

Degradation-Aware Cooperative Multi-Modal GNSS-Denied Localization Leveraging LiDAR-Based Robot Detections

Václav Pritzl*, Xianjia Yu**, Tomi Westerlund**, Petr Štěpán*, and Martin Saska*

Abstract—Accurate long-term localization using onboard sensors is crucial for robots operating in Global Navigation Satellite System (GNSS)-denied environments. While complementary sensors mitigate individual degradations, carrying all the available sensor types on a single robot significantly increases the size, weight, and power demands. Distributing sensors across multiple robots enhances the deployability but introduces challenges in fusing asynchronous, multi-modal data from independently moving platforms. We propose a novel adaptive multi-modal multi-robot cooperative localization approach using a factor-graph formulation to fuse asynchronous Visual-Inertial Odometry (VIO), LiDAR-Inertial Odometry (LIO), and 3D inter-robot detections from distinct robots in a loosely-coupled fashion. The approach adapts to changing conditions, leveraging reliable data to assist robots affected by sensory degradations. A novel interpolation-based factor enables fusion of the unsynchronized measurements. LIO degradations are evaluated based on the approximate scan-matching Hessian. A novel approach of weighting odometry data proportionally to the Wasserstein distance between the consecutive VIO outputs is proposed. A theoretical analysis is provided, investigating the cooperative localization problem under various conditions, mainly in the presence of sensory degradations. The proposed method has been extensively evaluated on real-world data gathered with heterogeneous teams of an Unmanned Ground Vehicle (UGV) and Unmanned Aerial Vehicles (UAVs), showing that the approach provides significant improvements in localization accuracy in the presence of various sensory degradations.

Index Terms—Multi-Robot Systems, Cooperative Localization, Multimodal Localization, Factor Graph, Unmanned Aerial Vehicle, Unmanned Ground Vehicle, LiDAR

SUPPLEMENTARY MATERIAL

Video: <https://mrs.fel.cvut.cz/coop-mm-localization>

I. INTRODUCTION

IN Global Navigation Satellite System (GNSS)-denied environments, fusing different localization modalities is crucial to provide robustness to various environmental challenges [1]. Visual-based localization requires cheap and light-weight sensors, but it is sensitive to illumination changes and texture-less environments. Light Detection and Ranging (LiDAR)-based

This work was supported by CTU grant no SGS23/177/OHK3/3T/13, by the Czech Science Foundation (GAČR) under research project No. 23-07517S, and by the European Union under the project Robotics and advanced industrial production (reg. no. CZ.02.01.01/00/22_008/0004590).

*The authors are with the Multi-robot Systems Group, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic {vaclav.pritzl, petr.stepan, martin.saska}@fel.cvut.cz

**The authors are with Turku Intelligent Embedded and Robotic Systems (TIERS) Lab, University of Turku, 20520 Turku, Finland {xianjia.yu,tovewe}@utu.fi

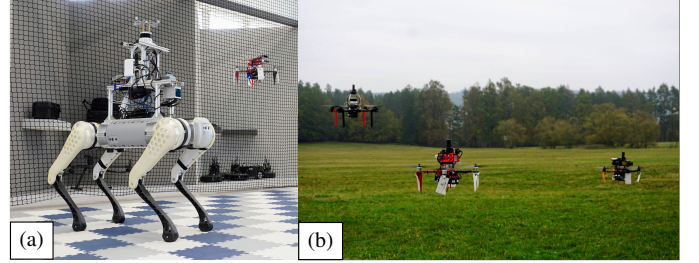


Figure 1. (a) A LiDAR-equipped UGV with a single camera-carrying UAV and (b) one LiDAR-equipped UAV cooperating with two camera-equipped UAVs.

localization exhibits better accuracy and works in challenging lighting conditions, but requires heavy power-consuming sensors, significant geometric structure in the environment, and presence of objects in the relatively small range of 3D LiDARs.

Multi-robot localization methods enable the distribution of different sensors on different robots in different parts of the environment. In such a case, all robots do not need to carry all the sensors to be able to fully function in every specific scenario. A LiDAR-carrying robot can provide accurate localization in an area with sufficient geometric structure, while a camera-equipped robot can take care of the localization in a place without geometric features but with enough visual texture. Moreover, multi-robot localization enables the robot team to cooperate to fulfill the desired task, such as cooperative mapping, sensing, inspection, or search and rescue.

Multi-modal multi-robot localization is crucial for enabling truly heterogeneous robot teams to operate in various environmental conditions (see Fig. 1), but it comes with a set of its own unique challenges. In a multi-modal multi-robot localization scenario, we need to fuse data from sensors carried by multiple different robots, which are independently moving with respect to each other and communicating over a network with possibly limited bandwidth. The data are asynchronous, and the sensory degradations can cause certain parts of the estimated state to be unobservable.

Multi-modal localization methods can be loosely-coupled, processing the modalities separately into distinct estimates and fusing them afterwards, or tightly-coupled, jointly processing the modalities at the feature / sensory data level. While tightly-coupled methods are generally more accurate, applying such methods to multi-robot scenarios is challenging due to the necessity to transmit more data between the robots and due to the data being obtained from mutually distant and moving

Table I
MATHEMATICAL NOTATION AND NOMENCLATURE.

Symbol	Meaning
a	scalar value a
$\mathbf{a}$	vector $\mathbf{a}$
$\mathbf{A}$	matrix $\mathbf{A}$
$\mathcal{A}$	set $\mathcal{A}$
A	reference frame / robot A
$SE(3)$	special Euclidean group
$SO(3)$	special orthogonal group
$\mathbf{T}_B^A$	$SE(3)$ pose of B in frame A , matrix transforming vectors from B to A
$\hat{\mathbf{T}}_B^A$	estimated $SE(3)$ pose of B in frame A
$\bar{\mathbf{T}}$	pose obtained by applying tangent-space perturbation to pose $\mathbf{T}$
$\mathbf{R}_B^A$	$SO(3)$ rotation of B in frame A
$\mathbf{R}_z(\theta)$	rotation matrix about the z -axis by angle θ
$\mathbf{R}_A^W(\alpha, \beta)$	constraint on the roll α and pitch β of A in frame W
$\mathbf{t}_B^A$	translation of B in frame A
$\mathbf{d}_B^A$	detection of B in frame A
ξ_A	perturbation vector on the tangent space associated with the pose $\mathbf{T}_A^W$; applied on the right side of $\mathbf{T}_A^W$; orientation is expressed before translation
$\text{Exp} : \mathbb{R}^6 \rightarrow SE(3)$	exponential map, mapping from the local vector space to the Lie group $SE(3)$
$\text{Log} : SE(3) \rightarrow \mathbb{R}^6$	logarithm map, mapping from the Lie group $SE(3)$ to local vector space
Σ	covariance matrix
$\mathcal{N}(\mu, \Sigma)$	multivariate normal distribution with mean μ and covariance matrix Σ
$W_2(\Sigma_1, \Sigma_2)$	2-Wasserstein distance between Σ_1 and Σ_2
$\text{Tr} \mathbf{A}$	trace of matrix $\mathbf{A}$
$[\mathbf{T}_B^A]_{\text{tr}}$	translational component of pose $\mathbf{T}_B^A$
$\mathbf{J}$	Jacobian matrix
$[\mathbf{x}]_{\times}$	skew-symmetric matrix constructed from $\mathbf{x}$
$\mathbf{J}_r _{\mathbf{x}}$	right Jacobian of $SE(3)$ evaluated at $\mathbf{x}$
$\text{Ad}_{\mathbf{T}}$	adjoint of $\mathbf{T}$
$\mathbf{I}_6$	identity matrix of size 6×6
$[\mathbf{a}]_x$	x -component of the vector $\mathbf{a}$
$\mathbf{a}^\perp$	vector orthogonal to vector $\mathbf{a}$

locations.

In contrast, a loosely-coupled fusion scheme requires comparatively lower communication bandwidth and enables modularity, providing the option to easily swap between the specific algorithms processing each sensory modality. However, such a loosely-coupled fusion method needs to be able to efficiently evaluate the reliability of each input to adaptively fuse them to provide an accurate location estimate at all times.

A. Problem Statement

We tackle the problem of adaptive loosely-coupled multi-modal multi-robot localization. The problem is illustrated in Fig. 2. Table I describes the mathematical notation and nomenclature used throughout the paper. We focus on the case of cooperative localization in a team of one LiDAR-carrying

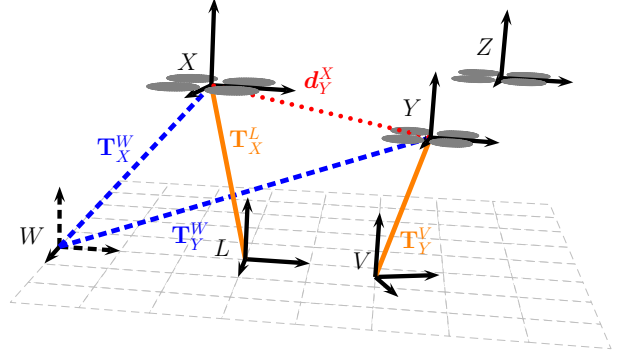


Figure 2. Robot X is localized in local frame L using a LIO algorithm. Robot Y is localized in a local frame V using a VIO algorithm. Robot X detects the relative 3D position of robot Y . Robot Z represents another VIO-utilizing robot detected by robot X . W denotes the world reference frame of the cooperative localization algorithm. All the reference frames are gravity-aligned. Blue dashed lines represent the estimated variables $\mathbf{T}_X^W, \mathbf{T}_Y^W$. Orange solid lines represent poses $\mathbf{T}_X^L, \mathbf{T}_Y^V$ provided by the individual odometry algorithms. The red dotted line represents the relative detection $\mathbf{d}_Y^X$ of robot Y by robot X .

robot and one or more camera-equipped robots. A LiDAR-carrying robot X is localized in local frame L using a LiDAR-Inertial Odometry (LIO) algorithm, providing pose $\mathbf{T}_X^L$. A camera-equipped robot Y is localized in local frame V using a Visual-Inertial Odometry (VIO) algorithm, providing pose $\mathbf{T}_Y^V$. Robot X detects the 3D position $\mathbf{d}_Y^X$ of robot Y w.r.t. the gravity-aligned body frame of robot X . We assume that all the frames of reference are aligned to the gravity vector as all robots are equipped with Inertial Measurement Units (IMUs), enabling the estimation of the gravity vector direction.

The task tackled in this paper is to utilize the LIO / VIO data from separate robots and inter-robot detections to periodically estimate the poses $\mathbf{T}_X^W, \mathbf{T}_Y^W$ of each robot in the common reference frame W , while minimizing localization error with respect to the true robot poses.

The system times of all robots are assumed to be synchronized throughout the estimation process. The odometry data and detections are produced asynchronously. The inter-robot detections are anonymous, requiring the approach to solve the association problem. No global measurements, i.e., no GNSS data, are available to the algorithm. The robots are capable of explicit communication.

The LIO algorithm is based on solving the scan matching problem, optimizing a point-to-plane / point-to-line cost function, enabling the option to analyze the approximate Hessian to evaluate scan matching degeneracy. The VIO is based on a Kalman filter, outputting uncertainty estimates in the form of covariance matrices.

We focus on the specific case of LIO and VIO-utilizing robots to highlight the significant applicability of the approach to heterogeneous robot teams, but it is worth mentioning that the proposed method is applicable to other odometry sources as well, assuming that they provide similar capabilities of evaluating their uncertainty. Similarly, though we focus on LiDAR-based detections between the robots, the method is applicable to other relative localization methods capable of producing accurate 3D detections of neighboring robots.

B. Related Work

1) *Multi-Modal Localization on a Single Robot:* Most single-robot fusion approaches are not easily applicable to a multi-robot scenario. Many approaches operate in a coarse-to-fine manner, obtaining the estimated pose of the robot body with one modality and using the result as a prior to refine the estimate with another modality [2], [3]. In [4], the authors proposed an IMU-centric multi-modal odometry based on a factor graph. In their approach, they constrained the IMU prediction using relative pose factors and detected possible sensory degenerations by analyzing the information matrices of the respective optimization problems for each modality. Many single-robot multi-modal approaches use a tightly-coupled scheme [5]–[11], fusing LiDAR, visual, and inertial measurements in a single estimator without dividing the process into completely separate modality-specific subsystems. In [12], the authors utilize intermittent VIO data when LiDAR degeneracy is detected, while aiding VIO by using LiDAR points as additional features. In [13], a factor-graph-based LiDAR-visual-inertial odometry was proposed, fusing data from different modalities as relative poses and associating depth information from the LiDAR data to the visual information. Despite the advantages of these methods in a single robot scenario, a coarse-to-fine approach, tight-coupling of sensor data, associating depth from LiDAR data with visual features, or centering the approach around a single IMU odometry are not easily applicable to a multi-robot scenario, where each sensory modality is placed on a separate robot, the robots need to wirelessly communicate their data, and the changing transformation between the robots is obtained through noisy detections. Therefore, we focus on a parallel loosely-coupled architecture, running completely separate odometry algorithms on each robot, converting them to relative poses, and weighting the individual poses based on their estimated reliability.

2) *Evaluating the Uncertainty of a Sensory Modality:* To achieve accurate and robust performance, loosely-coupled fusion methods must be able to detect sensory degradations and adapt the fusion process accordingly. For LiDAR-fusing methods based on minimizing the point-to-plane cost function, the standard solution is to analyze the eigenvalues of the approximate Hessian of the optimization problem and perform thresholding or calculate a covariance matrix from the approximate Hessian [2], [3], [14]–[16]. In our method, we obtain a binary evaluation of the reliability of LiDAR-based localization by thresholding the minimum eigenvalue of the approximate Hessian matrix.

For visual-based methods, the usual approaches for evaluating uncertainty and detecting sensory degradation include analyzing the number of visual features [13], [17], thresholding the estimated IMU biases [13], analyzing the determinant of the covariance matrix [3] in Kalman filter-based algorithms, thresholding on the maximum change of odometry output with respect to IMU odometry [13], or analyzing the information matrix of the optimization problem [4]. However, even a small amount of visual features may provide sufficient constraints, and thus, the number of features may not sufficiently correlate with the localization error. Thresholding IMU biases

may provide enough information about the complete loss of localization, but not about finer uncertainty changes. In a Kalman filter-based VIO, analyzing the covariance matrices is the obvious choice. We need to obtain the uncertainty of the relative poses between consecutive filter outputs. However, the cross-covariances between the consecutive filter outputs are unknown. Ignoring cross-covariances when calculating relative pose covariance would lead to degradation of the uncertainty [18].

