\mathsf{CO_i}}, \hat{\boldsymbol{\Sigma}}_{\mathbf{p},\mathsf{CO_i}})$ and $\mathcal{N}(\hat{\boldsymbol{\vartheta}}_{\mathsf{CO_i}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\vartheta},\mathsf{CO_i}})$. The predicted aleatoric measurement covariance matrices can be used as our state estimator's measurement noise covariance matrix.

As the measurement noise covariance is expressed in the camera's frame and due to our direct filter formulation, the need for inversion of the measurement covariance matrix is eliminated. The inverse measurement covariance matrix depends on the predicted rotation:

$$\boldsymbol{\Sigma}_{\mathbf{y},\mathsf{O_iC}} = \hat{\mathbf{R}}_{\mathsf{O_iC}} \boldsymbol{\Sigma}_{\mathbf{y},\mathsf{CO_i}} \hat{\mathbf{R}}_{\mathsf{O_iC}}^T, \quad \mathbf{y} \in \{\mathbf{p}, \boldsymbol{\vartheta}\}. \tag{37}$$

Similar to the inversion of the position measurement, erroneous rotation measurements negatively impact the EKF steps involving the measurement noise covariance matrix. Once again, our direct filter formulation enables the individual consideration of the position and rotation measurement, also for outlier measurement rejection. With partial measurement rejection in place, the EKF can incorporate more information into the estimation process.

We expand `AOR` for partial measurement rejection by comparing the positional and rotational uncertainties to a fixed position and rotation uncertainty threshold. If the predicted uncertainty of a single component is above the threshold, the respective position or rotation measurement is rejected. Independent of the outlier rejection method, the rejection of the full or partial measurement requires the appropriate removal and restacking of the residual, Jacobians, and measurement noise covariance matrix, see Eqs. (15), (31) and (36).

Fig. 3. `Left`: Example of a synthetic training image containing the subset of the YCB-V objects and distractor objects. `Right`: Example image from an evaluation trajectory with a challenging object constellation and occlusion.

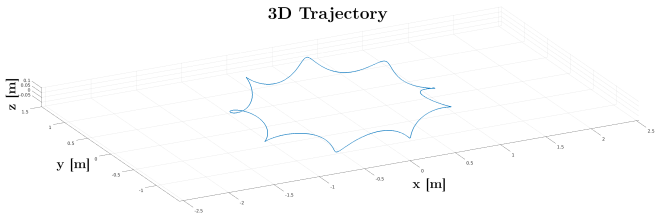

Fig. 4. Example trajectory used for evaluation with varying distances and viewing angles of the objects.

IV. Experiments & Results

In this section, the conducted experiments are presented and the corresponding results are discussed. Before introducing the 6-DoF pose prediction framework, we go into the dataset used for the experiments. Afterwards, we show the strengths of our filter formulation in comparison to the EKF presented in [2]. Finally, the influence of aleatoric uncertainty and partial measurement rejection for the object-relative state estimation task is highlighted.

A. Dataset

The experiments focus on a subset of the commonly used YCB-V object set [18]. This subset covers a wide range of object attributes, ranging from well-textured (cracker box, sugar box, power drill) to having ambiguous viewpoints (mustard bottle, bleach cleanser) to being almost symmetrical (scissors, mug). This work focuses on synthetic data as it offers controlled conditions with perfect annotation to perform meaningful analysis of the proposed approach.

NVIDIA Omniverse IsaacSim allows for the efficient generation and annotation of photorealistic RGB images, given the already included 3D models of the objects. To train and validate the 6-DoF object pose estimator, we generate 100,000 and 3000 synthetic images, respectively. To evaluate the object-relative state estimator, we generate ten diverse physically feasible trajectories with different objects and constellations, and varying distances to the objects. The data includes synthetic IMU data at a rate of 200 Hz and corresponding RGB images recorded with 20 FPS. Example synthetic images are shown in Fig. 3.

