AQUA-SLAM: Tightly Coupled Underwater Acoustic-Visual-Inertial SLAM With Sensor Calibration

Shida Xu, Kaicheng Zhang, and Sen Wang

***Abstract*—Underwater environments pose significant challenges for visual simultaneous localization and mapping (SLAM) systems due to limited visibility, inadequate illumination, and sporadic loss of structural features in images. Addressing these challenges, this article introduces a novel, tightly coupled acoustic-visual-inertial SLAM approach, termed AQUA-SLAM, to fuse a Doppler velocity log (DVL), a stereo camera, and an inertial measurement unit (IMU) within a graph optimization framework. Moreover, we propose an efficient sensor calibration technique, encompassing the multisensor extrinsic calibration (among the DVL, camera, and IMU) and the DVL transducer misalignment calibration, with a fast linear approximation procedure for real-time online execution. The proposed methods are extensively evaluated in a tank environment with ground truth, and validated for offshore applications in the North Sea. The results demonstrate that our method surpasses current state-of-the-art underwater and visual-inertial SLAM systems in terms of localization accuracy and robustness. The proposed system will be made open-source for the community.**

***Index Terms*—Doppler velocity log (DVL), DVL calibration, extrinsic calibration, simultaneous localization and mapping (SLAM), underwater localization.**

I. Introduction

AUTONOMOUS underwater vehicles (AUVs) are critical tools in offshore applications and ocean science, offering the capability to operate autonomously in challenging and often hazardous underwater environments. These vehicles are indispensable for tasks, such as seabed mapping, pipeline and cable inspections, biological and environmental monitoring, and the maintenance of underwater infrastructure. A key application area is the detailed visual inspection of the subsea structures, including offshore wind turbine foundations, where precise localization and mapping are paramount for effective operation. Considering cameras are widely equipped on underwater robots, visual simultaneous localization and mapping (SLAM) techniques emerge as the natural solutions.

Received 15 November 2024; accepted 25 February 2025. Date of publication 24 March 2025; date of current version 21 April 2025. This article was recommended for publication by Associate Editor M. Ghaffari and Editor J. Civera upon evaluation of the reviewers' comments. *(Corresponding author: Sen Wang.)*

The authors are with the Department of Electrical and Electronic Engineering and I-X, Imperial College London, SW7 2AZ London, U.K. (e-mail: s.xu23@imperial.ac.uk; k.zhang23@imperial.ac.uk; sen.wang@imperial.ac.uk).

Digital Object Identifier 10.1109/TRO.2025.3554396

Yet, underwater environments pose substantial unique challenges to visual SLAM techniques. The rapid attenuation of light energy in water severely limits the visibility of optical camera sensors, especially in murky water conditions. Moreover, underwater vision often suffers from poor lighting and blizzards of "marine snow" caused by small particles of organic matter in the water, severely reducing image quality with increased motion blur and dynamic image regions. In addition, ocean current disturbances to the robots can frequently push them away, which causes underwater structures to be occasionally outside the camera's field of view leading to intermittent loss of visual tracking. Therefore, although visual SLAM techniques have recently made tremendous progress in the terrestrial settings [1], [2], [3], their performance and robustness are inevitably compromised in underwater due to the complex and dynamic nature of the aquatic environments.

Fusing visual SLAM with an inertial measurement unit (IMU), known as visual-inertial SLAM (VI-SLAM) [4], [5], can alleviate some of the challenges arising from transient, noise-affected visual inputs from an optical camera, such as momentary motion blur. Therefore, the accuracy and robustness of underwater SLAM systems, particularly against short-term visual disruptions, can be substantially enhanced [6]. However, most of the challenges for underwater vision, such as the limited visibility and the "marine snow," are long-term effects that last at least from tens of seconds to a few minutes before being mitigated. VI-SLAM also encounters its own set of problems underwater. The low signal-to-noise ratio in accelerometers' measurements and the need for double integration to estimate translation greatly amplify inherent noises, making VI-SLAM systems particularly vulnerable to unreliability in visually challenging underwater conditions. Meanwhile, linear acceleration measurements are coupled with gravitational forces and IMU biases. Therefore, a proper initialization process is required to estimate gravity direction and biases. However, this process often depends on optimal visual conditions and fully excited motion, which are difficult in underwater environments and for underwater vehicles moving against water.

To address these problems, this article focuses on the sensor fusion of a stereo camera, an IMU and a Doppler velocity log (DVL), an underwater acoustic sensor measuring a linear velocity to the seabed [7] (see details on DVL in Section IV-A). Current state-of-the-art approaches on fusing underwater localization sensors in the form of cameras, inertial and DVL

sensors include: 1) DVL integration with cameras without IMU [8], 2) filtering-based methods [9], 3) DVL integration with LiDAR and visual SLAM for marine surface scenarios [10], and 4) loosely coupled integration of cameras, gyroscope, and DVL, e.g., [11]. To the best of the author's knowledge, a SLAM system with rigorous modeling of DVL and tightly-coupled fusion of DVL, camera, and IMU data within a graph optimization framework, specifically designed for underwater scenarios, has not been previously explored in the literature. Specifically, compared to our previous work [11], [12] which loosely fused cameras, DVL and gyroscope data, the new rigorous DVL modeling, tightly coupled formulation, and accelerometer integration significantly improve performance. Furthermore, these enhancements enable more accurate and efficient extrinsic calibration and facilitate the calibration of DVL transducer orientation, which was not possible in the previous approach.

Therefore, this article first models DVL's transducer and velocity measurements rigorously, and derives a tightly coupled acoustic-visual-inertial graph optimization for underwater SLAM. By making full use of the complementary strengths of these three sensing modalities, the approach aims to create a robust SLAM system capable of overcoming the challenges posed by underwater environments: the DVL provides reliable velocity measurements in underwater environments confining accelerometer-caused velocity drifts and enables dead-reckoning in visually degraded scenarios; the stereo camera offers high-accuracy localization capacity under good visual conditions and loop-closure for long-term drift correction; the IMU delivers reliable short-term motion estimation and renders absolute roll and pitch angles observable.

Second, for the multisensor extrinsic calibration of DVL, camera, and IMU, existing calibration mechanisms, such as hand–eye calibration [13], are suboptimal in underwater scenarios for two primary reasons: 1) they often necessitate a pre-established configuration with a calibration pattern being consistently visible—a challenging or even unviable requirement in underwater environments; 2) the calibration accuracy of these mechanisms is often compromised, attributable to the loosely correlated nature of the process and the inherent degradation of underwater image quality. Furthermore, the extrinsic parameters are susceptible to variation over extended periods, such as several weeks, due to the continuous exposure to the water drag forces and wave/current impacts. For users who have no dedicated expertise or underwater facilities to carry out sensor calibration, an automatic online calibration system only relying on features of surrounding scenes is more appealing.

Misalignment calibration of DVL's transducers is also of paramount importance to avoid the errors in velocity measurements. Recent works [14], [15] have attempted to calibrate a DVL sensor in the context of strapdown inertial navigation systems (SINS), concentrating on their extrinsic calibration. Notably, there is a lack of research exploring vision-facilitated calibration of DVL transducers.

To address these problems on the multisensor calibration and the DVL misalignment calibration, this article proposes a novel online sensor calibration algorithm that is designed to calibrate the extrinsic parameters between DVL, camera, and IMU, and to correct the alignment of DVL's transducers.

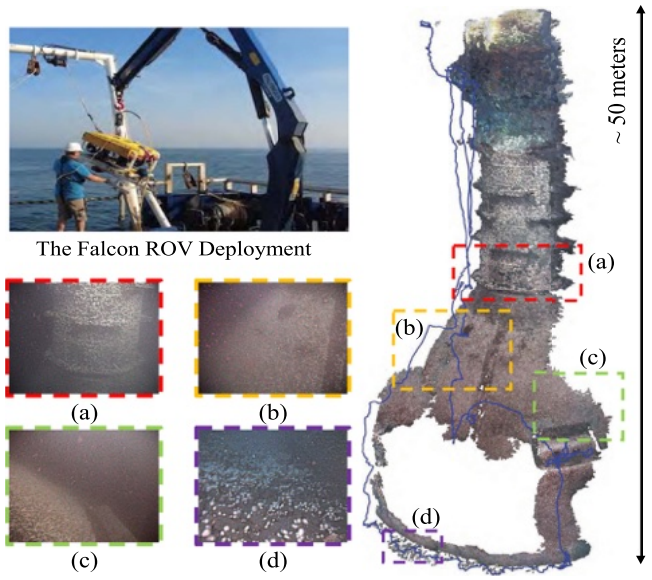


Fig. 1. Estimated trajectory (—) and dense 3-D reconstruction of an offshore structure using the proposed AQUA-SLAM algorithm. Images (a)–(d) show the challenging underwater conditions for a SLAM system using a camera.

A. Contributions

The main contributions of this work include the following.