In this work, we propose a novel approach of weighting relative poses calculated from the outputs of a Kalman filter-based algorithm proportionally to the Wasserstein distance between the probability distributions of consecutive filter outputs. Such an approach is compatible with any Kalman filter-based method without requiring any modifications to the odometry method itself.

3) *Multi-Robot Cooperative Localization:* Multi-robot Simultaneous Localization and Mapping (SLAM) algorithms usually rely on inter- and intra-robot loop closures, obtaining relative transformations between the different robot frames by aligning mutually observed features, viewpoints, or partial maps. The architecture usually consists of a front-end odometry combined with a back-end pose graph optimization running on a server or running on each robot in a decentralized fashion. Most of such approaches focus on a single exteroceptive modality. In [19]–[21], such methods are proposed for 3D LiDAR sensors. In [22], pairwise Ultra-wideband (UWB) measurements are fused with the LiDAR odometry and LiDAR-based loop closures. In [23]–[25], visual-based multi-robot SLAM methods are proposed, combining VIO front-ends with visual-based place recognition or map-based loop closures. The authors of [26] fused VIO ego-motion estimates with UWB ranges, visual detections of cooperating UAVs, and map-based loop closures. However, all of these methods are limited in the sense that they utilize only a single main sensory modality, either visual or LiDAR data. Such an approach lacks the adaptiveness that can be provided by a multi-modal method. In [27], a collaborative localization and mapping method for a ground-aerial team is proposed, but each robot in the team needs to be equipped with both a LiDAR and a camera. In [28], a SLAM method for a UAV swarm fusing both LiDAR and visual data is proposed. However, all robots need to carry both LiDAR and camera sensors, as the inter-robot loop closures are only obtained for each modality separately. Obtaining loop closures from a combination of different modalities is complicated, as the sensors inherently capture different types of information. Furthermore, place-recognition-based approaches require significant communication bandwidth to perform the loop closures.

Utilizing direct detections of cooperating robots enables straightforward relative localization without requiring the robots to use the same sensory modalities and minimizes the communication requirements. In [26], the visual detections were used in addition to loop closures and UWB measurements. In [17], ultraviolet markers were utilized for camera-based detection of cooperating UAVs in a multi-robot state estimation method. In [29], the authors proposed a tightly-coupled decentralized LiDAR-based odometry for a swarm of

UAVs utilizing LiDAR-based detections of reflective markers on board the UAVs. However, the aforementioned all relied primarily on a single sensory modality, either LiDAR or camera-based. In [30], LiDAR-based detections were utilized for relative localization in a heterogeneous team of a LiDAR-equipped and camera-equipped UAV. However, the fusion approach did not deal with variable uncertainties of the localization inputs and always treated the LiDAR-based localization as perfectly reliable. Such an assumption is not valid in many real-world environments. In [31], the authors proposed an active collaborative localization method for a UGV-UAV team. The UGVs were detected from cameras on board the UAVs and used as landmarks to improve the localization of the UAVs. The work primarily dealt with the optimal placement of the UGVs in the environment. However, the UGVs were only utilized as static landmarks, and the work did not deal with improving localization performance in the other way around, i.e., improving localization of the UGVs by the UAVs.

In our work, we utilize direct 3D detections of the cooperating robots, obtained from LiDAR data. In contrast to the state-of-the-art methods, we focus on a scenario where the multi-robot team combines multiple different sensory modalities, but each robot is equipped with only one distinct exteroceptive sensor type, e.g., one robot is only equipped with a LiDAR, while another robot is only equipped with a camera. Our approach adapts to changing sensory degradations in the environment and does not rely on the assumption that a single modality is always reliable or always more accurate than the other one.

C. Contributions

The contributions of this work are summarized as:

- A novel adaptive loosely-coupled multi-modal multi-robot fusion method for cooperative localization. The method utilizes direct 3D detections of cooperating UAVs, enabling us to efficiently fuse localization outputs from different sensory modalities without the need to perform place recognition on the sensory data. The detections are represented by a novel interpolation-based quaternary factor, enabling efficient fusion of data from unsynchronized sources.
- A novel approach of weighting relative pose factors proportionally to the Wasserstein distance of consecutive outputs of a Kalman filter-based algorithm, enabling the method to utilize information about the variable uncertainty of the odometry without degrading the estimate due to unknown cross-covariances.
- Theoretical observability analysis of the problem of cooperative localization based on odometry estimates and relative 3D detections with respect to various assumptions and odometry degradations.

The proposed method was extensively evaluated on real-world data gathered with a UGV-UAV and UAV-only team. The accuracy of the approach was quantitatively evaluated with respect to motion-capture and Real-Time Kinematic (RTK) ground truth.

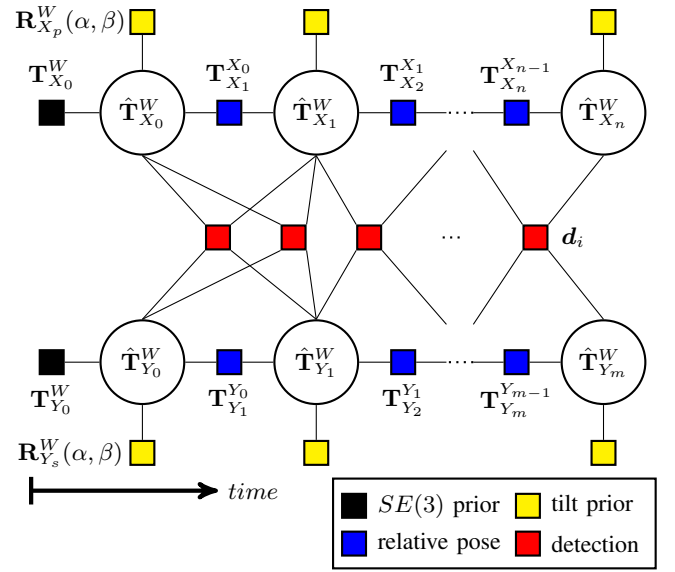


Figure 3. Factor graph representation of the cooperative localization problem for two robots. Circles represent the estimated variables. Squares represent the factors, i.e., the measurements and prior information. Relative poses of robot X are obtained from a LIO algorithm, while relative poses of robot Y are provided by a VIO algorithm.

II. MULTI-ROBOT LOCALIZATION METHOD

We formulate the cooperative localization problem as a factor graph [32] with the estimated robot poses, priors, and measurements represented by multivariate Gaussian distributions. As the distinct measurement sources are unsynchronized, we utilize a novel interpolation-based factor representing the inter-robot detections. The odometry measurements are adaptively weighted based on the reliability of each robot's odometry.

A. Factor Graph-Based Formulation

The factor graph for the cooperative localization problem for two robots is illustrated in Fig. 3. The problem is modeled by a set of estimated $SE(3)$ variables $\hat{\mathcal{X}} = \{\hat{\mathbf{T}}_{X_0}^W, \hat{\mathbf{T}}_{X_1}^W, \dots, \hat{\mathbf{T}}_{X_n}^W, \hat{\mathbf{T}}_{Y_0}^W, \hat{\mathbf{T}}_{Y_1}^W, \dots, \hat{\mathbf{T}}_{Y_m}^W\}$ representing the poses of the robots at specific moments in time and a set of factors representing the measurements. The factors include prior factors, the $SE(3)$ relative pose factors representing the relative movement of each robot obtained from an odometry algorithm, and the $\mathbb{R}^3$ inter-robot position detection factors. Additional robots detected by robot X would be incorporated analogously, by connecting the corresponding set of estimated robot poses to the poses of robot X through the 3D position detection factors. Let $\mathcal{L} = \{\mathbf{T}_{X_{k+1}}^{X_k} \mid k = 0, \dots, n-1\}$ be the set of n LIO relative poses of robot X , $\mathcal{V} = \{\mathbf{T}_{Y_{l+1}}^{Y_l} \mid l = 0, \dots, m-1\}$ be the set of m VIO relative poses of robot Y , and $\mathcal{D} = \{d_i \mid i = 0, \dots, p\}$ be the set of p inter-robot detections. Let $\mathcal{X}_i, \mathcal{X}_j, \mathcal{X}_k, \mathcal{X}_l, \mathcal{X}_q$ be the subsets of variables connected to the i -th inter-robot detection factor, j -th $SE(3)$ prior factor, k -th LIO relative pose factor, l -th VIO relative pose factor, and q -th tilt prior factor, respectively. Let $\mathcal{L}_k, \mathcal{V}_l$ be the relative pose measurement corresponding to the k -th LIO and l -th VIO relative pose factors, respectively. The estimation

problem can be equivalently expressed as a weighted nonlinear least squares problem as

$$\begin{aligned} \hat{\mathcal{X}} = \arg \min_{\mathcal{X}} & \left(\sum_j \left\| e_j^{\text{prior}}(\mathcal{X}_j) \right\|_{\Sigma_j^{\text{prior}}}^2 + \right. \\ & + \sum_q \left\| e_q^{\text{tilt}}(\mathcal{X}_q) \right\|_{\Sigma_q^{\text{tilt}}}^2 + \sum_k \left\| e_k^{\text{LIO}}(\mathcal{X}_k, \mathcal{L}_k) \right\|_{\Sigma_k^{\text{LIO}}}^2 + \\ & \left. + \sum_l \left\| e_l^{\text{VIO}}(\mathcal{X}_l, \mathcal{V}_l) \right\|_{\Sigma_l^{\text{VIO}}}^2 + \sum_i \left\| e_i^{\text{det}}(\mathcal{X}_i, \mathcal{D}_i) \right\|_{\Sigma_i^{\text{det}}}^2 \right), \end{aligned} \quad (1)$$

where e_j^{prior} is the error function of the $SE(3)$ prior factors, e_q^{tilt} is the error function of the zero-roll-pitch prior, e_k^{LIO} is the error function of the LIO relative poses of robot X , e_l^{VIO} is the error function of the VIO relative poses of robot Y , and e_i^{det} is the error function of the 3D inter-robot detections between robot X and robot Y . Each error is weighted by the corresponding covariance matrix representing the Gaussian noise associated with the measurement. The random $SE(3)$ variables [33] are defined with the tangent-space perturbations of the noise-free pose $\mathbf{T}$ applied on the right side:

$$\tilde{\mathbf{T}} = \mathbf{T} \text{Exp}(\boldsymbol{\xi}), \mathbf{T} \in SE(3), \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \quad (2)$$

where $\text{Exp}()$ is the exponential map of $SE(3)$, mapping elements of the vector space $\mathbb{R}^6$ to the Lie group $SE(3)$, and $\boldsymbol{\xi}$ is the perturbation vector in the tangent space with the convention of expressing orientation before translation to be compatible with the GTSAM library. Assuming that $\mathbf{T}$ represents an $SE(3)$ pose in the world frame, the corresponding covariance matrix $\boldsymbol{\Sigma}$ is then defined in the body frame of $\mathbf{T}$.

To be able to incrementally build and solve the factor graph in real time, the graph is solved in a sliding window, utilizing a fixed-lag smoother based on the iSAM2 [34]. Due to the assumption that all the reference frames are gravity-aligned, we constrain the roll and pitch angles of all poses to zero using a prior factor on each estimated variable.

B. LIO Relative Factor

The relative pose is calculated from two consecutive outputs of the LIO algorithm as

$$\mathbf{T}_{X_{k+1}}^{X_k} = (\mathbf{T}_{X_k}^L)^{-1} \mathbf{T}_{X_{k+1}}^L. \quad (3)$$

We assume that the LIO utilizes a scan matching algorithm minimizing a point-to-line or point-to-plane cost function, and we utilize binary detection of LIO degeneration based on the minimum eigenvalue of the approximate Hessian evaluated by the optimization method. For a detailed description of the scan matching degeneracy detection, see [14]. We consider the current LIO output degenerated if

$$\min(\text{eig}(\mathbf{A}^T \mathbf{A})) < \lambda_{\text{thr}}, \quad (4)$$

where $\mathbf{A}$ is the Jacobian of the scan-matching cost function evaluated at the current linearization point and λ_{thr} is a predefined threshold. We assume that when the LIO is not degenerated, it is more reliable than the VIO method, and when the LIO is degenerated, the LIO position and yaw output

are unusable. Therefore, we set the covariance matrix of the LIO relative pose as

$$\boldsymbol{\Sigma}_k^{\text{LIO}} = \text{diag}(\sigma_\alpha^2, \sigma_\beta^2, {}^{\text{lo}}\sigma_\gamma^2, {}^{\text{lo}}\sigma_{\text{pos}}^2, {}^{\text{lo}}\sigma_{\text{pos}}^2, {}^{\text{lo}}\sigma_{\text{pos}}^2) \Delta t, \quad (5)$$

when both of the LIO poses $\mathbf{T}_{X_k}^L, \mathbf{T}_{X_{k+1}}^L$ are reliable and as

$$\boldsymbol{\Sigma}_k^{\text{LIO}} = \text{diag}(\sigma_\alpha^2, \sigma_\beta^2, {}^{\text{hi}}\sigma_\gamma^2, {}^{\text{hi}}\sigma_{\text{pos}}^2, {}^{\text{hi}}\sigma_{\text{pos}}^2, {}^{\text{hi}}\sigma_{\text{pos}}^2) \Delta t, \quad (6)$$

if either of the poses $\mathbf{T}_{X_k}^L, \mathbf{T}_{X_{k+1}}^L$ is degenerated. Δt is the difference of timestamps of poses $\mathbf{T}_{X_k}^L, \mathbf{T}_{X_{k+1}}^L$. σ_α and σ_β are equal for both cases, as we constrain the roll and pitch angles to be constantly zero. σ_γ and σ_{pos} are the respective standard deviations corresponding to the yaw angle and to the positional part. The standard deviations are selected so that

$${}^{\text{lo}}\sigma_\gamma \ll {}^{\text{hi}}\sigma_\gamma, \quad {}^{\text{lo}}\sigma_{\text{pos}} \ll {}^{\text{hi}}\sigma_{\text{pos}} \quad (7)$$

and based on analyzing the average error of the LIO relative poses w.r.t. ground-truth measurements in real-world data.