B. DL 6-DoF Object Pose Prediction Framework

We follow previous work on AI-based object-relative state estimation [3] and use PoET [19] for 6-DoF object pose and aleatoric uncertainty prediction. The object detection backbone is a Scaled-YOLOv4 [20], and the transformer consists of five encoder and decoder layers and 16 attention heads. The translation, rotation, and aleatoric uncertainty heads are simple multi-layer perceptrons with three layers and an output dimension of three. We first train PoET for 50 epochs before calibrating the aleatoric uncertainty heads for 10 additional epochs, while freezing the remainder of the network.

TABLE I

Noise Analysis for the direct measurement filter formulation. We report the RMSE in [m]

$\boldsymbol{\Sigma}_{\mathbf{p}}/\boldsymbol{\Sigma}_{\vartheta}$	1°	5°	10°	20°
1cm	0.073 ± 0.030	0.285 ± 0.123	0.377 ± 0.183	0.370 ± 0.204
5cm	0.111 ± 0.033	0.327 ± 0.151	0.475 ± 0.283	0.719 ± 0.422
10cm	0.179 ± 0.061	0.343 ± 0.115	0.481 ± 0.270	0.789 ± 0.529
30cm	0.477 ± 0.191	0.622 ± 0.198	0.737 ± 0.267	1.075 ± 0.496

TABLE II

Noise Analysis for the inverse measurement filter formulation. We report the RMSE in [m]

$\boldsymbol{\Sigma}_{\mathbf{p}}/\boldsymbol{\Sigma}_{\vartheta}$	1°	5°	10°	20°
1cm	0.082 ± 0.028	0.944 ± 0.893	2.197 ± 0.419	2.538 ± 0.449
5cm	0.113 ± 0.033	0.420 ± 0.163	1.686 ± 0.886	2.480 ± 0.449
10cm	0.180 ± 0.061	0.410 ± 0.130	1.063 ± 0.637	2.290 ± 0.550
30cm	0.477 ± 0.191	0.630 ± 0.190	0.979 ± 0.401	1.933 ± 0.578

C. Filter Comparison

In order to compare the proposed object-relative filter formulation to the inverse formulation presented in [2], we implement both EKFs using MaRS [21] and report their position root mean square error (RMSE) in meters. As pointed out in Section III-B, the main benefit of our approach is the decoupling of position and rotation of the relative 6-DoF pose measurement. Moreover, by not relying on inverting the full measurement, the position measurement is not negatively impacted by wrong rotation measurements, as shown in Fig. 1.

For an unbiased comparison, we generate perfect synthetic IMU and relative pose measurement data and perturb the translation and rotation with Gaussian noise with different standard deviations. The following noise values are chosen:

$$\sigma_x = \sigma_y = \sigma_z \in \{0.01, 0.05, 0.1, 0.2, 0.3\}[\mathrm{m}]$$
$$\sigma_{\vartheta_1} = \sigma_{\vartheta_2} = \sigma_{\vartheta_3} \in \{0.0175, 0.0875, 0.175, 0.35\}[\mathrm{rad}] \quad .$$

For each noise value pair, 100 Monte Carlo simulations are performed. Please note that the EKF is provided with a measurement noise covariance corresponding to the perturbation noise. The mean position RMSE and standard deviation are reported in Tables I and II.

In the presence of minimal rotational perturbation, i.e., 1°, both filter formulations achieve an almost identical performance independent of the translation perturbation. Given perfect measurement data with no perturbation, both filters achieve the same RMSE of 1.3cm. Increasing the rotation perturbation gradually worsens the performance of both approaches. However, the inverse filter formulation is more affected. This highlights that removing the inversion