1) A novel underwater acoustic-VI-SLAM algorithm, termed AQUA-SLAM, to tightly fuse DVL, camera, and IMU sensors within a graph optimization framework. To the best of the authors' knowledge, this is the first tightly coupled graph-based SLAM system designed for underwater environments that integrates DVL, IMU, and camera data in a graph-based optimization framework.
2) An efficient online sensor calibration algorithm for both DVL-camera-IMU extrinsic calibration and DVL transducer misalignment calibration, with a rapid linear approximation method designed to enable its real-time execution.
3) Extensive real-world experiments conducted in a water tank, along with offshore validation in the North Sea, demonstrating that our proposed method outperforms the state-of-the-art underwater and VI-SLAM systems, and is viable for offshore applications (see result in Fig. 1).

Our source code implementation will be released.[1]

The rest of the article is organized as follows. Section II reviews the literature on underwater visual SLAM and the sensor calibration, followed by a problem formulation in Section III. Sections IV and V describe the proposed AQUA-SLAM and sensor calibration algorithms, respectively. System implementation is detailed in Section VI. Experiment evaluation is presented in Section VII. Finally, Section VIII concludes this article.

II. Related Work

In this section, we review three topics related to our work: underwater visual SLAM, underwater vision-based extrinsic sensor calibration, and DVL calibration.

[1][Online]. Available: http://github.com/SenseRoboticsLab/AQUA-SLAM

A. Underwater Visual SLAM

1) Methods Using DVL: In the early years, most research on underwater visual SLAM modeled DVL measurements as odometry, rather than directly incorporating the sensor's raw data. Eustice et al. [16] introduced a sensor fusion framework integrating navigation data with 5 Degrees-of-Freedom (DoF) relative pose measurements for vehicle motion in underwater environments, utilizing an augmented state Kalman filter. Ozog and Eustice [17] proposed a SLAM method employing a sparse point cloud derived from a DVL, based on a piecewise-planar model. The underwater visual SLAM system proposed in [18] was based on a pose graph framework with DVL modeled as odometry constraints for hull inspection. This approach was extended to employ piecewise-planar panels for 3-D reconstruction of curved ship hull surfaces [19]. Fiducial markers were also incorporated into a visual SLAM framework alongside DVL, IMU, and depth sensor in [20].

Visual SLAM system integrating with DVL has gained more attention recently. A DVL, a stereo camera, and a gyroscope were fused in a loosely coupled fashion in [11] and [12], enabling reasonable pose estimation in challenging underwater environments. However, its DVL was still used in a loosely coupled manner. Meanwhile, it did not incorporate an accelerometer and assumed zero bias for the gyroscope, resulting in unbounded roll and pitch estimates. Thoms et al. [10] tightly integrated DVL into a LiDAR-VI-SLAM system for autonomous surface vehicles (ASV). However, their method is designed for USVs operating on 2-D water surface and may not be directly applicable to underwater environments. A tightly coupled multisensor fusion framework for camera, IMU, DVL, and a depth sensor, based on the multistate constraint Kalman filter, was proposed in [9]. It is a filtering based method, different from our graph optimization-based method in this work. Huang et al. [8] proposed a tightly coupled visual-DVL fusion method, which integrates the velocity measurements from a DVL into a visual odometry for improved localization accuracy. However, the lack of IMU integration might compromise the robustness of orientation estimation in challenging visual conditions. Importantly, none of these works addressed the sensor calibration problem, simultaneously.

In contrast, the approach proposed in this article focuses on a tightly coupled integration of DVL, IMU, and camera data, formulating the problem as a graph optimization to achieve accurate and robust localization and mapping in challenging underwater scenarios. To the best of the author's knowledge, this is the first tightly coupled graph-based SLAM system that integrates DVL, IMU, and camera data in a unified framework for underwater environments. Furthermore we model the sensor calibration problem simultaneously as an online calibration module, which is essential for multisensor fusion.

2) Methods Without Using DVL: Recent works attempted to enhance the accuracy and robustness of underwater visual SLAM by integrating other sensing modalities. Rahman et al. [21] proposed the sonar visual inertial (SVIN) SLAM system, which integrated a downward-facing mechanical scanning sonar, a stereo camera, and an IMU under a tightly coupled framework based on OKVIS [22]. They extended this work to SVIN2 [6], [23], which additionally included a water-pressure depth sensor and a loop closure module. More recently, Joshi et al. [24] proposed a state estimation switching strategy that can detect failures in SVIN2 and seamlessly transition to a model-based approach using the robot's kinematics and proprioceptive sensors to maintain a pose estimation. However, none of these methods investigate the incorporation of DVL data.

B. Underwater Vision-Based Extrinsic Calibration

As previously discussed, extrinsic sensor calibration is vital for multisensor fusion, especially for underwater scenarios. Camera-IMU extrinsic calibration for underwater environments was studied in [25] and [26] when sonar data was available. However, camera-DVL calibration is rarely explored, with [20] being one of the only few existing works. However, its reliance on a marinized panel of fiducial markers and presetup made it impractical or time-consuming in offshore underwater settings. Our proposed calibration algorithm, in contrast, simply utilizes the natural scene features for automatic online sensor calibration, being seamlessly integrated with our proposed SLAM algorithm.

C. DVL Calibration

Regarding DVL calibration, existing research predominantly focuses on SINS-DVL systems rather than vision-DVL systems. Xu and Guo [14] introduced an EKF-based method for calibrating DVL installation and scale factor using the Special Orthogonal Group. Li et al. [27] presented a DVL calibration algorithm employing particle swarm optimization, transforming the calibration issue into a Wahba problem. Luo et al. [28] proposed a SINS-DVL calibration system capable of calibrating various errors and misalignments, utilizing observability-based trajectory design for parameter observability. However, none of these addressed DVL transducer misalignment calibration, which can have considerable impacts on its velocity measurements and its fusion with other sensors.

III. Preliminaries

We use these following notations throughout this article. A scalar is lowercase italics a, a vector is lowercase bold Roman $\mathbf{a}$, a matrix is uppercase bold Roman $\mathbf{A}$, a coordinate frame is typewriter $\mathtt{A}$, and a set is calligraphic typeface $\mathcal{A}$.

A. Riemannian Geometry Representation

The group of 3-D rotation matrices is described by Special Orthogonal Group $\mathrm{SO}(3)$ as $\mathrm{SO}(3) \doteq \{\mathbf{R} \in \mathbb{R}^{3\times 3} \mid \mathbf{R}\mathbf{R}^T = \mathbf{I}, det(\mathbf{R}) = 1\}$. Its tangent space is denoted as $\mathfrak{so}(3)$, which can be represented as a 3×3 skew-symmetric matrix from a hat operator $(\cdot)^\wedge$ on a 3×1 vector, i.e., $\mathfrak{so}(3) \doteq \{\boldsymbol{\phi}^\wedge \in \mathbb{R}^{3\times 3} \big| \boldsymbol{\phi} \in \mathbb{R}^3\}$.

The group of 3-D rigid motion belongs to the Special Euclidean Group $\mathrm{SE}(3) \doteq \{(\mathbf{R}, \mathbf{t}) \big| \mathbf{R} \in \mathrm{SO}(3), \mathbf{t} \in \mathbb{R}^3\}$ whose corresponding transformation matrix is defined as $\mathbf{T}$. Following

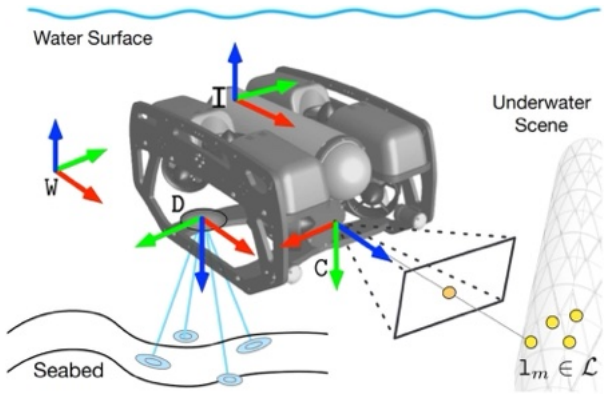


Fig. 2. Coordinate frames. The world frame W's z-axis is aligned with the gravity vector. The DVL frame D, the IMU frame I, and the camera frame C are rigidly fixed on the robot. The DVL sensor measures linear velocity with respect to the seabed. 3-D visual landmarks $\mathcal{L}$ are estimated from scenes.

the notations suggested in [29], a 3-D transformation from a coordinate frame B to a coordinate frame A is defined as

$$\mathbf{T}_{\mathtt{AB}} \doteq \begin{bmatrix} \mathbf{R}_{\mathtt{AB}} & {}_{\mathtt{A}}\mathbf{p}_{\mathtt{AB}} \\ \mathbf{0} & 1 \end{bmatrix} \in \mathrm{SE}(3)$$

where $\mathbf{R}_{\mathtt{AB}} \in \mathrm{SO}(3)$ describes its 3-D rotation matrix and ${}_{\mathtt{A}}\mathbf{p}_{\mathtt{AB}} \in \mathbb{R}^3$ is its 3-D translation expressed in frame A.

B. Definition of Coordinate Frames

Fig. 2 specifies the coordinate frames relevant to this work, including the static world frame W, the IMU frame I, the camera frame C, and the DVL frame D. The z-axis of the world frame is aligned with the gravity direction. A time-dependent moving coordinate frame is specified with a subscript, e.g., $\mathtt{C}_i$ means the camera coordinate frame C of keyframe i.