C. VIO Relative Factor

The relative pose is calculated from two consecutive outputs of the VIO algorithm as

$$\mathbf{T}_{Y_{l+1}}^{Y_l} = (\mathbf{T}_{Y_l}^V)^{-1} \mathbf{T}_{Y_{l+1}}^V. \quad (8)$$

We assume that the VIO algorithm is based on a Kalman filter and encodes uncertainty information in the covariance matrix of each output. The consecutive local poses are described by the jointly Gaussian probability distribution

$$\mathcal{N} \left(\begin{bmatrix} \mathbf{T}_{Y_l}^V \\ \mathbf{T}_{Y_{l+1}}^V \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{l,l} & \boldsymbol{\Sigma}_{l,l+1} \\ \boldsymbol{\Sigma}_{l+1,l} & \boldsymbol{\Sigma}_{l+1,l+1} \end{bmatrix} \right). \quad (9)$$

Assuming that all the covariance matrices are defined in the same reference frame, the covariance matrix of the relative pose can be calculated as

$$\boldsymbol{\Sigma}_{\text{rel}} = \boldsymbol{\Sigma}_{l,l} + \boldsymbol{\Sigma}_{l+1,l+1} - \boldsymbol{\Sigma}_{l,l+1} - \boldsymbol{\Sigma}_{l+1,l}, \quad (10)$$

where $\boldsymbol{\Sigma}_{l,l}$ and $\boldsymbol{\Sigma}_{l+1,l+1}$ are the covariance matrices estimated by the Kalman filter at time steps l and $l+1$, respectively. The cross-covariance matrix $\boldsymbol{\Sigma}_{l,l+1}$ between the consecutive Kalman filter estimates is generally unknown, but ignoring the cross-covariance would lead to degradation of the estimate [18]. Therefore, we approximate the relative pose covariance matrix by a matrix proportional to the 2-Wasserstein distance [35]–[37] between the covariance matrices of the consecutive filter outputs. For the covariance matrices $\boldsymbol{\Sigma}_{l,l}, \boldsymbol{\Sigma}_{l+1,l+1}$, the squared 2-Wasserstein distance between them is the solution of the problem

$$W_2(\boldsymbol{\Sigma}_{l,l}, \boldsymbol{\Sigma}_{l+1,l+1})^2 = \min_{\boldsymbol{\Sigma}_{l,l+1}} \text{Tr}(\boldsymbol{\Sigma}_{l,l} + \boldsymbol{\Sigma}_{l+1,l+1} - 2\boldsymbol{\Sigma}_{l,l+1}) \quad (11a)$$

$$\text{s.t.} \quad \begin{bmatrix} \boldsymbol{\Sigma}_{l,l} & \boldsymbol{\Sigma}_{l,l+1} \\ \boldsymbol{\Sigma}_{l+1,l} & \boldsymbol{\Sigma}_{l+1,l+1} \end{bmatrix} \geq 0. \quad (11b)$$

The Wasserstein distance is a statistical distance and a metric on the space of covariance matrices, and its calculation assumes maximal correlation between the two covariance matrices and thus provides a lower bound on the trace of the

relative pose covariance. The 2-Wasserstein distance between the two Gaussians can be calculated as

$$W_2(\Sigma_{l,l}, \Sigma_{l+1,l+1}) = \sqrt{\text{Tr} \left[\Sigma_{l,l} + \Sigma_{l+1,l+1} - 2(\Sigma_{l,l}^{\frac{1}{2}} \Sigma_{l+1,l+1} \Sigma_{l,l}^{\frac{1}{2}})^{\frac{1}{2}} \right]}. \quad (12)$$

To utilize this calculation, the covariance matrices $\Sigma_{l,l}, \Sigma_{l+1,l+1}$ need to be defined in the same reference frame. In our case, the covariance matrices outputted by the VIO algorithm are defined on the $SO(3) \times \mathbb{R}^3$ manifold. In our approach, we utilize only the positional part of the covariance matrices; therefore, we can directly calculate the Wasserstein distance between the positional covariance matrices with no transformation necessary. We assume that the error of the relative pose measurement is proportional to the Wasserstein distance between the two covariance matrices. However, the Wasserstein distance provides only a lower bound on the trace of the relative pose covariance matrix, and using it directly as the trace of the relative pose covariance matrix would result in an overconfident estimate. Furthermore, we do not assume that the covariance matrix always provides a consistent estimate of the uncertainty, i.e., the incoming covariance matrices $\Sigma_{l,l}, \Sigma_{l+1,l+1}$ may be underconfident or overconfident. Therefore, we scale it by a constant factor as

$${}^\nu \sigma_{\text{pos}} = \mu W_2(\Sigma_{l,l}, \Sigma_{l+1,l+1}). \quad (13)$$

The scaling factor μ depends on the system model of the VIO algorithm and the odometry sampling rate and is selected based on analyzing the average error of the VIO output with respect to ground-truth measurements on real-world data. Next, we bound the standard deviation to always stay between predefined thresholds as

$${}^\nu \bar{\sigma}_{\text{pos}} = \max({}^{\text{vmin}} \sigma_{\text{pos}} \sqrt{\Delta t}, \min({}^{\text{vmax}} \sigma_{\text{pos}} \sqrt{\Delta t}, {}^\nu \sigma_{\text{pos}})). \quad (14)$$

The thresholds are selected such that

$${}^{\text{lo}} \sigma_{\text{pos}} \leq {}^{\text{vmin}} \sigma_{\text{pos}} < {}^{\text{vmax}} \sigma_{\text{pos}} \leq {}^{\text{hi}} \sigma_{\text{pos}}. \quad (15)$$

The standard deviation of the yaw measurement is calculated as

$${}^\nu \bar{\sigma}_\gamma = \begin{cases} {}^\nu \sigma_\gamma \Delta t & \text{if } {}^\nu \sigma_{\text{pos}} \leq {}^{\text{vmax}} \sigma_{\text{pos}} \\ \nu {}^\nu \sigma_\gamma \Delta t & \text{if } {}^\nu \sigma_{\text{pos}} > {}^{\text{vmax}} \sigma_{\text{pos}} \end{cases}, \quad (16)$$

where ${}^\nu \sigma_\gamma$ is an empirically-selected parameter and ν is a predefined scaling factor inflating the yaw sigma if the positional measurement is deemed unreliable due to high positional standard deviation. The covariance matrix of the relative pose measurement is set as

$$\Sigma_l^{\text{VIO}} = \text{diag}(\sigma_\alpha^2, \sigma_\beta^2, {}^\nu \bar{\sigma}_\gamma^2, {}^\nu \bar{\sigma}_{\text{pos}}^2, {}^\nu \bar{\sigma}_{\text{pos}}^2, {}^\nu \bar{\sigma}_{\text{pos}}^2). \quad (17)$$

D. Interpolation-Based Detection Factor

The relative robot detection is represented by a quaternary factor connecting the temporally-adjacent variables $\mathbf{T}_{X_k}^W, \mathbf{T}_{X_{k+1}}^W, \mathbf{T}_{Y_l}^W, \mathbf{T}_{Y_{l+1}}^W$. The factor utilizes the assumption

of constant velocity between consecutive variables. The error function is defined as

$$\mathbf{e}_i^{\text{det}} = \left[(\mathbf{T}_{X_{\text{int}}}^W)^{-1} \mathbf{T}_{Y_{\text{int}}}^W \right]_{\text{tr}} - \mathbf{d}_i, \quad (18)$$

where $\mathbf{d}$ is the detection, $[\cdot]_{\text{tr}}$ selects the translational part of the $SE(3)$ matrix, and $\mathbf{T}_{X_{\text{int}}}^W, \mathbf{T}_{Y_{\text{int}}}^W$ are obtained by linear interpolation on the $SE(3)$ manifold as

$$\mathbf{T}_{X_{\text{int}}}^W = \mathbf{T}_{X_k}^W \text{Exp} \left(\tau_X \text{Log} \left((\mathbf{T}_{X_k}^W)^{-1} \mathbf{T}_{X_{k+1}}^W \right) \right), \quad (19a)$$

$$\mathbf{T}_{Y_{\text{int}}}^W = \mathbf{T}_{Y_l}^W \text{Exp} \left(\tau_Y \text{Log} \left((\mathbf{T}_{Y_l}^W)^{-1} \mathbf{T}_{Y_{l+1}}^W \right) \right), \quad (19b)$$

$$\tau_X = \frac{t_{\text{det}} - t_{X_k}}{t_{X_{k+1}} - t_{X_k}}; \quad \tau_Y = \frac{t_{\text{det}} - t_{Y_l}}{t_{Y_{l+1}} - t_{Y_l}}, \quad (19c)$$

where t_{det} is the time of the detection, t_A represents time of variable $\mathbf{T}_A^W$, Exp denotes the $SE(3)$ exponential map, and Log is the $SE(3)$ logarithm map.

Let us denote ξ_A the perturbation vector on the tangent space associated with the pose $\mathbf{T}_A^W$. The analytic Jacobian matrix of the detection factor error function at $\mathbf{T}_{X_k}^W, \mathbf{T}_{X_{k+1}}^W, \mathbf{T}_{Y_l}^W, \mathbf{T}_{Y_{l+1}}^W$ is formulated as

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{e}_{\text{det}}}{\partial \xi_{X_k}} & \frac{\partial \mathbf{e}_{\text{det}}}{\partial \xi_{X_{k+1}}} & \frac{\partial \mathbf{e}_{\text{det}}}{\partial \xi_{Y_l}} & \frac{\partial \mathbf{e}_{\text{det}}}{\partial \xi_{Y_{l+1}}} \end{bmatrix}, \quad (20)$$

where the submatrices are calculated using the chain rule as

$$\frac{\partial \mathbf{e}_{\text{det}}}{\partial \xi_{X_k}} = \frac{\partial \mathbf{e}_{\text{det}}}{\partial \xi_{X_{\text{int}}}} \frac{\partial \xi_{X_{\text{int}}}}{\partial \xi_{X_k}}, \quad \frac{\partial \mathbf{e}_{\text{det}}}{\partial \xi_{X_{k+1}}} = \frac{\partial \mathbf{e}_{\text{det}}}{\partial \xi_{X_{\text{int}}}} \frac{\partial \xi_{X_{\text{int}}}}{\partial \xi_{X_{k+1}}}, \quad (21a)$$

$$\frac{\partial \mathbf{e}_{\text{det}}}{\partial \xi_{Y_l}} = \frac{\partial \mathbf{e}_{\text{det}}}{\partial \xi_{Y_{\text{int}}}} \frac{\partial \xi_{Y_{\text{int}}}}{\partial \xi_{Y_l}}, \quad \frac{\partial \mathbf{e}_{\text{det}}}{\partial \xi_{Y_{l+1}}} = \frac{\partial \mathbf{e}_{\text{det}}}{\partial \xi_{Y_{\text{int}}}} \frac{\partial \xi_{Y_{\text{int}}}}{\partial \xi_{Y_{l+1}}}. \quad (21b)$$

The partial derivatives of the detection error function with respect to perturbations of the interpolated poses $\mathbf{T}_{X_{\text{int}}}^W$ and $\mathbf{T}_{Y_{\text{int}}}^W$ are formulated as

$$\frac{\partial \mathbf{e}_{\text{det}}}{\partial \xi_{X_{\text{int}}}} = \left[(\mathbf{R}_{X_{\text{int}}}^W)^T [t_{Y_{\text{int}}}^W - t_{X_{\text{int}}}^W]_{\times} \mathbf{R}_{X_{\text{int}}}^W \quad -\mathbf{I}_3 \right], \quad (22a)$$

$$\frac{\partial \mathbf{e}_{\text{det}}}{\partial \xi_{Y_{\text{int}}}} = \begin{bmatrix} \mathbf{0}_3 & (\mathbf{R}_{Y_{\text{int}}}^W)^T \mathbf{R}_{Y_{\text{int}}}^W \end{bmatrix}, \quad (22b)$$

where $[a]_{\times}$ denotes a skew-symmetric matrix constructed from vector $\mathbf{a}$. The partial derivatives of the interpolated poses with respect to perturbations of the estimated variables are

$$\frac{\partial \xi_{A_{\text{int}}}}{\partial \xi_{A_k}} = \mathbf{Ad}_{\mathbf{T}_{A_k}^{A_{\text{int}}}} - \tau_A \mathbf{J}_{\text{Exp}}^A \mathbf{J}_{\text{Log}}^A \mathbf{Ad}_{\mathbf{T}_{W_{k+1}}^{A_k} \mathbf{T}_{A_k}^W}, \quad (23a)$$

$$\frac{\partial \xi_{A_{\text{int}}}}{\partial \xi_{A_{k+1}}} = \tau_A \mathbf{J}_{\text{Exp}}^A \mathbf{J}_{\text{Log}}^A, \quad (23b)$$

where $\mathbf{Ad}_{\mathbf{T}}$ denotes the adjoint of $\mathbf{T}$ and the specific partial derivative can be obtained by substituting A_k, A_{k+1} with X_k, X_{k+1} or Y_l, Y_{l+1} . The Jacobians of the exponential and logarithm maps are formulated as

$$\mathbf{J}_{\text{Exp}}^A = \mathbf{J}_r \Big|_{\tau_A \text{Log} \left((\mathbf{T}_{A_k}^W)^{-1} \mathbf{T}_{A_{k+1}}^W \right)}, \quad (24a)$$

$$\mathbf{J}_{\text{Log}}^A = \mathbf{J}_r^{-1} \Big|_{(\mathbf{T}_{A_k}^W)^{-1} \mathbf{T}_{A_{k+1}}^W}, \quad (24b)$$

where $\mathbf{J}_r, \mathbf{J}_r^{-1}$ denote the right Jacobian of $SE(3)$ and its in-

verse, evaluated at the corresponding linearization points [38]. The detection covariance Σ^{det} is set as

$$\Sigma_i^{\text{det}} = \text{diag}(\sigma_{\text{det}}^2, \sigma_{\text{det}}^2, \sigma_{\text{det}}^2) \quad (25)$$

where σ_{det} are constant values, which are empirically tuned on real-world data to reflect the noise of the detections and the inaccuracies caused by the constant-velocity assumption used by the factor.