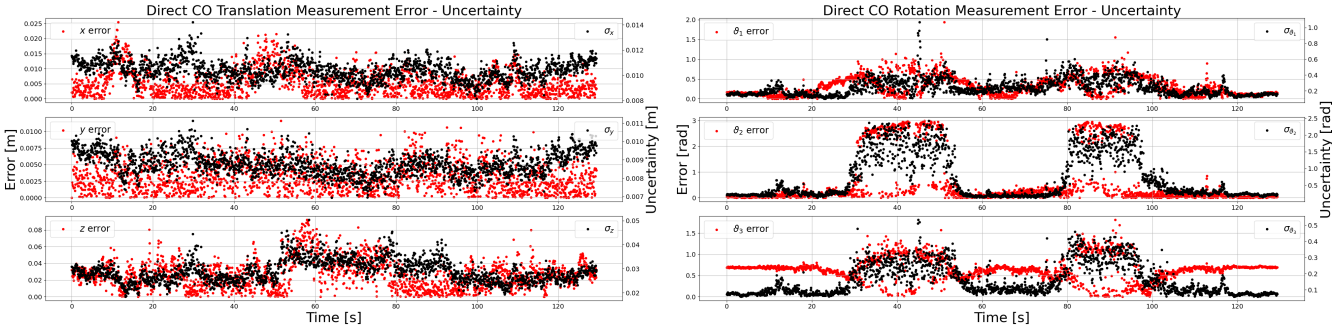

Fig. 5. **Direct Measurement (ours)**: Comparison of the absolute translation error (`left`, `red`) and rotation error (`right`, `red`) to the estimated aleatoric uncertainty (`black`) across the whole trajectory for the mug. Note the different scales in the plots' axes and across the plots.

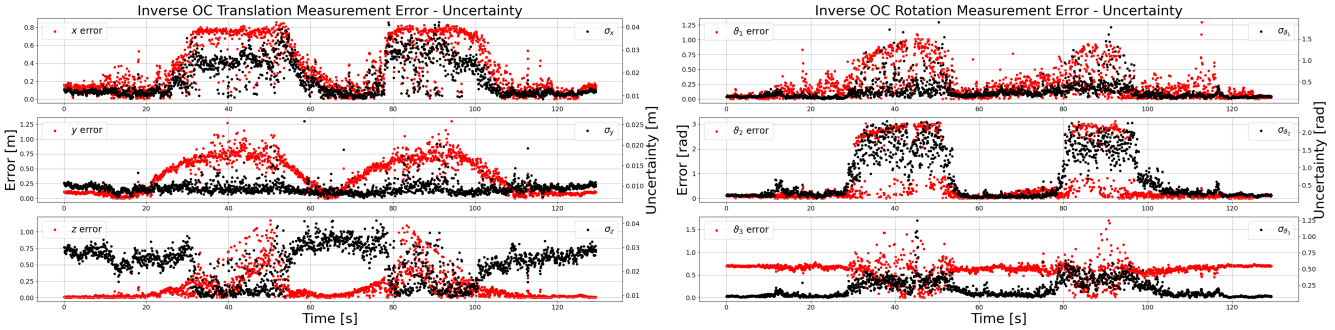

Fig. 6. **Inverse Measurement ([2])**: Comparison of the absolute translation error (`left`, `red`) and rotation error (`right`, `red`) to the estimated aleatoric uncertainty (`black`) across the whole trajectory for the mug. Note the different scales in the plots' axes and across the plots.

of the 6-DoF drastically benefits the state estimation performance as it reduces the influence of the rotation component. Interestingly, while increased translation measurement perturbation worsens our state estimator's performance, it has a soothing effect for the filter using the inverse measurement. This is mainly due to the increased measurement noise, which smooths out the measurements. The influence of the inversion of the measurement and its noise covariance matrix is further visualized in Figs. 5 and 6. The plots show the predicted 6-DoF pose measurements and corresponding aleatoric uncertainties for the mug. When viewed from an ambiguous viewpoint, i.e., the occlusion of the handle ($\sim$40s & $\sim$90s), the rotational error and the aleatoric uncertainty are increased. While the direct translation measurement is unaffected, the rotational errors are additionally reflected in the inverse translation measurement. Besides the inverted position covariance matrix not matching the error characteristics, the numeric values do not adequately capture the error. This leads to either including outlier measurements with high confidence into the state estimation process or to the rejection of originally accurate measurements. Inverting the measurement noise covariance matrix, according to Eq. (37), will introduce cross-covariances between the individual position and rotation components, but not between the position and rotation measurements.