C. State Estimation

1) State Definition: The state of keyframe i is defined as

$$\mathbf{x}_i \doteq [\mathbf{R}_i, \mathbf{p}_i, \mathbf{v}_i, \mathbf{b}_i^g, \mathbf{b}_i^a] \in \mathrm{SO}(3) \times \mathbb{R}^{12}$$

where $(\mathbf{R}_i, \mathbf{p}_i) \in \mathrm{SE}(3)$ denotes the 3-D camera pose $\mathbf{T}_i$ in $\mathtt{C}_0$ which stands for the initial camera frame, i.e., $(\mathbf{R}_{\mathtt{C}_0\mathtt{C}_i}, {}_{\mathtt{C}_0}\mathbf{p}_{\mathtt{C}_0\mathtt{C}_i})$, while $\mathbf{v}_i \doteq {}_{\mathtt{D}_i}\mathbf{v} \in \mathbb{R}^3$ represents the linear velocity in D. $\mathbf{b}_i^g$ and $\mathbf{b}_i^a$ are the IMU gyroscope and accelerometer biases. Considering all keyframes $\mathcal{K}_n$ up to n, the set of historical keyframe states, landmarks and calibration parameters is defined as

$$\mathcal{X}_n \doteq \{\mathbf{x}_i, \mathcal{L}_i, \mathcal{E}, \mathbf{R}_{\mathtt{WI}_0}\}, \quad i \in \mathcal{K}_n$$

where $\mathcal{L}_i \doteq \{{}_{\mathtt{C}_0}\mathbf{l}_{\mathtt{C}_0\mathtt{l}_m}\}, m \in \mathcal{M}_i$ is the set of 3-D locations of landmarks visible in keyframe i, $\mathcal{E} \doteq \{\mathbf{T}_{\mathtt{ID}}, \mathbf{T}_{\mathtt{DC}}\}$ includes the fixed extrinsic parameters between the sensors (IMU, DVL, and camera), and $\mathbf{R}_{\mathtt{WI}_0}$ is the initial orientation of the IMU (accelerometer) in W standing for the gravity direction. Notably, the extrinsic parameter $\mathbf{T}_{\mathtt{IC}}$ can be derived by the transformation composition using $\mathcal{E}$

$$\mathbf{T}_{\mathtt{IC}} \doteq \mathbf{T}_{\mathtt{ID}}\mathbf{T}_{\mathtt{DC}}. \tag{1}$$

Therefore, we do not define it explicitly in the state.

2) Measurement Definitions: Measurements provided by a DVL, an IMU, and a camera between keyframes i and j are defined as follows.

a) DVL: The raw measurements from the DVL are obtained by decoding the acoustic signals emanated from its transducers. By utilizing the Doppler shift principle, these measurements facilitate the calculation of velocity. The collection of the DVL measurements is denoted as $\mathcal{D}_{i,j}$.

b) IMU: The IMU measurements $\mathcal{I}_{i,j}$ are a set of rotational velocity $\tilde{\boldsymbol{\omega}}$ and linear acceleration $\tilde{\mathbf{a}}$.

c) Camera: The camera yields a pair of stereo images for ith keyframe $\mathcal{C}_i$, from which landmarks are extracted.

To summarize, the set of all sensor measurements up to keyframe n is $\mathcal{Z}_n \doteq \{\mathcal{D}_{i,j}, \mathcal{I}_{i,j}, \mathcal{C}_i\}, i, j \in \mathcal{K}_n$. The detailed DVL, IMU, and camera measurement models will be discussed in Sections IV-A–IV-C, respectively.

3) Maximum a Posteriori (MAP) Estimation: Given the full set of measurements $\mathcal{Z}_n$, our goal is to estimate the optimal $\mathcal{X}_n$ by maximizing the posterior probability

$$\begin{aligned} \mathcal{X}_n^* &= \underset{\mathcal{X}_n}{\operatorname{argmax}}\ p(\mathcal{X}_n|\mathcal{Z}_n) \\ &= \underset{\mathcal{X}_n}{\operatorname{argmax}}\ p(\mathcal{X}_0) \prod_{i,j\in\mathcal{K}_n} p(\mathcal{D}_{i,j}|\mathcal{X}_i, \mathcal{X}_j) p(\mathcal{I}_{i,j}|\mathcal{X}_i, \mathcal{X}_j) \\ &\quad \prod_{i\in\mathcal{K}_n} p(\mathcal{C}_i|\mathcal{X}_i). \end{aligned}$$

Since the measurements are usually assumed with Gaussian noises, this can be reformulated as a nonlinear least-squares problem

$$\begin{aligned} \mathcal{X}_n^* = \underset{\mathcal{X}_n}{\operatorname{argmin}}\ & \|\mathbf{r}_{x_0}\|_{\Sigma_0}^2 + \sum_{i,j\in\mathcal{K}_n} \Big(\|\mathbf{r}_D\left(h_D\left(\mathcal{X}_i, \mathcal{X}_j\right), \mathcal{D}_{i,j}\right)\|_{\mathbf{\Sigma}_D}^2 \\ & + \|\mathbf{r}_I(h_I(\mathcal{X}_i, \mathcal{X}_j), \mathcal{I}_{i.j})\|_{\mathbf{\Sigma}_I}^2 \Big) + \sum_{i\in\mathcal{K}_n} \|\mathbf{r}_C(h_C(\mathcal{X}_i), \mathcal{C}_i)\|_{\mathbf{\Sigma}_C}^2 \end{aligned} \tag{2}$$

The notation $\|\cdot\|_{\Sigma}^2$ represents the Mahalanobis distance, with Σ denoting a covariance. The prior residual $\mathbf{r}_{x_0}$ computes the discrepancy between the initial state and the prior information. The three residuals $\mathbf{r}_D$, $\mathbf{r}_I$, and $\mathbf{r}_C$ represent the DVL, IMU, and camera residuals, respectively. Each of these residuals quantifies the errors between the sensor measurements and the predicted measurements from the corresponding measurement models $h_D(\cdot)$, $h_I(\cdot)$, and $h_C(\cdot)$, which are to be detailed now.

IV. Acoustic-Visual-Inertial Underwater SLAM

In this section, the sensor measurement models and the residual terms in (2) are derived for the proposed acoustic-visual-inertial underwater SLAM system, given the DVL, IMU, and camera measurements between image keyframes i and j as shown in Fig. 3. The extrinsic parameters $\mathcal{E}$, which are estimated using the method proposed in Section V, are assumed known in this section.

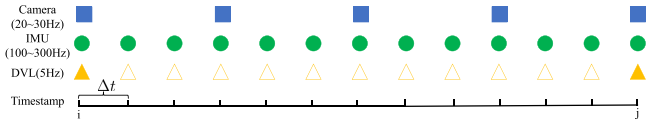


Fig. 3. Sensor measurements from camera, IMU, and DVL. Dash yellow triangles represent constant velocity taken from the last DVL measurement.

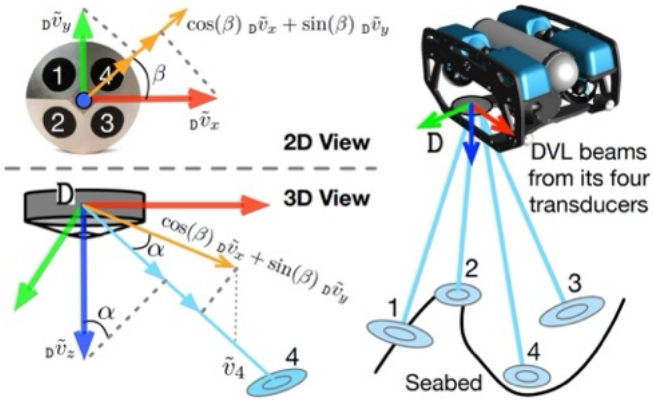


Fig. 4. DVL transducer measurements with 2-D and 3-D views. The DVL has four transducers facing different directions. Transducer 4 is shown as an example.

A. DVL Measurement Model and its Residuals

A DVL sensor has the capability to measure linear velocity with respect to the seabed. Typically, it encompasses four transducers, as the example in Fig. 4. Each transducer is oriented towards the seabed and continuously emits the acoustic signals. These signals, upon reflection from the seabed, are sampled to measure Doppler shifts and then the 1-D velocity along the direction of each transducer. By aggregating each individual transducer's velocity, the overall velocity in the 3-D space can be obtained [30]. It is a common practice to mount the DVL at the base of an underwater vehicle facing the seabed.