The detections are anonymous, i.e., the detections contain no information about which robot is detected and may contain false positives. Therefore, it is necessary to associate each incoming detection to a tracked robot before the detection is inserted into the graph. To perform the association, the measurement distance of the detection is calculated based on eq. (18) and the detection is associated to the closest robot if the measurement distance is below a predefined threshold.

E. Graph Initialization

To initialize the factor graph, we need an initial guess for the pose of each robot. We assume that there is no prior information about the position and yaw orientation of the robots. The detections are anonymous, i.e., the detections contain no information about which robot is detected and may contain false positives. Therefore, we obtain the initial pose of each robot by aligning its movement trajectory observed in the detections and in the odometry data.

We assume that the detector provides tracking of the detected object, i.e., we can construct a buffer of detections corresponding to a single object, but it is unknown which buffer corresponds to which robot. For each robot and for each detection buffer, we construct a sliding window of detections and corresponding robot poses, and solve the problem of alignment of the two point patterns as

$$\hat{t}_V^L, \hat{\theta} = \arg \min_{t_V^L, \theta} \sum_i \|\mathbf{R}_V^L(\theta) \mathbf{p}_{Y_i}^V + t_V^L - \mathbf{d}_{Y_i}^L\|^2, \quad (26a)$$

$$t_V^L \in \mathbb{R}^3, \theta \in [-\pi, \pi], \quad (26b)$$

$$\mathbf{T}_V^L = \begin{bmatrix} \mathbf{R}_z(\theta) & t_V^L \\ \mathbf{0}^T & 1 \end{bmatrix} \in SE(3). \quad (26c)$$

To obtain correspondences at the same time steps, $\mathbf{p}_{Y_i}^V$ is calculated by linear interpolation of the temporally-adjacent positions to the time of the detection $\mathbf{d}_{Y_i}^L$. The problem can be solved analytically using the Kabsch-Umeyama algorithm [39], solving only for the rotation and translation. The initialization is considered successful if the cost of (26a) falls below a predefined threshold and the trajectories are sufficiently large, as the alignment problem degrades as the robot trajectories degrade to a single point, and noise in the data might compromise the result.

To initialize the factor graph, we anchor the pose of the robot X at the origin

$$\mathbf{T}_{X_0}^W = \mathbf{I}_4 \quad (27)$$

and calculate prior poses of the remaining robots based on the transformations obtained from trajectory alignment as

$$\mathbf{T}_{Y_0}^W = \mathbf{T}_V^L \mathbf{T}_{Y_0}^V. \quad (28)$$

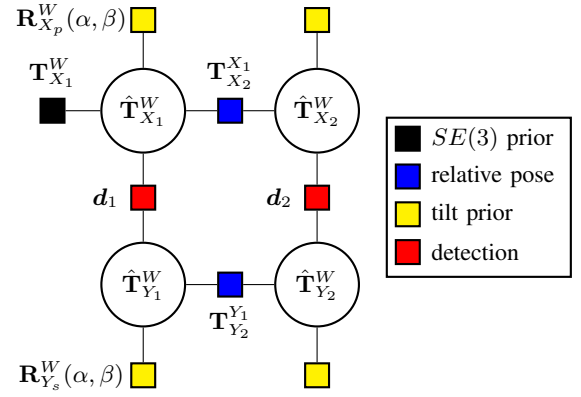


Figure 4. Simplified factor graph used in the observability analysis of the cooperative localization problem. The graph represents two poses of robot X and two poses of robot Y . Detections are synchronized with the odometry measurements, removing the need for interpolation. In the visualized situation, the factor graph is fully constrained.

The covariance matrices of the prior poses are empirically selected to reflect a strong constraint on anchoring $\mathbf{T}_{X_0}^W$ at origin and a comparatively larger uncertainty of the prior pose $\mathbf{T}_{Y_0}^W$, which can be corrected over the course of the estimation process.

To perform the initialization, each detected robot needs to separately perform an initialization maneuver so that the trajectory alignment problem can be solved. As the estimated transformation $\mathbf{T}_V^L$ has only 4 Degrees of Freedom (DOFs), this initialization maneuver can be a short flight in a straight line.

III. OBSERVABILITY ANALYSIS

To analyze the observability of the estimation problem with respect to various situations and sensory degradations, we construct the Jacobian matrix of the stacked measurement model of a simplified factor graph representing two poses of robot X and two poses of robot Y (see Fig. 4). By analyzing the rank of the resulting Jacobian matrix, we can determine whether the stacked measurements are sufficient for fully constraining the estimated variables. Furthermore, by analyzing the nullspace of the Jacobian matrix, we can obtain the unobservable directions, i.e., the subspace of the values of the estimated variables whose changes result in no change of the cost function.

For clarity of the analysis, we consider robot detections to be synchronized with the odometry measurements, thus removing the need for pose interpolation, and we assume that the roll and pitch angles of all poses are zero, with the orientation depending on yaw only. We construct the following Jacobian matrices from partial derivatives of the cost function (1) w.r.t. estimated robot poses:

$$\mathbf{J}^{\text{LIO}} = \begin{bmatrix} \frac{\partial \mathbf{e}^{\text{LIO}}}{\partial \boldsymbol{\xi}_{X_1}} & \frac{\partial \mathbf{e}^{\text{LIO}}}{\partial \boldsymbol{\xi}_{X_2}} & \mathbf{0}_{6 \times 6} & \mathbf{0}_{6 \times 6} \end{bmatrix} \in \mathbb{R}^{6 \times 24}, \quad (29a)$$

$$\mathbf{J}^{\text{VIO}} = \begin{bmatrix} \mathbf{0}_{6 \times 6} & \mathbf{0}_{6 \times 6} & \frac{\partial \mathbf{e}^{\text{VIO}}}{\partial \boldsymbol{\xi}_{Y_1}} & \frac{\partial \mathbf{e}^{\text{VIO}}}{\partial \boldsymbol{\xi}_{Y_2}} \end{bmatrix} \in \mathbb{R}^{6 \times 24}, \quad (29b)$$

$$\mathbf{J}_1^{\text{det}} = \begin{bmatrix} \frac{\partial \mathbf{e}_1^{\text{det}}}{\partial \boldsymbol{\xi}_{X_1}} & \mathbf{0}_{3 \times 6} & \frac{\partial \mathbf{e}_1^{\text{det}}}{\partial \boldsymbol{\xi}_{Y_1}} & \mathbf{0}_{3 \times 6} \end{bmatrix} \in \mathbb{R}^{3 \times 24}, \quad (29c)$$

$$\mathbf{J}_2^{\text{det}} = \begin{bmatrix} \mathbf{0}_{3 \times 6} & \frac{\partial \mathbf{e}_2^{\text{det}}}{\partial \boldsymbol{\xi}_{X_2}} & \mathbf{0}_{3 \times 6} & \frac{\partial \mathbf{e}_2^{\text{det}}}{\partial \boldsymbol{\xi}_{Y_2}} \end{bmatrix} \in \mathbb{R}^{3 \times 24}, \quad (29d)$$

$$\mathbf{J}_X^{\text{prior}} = \begin{bmatrix} \mathbf{I}_{6 \times 6} & \mathbf{0}_{6 \times 6} & \mathbf{0}_{6 \times 6} & \mathbf{0}_{6 \times 6} \end{bmatrix} \in \mathbb{R}^{6 \times 24}, \quad (29e)$$

$$\mathbf{J}_Y^{\text{prior}} = \begin{bmatrix} \mathbf{0}_{6 \times 6} & \mathbf{I}_{6 \times 6} & \mathbf{0}_{6 \times 6} & \mathbf{0}_{6 \times 6} \end{bmatrix} \in \mathbb{R}^{6 \times 24}, \quad (29f)$$

$$\mathbf{J}_{\text{tilt}} = \begin{bmatrix} \mathbf{I}_{2 \times 2} & \mathbf{0}_{2 \times 4} \end{bmatrix} \in \mathbb{R}^{2 \times 6}, \quad (29g)$$

$$\mathbf{J}_4^{\text{tilt}} = \begin{bmatrix} \mathbf{J}_{\text{tilt}} & \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} \\ \mathbf{0}_{2 \times 6} & \mathbf{J}_{\text{tilt}} & \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} \\ \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} & \mathbf{J}_{\text{tilt}} & \mathbf{0}_{2 \times 6} \\ \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} & \mathbf{J}_{\text{tilt}} \end{bmatrix} \in \mathbb{R}^{8 \times 24}, \quad (29h)$$

corresponding to Jacobian of the LIO odometry factor, VIO odometry factor, detection between $\mathbf{T}_{X_1}^W$ and $\mathbf{T}_{Y_1}^W$, detection between $\mathbf{T}_{X_2}^W$ and $\mathbf{T}_{Y_2}^W$, $SE(3)$ priors on poses $\mathbf{T}_{X_1}^W$ and $\mathbf{T}_{Y_1}^W$, and prior on the roll and pitch angles of all poses, respectively. The values of the partial derivatives are formulated as

$$\frac{\partial \mathbf{e}^{\text{LIO}}}{\partial \boldsymbol{\xi}_{X_1}} = \begin{bmatrix} -(\mathbf{R}_{X_2}^W)^T \mathbf{R}_{X_1}^W & \mathbf{0}_{3 \times 3} \\ (\mathbf{R}_{X_2}^W)^T [\mathbf{t}_{X_2}^W - \mathbf{t}_{X_1}^W]_{\times} \mathbf{R}_{X_1}^W & -(\mathbf{R}_{X_2}^W)^T \mathbf{R}_{X_1}^W \end{bmatrix} \quad (30a)$$

$$\frac{\partial \mathbf{e}^{\text{LIO}}}{\partial \boldsymbol{\xi}_{X_2}} = \mathbf{I}_{6 \times 6} \quad (30b)$$

$$\frac{\partial \mathbf{e}_1^{\text{det}}}{\partial \boldsymbol{\xi}_{X_1}} = \begin{bmatrix} (\mathbf{R}_{X_1}^W)^T [\mathbf{t}_{Y_1}^W - \mathbf{t}_{X_1}^W]_{\times} \mathbf{R}_{X_1}^W & -\mathbf{I}_3 \end{bmatrix}, \quad (30c)$$

$$\frac{\partial \mathbf{e}_1^{\text{det}}}{\partial \boldsymbol{\xi}_{Y_1}} = \begin{bmatrix} \mathbf{0}_3 & (\mathbf{R}_{X_1}^W)^T \mathbf{R}_{Y_1}^W \end{bmatrix}, \quad (30d)$$

with $\frac{\partial \mathbf{e}^{\text{VIO}}}{\partial \boldsymbol{\xi}_{Y_1}}$, $\frac{\partial \mathbf{e}^{\text{VIO}}}{\partial \boldsymbol{\xi}_{Y_2}}$, $\frac{\partial \mathbf{e}_2^{\text{det}}}{\partial \boldsymbol{\xi}_{X_2}}$, and $\frac{\partial \mathbf{e}_2^{\text{det}}}{\partial \boldsymbol{\xi}_{Y_2}}$ formulated analogously by substituting the corresponding variables into the partial derivative equations.

A. Full-Rank Jacobian

The Jacobian of the full factor graph with both odometry measurements, both detections, prior on pose $\mathbf{T}_{X_1}^W$, and prior on the roll and pitch angles is

$$\mathbf{J}^{\text{full}} = [\mathbf{J}^{\text{LIO}}, \mathbf{J}^{\text{VIO}}, \mathbf{J}_1^{\text{det}}, \mathbf{J}_2^{\text{det}}, \mathbf{J}_X^{\text{prior}}, \mathbf{J}_4^{\text{tilt}}]^T \quad (31)$$

and its rank is

$$\text{rank}(\mathbf{J}^{\text{full}}) = 24, \quad (32)$$

demonstrating full observability of the estimation problem.

B. No Global $SE(3)$ Prior

By removing the $SE(3)$ prior on pose $\mathbf{T}_{X_1}^W$, we obtain

$$\mathbf{J}_{\text{prior}}^{\text{no}} = [\mathbf{J}^{\text{LIO}}, \mathbf{J}^{\text{VIO}}, \mathbf{J}_1^{\text{det}}, \mathbf{J}_2^{\text{det}}, \mathbf{J}_4^{\text{tilt}}]^T, \quad (33a)$$

$$\text{rank}(\mathbf{J}_{\text{prior}}^{\text{no}}) = 20, \quad (33b)$$

exhibiting four unobservable DOFs corresponding to the translation and yaw orientation w.r.t. the global frame of reference. As the graph does not contain any global measurements and the $SE(3)$ prior is utilized only to anchor the graph at an initial pose and to make the problem solvable, over time, the estimated poses will exhibit long-term drift with respect to the global reference frame.

C. No Roll and Pitch Prior

By removing the roll and pitch prior from Jacobian $\mathbf{J}^{\text{full}}$, we obtain

$$\mathbf{J}_{\text{rpprior}}^{\text{no}} = [\mathbf{J}^{\text{LIO}}, \mathbf{J}^{\text{VIO}}, \mathbf{J}_1^{\text{det}}, \mathbf{J}_2^{\text{det}}, \mathbf{J}_X^{\text{prior}}]^T, \quad (34a)$$

$$\text{rank}(\mathbf{J}_{\text{rpprior}}^{\text{no}}) = 23. \quad (34b)$$

To obtain the unobservable direction, we calculate the nullspace of the Jacobian as

$$\text{Null}(\mathbf{J}_{\text{rpprior}}^{\text{no}}) = \text{span} \left\{ \begin{bmatrix} \mathbf{0}_{12 \times 1} \\ \mathbf{R}_{Y_1}^{Y_1} (\mathbf{t}_{Y_1}^W - \mathbf{t}_{Y_2}^W) \\ \mathbf{0}_{3 \times 1} \\ \mathbf{R}_{Y_2}^{Y_2} (\mathbf{t}_{Y_1}^W - \mathbf{t}_{Y_2}^W) \\ \mathbf{0}_{3 \times 1} \end{bmatrix} \right\}, \quad (35)$$

showing that when robot Y is moving in a straight line, its 3D orientation cannot be fully determined without additional information. By incorporating the assumption of gravity-alignment of all reference frames in the form of the roll-pitch prior, our approach circumvents this unobservability. Note that the unobservable direction is expressed in the body frame of the robot poses.