D. Partial Measurement Rejection

To highlight the benefit of aleatoric uncertainty and partial measurement rejection for object-relative state estimation, we compare different combinations of measurement noise covariances and outlier rejection methods. In each case, the predicted relative pose measurements of PoET are used. The performance of the state estimator is measured in terms of the RMSE in position and orientation, the maximum position error, and the average normalized estimation error squared (ANEES) for the position and orientation. For better readability, the ANEES is further normalized by the degrees of freedom of the variable, i.e., optimal values are close to 1.

First, we use a fixed measurement noise covariance matrix (`F`). It is determined by calculating the average translation and rotation error across all objects for the validation dataset, resulting in the following measurement noise values:

$$\sigma_x = \sigma_y = \sigma_z = 0.04[\text{m}]$$
$$\sigma_{\vartheta_1} = \sigma_{\vartheta_2} = \sigma_{\vartheta_3} = 0.628[\text{rad}] \quad .$$

Second, we use the per-image and -object predicted aleatoric uncertainty as dynamic measurement noise covariance (`U`). In terms of outlier rejection, we employ the χ^2-test for statistical outlier rejection, aleatoric uncertainty-based outlier rejection (`AOR`), and their counterparts for partial measurement rejection (χ^2`P`, `AORP`). When determining suitable thresholds for `AOR(P)`, several aspects need to be taken into consideration, such as applicability to different objects and scenarios, and finding the balance between incorporating outlier measurements and rejecting too many measurements, favoring dead reckoning. The experiments conducted in Section IV-C show that measurements with a perturbation noise of 0.1m and 0.175rad already lead to a deteriorating performance. Hence, we choose these values as our thresholds for

AORP. As AOR rejects the whole measurement, we choose more conservative threshold values, i.e., 0.15m and 0.35rad, to prevent the rejection of too many measurements and dead reckoning. Please note that the combination of F and AOR(P) is not investigated as thresholding would either accept or reject all measurements. The results for the ten trajectories and the mean are reported in Tables III to VII. Bold values indicate the best-performing uncertainty and outlier rejection combination, while a dash (—) indicates that the state estimator diverged for this particular trajectory. These runs are excluded from the mean calculation, skewing the results for the specific combination, and are marked in italic.

On average, using aleatoric uncertainty (U) outperforms its fixed measurement noise covariance matrix (F) counterparts. Besides improved estimation error metrics, aleatoric uncertainty leads to a more consistent estimator, as indicated by the normalized ANEES values. The state estimator's covariance better captures the estimator's error. Partial aleatoric uncertainty-based outlier rejection (U + AORP) outperforms all other methods across every metric. For example, visualized in Fig. 5, the symmetry of the mug causes erroneous rotation predictions, thus motivating partial measurement rejection. Assuming a single object scenario, rejection of the full object-relative measurement leads to prolonged segments of only IMU propagation, leading to dead reckoning and ultimately divergence. However, by only rejecting the rotation measurement, we can still perform EKF updates and include more information in the estimation process. Comparing AORP to AOR highlights that partial measurement rejection prevents the filter from diverging. Disregarding trajectory 3, χ^2P leads to improved performance over χ^2, independent of the underlying measurement noise covariance method. This underlines the need for our newly proposed filter formulation using direct measurements to treat the position and rotation measurements individually and to perform partial measurement outlier rejection. The results show the importance of outlier rejection for AI-based object-relative pose measurements for improved performance and reduced divergence. In contrast to χ^2-based outlier rejection, the AOR approach is likelier to fail due to its thresholding, as an unsuitable choice leads to increased measurement rejections.