1) Velocity Measurement From an Individual Transducer: Each transducer of the DVL operates as both a transmitter and a receiver, able to emit and receive the acoustic signals. It is hypothesized that its transmitted signal possesses a frequency f_t, while the received signal has a frequency f_r. Therefore, the frequency shift is $\Delta f = f_r - f_t$. According to the Doppler shift, the velocity along the radial direction of the nth transducer is $v_n = 2c_s\Delta f/f_t$, where c_s is the sound speed in water [30]. Therefore, the 1-D velocity measurement $\tilde{v}_n$ of an individual transducer along its radial direction is assumed below with Gaussian noise

$$\tilde{v}_n = v_n + \eta^D, \quad n \in \{1,2,3,4\}$$

where $\eta^D \sim \mathcal{N}(0, \sigma^D)$

2) DVL Velocity Measurement and its Model: The DVL velocity measurement, associated with keyframe i in the DVL frame $\mathtt{D}_i$, is denoted as ${}_{\mathtt{D}_i}\tilde{\mathbf{v}} \in \mathbb{R}^3$. It is correlated with the individual transducer's velocity $\tilde{v}_n$ through

$$\tilde{v}_n = \mathbf{e}_n \cdot {}_{\mathtt{D}_i}\tilde{\mathbf{v}}, \quad n \in \{1,2,3,4\}$$

where $\mathbf{e}_n \in \mathbb{R}^{1\times 3}$ signifies the orientation-dependent factor of the transducer n and projects ${}_{\mathtt{D}_i}\tilde{\mathbf{v}}$ to the radial direction of transducer n, as shown in Fig. 4. Assume that the radial direction of each transducer is rotated by an angle α from horizontal plane and a yaw β with respect to the DVL frame $\mathtt{D}_i$. Then, the z velocity ${}_{\mathtt{D}}\tilde{v}_z$ of ${}_{\mathtt{D}_i}\tilde{\mathbf{v}}$ can be projected onto the transducer direction by $\sin(\alpha)$. For the x and y velocities, ${}_{\mathtt{D}}\tilde{v}_x$ and ${}_{\mathtt{D}}\tilde{v}_y$ of ${}_{\mathtt{D}_i}\tilde{\mathbf{v}}$, it is necessary to establish their components along the transducer's x–y projection (see 2-D view in Fig. 4). This aggregated x–y plane velocity is further projected to the transducer direction. Taking transducer 4 as an example, we can determine

$$\tilde{v}_4 = \mathbf{e}_4 \cdot {}_{\mathtt{D}_i}\tilde{\mathbf{v}} = [\cos(\beta)\cos(\alpha) \;\; \sin(\beta)\cos(\alpha) \;\; \sin(\alpha)] \; {}_{\mathtt{D}_i}\tilde{\mathbf{v}}.$$

Since the transducer often shares the same amount of rotation by design, we can similarly obtain the projection vectors $\mathbf{e}_n$ for other transducers as

$$\begin{aligned}
\mathbf{e}_1 &= [-\cos(\beta)\cos(\alpha) \quad \sin(\beta)\cos(\alpha) \;\; \sin(\alpha)] \\
\mathbf{e}_2 &= [-\cos(\beta)\cos(\alpha) \quad -\sin(\beta)\cos(\alpha) \;\; \sin(\alpha)] \\
\mathbf{e}_3 &= [\cos(\beta)\cos(\alpha) \quad -\sin(\beta)\cos(\alpha) \;\; \sin(\alpha)] \\
\mathbf{e}_4 &= [\cos(\beta)\cos(\alpha) \quad \sin(\beta)\cos(\alpha) \;\; \sin(\alpha)].
\end{aligned} \tag{3}$$

After vectorizing the velocity measurements from all the transducers, we can be obtain

$$[\tilde{v}_1 \; \tilde{v}_2 \; \tilde{v}_3 \; \tilde{v}_4]^T = \mathbf{E} \cdot {}_{\mathtt{D}_i}\tilde{\mathbf{v}} \tag{4}$$

where $\mathbf{E} \doteq [\mathbf{e}_1;\, \mathbf{e}_2;\, \mathbf{e}_3;\, \mathbf{e}_4] \in \mathbb{R}^{4\times 3}$. Therefore, ${}_{\mathtt{D}_i}\tilde{\mathbf{v}}$ can be derived from (4) which contains four equations and three unknowns. It presents an overdetermined problem with linear equations. Therefore, the closed-form solution of ${}_{\mathtt{D}_i}\tilde{\mathbf{v}}$ from the individual transducer velocity measurements is

$${}_{\mathtt{D}_i}\tilde{\mathbf{v}} = (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T[\tilde{v}_1 \; \tilde{v}_2 \; \tilde{v}_3 \; \tilde{v}_4]^T. \tag{5}$$

Since we assume each individual $\tilde{v}_n$ is corrupt by Gaussian noise, after a linear transformation ${}_{\mathtt{D}_i}\tilde{\mathbf{v}}$ still follow the Gaussian distribution:

$${}_{\mathtt{D}_i}\mathbf{v} = {}_{\mathtt{D}_i}\tilde{\mathbf{v}} - \boldsymbol{\eta}^D$$

where $\boldsymbol{\eta}^D \sim \mathcal{N}(0, \boldsymbol{\Sigma}^D)$.

The DVL velocity measurement model for keyframes i and j can be expressed as

$$h_{D_v}(\mathcal{X}_i) \doteq {}_{\mathtt{D}_i}\tilde{\mathbf{v}} - \boldsymbol{\eta}^D, \quad h_{D_v}(\mathcal{X}_j) \doteq {}_{\mathtt{D}_j}\tilde{\mathbf{v}} - \boldsymbol{\eta}^D. \tag{6}$$

3) DVL Translation Measurement Model: The DVL measurements, in conjunction with the orientation estimates using the gyroscope measurements, can constrain the translation estimate between keyframes i and j.

It is posited that the velocity remains consistent between two consecutive DVL measurements. Upon receiving a gyroscope measurement, the DVL velocity is integrated into the translation. When determining the translation from keyframe i to keyframe j, denoted as ${}_{\mathtt{C}_0}\mathbf{p}_{\mathtt{D}_i\mathtt{D}_j}$, it is essential to aggregate all the DVL translations occurring between these two keyframes. This leads to the subsequent relationship

$${}_{\mathtt{C}_0}\mathbf{p}_{\mathtt{C}_0\mathtt{D}_j} = {}_{\mathtt{C}_0}\mathbf{p}_{\mathtt{C}_0\mathtt{D}_i} + {}_{\mathtt{C}_0}\mathbf{p}_{\mathtt{D}_i\mathtt{D}_j} \tag{7}$$

where the terms ${}_{\mathtt{C}_0}\mathbf{p}_{\mathtt{C}_0\mathtt{D}_i}$ and ${}_{\mathtt{C}_0}\mathbf{p}_{\mathtt{C}_0\mathtt{D}_j}$ are defined as the DVL positions at keyframes i and j, respectively. They can be represented as ${}_{\mathtt{C}_0}\mathbf{p}_{\mathtt{C}_0\mathtt{D}_i} \doteq \mathbf{p}_i - \mathbf{R}_i\mathbf{R}^T_{\mathtt{DC}}\mathbf{p}_{\mathtt{DC}}$ and ${}_{\mathtt{C}_0}\mathbf{p}_{\mathtt{C}_0\mathtt{D}_j} \doteq \mathbf{p}_j - \mathbf{R}_j\mathbf{R}^T_{\mathtt{DC}}\mathbf{p}_{\mathtt{DC}}$, respectively. In addition, the term ${}_{\mathtt{C}_0}\mathbf{p}_{\mathtt{D}_i\mathtt{D}_j}$ denotes the integration of DVL translations. This is formally expressed as ${}_{\mathtt{C}_0}\mathbf{p}_{\mathtt{D}_i\mathtt{D}_j} \doteq \sum_{k=i}^{j-1}\mathbf{R}_i\mathbf{R}^T_{\mathtt{IC}}\Delta\mathbf{R}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}}\,({}_{\mathtt{D}_i}\tilde{\mathbf{v}} - \boldsymbol{\eta}^D)\Delta t$ where Δt is the time interval between the gyroscope (IMU) measurements, and the variable $\Delta\mathbf{R}_{\mathtt{I}_i\mathtt{I}_k}$ represents the gyroscope relative incremental defined in (14).

Notably, the term ${}_{\mathtt{C}_0}\mathbf{p}_{\mathtt{D}_i\mathtt{D}_j}$ is interrelated with the state $\mathbf{R}_i$, which can cause repeated DVL translation integration during optimization iterations. To address this problem, we propose a DVL preintegration term decoupling the state from the DVL translation integration. Hence, we reformulate (7) as

$$\underbrace{\mathbf{R}_{\mathtt{ID}}\left({}_{\mathtt{D}}\mathbf{p}_{\mathtt{DC}} - \mathbf{R}_{\mathtt{DC}}\mathbf{R}_i^T\mathbf{R}_j\mathbf{R}^T_{\mathtt{DC}}\mathbf{p}_{\mathtt{DC}} + \mathbf{R}_{\mathtt{DC}}(\mathbf{R}_i^T\mathbf{p}_j - \mathbf{R}_i^T\mathbf{p}_i)\right)}_{h_{D_t}(\mathcal{X}_i,\mathcal{X}_j)}$$
$$= \Delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j} - \delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j} \tag{8}$$

where $h_{D_t}(\mathcal{X}_i,\mathcal{X}_j)$ stands for the DVL translation measurement model. $\Delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j} \doteq \sum_{k=i}^{j-1}\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}}\,{}_{\mathtt{D}_i}\tilde{\mathbf{v}}\Delta t$ represents the DVL translation preintegration and $\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}$ stands for the gyroscope preintegration from keyframe i to k defined at (14). $\delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j} \doteq \sum_{k=i}^{j-1} -\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}(\mathbf{R}_{\mathtt{ID}}\,{}_{\mathtt{D}_i}\tilde{\mathbf{v}})^\wedge\delta\hat{\boldsymbol{\phi}}_{\mathtt{I}_i\mathtt{I}_k}\Delta t + \Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}}\,\boldsymbol{\eta}^D\Delta t$ stands for the Gaussian noise. Please see Appendices A1 and A2 for detailed derivation of this DVL preintegration.