D. No Movement of Robot Y

By setting the pose $\mathbf{T}_{Y_2}^W$ equal to pose $\mathbf{T}_{Y_1}^W$, we obtain a situation with no movement of the robot Y :

$$\mathbf{J}_{\text{move}}^{\text{no}} = \mathbf{J}^{\text{full}}|_{\mathbf{T}_{Y_1}^W = \mathbf{T}_{Y_2}^W}, \quad (36a)$$

$$\text{rank}(\mathbf{J}_{\text{move}}^{\text{no}}) = 23, \quad (36b)$$

$$\text{Null}(\mathbf{J}_{\text{move}}^{\text{no}}) = \text{span} \left\{ \begin{bmatrix} \mathbf{0}_{14 \times 1} \\ 1 \\ \mathbf{0}_{5 \times 1} \\ 1 \\ \mathbf{0}_{3 \times 1} \end{bmatrix} \right\}, \quad (36c)$$

showing that without the movement of robot Y , we cannot fully determine the orientation of robot Y , as the relative orientation between the odometry frame of robot Y and the odometry frame of robot X is unobservable.

E. LIO Degradation

In case of a detected LIO degradation, the Σ^{LIO} covariance in eq. (1) will be greatly increased, minimizing the influence of the LIO term on the objective function. Therefore, we remove the corresponding line from the Jacobian $\mathbf{J}^{\text{full}}$. By removing the part corresponding to the LIO measurement from the full Jacobian and analyzing the result, we can ascertain which subspace of the estimated variables is not constrained by other measurements and thus cannot be corrected in the event of LIO degradation. We assume that the degradation happens after a period of normal operation, and the graph contains valid prior estimates of both robot poses. We construct the Jacobian of a factor graph with LIO degradation as

$$\mathbf{J}_{\text{deg}}^{\text{LIO}} = [\mathbf{J}^{\text{VIO}}, \mathbf{J}_1^{\text{det}}, \mathbf{J}_2^{\text{det}}, \mathbf{J}_X^{\text{prior}}, \mathbf{J}_Y^{\text{prior}}, \mathbf{J}_4^{\text{tilt}}]^T, \quad (37a)$$

$$\text{rank}(\mathbf{J}_{\text{deg}}^{\text{LIO}}) = 23, \quad (37b)$$

$$\text{Null}(\mathbf{J}_{\text{deg}}^{\text{LIO}}) = \text{span} \left\{ \begin{bmatrix} \mathbf{0}_{8 \times 1} \\ 1 \\ \left[\mathbf{R}_W^{X_2} (\mathbf{t}_{X_2}^W - \mathbf{t}_{Y_2}^W)^\perp \right]_x \\ \left[\mathbf{R}_W^{X_2} (\mathbf{t}_{X_2}^W - \mathbf{t}_{Y_2}^W)^\perp \right]_y \\ \mathbf{0}_{13 \times 1} \end{bmatrix} \right\}, \quad (37c)$$

indicating that in such a case, there is insufficient information to distinguish between the change in the yaw of the robot X and its movement in the direction perpendicular to the line connecting the robots. The notation $\mathbf{x}^\perp$ represents a vector perpendicular to vector $\mathbf{x}$ and the subscripts $[\mathbf{a}]_x$ and $[\mathbf{a}]_y$ represent the x - or y -component of vector $\mathbf{a}$. By inserting an additional detected robot Z , we obtain

$$\mathbf{J}_{\text{deg}}^{\text{LIO}'} = \begin{bmatrix} \mathbf{J}_{\text{deg}}^{\text{LIO}} & \mathbf{0}_{32 \times 6} & \mathbf{0}_{32 \times 6} \\ \mathbf{0}_{6 \times 6} & \mathbf{0}_{6 \times 6} & \mathbf{0}_{6 \times 6} & \mathbf{0}_{6 \times 6} & \frac{\partial \mathbf{e}_3^{\text{det}}}{\partial \xi_{X_1}} & \mathbf{0}_{3 \times 6} & \mathbf{0}_{3 \times 6} & \mathbf{0}_{3 \times 6} & \frac{\partial \mathbf{e}_3^{\text{det}}}{\partial \xi_{Z_1}} & \mathbf{0}_{3 \times 6} \\ \frac{\partial \mathbf{e}_3^{\text{det}}}{\partial \xi_{X_1}} & \mathbf{0}_{3 \times 6} & \mathbf{0}_{3 \times 6} & \mathbf{0}_{3 \times 6} & \frac{\partial \mathbf{e}_3^{\text{det}}}{\partial \xi_{Z_1}} & \mathbf{0}_{3 \times 6} & \mathbf{0}_{3 \times 6} & \mathbf{0}_{3 \times 6} & \frac{\partial \mathbf{e}_3^{\text{det}}}{\partial \xi_{Z_2}} & \mathbf{0}_{3 \times 6} \\ \mathbf{0}_{3 \times 6} & \frac{\partial \mathbf{e}_4^{\text{det}}}{\partial \xi_{X_2}} & \mathbf{0}_{3 \times 6} & \mathbf{0}_{3 \times 6} & \mathbf{0}_{3 \times 6} & \frac{\partial \mathbf{e}_4^{\text{det}}}{\partial \xi_{Z_2}} & \mathbf{0}_{3 \times 6} & \mathbf{0}_{3 \times 6} & \mathbf{0}_{3 \times 6} & \mathbf{0}_{3 \times 6} \\ \mathbf{0}_{6 \times 6} & \mathbf{0}_{6 \times 6} & \mathbf{0}_{6 \times 6} & \mathbf{0}_{6 \times 6} & \mathbf{I}_{6 \times 6} & \mathbf{0}_{6 \times 6} & \mathbf{0}_{6 \times 6} & \mathbf{0}_{6 \times 6} & \mathbf{0}_{6 \times 6} & \mathbf{0}_{6 \times 6} \\ \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} & \mathbf{J}_{\text{tilt}} & \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} \\ \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} & \mathbf{J}_{\text{tilt}} & \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} & \mathbf{0}_{2 \times 6} \end{bmatrix}, \quad (38a)$$

$$\text{rank}(\mathbf{J}_{\text{deg}}^{\text{LIO}'}) = 36. \quad (38b)$$

The resulting Jacobian is full-rank, showing that the problem becomes fully observable when two distinct robots with reliable pose estimates are detected.

F. VIO Degradation

In case of a VIO degradation, the influence of the VIO term from eq. (1) is minimized. Therefore, we obtain

$$\mathbf{J}_{\text{deg}}^{\text{VIO}} = [\mathbf{J}^{\text{LIO}}, \mathbf{J}_1^{\text{det}}, \mathbf{J}_2^{\text{det}}, \mathbf{J}_X^{\text{prior}}, \mathbf{J}_Y^{\text{prior}}, \mathbf{J}_4^{\text{tilt}}]^\top, \quad (39a)$$

$$\text{rank}(\mathbf{J}_{\text{deg}}^{\text{VIO}}) = 23, \quad (39b)$$

$$\text{Null}(\mathbf{J}_{\text{deg}}^{\text{VIO}}) = \text{span} \left\{ \begin{bmatrix} \mathbf{0}_{20 \times 1} \\ 1 \\ \mathbf{0}_{3 \times 1} \end{bmatrix} \right\}, \quad (39c)$$

indicating that the yaw of pose $\mathbf{T}_{Y_2}^W$ is unobservable from the available measurements. Such a situation cannot be improved by adding additional robots, as the detections do not provide any orientation information about the robot Y , and the robot Y itself cannot detect other robots.

IV. EXPERIMENTAL VERIFICATION

The proposed cooperative localization approach was quantitatively evaluated on custom datasets gathered in an indoor environment with motion-capture ground-truth data and in a large-scale outdoor environment with RTK ground-truth data.

In all evaluations, the LIO-SAM algorithm [40] was used as the LIO of robot X . The LIO-SAM algorithm was modified to work with the six-axis IMU of the Ouster 3D LiDAR; thus, the LIO does not have information about the global yaw orientation. The internal threshold of LIO-SAM for considering the scan matching problem degenerated was empirically tuned for use with the employed 3D LiDAR. The OpenVINS

Table II
PARAMETER VALUES USED IN THE EXPERIMENTS. ALL THE PRESENTED RESULTS WERE OBTAINED WITH THE SAME SET OF PARAMETERS. μ_Y AND μ_Z REPRESENT DIFFERENT VALUES OF μ FOR UAVS Y AND Z SELECTED DUE TO THE UAVS' DIFFERENT CONSTRUCTION RESULTING IN DIFFERENT ODOMETRY ACCURACY.

Parameter	Value
odometry rate	2 Hz
detection rate	10 Hz
smoother window length	30 s
λ_{thr}	430
σ_{det}	0.13 m
$lo \sigma_{\text{pos}}$	0.01 m
$hi \sigma_{\text{pos}}$	5.0 m
$lo \sigma_\gamma$	0.001 rad
$hi \sigma_\gamma$	1.0 rad
$vmin \sigma_{\text{pos}}$	0.1 m
$vmax \sigma_{\text{pos}}$	5.0 m
ν	20
μ_Y	260
μ_Z	500
$v \sigma_\gamma$	0.01 rad

algorithm [41] was utilized as the VIO of robot Y . Parameters of OpenVINS were tuned for reliable performance on the real-world data. Note that the parameters of OpenVINS were not the same on all datasets due to the use of different cameras in the indoor and outdoor experiments, different constructions of the camera-equipped UAVs, and different amounts of visual features in the environments. For obtaining the relative detections of robot positions, we placed reflective markers on the UAVs representing the detected robots Y and Z and we utilized a 3D LiDAR-based detector inspired by the tracking module of [42] with the detections initialized based on the reflective markers and using centroids of clusters of LiDAR points as the robot detections. The system time of the robots was synchronized over a wireless network using the chrony implementation of the Network Time Protocol (NTP). The sensor data were recorded on board the robots and processed offline. Table II shows data rates and parameters used by the proposed algorithm. These parameter values were the same for all experiments.

When processing the ground-truth measurements and evaluating the accuracy of a robot trajectory, we followed the approach proposed in [43]. For each individual estimated trajectory, we calculated ground-truth data interpolated to the times of the estimate, aligned the estimated data to the ground-truth data using the Kabsch-Umeyama algorithm [39], and calculated the Absolute Trajectory Error (ATE) as the Root Mean Squared Error (RMSE) of the estimated position/yaw orientation. In all plots of the estimated data, the estimates were plotted in the ground-truth reference frame after the alignment.

A. Real-World UGV-UAV Indoor Experiments

As the LiDAR-carrying robot X , we utilized the Unitree B1 quadruped carrying the Ouster OS0-128 LiDAR. As the camera-equipped robot Y , we used a small custom-made UAV

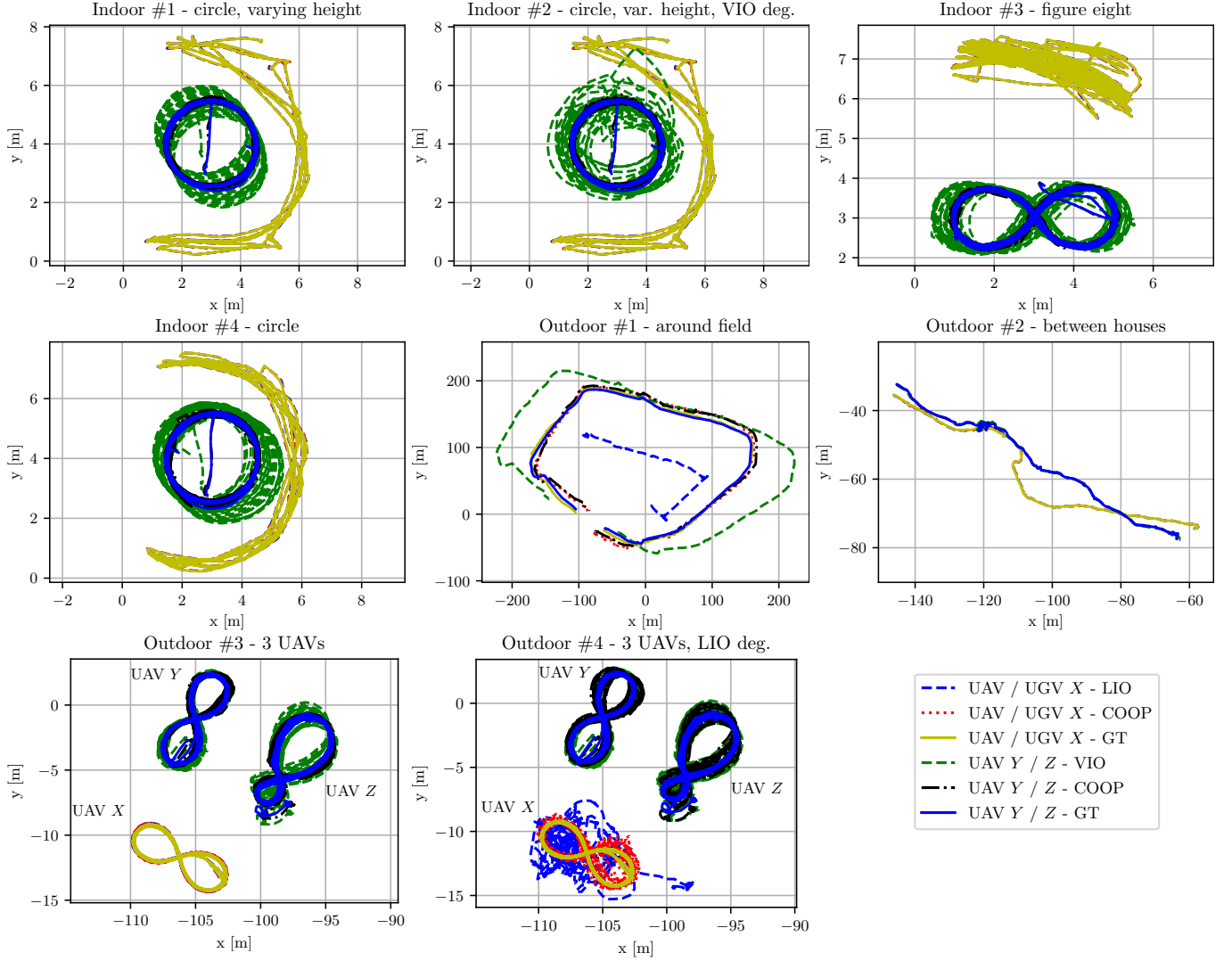


Figure 5. xy -plots of odometry data (LIO, VIO), the proposed cooperative localization method (COOP), and ground truth (GT) from the experimental evaluation. All estimated trajectories were aligned to the ground-truth data for evaluation.

based on the DJI F330 frame, equipped with the RealSense T265 tracking camera and with reflective markers placed on its legs (see Fig. 1). Ground-truth measurements were obtained using the OptiTrack motion capture system.