V. Conclusion

In this paper, we presented a novel formulation for EKF-based object-relative state estimation and highlighted the benefits of DL-based, dynamic aleatoric uncertainty for state estimation. By deriving the update equations for the direct 6-DoF pose measurement, we can consider the position and rotation measurement independently, reducing the deteriorating behavior of erroneous rotation measurements and enabling partial measurement rejection. Compared to the inverse filter formulation, the Monte Carlo simulations showed the improved performance of our approach, while indicating that both filter formulations perform equivalently well in the absence of noise. Furthermore, we cemented using aleatoric uncertainty as a dynamic measurement noise covariance for DL-based 6-DoF object pose measurements.

TABLE III
RMSE Position [m]

	F	+ χ^2	+ χ^2P	U	+ χ^2	+ χ^2P	+ AOR	+ AORP
1	—	0.082	0.076	0.056	0.058	0.058	**0.051**	0.061
2	0.747	0.116	0.111	0.295	0.091	0.088	—	**0.079**
3	—	0.118	0.112	—	**0.093**	—	0.141	0.098
4	0.103	**0.086**	0.087	0.109	0.108	0.107	0.115	0.104
5	0.553	0.119	0.123	0.281	0.113	0.109	0.154	**0.090**
6	0.741	0.195	0.192	0.595	0.204	0.207	—	**0.114**
7	—	0.130	0.133	0.175	0.166	0.158	0.161	**0.082**
8	—	0.132	0.140	—	0.137	0.140	0.159	**0.123**
9	0.728	0.726	0.582	0.622	0.444	**0.101**	—	0.123
10	0.770	0.230	0.200	—	0.119	**0.116**	0.281	0.195
Mean	*0.607*	0.193	0.176	*0.305*	0.153	*0.120*	*0.152*	**0.107**

TABLE IV
RMSE Orientation [°]

	F	+ χ^2	+ χ^2P	U	+ χ^2	+ χ^2P	+ AOR	+ AORP
1	—	4.30	4.06	3.28	3.29	3.23	**3.00**	3.35
2	44.28	6.4	6.17	16.43	5.02	4.83	—	**4.21**
3	—	6.69	6.26	—	**5.30**	—	7.97	5.53
4	5.64	**4.59**	4.64	6.08	5.99	5.97	6.41	5.77
5	29.53	5.84	5.96	14.12	5.02	4.84	7.04	**3.93**
6	38.32	9.37	9.19	30.59	9.84	9.98	—	**5.34**
7	—	6.61	6.81	10.08	9.1	8.69	8.88	**4.55**
8	—	5.33	5.75	—	5.63	5.74	7.27	**5.11**
9	43.63	42.49	34.04	36.69	25.70	**5.17**	—	6.32
10	49.43	13.07	11.16	—	6.28	**6.09**	15.49	10.98
Mean	*35.14*	10.47	9.40	*16.75*	8.12	*6.06*	*8.01*	**5.51**

TABLE V
Maximum Position Error [m]

	F	+ χ^2	+ χ^2P	U	+ χ^2	+ χ^2P	+ AOR	+ AORP
1	—	0.242	0.225	0.339	0.140	0.139	0.148	**0.133**
2	2.677	0.253	0.242	1.412	0.224	0.218	—	**0.192**
3	—	0.245	0.269	—	0.217	—	0.640	**0.241**
4	0.297	**0.219**	0.223	0.240	0.239	0.237	0.277	0.271
5	1.827	0.214	0.215	0.712	0.217	0.212	0.337	**0.190**
6	1.846	0.527	0.505	1.87	0.406	0.414	—	**0.257**
7	—	0.301	0.316	0.728	0.305	0.310	0.277	**0.169**
8	—	0.334	0.370	—	0.297	0.306	0.293	**0.268**
9	2.231	1.309	1.184	1.253	0.743	0.317	—	**0.262**
10	2.712	0.525	0.591	—	0.372	**0.273**	0.938	0.442
Mean	*1.932*	0.417	0.414	*0.936*	0.316	*0.270*	*0.416*	**0.242**