4) Residual Derivation: The explicit formulation of the DVL residual in (2) can be derived as

$$\mathbf{r}_D \doteq \begin{bmatrix} \mathbf{r}_v(h_{D_v}(\mathcal{X}_i),\mathcal{D}_i) \\ \mathbf{r}_v(h_{D_v}(\mathcal{X}_j),\mathcal{D}_j) \\ \mathbf{r}_t(h_{D_t}(\mathcal{X}_i,\mathcal{X}_j),\mathcal{D}_{i,j}) \end{bmatrix} \tag{9}$$

where $\mathbf{r}_v(\cdot)$ constrains the velocities at keyframe i and j, while $\mathbf{r}_t(\cdot)$ constrains the relative translation between the keyframes. In accordance with (5) and (6), the DVL velocity residual can be defined as follows:

$$\mathbf{r}_v(h_{D_v}(\mathcal{X}_i),\mathcal{D}_i) \doteq {}_{\mathtt{D}_i}\tilde{\mathbf{v}} - \mathbf{v}_i$$
$$\mathbf{r}_v(h_{D_v}(\mathcal{X}_j),\mathcal{D}_j) \doteq {}_{\mathtt{D}_j}\tilde{\mathbf{v}} - \mathbf{v}_j.$$

According to (8), the DVL translation residual can be obtained as

$$\mathbf{r}_t(h_{D_t}(\mathcal{X}_i,\mathcal{X}_j),\mathcal{D}_{i,j}) \doteq \Delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j} - h_{D_t}(\mathcal{X}_i,\mathcal{X}_j). \tag{10}$$

B. IMU Measurement Model and its Residuals

1) Gyroscope Model: The gyroscope measures the instantaneous angular velocity of $\mathtt{I}$ relative to $\mathtt{W}$ expressed in the IMU coordinate frame $\mathtt{I}$. We assume that the measurement ${}_{\mathtt{I}}\tilde{\boldsymbol{\omega}}_{\mathtt{WI}} \in \mathbb{R}^3$ is affected by a zero-mean Gaussian white noise $\boldsymbol{\eta}^g$ and random walk noise $\mathbf{b}^g$

$${}_{\mathtt{I}}\tilde{\boldsymbol{\omega}}_{\mathtt{WI}} = {}_{\mathtt{I}}\boldsymbol{\omega}_{\mathtt{WI}} + \mathbf{b}^g + \boldsymbol{\eta}^g$$

where $\boldsymbol{\eta}^g \sim \mathcal{N}(0,\boldsymbol{\Sigma}^g)$

Given $\mathcal{X}_i$ at keyframe i and the extrinsic parameters $\mathcal{E}$, along with the series of the gyroscope measurements between keyframes i and j, the gyroscope orientation at keyframe j, $\mathbf{R}_{\mathtt{WI}_j}$, is computed as follows:

$$\mathbf{R}_{\mathtt{WI}_j} = \underbrace{\mathbf{R}_{\mathtt{WI}_i}\prod_{k=i}^{j-1}\mathrm{Exp}\left(({}_{\mathtt{I}_k}\tilde{\boldsymbol{\omega}}_{\mathtt{WI}_k} - \mathbf{b}_i^g - \boldsymbol{\eta}^g)\Delta t\right)}_{\mathbf{R}_{\mathtt{WI}_k}} \tag{11}$$

where $\mathbf{R}_{\mathtt{WI}_j} \doteq \mathbf{R}_{\mathtt{WI}_0}\mathbf{R}_{\mathtt{IC}}\mathbf{R}_j\mathbf{R}^T_{\mathtt{IC}}$ and $\mathbf{R}_{\mathtt{WI}_i} \doteq \mathbf{R}_{\mathtt{WI}_0}\mathbf{R}_{\mathtt{IC}}\mathbf{R}_i\mathbf{R}^T_{\mathtt{IC}}$ are the IMU orientations at keyframe j and i, respectively, $\mathrm{Exp}(\cdot):\mathbb{R}^3 \to \mathrm{SO}(3)$ stands for the exponential map from a vectorized $\mathfrak{so}(3)$ to $\mathrm{SO}(3)$, and $\mathbf{b}_i^g$ stands for the gyroscope bias at i.

2) Accelerometer Model: The accelerometer measures linear acceleration with respect to the IMU frame $\mathtt{I}$, where the measurement ${}_{\mathtt{I}}\tilde{\mathbf{a}} \in \mathbb{R}^3$ is consistently influenced by the gravity. We assume that it is also affected by zero-mean Gaussian white noise $\boldsymbol{\eta}^a$ and a random walk noise $\mathbf{b}^a$

$${}_{\mathtt{I}}\tilde{\mathbf{a}} = \mathbf{R}^T_{\mathtt{WI}}({}_{\mathtt{W}}\mathbf{a} - {}_{\mathtt{W}}\mathbf{g}) + \mathbf{b}^a + \boldsymbol{\eta}^a$$

where ${}_{\mathtt{W}}\mathbf{g} \doteq [0,0,-g]^T$ is the gravity vector in the world frame, and g is its magnitude.

Similarly, given the state $\mathcal{X}_i$ and the extrinsic parameters $\mathcal{E}$ and a set of accelerometer measurements between keyframe i and j, the linear velocity in the world frame $\mathtt{W}$ at j is

$${}_{\mathtt{W}_j}\mathbf{v} = \underbrace{{}_{\mathtt{W}_i}\mathbf{v} + \sum_{k=i}^{j-1}({}_{\mathtt{W}}\mathbf{g}\Delta t + \mathbf{R}_{\mathtt{WI}_k}({}_{\mathtt{I}_k}\tilde{\mathbf{a}} - \mathbf{b}_i^a - \boldsymbol{\eta}^a)\Delta t)}_{{}_{\mathtt{W}_k}\mathbf{v}} \tag{12}$$

where ${}_{\mathtt{W}_j}\mathbf{v} \doteq \mathbf{R}_{\mathtt{WI}_0}\mathbf{R}_{\mathtt{IC}}\mathbf{R}_j\mathbf{R}^T_{\mathtt{DC}}\mathbf{v}_j$ and ${}_{\mathtt{W}_i}\mathbf{v} \doteq \mathbf{R}_{\mathtt{WI}_0}\mathbf{R}_{\mathtt{IC}}\mathbf{R}_i\mathbf{R}^T_{\mathtt{DC}}\mathbf{v}_i$ are the linear velocities in the frame $\mathtt{W}$ at keyframes j and i respectively, and $\mathbf{b}_i^a$ stands for the accelerometer bias at i. In addition, the position at keyframe j is

$${}_{\mathtt{W}}\mathbf{p}_{\mathtt{WI}_j} = {}_{\mathtt{W}}\mathbf{p}_{\mathtt{WI}_i} + \sum_{k=i}^{j-1}\left({}_{\mathtt{W}_k}\mathbf{v}\Delta t + \frac{1}{2}{}_{\mathtt{W}}\mathbf{g}\Delta t^2 + \frac{1}{2}\mathbf{R}_{\mathtt{WI}_k}({}_{\mathtt{I}_k}\tilde{\mathbf{a}} - \mathbf{b}_i^a - \boldsymbol{\eta}^a)\Delta t^2\right) \tag{13}$$

where ${}_{\mathtt{W}}\mathbf{p}_{\mathtt{WI}_j} \doteq \mathbf{R}_{\mathtt{WI}_0}\mathbf{R}_{\mathtt{IC}}\mathbf{R}_{j\mathtt{C}}\mathbf{p}_{\mathtt{CI}} + \mathbf{R}_{\mathtt{WI}_0}\mathbf{R}_{\mathtt{IC}}\mathbf{p}_j$ and ${}_{\mathtt{W}}\mathbf{p}_{\mathtt{WI}_i} \doteq \mathbf{R}_{\mathtt{WI}_0}\mathbf{R}_{\mathtt{IC}}\mathbf{R}_{i\mathtt{C}}\mathbf{p}_{\mathtt{CI}} + \mathbf{R}_{\mathtt{WI}_0}\mathbf{R}_{\mathtt{IC}}\mathbf{p}_i$ are the positions of the IMU in $\mathtt{W}$ at keyframes j and i, respectively.