We performed four different experiments. The localization errors from these experiments are listed in Table III. The xy -plots of data obtained in these experiments are shown in Fig. 5, and the altitude and yaw plots are shown in Fig. 6. The experiments were performed in a feature-rich environment, where the LIO was considered reliable the entire time.

In the *Indoor #1* experiment, the UAV Y was flying in a circle with varying altitude, while the UGV X was walking around the circle. For the UGV, the error of the proposed cooperative localization method (COOP) was the same as the error of the LIO, as the LIO was considered reliable the entire time. For the UAV, both the positional and rotational errors were significantly reduced.

In the *Indoor #2* experiment, we utilized the same data as in the previous case, but we artificially created strong VIO degradation in part of the trajectory. For half of the UAV's

trajectory (when $x > 3$), the camera images were replaced with completely black images. In the degraded part of the trajectory, the VIO needed to rely on IMU data only and therefore exhibited significant drift, mainly in the z -axis. For the UGV, the COOP error stayed approximately the same as the LIO error. For the UAV, the VIO error was corrected by the COOP method, with the 3D ATE decreasing from 4.671 m to 0.101 m.

In the *Indoor #3* and *Indoor #4* experiments, the UAV X was flying in a figure-eight and circular trajectory, respectively, with the UGV walking next to it. For the UGV, the COOP error again stayed approximately the same as the error of LIO. For the UAV, the positional error was significantly reduced. In the figure-eight trajectory, the UAV's VIO rotational error was already very low, and the COOP rotational error stayed approximately the same. In the circular trajectory, the UAV's rotational error was significantly reduced.

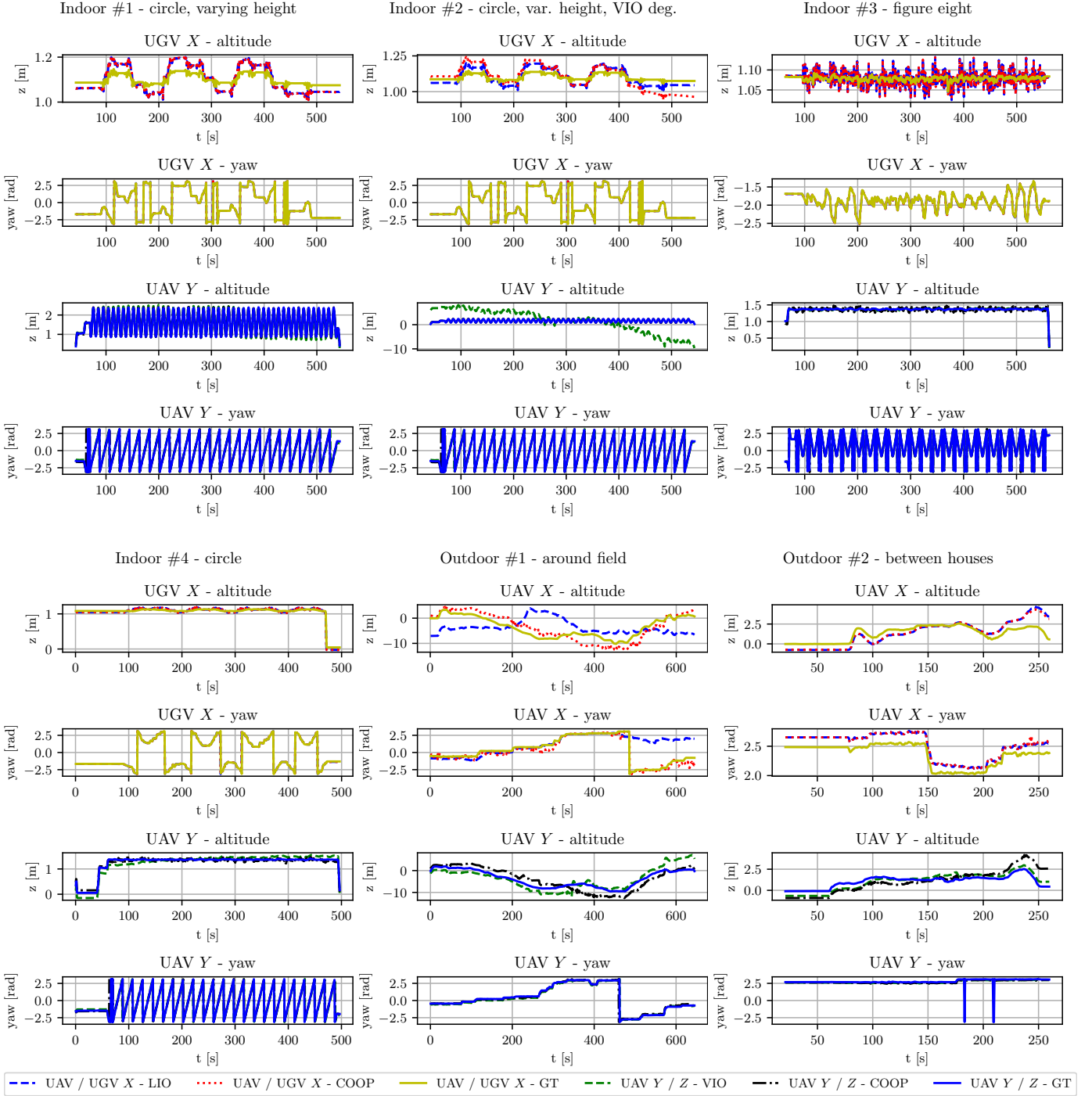


Figure 6. Altitude and yaw plots of odometry data (LIO, VIO), the proposed cooperative localization method (COOP), and ground-truth data (GT).

B. Real-World Multi-UAV Outdoor Experiments

A custom-made UAV based on the X500 frame was used as the robot *X* [44]. The UAV carried the Ouster OS0-128 LiDAR. As the camera-equipped robots *Y* and *Z*, we utilized two smaller UAVs equipped with the mvBlueFOX-MLC200wG cameras and the ICM-42688-P IMU, providing data for the VIO (see Fig. 1). The smaller UAVs each had a slightly different construction, resulting in different amounts of propeller-induced vibrations acting upon the IMU, thus resulting in different VIO accuracy for each UAV. Ground-truth positional

measurements were provided by the Holybro H-RTK F9P modules. Ground-truth measurements of yaw were obtained from the onboard flight controller units processing external magnetometer data from the F9P modules. The localization errors from the outdoor experiments are listed in Table III, the *xy*-plots are shown in Fig. 5, and the altitude and yaw plots are shown in Fig. 6 and Fig. 7.

Four different experiments were made. In the *Outdoor #1*, the UAVs *X* and *Y* flew around a large-scale open field (see Fig. 8). The mutual distance between the UAVs ranged

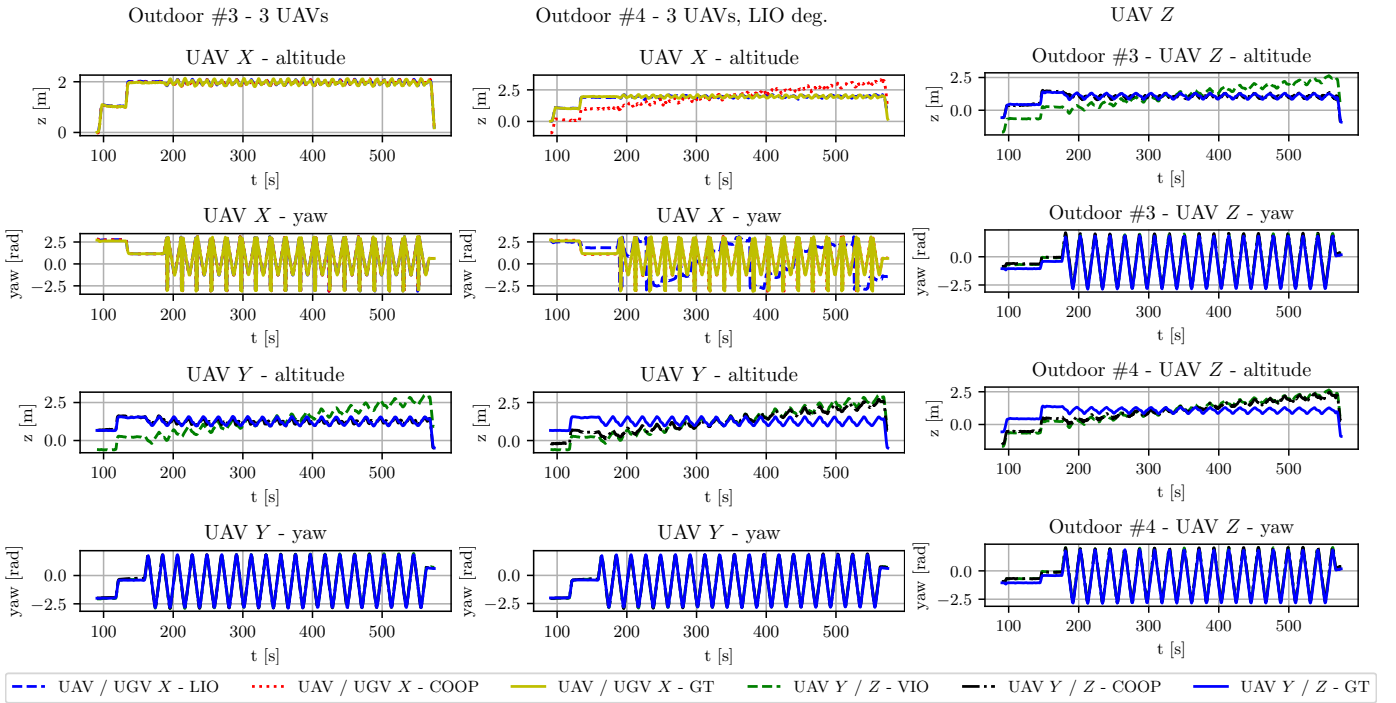


Figure 7. Continuation of the altitude and yaw plots of odometry data (LIO, VIO), the proposed cooperative localization method (COOP), and ground truth data (GT).

between 2.9m and 24.9m with the mean mutual distance of 9.9m. The experiment was characterized by large portions of LIO degeneracy. The LIO was reliable only for limited times, mainly at the right and top parts of the trajectory, when the UAVs flew close to the trees next to the field. The rest of the time, the LiDAR data contained only the featureless ground plane with no geometric features usable to sufficiently constrain the scan matching problem. Although the VIO of UAV Y was usable the entire time, it exhibited significant long-term drift, significantly increasing the scale of the entire VIO trajectory (see Fig. 5 - Outdoor #1). The proposed cooperative localization method managed to combine the data, significantly decreasing the localization errors for both UAVs. The localization error of UAV X decreased from 74.591 m to 10.373 m in position and from 1.219 rad to 0.360 rad in yaw. The error of UAV Y decreased from 46.458 m to 10.692 m in 3D position and from 0.131 rad to 0.092 rad in yaw.

In the *Outdoor #2* experiment, the UAVs flew in an environment between several buildings, where the LIO was reliable for the entire time. The COOP error of UAV X stayed approximately the same as the LIO error. The COOP error of UAV Y significantly decreased in 2D, but slightly increased in 3D due to the altitude estimation error of the LIO at the end of the trajectory.

In the *Outdoor #3* experiment, one LiDAR-equipped UAV X and two camera-equipped UAVs Y and Z were utilized. The UAVs were flying in figure-eight trajectories, and the LIO of UAV X was reliable the entire time. For the UAV X, the localization error of COOP stayed approximately the same as the LIO localization error. For the UAV Y, the error of COOP was similar to the VIO error with significant error decrease in

the z -axis. For the UAV Z, there was a significant decrease in both the 2D and 3D error.

The *Outdoor #4* experiment utilized the same raw data as the previous experiment. In this experiment, we artificially created LIO degradation. After the UAVs started following their trajectories, the point clouds incoming into the LIO algorithm were modified by removing all points above the ground plane, thus removing features usable for scan matching. This feature removal resulted in the LIO data being useless in the xy -plane (see Fig. 5 - Outdoor #4), as well as in yaw (see Fig. 7 - Outdoor #4). Utilizing the VIO data from the remaining two UAVs and their detections, the proposed cooperative localization method was able to correctly estimate the pose of UAV X and reconstruct its figure-eight trajectory. The localization error of UAV X was reduced from 4.077 m to 0.875 m in 3D position and from 1.668 rad to 0.067 m in yaw orientation. It is worth mentioning that the use of a single detected UAV was insufficient for correctly estimating the changing pose of UAV X in the presence of such strong LIO degradation. This corresponds to the findings of the observability analysis in sec. III-E, which showed that in the case of a LIO degeneracy and a single detected robot, there is one unobservable DOF, while in the case of two detected robots, this unobservability is mitigated.

C. Evaluating the Use of Wasserstein Distance

We want to verify the assumption that the Wasserstein distance between the covariance matrices of two consecutive VIO positions correlates with the real-world localization error and validate the decision to set the standard deviations of the VIO relative position measurements proportional to the

Table III

POSITIONAL AND ROTATIONAL LOCALIZATION ERRORS FROM THE EXPERIMENTAL EVALUATION. LOCALIZATION ERRORS OF THE PROPOSED COOPERATIVE LOCALIZATION METHOD (COOP) AND OF THE RESPECTIVE INDIVIDUAL ODOMETRY METHODS (LIO, VIO) ARE COMPARED. THE MINIMAL ERRORS FOR EACH ROBOT IN EACH DATASET ARE IN BOLD.