TABLE VI
Normalized ANEES for Position

	F	+ χ^2	+ χ^2P	U	+ χ^2	+ χ^2P	+ AOR	+ AORP
1	—	3.16	**2.84**	6.52	6.54	6.60	4.53	7.04
2	174.3	2.35	**2.23**	51.00	4.11	3.93	—	3.64
3	—	**1.83**	2.42	—	4.00	—	9.77	4.21
4	2.28	**1.57**	1.62	6.61	6.42	6.38	6.54	2.80
5	55.00	2.62	2.97	22.31	4.16	4.12	2.63	**2.14**
6	96.52	2.78	2.71	151.0	7.41	7.71	—	**1.01**
7	—	**4.84**	5.20	33.65	31.57	29.27	31.60	6.44
8	—	4.67	5.29	—	6.21	6.68	4.72	**4.46**
9	346.1	214.7	126.9	402.9	125.9	**7.55**	—	10.90
10	325.8	14.75	13.28	—	**6.43**	7.64	8.26	17.13
Mean	*166.7*	25.33	16.54	*96.28*	20.28	*8.88*	*6.54*	**5.98**

TABLE VII
Normalized ANEES for Orientation

	F	+ χ^2	+ χ^2P	U	+ χ^2	+ χ^2P	+ AOR	+ AORP
1	—	2.18	**1.91**	6.33	5.73	5.65	4.29	5.74
2	110.0	1.51	**1.41**	54.33	3.43	3.19	—	2.27
3	—	**1.27**	1.61	—	3.47	—	7.54	3.24
4	1.27	0.79	**0.81**	5.11	4.94	4.93	5.03	2.03
5	56.67	1.23	1.29	23.49	2.41	2.30	1.95	**1.08**
6	43.12	1.49	**1.43**	258.7	4.56	4.64	—	0.34
7	—	**2.83**	3.14	28.03	26.35	24.07	26.27	5.26
8	—	0.70	0.83	—	2.78	3.00	3.15	**0.87**
9	317.3	171.0	104.54	433.1	111.4	**5.30**	—	7.51
10	1265	10.60	8.56	495.2	**4.35**	5.06	11.93	12.01
Mean	*299.0*	19.36	12.55	*163.0*	16.94	*6.46*	*8.59*	**4.04**

Besides improving the state estimator's performance and consistency, it removes the need for the time-intensive engineering of a fixed measurement noise covariance that can not capture the situational error characteristics of the AI-based object pose predictor. Combining our novel filter formulation with aleatoric uncertainty-based outlier rejection for partial measurements (AORP) leads to drastic improvements. Future work will investigate the possibility of rejecting individual components of the 6-DoF pose measurement and automatic methods for determining suitable AOR(P) thresholds.

APPENDIX

Following the definitions and notation style from [16], we define $\mathbf{Q}_\theta, \mathbf{R}_\phi, \mathbf{S}_\eta \in SO(3)$ and $\mathbf{v} \in \mathbb{R}^3$. The following equality holds:

$$\mathbf{R}_\phi Exp(\delta\theta)\mathbf{R}_\phi^T = Exp(\mathbf{R}_\phi \delta\theta) \quad . \tag{38}$$

The derivatives can then be derived as:

$$\frac{\delta \mathbf{Q}\mathbf{R}^T\mathbf{S}}{\delta \mathbf{R}} = \frac{\delta \mathbf{Q}_\theta \mathbf{R}_\phi^T \mathbf{S}_\eta}{\delta \phi} \tag{39}$$

$$= \lim_{\delta\phi\to 0} \frac{1}{\delta\phi}((\mathbf{Q}_\theta(\mathbf{R}_\phi Exp(\delta\phi))^T\mathbf{S}_\eta) \ominus (\mathbf{Q}_\theta\mathbf{R}_\phi^T\mathbf{S}_\eta)) \tag{40}$$

$$= \lim_{\delta\phi\to 0} \frac{1}{\delta\phi} Log[(\mathbf{Q}_\theta\mathbf{R}_\phi^T\mathbf{S}_\eta)^T\mathbf{Q}_\theta(\mathbf{R}_\phi Exp(\delta\phi))^T\mathbf{S}_\eta] \tag{41}$$