As thoroughly investigated in [31], defining the IMU residual in accordance with (11)–(13) would necessitate multiple times of integration during the optimization iterations. This can lead to considerable time expenditure. Therefore, a further reformulation of (11)–(13) is applied by following the IMU preintegration proposed in [31]. First, (11) can be reformulated as

$$\underbrace{\mathbf{R}_{\mathtt{IC}}\mathbf{R}_i^T\mathbf{R}_j\mathbf{R}^T_{\mathtt{IC}}}_{h_{I_r}(\mathcal{X}_i,\mathcal{X}_j)} = \underbrace{\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_j}\mathrm{Exp}(-\delta\hat{\boldsymbol{\phi}}_{\mathtt{I}_i\mathtt{I}_j})}_{\Delta\mathbf{R}_{\mathtt{I}_i\mathtt{I}_j}} \tag{14}$$

where $h_{I_r}(\mathcal{X}_i,\mathcal{X}_j)$ is the gyroscope measurement model and $\Delta\mathbf{R}_{\mathtt{I}_i\mathtt{I}_j}$ represents the incremental gyroscope relative motion. $\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_j} \doteq \prod_{k=i}^{j-1}\mathrm{Exp}(({}_{\mathtt{I}_k}\tilde{\boldsymbol{\omega}}_{\mathtt{WI}_k} - \mathbf{b}_i^g)\Delta t)$ stands for the gyroscope preintegration from keyframe i to j and $\mathrm{Exp}(-\delta\hat{\boldsymbol{\phi}}_{\mathtt{I}_i\mathtt{I}_j}) \doteq$

$\prod_{k=i}^{j-1} \mathrm{Exp}(-\Delta\hat{\mathbf{R}}_{\mathtt{I}_{k+1}\mathtt{I}_j}^T J_r({}_{\mathtt{I}_k}\tilde{\boldsymbol{\omega}}_{\mathtt{WI}_k} - \mathbf{b}_i^g)\boldsymbol{\eta}^g \Delta t)$ is the noise term and $J_r(\boldsymbol{\phi}) \doteq \mathbf{I} - \frac{1-cos(\|\boldsymbol{\phi}\|)}{\|\boldsymbol{\phi}\|^2}\boldsymbol{\phi}^\wedge + \frac{\|\boldsymbol{\phi}\|-sin(\|\boldsymbol{\phi}\|)}{\|\boldsymbol{\phi}^3\|}(\boldsymbol{\phi}^\wedge)^2$ is the right Jacobian of SO(3) [31]. Moreover, (12) can be reformulated as

$$\underbrace{\mathbf{R}_{\mathtt{IC}}\mathbf{R}_i^T\left(\mathbf{R}_j\mathbf{R}_{\mathtt{DC}}^T\mathbf{v}_j - \mathbf{R}_i\mathbf{R}_{\mathtt{DC}}^T\mathbf{v}_i - (\mathbf{R}_{\mathtt{WI}_0}\mathbf{R}_{\mathtt{IC}})^T{}_{\mathtt{W}}\mathbf{g}\Delta t_{ij}\right)}_{h_{I_v}(\mathcal{X}_i,\mathcal{X}_j)} = \underbrace{\Delta\hat{\mathbf{v}}_{\mathtt{I}_i\mathtt{I}_j} - \delta\hat{\mathbf{v}}_{\mathtt{I}_i\mathtt{I}_j}}_{\Delta\mathbf{v}_{\mathtt{I}_i\mathtt{I}_j}} \tag{15}$$

where $h_{I_v}(\mathcal{X}_i,\mathcal{X}_j)$ is the accelerometer velocity measurement model and $\Delta\mathbf{v}_{\mathtt{I}_i\mathtt{I}_j}$ is the relative linear velocity incremental. $\Delta\hat{\mathbf{v}}_{\mathtt{I}_i\mathtt{I}_j} \doteq \sum_{k=i}^{j-1}\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}({}_{\mathtt{I}_k}\tilde{\mathbf{a}} - \mathbf{b}_i^a)\Delta t$ is the velocity preintegration of the accelerometer measurements from i to j and $\delta\hat{\mathbf{v}}_{\mathtt{I}_i\mathtt{I}_j} \doteq \sum_{k=i}^{j-1}[-\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}({}_{\mathtt{I}_k}\tilde{\mathbf{a}} - \mathbf{b}_i^a)^\wedge\delta\hat{\boldsymbol{\phi}}_{\mathtt{I}_i\mathtt{I}_k}\Delta t + \Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\boldsymbol{\eta}^a\Delta t]$ is the Gaussian noise [31]. In addition, (13) can be reformulated as

$$\underbrace{\begin{aligned}&\mathbf{R}_{\mathtt{IC}}\mathbf{R}_i^T\left(\mathbf{R}_j{}_{\mathtt{C}}\mathbf{p}_{\mathtt{CI}} + \mathbf{p}_j - \mathbf{R}_i{}_{\mathtt{C}}\mathbf{p}_{\mathtt{CI}} - \mathbf{p}_i\right.\\ &\left.-\mathbf{R}_i\mathbf{R}_{\mathtt{IC}}^T\mathbf{v}_i\Delta t_{ij} - \tfrac{1}{2}(\mathbf{R}_{\mathtt{WI}_0}\mathbf{R}_{\mathtt{IC}})^T{}_{\mathtt{W}}\mathbf{g}\Delta t_{ij}^2\right)\end{aligned}}_{h_{I_t}(\mathcal{X}_i,\mathcal{X}_j)} = \underbrace{\Delta_{\mathtt{I}_i}\hat{\mathbf{p}}_{\mathtt{I}_i\mathtt{I}_j} - \delta_{\mathtt{I}_i}\hat{\mathbf{p}}_{\mathtt{I}_i\mathtt{I}_j}}_{\Delta_{\mathtt{I}_i}\mathbf{p}_{\mathtt{I}_i\mathtt{I}_j}} \tag{16}$$

where $h_{I_t}(\mathcal{X}_i,\mathcal{X}_j)$ is the accelerometer translation model, $\Delta_{\mathtt{I}_i}\hat{\mathbf{p}}_{\mathtt{I}_i\mathtt{I}_j} \doteq \sum_{k=i}^{j-1}[\Delta\hat{\mathbf{v}}_{\mathtt{I}_i\mathtt{I}_k}\Delta t + \frac{1}{2}\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}({}_{\mathtt{I}_k}\tilde{\mathbf{a}} - \mathbf{b}_i^a)\Delta t^2]$ is the translation preintegration of accelerometer measurement, and $\delta_{\mathtt{I}_i}\hat{\mathbf{p}}_{\mathtt{I}_i\mathtt{I}_j} \doteq \sum_{k=i}^{j-1}[\delta\hat{\mathbf{v}}_{\mathtt{I}_i\mathtt{I}_k}\Delta t - \frac{1}{2}\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}({}_{\mathtt{I}_k}\tilde{\mathbf{a}} - \mathbf{b}_i^a)^\wedge\delta\hat{\boldsymbol{\phi}}_{\mathtt{I}_i\mathtt{I}_k}\Delta t^2 + \frac{1}{2}\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\boldsymbol{\eta}^a\Delta t^2]$ is the noise. Refer to [31] for the detailed derivation.

3) Residual Definition: Then, we can define the formula of the IMU residual in (2) as follows:

$$\mathbf{r}_I(h_I(\mathcal{X}_i,\mathcal{X}_j),\mathcal{I}_{i,j}) \doteq \begin{bmatrix}\mathbf{r}_r(h_{I_r}(\mathcal{X}_i,\mathcal{X}_j),\mathcal{I}_{i,j})\\ \mathbf{r}_v(h_{I_v}(\mathcal{X}_i,\mathcal{X}_j),\mathcal{I}_{i,j})\\ \mathbf{r}_t(h_{I_t}(\mathcal{X}_i,\mathcal{X}_j),\mathcal{I}_{i,j})\end{bmatrix}. \tag{17}$$

According to (14), the IMU rotation residual is defined as

$$\mathbf{r}_r(h_{I_r}(\mathcal{X}_i,\mathcal{X}_j),\mathcal{I}_{i,j}) \doteq \mathrm{Log}\left(h_{I_r}(\mathcal{X}_i,\mathcal{X}_j)\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}^T\right) \tag{18}$$

where $\mathrm{Log}(\cdot): \mathrm{SO}(3) \to \mathbb{R}^3$ stands for the logarithm map from SO(3) to a vectorized $\mathfrak{so}(3)$.