Dataset	Robot	Method	2D ATE [m]	3D ATE [m]	rot. ATE [rad]
Indoor #1 circle var. height	UGV X	LIO	0.029	0.047	0.013
		COOP	0.029	0.047	0.013
	UAV Y	VIO	0.483	0.489	0.155
		COOP	0.065	0.079	0.041
Indoor #2 circle var. height VIO deg.	UGV X	LIO	0.029	0.047	0.013
		COOP	0.030	0.068	0.013
	UAV Y	VIO	0.497	4.671	0.104
		COOP	0.076	0.101	0.043
Indoor #3 figure eight	UGV X	LIO	0.025	0.029	0.010
		COOP	0.026	0.029	0.010
	UAV Y	VIO	0.280	0.281	0.012
		COOP	0.076	0.089	0.013
Indoor #4 circle	UGV X	LIO	0.023	0.044	0.017
		COOP	0.027	0.046	0.017
	UAV Y	VIO	0.440	0.458	0.137
		COOP	0.084	0.098	0.048
Outdoor #1 around field	UAV X	LIO	74.386	74.591	1.219
		COOP	10.091	10.373	0.360
	UAV Y	VIO	46.391	46.458	0.131
		COOP	10.425	10.692	0.092
Outdoor #2 between houses	UAV X	LIO	0.145	0.972	0.154
		COOP	0.155	0.902	0.156
	UAV Y	VIO	0.553	0.715	0.112
		COOP	0.162	0.896	0.109
Outdoor #3 3 UAVs	UAV X	LIO	0.194	0.200	0.052
		COOP	0.198	0.204	0.051
	UAV Y	VIO	0.295	0.923	0.069
		COOP	0.315	0.333	0.063
	UAV Z	VIO	0.595	1.010	0.197
		COOP	0.311	0.328	0.239
Outdoor #4 3 UAVs LIO deg.	UAV X	LIO	4.076	4.077	1.668
		COOP	0.525	0.875	0.067
	UAV Y	VIO	0.294	0.923	0.069
		COOP	0.354	0.767	0.060
	UAV Z	VIO	0.595	1.010	0.197
		COOP	0.550	0.880	0.210

Wasserstein distance. For each experiment and each respective camera-equipped UAV, we calculated the Pearson correlation coefficient of the Wasserstein distance corresponding to each relative VIO position measurement and the norm of the 3D relative position error, calculated from the ground-truth data. Table IV shows the obtained correlation coefficients. A positive correlation between the Wasserstein distance and relative position error was found in all datasets. The largest correlations were observed in the *Indoor #2* and *Outdoor #1* experiments, which contained the largest variations in VIO accuracy. In the *Indoor #2* experiment, we repeatedly created VIO degradations by blocking the camera image, and in the *Outdoor #1* experiment, the number of visual features varied significantly while the UAV was flying around the field.

Fig. 9 shows a comparison of the VIO and LIO relative position errors and the σ_{pos} utilized by the proposed coop-

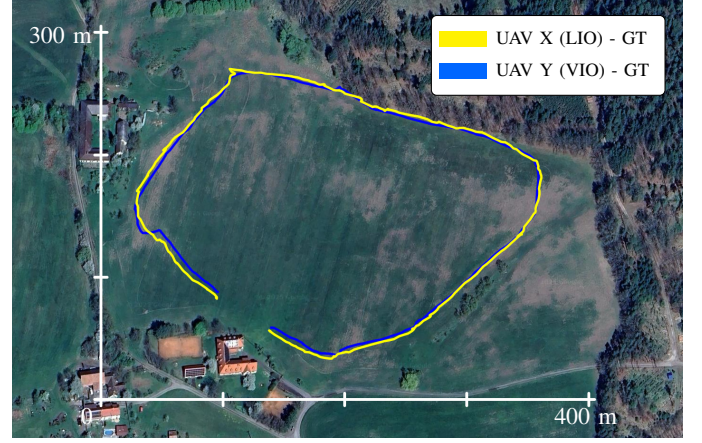


Figure 8. Ground-truth trajectories from the *Outdoor #1* experiment, where the two UAVs flew around an open field.

Table IV

PEARSON CORRELATION COEFFICIENTS OF WASSERSTEIN DISTANCES AND THE NORMS OF RELATIVE 3D POSITION ERROR OF THE VIO. THE VALUES FOR OUTDOOR #4 ARE EQUAL TO THE VALUES OF OUTDOOR #3, AS THE VIO DATA WERE EQUAL.

Dataset	Correlation coefficient
Indoor #1	0.221
Indoor #2	0.758
Indoor #3	0.590
Indoor #4	0.563
Outdoor #1	0.807
Outdoor #2	0.644
Outdoor #3 - UAV X	0.460
Outdoor #3 - UAV Y	0.597

erative localization method from the *Outdoor #1* experiment. This data showcases the necessity of having the measurement covariance matrices adaptively react to changing odometry accuracy. The LIO relative position error plot shows several moments in time when the LIO error was low, surrounded by periods of low accuracy. The σ_{pos} of UAV X in the bottom plot reflects this changing accuracy.

Similarly, the σ_{pos} of UAV Y reflects the changing localization accuracy of the VIO. Notably, there was a large increase of VIO localization error after $t = 320$ s with two sharp peaks at $t = 326$ s and $t = 342$ s. Failing to detect such an increase in localization error would lead to inconsistency between the measurements obtained from the detections and the VIO data and would result in large estimation errors. In our experiments, utilizing a static covariance matrix, which did not reflect this large localization error, resulted in loss of the estimate due to large localization error and subsequent failure to associate new detections to the estimate due to exceeding the maximum allowed detection distance. It is worth mentioning that such sharp peaks in localization error and in the Wasserstein distance are often produced after loss and subsequent re-detection of visual features. The employed VIO algorithm, OpenVINS, tracks a fixed number of SLAM features in its state vector throughout the estimation process.

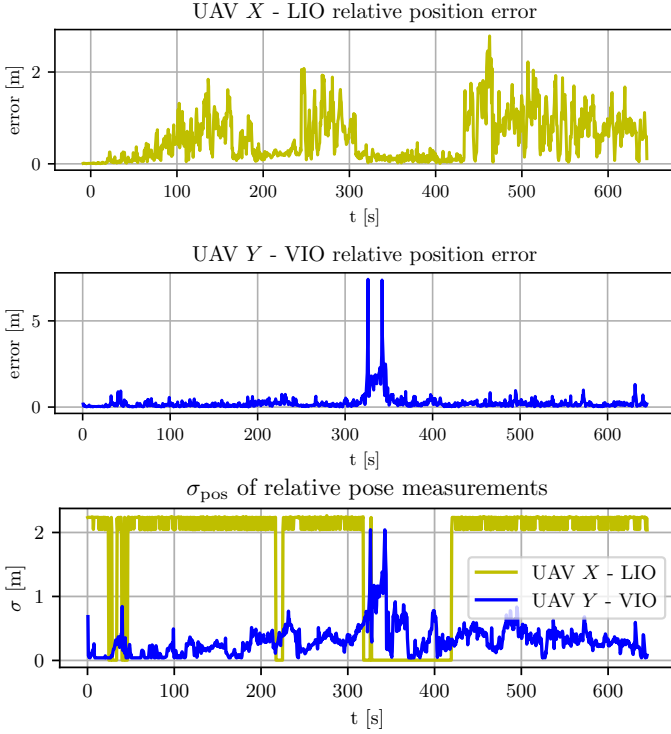


Figure 9. Comparison of the LIO relative position error, VIO relative position error, and the σ_{pos} values used in the covariance matrices of the proposed cooperative localization method from the *Outdoor #1* experiment.

After the features are lost, the correction caused by regaining the features can cause a sudden jump in the VIO estimate, accompanied by a sudden decrease in the covariance matrix of the VIO position. The Wasserstein distance can be used to conveniently quantify such a change in localization accuracy and lower the weight of the corresponding relative position measurement in the factor graph.

D. Investigating the Findings of the Observability Analysis

As established theoretically in sec. III, in the case of odometry degradations, the tackled cooperative localization problem exhibits unobservable directions. In case of LIO degradation, i.e., odometry degradation of the detecting robot X , there is one unobservable DOF in the direction of the yaw of robot X and the vector perpendicular to the line connecting the robots. In case of VIO degradation, i.e., odometry degradation of the detected robot Y , the yaw orientation of robot Y is unobservable. We explore these findings experimentally to verify them and to provide further insight into the problem.

1) *Unobservable DOF of the Detecting Robot X*: We utilized data from the experiment *Outdoor #3*, where one detecting UAV X and two detected UAVs Y and Z were following figure-eight trajectories. We overwrote the eigenvalues of the LIO's approximate Hessian to consider the LIO output degenerated between $t = 300$ s and $t = 400$ s. We compared three different situations. First, we utilized only a single detected UAV Y in the cooperative localization process. Second, we utilized only UAV Y and changed $^{\text{hi}}\sigma_{\gamma}$ to the same value as $^{\text{lo}}\sigma_{\gamma}$, forcing the algorithm to consider the yaw values

of the LIO measurements as reliable for the entire time. Third, we utilized both of the detected UAVs Y and Z .

Fig. 10 shows data from the analysis. To analyze the consistency of information in the factor graph, we composed each UAV detection with the pose of UAV X at the corresponding timestamp, and compared it with the estimated poses of the detected UAVs Y and Z , interpolated to the same timestamp. Furthermore, we plotted the estimated positions of UAV X at the same timestamp. The left plot in Fig. 10 shows that there is no significant discrepancy between the detections composed with the poses of UAV X and the interpolated poses of UAV Y , although the poses of UAV X exhibit significant drift. The position drift is caused by the unobservable direction, as the estimated poses of UAV X fulfill the optimization objective to minimize the error of the detection factors and the error of the VIO measurement factors, but at the time of the LIO degeneracy, the poses of the UAV X are not fully constrained, resulting in the observed drift.

In the central plot of Fig. 10, we constrained the yaw measurements from the LIO algorithm to always be considered reliable. Such a constraint eliminated the unobservable direction, and the position drift was mitigated. In the right plot of Fig. 10, we instead utilized two detected UAVs at the same time. The position drift of UAV X disappeared, because the unobservable direction was mitigated by utilizing multiple detected UAVs, confirming the findings of the observability analysis from sec. III-E.

Note that in the *Outdoor #1* experiment, which used a single detected UAV and contained large areas of LIO degradation, the drift caused by the unobservable direction was not as pronounced, as the UAVs were often changing their relative position, resulting in changes in the unobservability direction, and the errors caused by the unobservability were much smaller than the localization improvements provided by the cooperative localization method.

2) *Yaw Unobservability of the Detected Robot Y*: We utilized data from the *Indoor #1* experiment. Between $t = 300$ s and $t = 400$ s, we artificially inserted cumulative drift of the yaw orientation of UAV Y with a constant velocity of 0.05 rad/s. Fig. 11 shows a comparison of the resulting yaw of VIO, the yaw estimated by the cooperative localization method, and the ground-truth yaw. The yaw from VIO and from the cooperative localization method coincide the entire time, and there is an increasing offset from the ground-truth data. The offset stays constant even after the drift stops being added at $t = 400$ s because the yaw orientation of UAV Y is unobservable from other data than the VIO measurements. This behavior corresponds to the findings of the theoretical analysis in sec. III-F.

E. Discussion of Results

The experiments have confirmed the ability of the proposed cooperative localization method to adapt to changing conditions and significantly improve the localization accuracy in the presence of sensory degradations. The quantitative evaluation of experiments in sec. IV-A and sec. IV-B has shown the localization accuracy of the cooperative localization method to

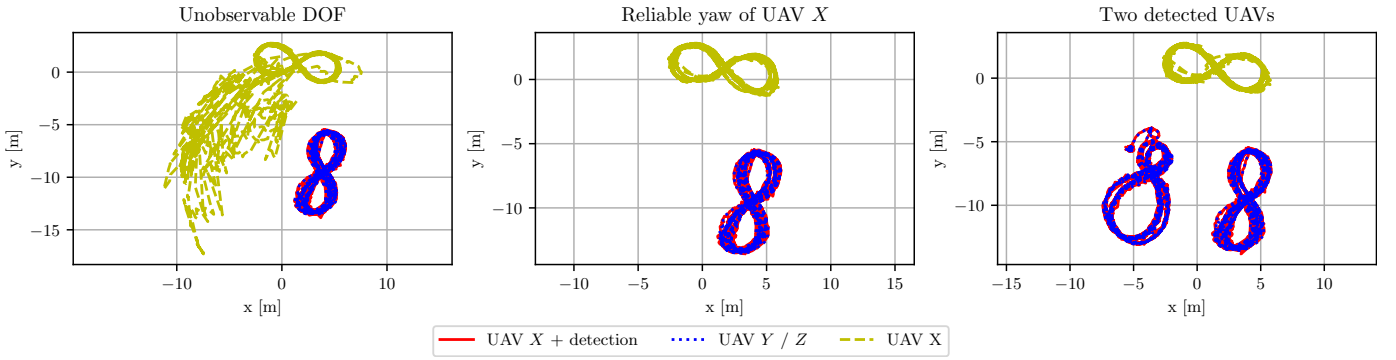


Figure 10. The effect of the unobservable direction during LIO degradation. The left plot shows the situation with one detected UAV and the unconstrained unobservable direction. The central plot shows the same situation when the yaw of UAV X is always considered reliable. The right plot shows the same situation as the left plot, but with two detected UAVs.