$$= \lim_{\delta\phi\to 0} \frac{1}{\delta\phi} Log[\mathbf{S}_\eta^T\mathbf{R}_\phi\mathbf{Q}_\theta^T\mathbf{Q}_\theta Exp(\delta\phi)^T\mathbf{R}_\phi^T\mathbf{S}_\eta] \tag{42}$$

$$= \lim_{\delta\phi\to 0} \frac{1}{\delta\phi} Log[\mathbf{S}_\eta^T\mathbf{R}_\phi Exp(-\delta\phi)\mathbf{R}_\phi^T\mathbf{S}_\eta] \tag{43}$$

$$= \lim_{\delta\phi\to 0} \frac{1}{\delta\phi} Log[Exp(-\mathbf{S}_\eta^T\mathbf{R}_\phi\delta\phi)] \tag{44}$$

$$= \lim_{\delta\phi\to 0} \frac{1}{\delta\phi} -\mathbf{S}_\eta^T\mathbf{R}_\phi\delta\phi = -\mathbf{S}_\eta^T\mathbf{R}_\phi \tag{45}$$

$$\frac{\partial \mathbf{Q}\mathbf{R}^T\mathbf{v}}{\partial \mathbf{R}} = \lim_{\delta\phi\to 0} \frac{\mathbf{Q}\mathbf{R}\{\phi+\delta\phi\}^T\mathbf{v} - \mathbf{Q}\mathbf{R}^T\mathbf{v}}{\delta\phi} \tag{46}$$

$$= \lim_{\delta\phi\to 0} \frac{1}{\delta\phi}(\mathbf{Q}(\mathbf{R}Exp(\mathbf{J}_r(\phi)\delta\phi)^T - \mathbf{Q}\mathbf{R}^T)\mathbf{v} \tag{47}$$

$$= \lim_{\delta\phi\to 0} \frac{1}{\delta\phi}(\mathbf{Q}Exp(-\mathbf{J}_r(\phi)\delta\phi)\mathbf{R}^T - \mathbf{Q}\mathbf{R}^T)\mathbf{v} \tag{48}$$

$$= \lim_{\delta\phi\to 0} \frac{1}{\delta\phi}\mathbf{Q}(Exp(-\mathbf{J}_r(\phi)\delta\phi) - \mathbf{I}_3)\mathbf{R}^T\mathbf{v} \tag{49}$$

$$\approx \lim_{\delta\phi\to 0} \frac{1}{\delta\phi}\mathbf{Q}(\mathbf{I}_3 - [\mathbf{J}_r(\phi)\delta\phi]_\times - \mathbf{I}_3)\mathbf{R}^T\mathbf{v} \tag{50}$$

$$= \lim_{\delta\phi\to 0} \frac{1}{\delta\phi} -\mathbf{Q}[\mathbf{J}_r(\phi)\delta\phi]_\times\mathbf{R}^T\mathbf{v} \tag{51}$$

$$= \lim_{\delta\phi\to 0} \frac{1}{\delta\phi}\mathbf{Q}[\mathbf{R}^T\mathbf{v}]_\times\mathbf{J}_r(\phi)\delta\phi \tag{52}$$

$$= \mathbf{Q}[\mathbf{R}^T\mathbf{v}]_\times\mathbf{J}_r(\phi) \tag{53}$$

REFERENCES

[1] T. Jantos, M. Scheiber, C. Brommer, E. Allak, S. Weiss, and J. Steinbrener, "Aivio: Closed-loop, object-relative navigation of uavs with ai-aided visual inertial odometry," *IEEE Robotics and Automation Letters*, vol. 9, no. 12, pp. 10 764–10 771, 2024.

[2] T. Jantos, C. Brommer, E. Allak, S. Weiss, and J. Steinbrener, "Ai-based multi-object relative state estimation with self-calibration capabilities," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 2789–2795.