The IMU velocity residual, based on (15), is

$$\mathbf{r}_v(h_{I_v}(\mathcal{X}_i,\mathcal{X}_j),\mathcal{I}_{i,j}) \doteq \Delta\hat{\mathbf{v}}_{\mathtt{I}_i\mathtt{I}_j} - h_{I_v}(\mathcal{X}_i,\mathcal{X}_j). \tag{19}$$

Similarly, from (16), the IMU translation residual can be derived as

$$\mathbf{r}_t(h_{I_t}(\mathcal{X}_i,\mathcal{X}_j),\mathcal{I}_{i,j}) \doteq \Delta_{\mathtt{I}_i}\hat{\mathbf{p}}_{\mathtt{I}_i\mathtt{I}_j} - h_{I_t}(\mathcal{X}_i,\mathcal{X}_j). \tag{20}$$

C. Camera Measurement Model and its Residuals

The camera measurement is a set of stereo image pairs that are used to extract image features and estimate 3-D landmarks. Given the landmarks $\mathcal{L}_i$ visible in keyframe i and the state $\mathcal{X}_i$, the camera projection model of the landmark ${}_{\mathtt{C}_0}\mathbf{l}_{\mathtt{C}_0\mathtt{l}_m}$ is

$$\Pi(\mathbf{T}_i^{-1} \otimes {}_{\mathtt{C}_0}\mathbf{l}_{\mathtt{C}_0\mathtt{l}_m}) \doteq \mathbf{u}_m$$

where $\Pi(\cdot)$ is the camera model that projects a 3-D map point from the local camera coordinate frame to the pixel coordinate frame, $\otimes$ is the transformation operation of SE(3) group over $\mathbb{R}^3$ elements, and $\mathbf{u}_m$ is the observed pixel position of the landmark on the image. Therefore, the camera measurement model in (2) is

$$h_C(\mathcal{X}_i) = \Pi(\mathbf{T}_i^{-1} \otimes {}_{\mathtt{C}_0}\mathbf{l}_{\mathtt{C}_0\mathtt{l}_m}).$$

We assume the camera measurement is affected by a Gaussian noise. Therefore, considering all the landmarks of keyframe i, the camera residual is formulated as

$$\mathbf{r}_C(h_C(\mathcal{X}_i),\mathcal{C}_i)) = \sum_{m\in\mathcal{M}_i}\Pi(\mathbf{T}_i^{-1} \otimes {}_{\mathtt{C}_0}\mathbf{l}_{\mathtt{C}_0\mathtt{l}_m}) - \mathbf{u}_m. \tag{21}$$

Refer to [1] for more details on the camera measurement model.

V. Online Sensor Calibration

The performance of the acoustic-VI-SLAM system proposed in the previous section can be affected by two primary sources of errors related to sensor calibration. The first source concerns the extrinsic sensor calibration, i.e., the transformations between the camera, the DVL, and the IMU. It is often challenging to manually measure these extrinsic parameters $\mathcal{E}$, particularly rotations, with a high accuracy. Since these parameters might also undergo minor variations over extended usage, the extrinsic calibration process needs to be carried out regularly. The second source of error arises from the misalignment of the DVL transducers. This misalignment can originate during the manufacturing process or be developed as the DVL sensor ages. Therefore, we propose online sensor calibration methods to alleviate or remove these two types of errors.

A. Extrinsic Calibration of DVL, Camera, and IMU Sensors

In the previous section, the extrinsic parameters $\mathcal{E}$ are presumed known. Now, they are considered unknown variables to be estimated. Therefore, we cannot directly leverage the SLAM methodology proposed in Section IV. A straightforward solution is to jointly estimate the SLAM states and the extrinsic parameters within the MAP framework in (2). However, this is impracticable since the optimization would be acutely sensitive to initial values and suffer from considerable local minima. To address these issues, we propose a systematic coarse-to-fine calibration methodology that has the following steps to progressively refine the estimates of the extrinsic parameters $\mathcal{E}$.

1) Vision-only bundle adjustment;
2) Initial estimation of extrinsic parameters;
3) Refined estimation of extrinsic parameters with gyroscope bias;
4) Initialization of gravity direction;
5) Full refinement;

1) Vision-Only Bundle Adjustment: The initial step is to estimate the camera poses and the 3-D landmark positions purely based on a short sequence of images (empirically 10–100

keyframes) for sensor calibration. The standard vision-based SLAM and local bundle adjustment method [1], [32] is used

$$\{\mathbf{T}_i^*, \mathcal{L}_i^*\} = \underset{\mathbf{T}_i, \mathcal{L}_i}{\operatorname{argmin}} \sum_{m \in \mathcal{M}_i} \|\Pi(\mathbf{T}_i^{-1} \otimes {}_{\mathtt{C}_0}\mathbf{l}_{\mathtt{C}_0\mathtt{l}_m}) - \mathbf{u}_m\|^2_{\Sigma_C}.$$

Since a stereo camera is used in our system, the absolute metric scale can be recovered. After solving the optimization, a set of refined camera poses and landmarks $\{\mathbf{T}_i^*$ and $\mathcal{L}_i^*\}$ are obtained. They are treated as known in the following calibration procedure to facilitate convergence and reliability. Notably, to ensure the success of this step, the short image sequence utilized needs to have reasonable image quality with sufficient visual textures for feature extraction and matching, and preferably the observability of the parameters is ensured from the robot motion [33].

2) Initial Estimation of Extrinsic Parameters: In the second phase, an initial estimation of $\mathcal{E}$ is computed by mitigating the influence of the IMU biases and the gravity. By leveraging the optimal camera poses $\{\mathbf{T}_i^*\}$ obtained in the last step, the optimization problem that employs the DVL translation residual in (10) and the IMU rotation residual in (18) is formulated as

$$\begin{aligned}\underset{\mathcal{E}}{\operatorname{argmin}} \quad & \sum_{i,j \in \mathcal{K}_n} \|\mathbf{r}_t(h_{D_t}(\mathcal{X}_i, \mathcal{X}_j), \mathcal{D}_{i,j})\|^2_{\Sigma_D} \\ & + \sum_{i,j \in \mathcal{K}_n} \|\mathbf{r}_r(h_{I_r}(\mathcal{X}_i, \mathcal{X}_j), \mathcal{I}_{i,j})\|^2_{\Sigma_I}.\end{aligned}$$

For this step, the gyroscope bias is assumed zero.

3) Refined Estimation of Extrinsic Parameters With Gyroscope Bias: This step incorporates the gyroscope bias into the optimization process to refine the $\mathcal{E}$ estimate. The same residual terms to the preceding stage are employed for the optimization problem but with the gyroscope biases included as variables to optimize

$$\begin{aligned}\mathcal{E}^*, \mathcal{B}^{g*} = \underset{\mathcal{E}, \mathcal{B}^g}{\operatorname{argmin}} \quad & \sum_{i,j \in \mathcal{K}_n} \|\mathbf{r}_t(h_{D_t}(\mathcal{X}_i, \mathcal{X}_j), \mathcal{D}_{i,j})\|^2_{\Sigma_D} \\ & + \sum_{i,j \in \mathcal{K}_n} \|\mathbf{r}_r(h_{I_r}(\mathcal{X}_i, \mathcal{X}_j), \mathcal{I}_{i,j})\|^2_{\Sigma_I}\end{aligned}$$

where $\mathcal{B}^g \doteq \{\mathbf{b}_i^a\}, i \in \mathcal{K}_n$ denotes the collection of gyroscope biases associated with the keyframes in the image sequence.

4) Initialization of Gravity Direction: To further enhance the accuracy of $\mathcal{E}$, the IMU velocity and translation residuals in (19) and (20) should be considered. For this purpose, the initialization of $\mathbf{R}_{\mathtt{WI}_0}$ is essential. Therefore, based on the optimal $\{\mathbf{T}_i^*, \mathcal{E}^*, \mathcal{B}^{g*}\}$, $\mathbf{R}_{\mathtt{WI}_0}$ is optimized by

$$\begin{aligned}\underset{\mathbf{R}_{\mathtt{WI}_0}}{\operatorname{argmin}} \quad & \sum_{i,j \in \mathcal{K}_n} \|\mathbf{r}_r(h_{I_r}(\mathcal{X}_i, \mathcal{X}_j), \mathcal{I}_{i,j})\|^2_{\Sigma_I} \\ & + \sum_{i,j \in \mathcal{K}_n} \|\mathbf{r}_v(h_{I_v}(\mathcal{X}_i, \mathcal{X}_j), \mathcal{I}_{i,j})\|^2_{\Sigma_I}.\end{aligned}$$

5) Full Refinement: Now, we are ready to perform full refinement of the extrinsic parameters by using the DVL translation residual and all the three IMU residual terms

$$\begin{aligned}\mathcal{E}^* = \underset{\mathcal{E}, \mathbf{R}_{\mathtt{WI}_0}, \mathcal{B}^g, \mathcal{B}^a}{\operatorname{argmin}} \quad & \sum_{i,j \in \mathcal{K}_n} \|\mathbf{r}_t(h_{D_t}(\mathcal{X}_i, \mathcal{X}_j), \mathcal{D}_{i,j})\|^2_{\Sigma_D} \\ & + \sum_{i,j \in \mathcal{K}_n} \|\mathbf{r}_I(h_I(\mathcal{X}_i, \mathcal{X}_j), \mathcal{I}_{i,j})\|^2_{\Sigma_I} \qquad (22)\end{aligned}$$

whose initialization takes the previous optimal values. $\mathcal{B}^a \doteq \{\mathbf{b}_i^a\}, i \in \mathcal{K}_n$ is the set of accelerometer biases.

Since the DVL translation residual incorporates the extrinsic parameter $\mathbf{R}_{\mathtt{ID}}$, any alteration in $\mathbf{R}_{\mathtt{ID}}$ necessitates the reintegration of $\Delta_{\mathtt{D}_i} \bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j}$ during the optimization process. This recurrent computation is computationally intensive and hinders online optimization. A solution to this problem is elaborated in Section V-C.