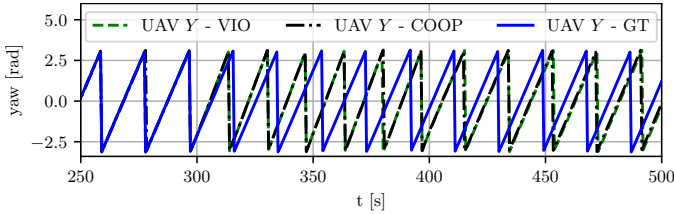


Figure 11. The effect of unobservable yaw of the detected UAV Y in the presence of pure yaw drift. We artificially inserted constant-velocity yaw drift between $t = 300$ s and $t = 400$ s.

be similar to the accuracy of the individual robot odometries when no sensory degradations are present and to significantly improve the localization performance in the presence of sensory degradations. Simultaneously, the method provides localization in a common reference frame, which is crucial for enabling any cooperation between the robots. As shown in sec. IV-C, the Wasserstein distance between the covariance matrices of consecutive VIO positions highly correlates with the real-world localization error, especially in the presence of visual degradations. Weighting the relative VIO measurements proportionally to the Wasserstein distance is thus crucial for the adaptivity of the algorithm. Finally, the findings of the theoretical analysis regarding the unobservable directions in the presence of odometry degradations were experimentally verified in sec. IV-D. The experiments have confirmed that there is one unobservable DOF in the presence of odometry degradation of the detecting robot, which can be mitigated by utilizing multiple detected robots. Finally, the experiments have demonstrated the unobservability of the yaw orientation of the detected UAV. To mitigate this unobservability, the method would need to fuse an additional source of information about the yaw of the detected robot, or the detected robot itself would need to have the ability to detect other robots.

V. CONCLUSIONS

A novel multi-modal multi-robot adaptive cooperative localization approach was proposed in this paper. The proposed method fuses LIO and VIO measurements from different robots with direct 3D detections between the robots in a loosely-coupled fashion using a factor graph-based formula-

tion. A novel interpolation-based quaternary factor enables efficient fusion of the data from unsynchronized sources of measurements. The approach adaptively reacts to the changing reliability of the measurements to enhance the localization performance in the presence of sensory degradations. The degradation of LIO measurements is detected by analyzing the eigenvalues of the approximate Hessian of the scan matching problem. The changing uncertainty of VIO measurements is evaluated in a novel approach based on the Wasserstein distance between the covariance matrices of consecutive VIO positions. Theoretical analysis was performed to analyze the observability of the tackled cooperative localization problem under various conditions, especially in the presence of odometry degradations. The proposed method and the findings of the theoretical analysis were extensively evaluated on real-world data gathered with a UGV-UAV and a UAV-only team. The accuracy of the approach was quantitatively evaluated with respect to motion capture and RTK ground truth. The experiments have shown that the proposed cooperative localization method provides significant improvements in localization accuracy in the presence of sensory degradations.

REFERENCES

- [1] K. Ebadi, L. Bernreiter, H. Biggie, G. Catt, Y. Chang, A. Chatterjee, C. E. Denniston, S.-P. Deschênes, K. Harlow, S. Khattak, L. Nogueira, M. Palieri, P. Petráček, M. Petrlík, A. Reinke, V. Krátký, S. Zhao, A.-a. Agha-mohammadi, K. Alexis, C. Heckman, K. Khosoussi, N. Kottege, B. Morrell, M. Hutter, F. Pauling, F. Pomerleau, M. Saska, S. Scherer, R. Siegwart, J. L. Williams, and L. Carlone, "Present and Future of SLAM in Extreme Environments: The DARPA SubT Challenge," *IEEE Transactions on Robotics*, vol. 40, pp. 936–959, 2024.
- [2] J. Zhang and S. Singh, "Laser-visual-inertial odometry and mapping with high robustness and low drift," *Journal of Field Robotics*, vol. 35, no. 8, pp. 1242–1264, 2018.
- [3] S. Khattak, H. Nguyen, F. Mascarchi, T. Dang, and K. Alexis, "Complementary Multi-Modal Sensor Fusion for Resilient Robot Pose Estimation in Subterranean Environments," in *2020 International Conference on Unmanned Aircraft Systems (ICUAS)*, Sep. 2020, pp. 1024–1029.
- [4] S. Zhao, H. Zhang, P. Wang, L. Nogueira, and S. Scherer, "Super Odometry: IMU-centric LiDAR-Visual-Inertial Estimator for Challenging Environments," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2021, pp. 8729–8736.
- [5] J. Graeter, A. Wilczynski, and M. Lauer, "LIMO: Lidar-Monocular Visual Odometry," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2018, pp. 7872–7879.

- [6] T. Shan, B. Englot, C. Ratti, and D. Rus, "LVI-SAM: Tightly-coupled Lidar-Visual-Inertial Odometry via Smoothing and Mapping," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, May 2021, pp. 5692–5698.
- [7] J. Lin and F. Zhang, "R³LIVE++: A Robust, Real-Time, Radiance Reconstruction Package With a Tightly-Coupled LiDAR-Inertial-Visual State Estimator," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 11 168–11 185, Dec. 2024.
- [8] W. Lee, P. Geneva, C. Chen, and G. Huang, "MINS: Efficient and Robust Multisensor-Aided Inertial Navigation System," *Journal of Field Robotics*, 2025.
- [9] Z. Yuan, J. Deng, R. Ming, F. Lang, and X. Yang, "SR-LIVO: LiDAR-Inertial-Visual Odometry and Mapping With Sweep Reconstruction," *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5110–5117, Jun. 2024.
- [10] H. Zhang, L. Du, S. Bao, J. Yuan, and S. Ma, "LVIO-Fusion: Tightly-Coupled LiDAR-Visual-Inertial Odometry and Mapping in Degenerate Environments," *IEEE Robotics and Automation Letters*, vol. 9, no. 4, pp. 3783–3790, Apr. 2024.
- [11] C. Zheng, W. Xu, Z. Zou, T. Hua, C. Yuan, D. He, B. Zhou, Z. Liu, J. Lin, F. Zhu, Y. Ren, R. Wang, F. Meng, and F. Zhang, "FAST-LIVO2: Fast, Direct LiDAR-Inertial-Visual Odometry," *IEEE Transactions on Robotics*, vol. 41, pp. 326–346, 2025.
- [12] J. Xu, T. Li, H. Wang, Z. Wang, T. Bai, and X. Hou, "Intermittent VIO-Assisted LiDAR SLAM Against Degeneracy: Recognition and Mitigation," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–13, 2025.
- [13] Y. Wang, W. Yao, B. Zhang, G. Sun, B. Zheng, and T. Cao, "MCLIVO: A low-drift LiDAR-inertial-visual odometry with multi-constrained optimization for planetary mapping," *Measurement*, vol. 240, p. 115551, Jan. 2025.
- [14] J. Zhang, M. Kaess, and S. Singh, "On degeneracy of optimization-based state estimation problems," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 809–816.
- [15] T. Tuna, J. Nubert, P. Pfreundschuh, C. Cadena, S. Khattak, and M. Hutter, "Informed, Constrained, Aligned: A Field Analysis on Degeneracy-aware Point Cloud Registration in the Wild," *IEEE Transactions on Field Robotics*, vol. 2, pp. 485–515, 2025.
- [16] T. Tuna, J. Nubert, Y. Nava, S. Khattak, and M. Hutter, "X-ICP: Localizability-Aware LiDAR Registration for Robust Localization in Extreme Environments," *IEEE Transactions on Robotics*, vol. 40, pp. 452–471, 2024.
- [17] J. Horyna, V. Krátký, V. Pritzl, T. Báča, E. Ferrante, and M. Saska, "Fast Swarming of UAVs in GNSS-Denied Feature-Poor Environments Without Explicit Communication," *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5284–5291, Jun. 2024.
- [18] J. G. Mangelson, M. Ghaffari, R. Vasudevan, and R. M. Eustice, "Characterizing the Uncertainty of Jointly Distributed Poses in the Lie Algebra," *IEEE Transactions on Robotics*, vol. 36, no. 5, pp. 1371–1388, Oct. 2020.
- [19] S. Zhong, Y. Qi, Z. Chen, J. Wu, H. Chen, and M. Liu, "DCL-SLAM: A Distributed Collaborative LiDAR SLAM Framework for a Robotic Swarm," *IEEE Sensors Journal*, vol. 24, no. 4, pp. 4786–4797, Feb. 2024.
- [20] Y. Huang, T. Shan, F. Chen, and B. Englot, "DiSCo-SLAM: Distributed Scan Context-Enabled Multi-Robot LiDAR SLAM With Two-Stage Global-Local Graph Optimization," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1150–1157, Apr. 2022.
- [21] Y. Chang, K. Ebadi, C. E. Denniston, M. F. Ginting, A. Rosinol, A. Reinke, M. Palieri, J. Shi, A. Chatterjee, B. Morrell, A.-a. Aghamohammadi, and L. Carlone, "LAMP 2.0: A Robust Multi-Robot SLAM System for Operation in Challenging Large-Scale Underground Environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9175–9182, Oct. 2022.
- [22] S. Zhong, H. Chen, Y. Qi, D. Feng, Z. Chen, J. Wu, W. Wen, and M. Liu, "CoLRIO: LiDAR-Ranging-Inertial Centralized State Estimation for Robotic Swarms," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, May 2024, pp. 3920–3926.
- [23] P. Schmuck, T. Ziegler, M. Karrer, J. Perraudin, and M. Chli, "COVINS: Visual-Inertial SLAM for Centralized Collaboration," in *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, Oct. 2021, pp. 171–176.
- [24] Y. Tian, Y. Chang, F. Herrera Arias, C. Nieto-Granda, J. P. How, and L. Carlone, "Kimera-Multi: Robust, Distributed, Dense Metric-Semantic SLAM for Multi-Robot Systems," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2022–2038, Aug. 2022.
- [25] J. Bird, J. Blumenkamp, and A. Prorok, "DVM-SLAM: Decentralized Visual Monocular Simultaneous Localization and Mapping for Multi-Agent Systems," Mar. 2025, arXiv:2503.04126 [cs].
- [26] H. Xu, Y. Zhang, B. Zhou, L. Wang, X. Yao, G. Meng, and S. Shen, "Omni-Swarm: A Decentralized Omnidirectional Visual-Inertial-UWB State Estimation System for Aerial Swarms," *IEEE Transactions on Robotics*, vol. 38, no. 6, pp. 3374–3394, 2022.
- [27] J. He, Y. Zhou, L. Huang, Y. Kong, and H. Cheng, "Ground and Aerial Collaborative Mapping in Urban Environments," *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 95–102, Jan. 2021.
- [28] P.-Y. Lajoie and G. Beltrame, "Swarm-SLAM: Sparse Decentralized Collaborative Simultaneous Localization and Mapping Framework for Multi-Robot Systems," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 475–482, Jan. 2024.
- [29] F. Zhu, Y. Ren, L. Yin, F. Kong, Q. Liu, R. Xue, W. Liu, Y. Cai, G. Lu, H. Li, and F. Zhang, "Swarm-LIO2: Decentralized Efficient LiDAR-Inertial Odometry for Aerial Swarm Systems," *IEEE Transactions on Robotics*, vol. 41, pp. 960–981, 2025.
- [30] V. Pritzl, M. Vrba, P. Štěpán, and M. Saska, "Fusion of Visual-Inertial Odometry with LiDAR Relative Localization for Cooperative Guidance of a Micro-Scale Aerial Vehicle," 2024, arXiv:2306.17544 [cs].
- [31] I. Spasojevic, X. Liu, A. Ribeiro, G. J. Pappas, and V. Kumar, "Active Collaborative Localization in Heterogeneous Robot Teams," in *Proceedings of Robotics: Science and Systems*, July 2023.
- [32] F. Dellaert and M. Kaess, "Factor Graphs for Robot Perception," *Foundations and Trends in Robotics*, vol. 6, no. 1-2, pp. 1–139, 2017.
- [33] T. D. Barfoot and P. T. Furgale, "Associating Uncertainty With Three-Dimensional Poses for Use in Estimation Problems," *IEEE Transactions on Robotics*, vol. 30, no. 3, pp. 679–693, Jun. 2014.
- [34] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. J. Leonard, and F. Dellaert, "iSAM2: Incremental smoothing and mapping using the Bayes tree," *The International Journal of Robotics Research*, vol. 31, no. 2, pp. 216–235, Feb. 2012.
- [35] I. Olkin and F. Pukelsheim, "The distance between two random vectors with given dispersion matrices," *Linear Algebra and its Applications*, vol. 48, pp. 257–263, Dec. 1982.
- [36] D. C. Dowson and B. V. Landau, "The Fréchet distance between multivariate normal distributions," *Journal of Multivariate Analysis*, vol. 12, no. 3, pp. 450–455, Sep. 1982.
- [37] R. Bhatia, T. Jain, and Y. Lim, "On the Bures–Wasserstein distance between positive definite matrices," *Expositiones Mathematicae*, vol. 37, no. 2, pp. 165–191, Jun. 2019.
- [38] J. Solà, J. Deray, and D. Atchuthan, "A micro Lie theory for state estimation in robotics," arXiv:1812.01537 [cs], 2021.
- [39] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 376–380, Apr. 1991.
- [40] T. Shan, B. Englot, D. Meyers, W. Wang, C. Ratti, and D. Rus, "LIO-SAM: Tightly-coupled Lidar Inertial Odometry via Smoothing and Mapping," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2020, pp. 5135–5142.
- [41] P. Geneva, K. Eickenhoff, W. Lee, Y. Yang, and G. Huang, "OpenVINS: A Research Platform for Visual-Inertial Estimation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, May 2020, pp. 4666–4672.
- [42] M. Vrba, V. Walter, V. Pritzl, M. Pliska, T. Báča, V. Spurný, D. Heřt, and M. Saska, "On Onboard LiDAR-Based Flying Object Detection," *IEEE Transactions on Robotics*, vol. 41, pp. 593–611, 2025.
- [43] Z. Zhang and D. Scaramuzza, "A Tutorial on Quantitative Trajectory Evaluation for Visual(-Inertial) Odometry," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2018.
- [44] D. Hert, T. Baca, P. Petracek, V. Kratky, R. Penicka, V. Spurný, M. Petrlik, M. Vrba, D. Zaitlik, P. Stoudek, V. Walter, P. Stepan, J. Horyna, V. Pritzl, M. Sramek, A. Ahmad, G. Silano, D. B. Licea, P. Stibinger, T. Nascimento, and M. Saska, "MRS Drone: A Modular Platform for Real-World Deployment of Aerial Multi-Robot Systems," *Journal of Intelligent & Robotic Systems*, vol. 108, no. 4, p. 64, Jul. 2023.