[3] T. Jantos, S. Weiss, and J. Steinbrener, "Aleatoric uncertainty from ai-based 6d object pose predictors for object-relative state estimation," *IEEE Robotics and Automation Letters*, pp. 1–8, 2025.

[4] M. Scheiber, A. Fornasier, C. Brommer, and S. Weiss, "Revisiting multi-gnss navigation for uavs–an equivariant filtering approach," in *2023 21st International Conference on Advanced Robotics (ICAR)*. IEEE, 2023, pp. 134–141.

[5] J. Michalczyk, M. Scheiber, R. Jung, and S. Weiss, "Radar-inertial odometry for closed-loop control of resource-constrained aerial platforms," in *2023 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, 2023, pp. 61–68.

[6] A. Fornasier, P. van Goor, E. Allak, R. Mahony, and S. Weiss, "Msceqf: A multi state constraint equivariant filter for vision-aided inertial navigation," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 731–738, 2024.

[7] T.-D. Do, N. Xuan-Mung, H. Jeong, Y.-S. Lee, C.-W. Sung, and S. K. Hong, "Vision-based autonomous perching of quadrotors on horizontal surfaces," in *2023 International Conference on System Science and Engineering (ICSSE)*. IEEE, 2023, pp. 352–357.

[8] L. Teixeira, F. Maffra, M. Moos, and M. Chli, "Vi-rpe: Visual-inertial relative pose estimation for aerial vehicles," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 2770–2777, 2018.

[9] K. Shen, Y. Zhuang, Y. Chen, S. Zuo, and T. Liu, "Aeronet: An efficient relative localization and object detection network for cooperative aerial-ground unmanned vehicles," *Pattern Recognition Letters*, vol. 171, pp. 28–37, 2023.

[10] J. Thomas, G. Loianno, K. Daniilidis, and V. Kumar, "Visual servoing of quadrotors for perching by hanging from cylindrical objects," *IEEE robotics and automation letters*, vol. 1, no. 1, pp. 57–64, 2015.

[11] G. Loianno, V. Spurny, J. Thomas, T. Baca, D. Thakur, D. Hert, R. Penicka, T. Krajnik, A. Zhou, A. Cho, *et al.*, "Localization, grasping, and transportation of magnetic objects by a team of mavs in challenging desert-like environments," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1576–1583, 2018.

[12] K. Máthé, L. Buşoniu, L. Barabás, C.-I. Iuga, L. Miclea, and J. Braband, "Vision-based control of a quadrotor for an object inspection scenario," in *2016 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE, 2016, pp. 849–857.

[13] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, *et al.*, "A survey of uncertainty in deep neural networks," *Artificial Intelligence Review*, vol. 56, no. Suppl 1, pp. 1513–1589, 2023.

[14] K. Zorina, V. Priban, M. Fourmy, J. Sivic, and V. Petrik, "Temporally consistent object 6d pose estimation for robot control," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 56–63, 2025.

[15] J. Han, L. L. Beyer, G. V. Cavalheiro, and S. Karaman, "Nvins: Robust visual inertial navigation fused with nerf-augmented camera pose regressor and uncertainty quantification," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[16] J. Sola, "Quaternion kinematics for the error-state kalman filter," *arXiv preprint arXiv:1711.02508*, 2017.

[17] S. Weiss, D. Scaramuzza, and R. Siegwart, "Monocular-slam–based navigation for autonomous micro helicopters in gps-denied environments," *Journal of Field Robotics*, vol. 28, no. 6, pp. 854–874, 2011.

[18] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," in *Robotics: Science and Systems (RSS)*, 2018.

[19] T. Jantos, M. A. Hamdad, W. Granig, S. Weiss, and J. Steinbrener, "PoET: Pose Estimation Transformer for Single-View, Multi-Object 6D Pose Estimation," in *Proceedings of the 6th Conference on Robot Learning*. PMLR, 2023.

[20] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 13 029–13 038.

[21] C. Brommer, R. Jung, J. Steinbrener, and S. Weiss, "MaRS: A modular and robust sensor-fusion framework," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 359–366, 2020.