B. DVL Misalignment Calibration

In Section IV-A2, it is assumed that each transducer is tilted by fixed angles α and β. For the DVL misalignment calibration, the α and β of each transducer are treated as unknown instead. Therefore, (3) can be reformulated as

$$\begin{aligned}\hat{\mathbf{e}}_1 &= [-\cos(\beta_1)\cos(\alpha_1) \quad \sin(\beta_1)\cos(\alpha_1) \quad \sin(\alpha_1)] \\ \hat{\mathbf{e}}_2 &= [-\cos(\beta_2)\cos(\alpha_2) \quad -\sin(\beta_2)\cos(\alpha_2) \quad \sin(\alpha_2)] \\ \hat{\mathbf{e}}_3 &= [\cos(\beta_3)\cos(\alpha_3) \quad -\sin(\beta_3)\cos(\alpha_3) \quad \sin(\alpha_3)] \\ \hat{\mathbf{e}}_4 &= [\cos(\beta_4)\cos(\alpha_4) \quad \sin(\beta_4)\cos(\alpha_4) \quad \sin(\alpha_4)] \qquad (23)\end{aligned}$$

where $\mathcal{O} \doteq \{\alpha_n, \beta_n\}, n \in \{1, 2, 3, 4\}$ is the DVL transducer alignment parameters to be estimated. The extrinsic parameters $\mathcal{E}$ are known during the DVL misalignment calibration.

Similar to the extrinsic calibration introduced in Section V-A, a coarse-to-fine strategy is adopted for the robust calibration of $\mathcal{O}$. It includes the following steps.

1) Vision-only bundle adjustment.
2) Optimization of DVL body velocity.
3) Optimization of DVL transducer alignment parameters.

1) Vision-Only Bundle Adjustment: This initial phase is identical to the vision-only bundle adjustment in Section V-A1. The optimal camera poses and landmark locations are also fixed in the following procedure.

2) Optimization of DVL Body Velocity: We then optimize the velocities based on the DVL translation residual in (10) with $h_{D_t}(\mathcal{X}_i, \mathcal{X}_j)$ being considered as a constant measurement. Therefore, the problem is reformulated as

$$\mathcal{V}^* = \underset{\mathcal{V}}{\operatorname{argmin}} \|h_{D_t}(\mathcal{X}_i, \mathcal{X}_j) - \sum_{k=i}^{j-1} \Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k} \mathbf{R}_{\mathtt{ID}\,\mathtt{D}_i} \mathbf{v} \Delta t\|^2_{\Sigma_D} \qquad (24)$$

where $\mathcal{V} \doteq \{{}_{\mathtt{D}_i}\mathbf{v}\}, i \in \mathcal{K}_n$ is the set of the DVL body velocity variables to optimize.

3) Optimization of DVL Transducer Alignment Parameters: Given the calculated DVL body velocities $\mathcal{V}^*$, the DVL transducer alignment parameters $\mathcal{O}$ is estimated by

$$\mathcal{O}^* = \underset{\mathcal{O}}{\operatorname{argmin}} \sum_{i \in \mathcal{K}_n} \sum_{n=1}^{4} \|{}_i\tilde{v}_n - \hat{\mathbf{e}}_n \cdot {}_{\mathtt{D}_i}\mathbf{v}^*\|^2_{\Sigma_d}$$

where n denotes the transducer index, ${}_i\tilde{v}_n$ refers to the velocity measurement of the nth individual transducer at keyframe i, and $\hat{\mathbf{e}}_n$ is defined in (23).

C. Rapid Linear Approximation Iteration

To enhance the computational efficiency of minimizing the DVL translation residual, a linear approximation method is introduced in the optimization iteration to rapidly approximate the full integration of $\Delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j}$ with linearization. When incremental updates are relatively minor, this approximation is used instead.

1) Approximation Iteration for Extrinsic Calibration: For the extrinsic calibration (22), the DVL translation preintegration $\Delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j}$ is linearized with respect to $\mathbf{R}_{\mathtt{ID}}$ as

$$\frac{\partial \Delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j}}{\partial \boldsymbol{\phi}_{\mathtt{ID}}} = \sum_{k=i}^{j-1} -\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}(\mathbf{R}_{\mathtt{ID}}\,{}_{\mathtt{D}_i}\mathbf{v})^{\wedge}\Delta t \tag{25}$$

where $\boldsymbol{\phi}_{\mathtt{ID}} \in \mathbb{R}^3$ represents the vectorized Lie algebra $\mathfrak{so}(3)$ of $\mathbf{R}_{\mathtt{ID}}$. See its full derivation in Appendix C. Therefore, the approximation of the DVL translation preintegration with an incremental update $\Delta\boldsymbol{\phi}_{\mathtt{ID}}$ can be computed by

$$\bar{\mathbf{P}}_{\boldsymbol{\phi}}(\Delta\boldsymbol{\phi}_{\mathtt{ID}}) \doteq \Delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j} + \frac{\partial \Delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j}}{\partial \boldsymbol{\phi}_{\mathtt{ID}}}\Delta\boldsymbol{\phi}_{\mathtt{ID}}$$

$\bar{\mathbf{P}}_{\boldsymbol{\phi}}(\cdot)$ maps the incremental update of $\mathbf{R}_{\mathtt{ID}}$ to the updated DVL translation integration. To balance accuracy and efficiency, the linearization is employed when the magnitude of $\Delta\boldsymbol{\phi}_{\mathtt{ID}}$ is smaller than a threshold $\sigma_{\boldsymbol{\phi}}$.

2) Approximation Iteration for Misalignment Calibration: For the optimization (24) in the DVL misalignment calibration, $\Delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j}$ is linearized with respect to ${}_{\mathtt{D}_i}\mathbf{v}$ by using

$$\frac{\partial \Delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j}}{\partial\, {}_{\mathtt{D}_i}\mathbf{v}} = \sum_{k=i}^{j-1} \Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}}\Delta t.$$

See its derivation in Appendix D. Therefore, the DVL translation preintegration can be approximated with an incremental update of ${}_{\mathtt{D}_i}\mathbf{v}$

$$\bar{\mathbf{P}}_{\mathbf{v}}(\Delta_{\mathtt{D}_i}\mathbf{v}) \doteq \Delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j} + \frac{\partial \Delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j}}{\partial_{\mathtt{D}_i}\mathbf{v}}\Delta_{\mathtt{D}_i}\mathbf{v} \tag{26}$$

where $\bar{\mathbf{P}}_{\mathbf{v}}(\cdot)$ defines the mapping from the incremental update of ${}_{\mathtt{D}_i}\mathbf{v}$ to the updated DVL translation integration. Similarly, this linear approximation is used only when the norm of $\Delta_{\mathtt{D}_i}\mathbf{v}$ is below a threshold $\sigma_{\mathbf{v}}$.

It is worth mentioning that the calibration accuracy is largely contingent upon the quality of poses estimated through the vision-only bundle adjustment, whose reliability can be compromised in underwater environments due to the challenges aforementioned. Therefore, it is recommended that calibration is conducted in areas with discernible visual features, ideally with sufficient motion to ensure the observability of the parameters to calibrate [33].

VI. System Implementation

We integrate the proposed acoustic-visual-inertial approach and the online sensor calibration method into the ORB-SLAM3 system [5], with a particular emphasis on incorporating DVL measurements to enhance accuracy and robustness. While the system retains the classic three-thread design (tracking, local mapping, and loop closure) and the multimap strategy At last, all of these modules have been upgraded. The system overview is shown in Fig. 5. Specifically, we have made the following upgrades to ORB-SLAM3.

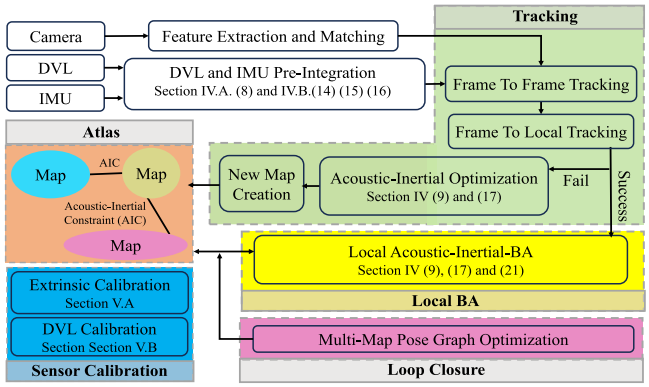


Fig. 5. Overview of system implementation.

A. Sensor Data Processing

The system takes stereo image pairs from a camera and performs stereo matching to provide the depth with an absolute scale. The DVL measurement is preintegrated using (8), and the IMU data are processed using (14)–(16).

B. Tracking

The tracking thread is mainly to track image frames using the predicted poses from the DVL and IMU measurements. When the frame-to-local tracking fails, e.g., due to poor image quality, acoustic-inertial optimization is performed for pose estimation based on the DVL, and IMU residuals in (9) and (17). Meanwhile, an acoustic-inertial constraint is derived and added between consecutive submaps in the Atlas module. Thanks to the acoustic-inertial optimization, pose tracking is more accurate and reliable in poor visual conditions.

C. Local BA

The local BA module maintains a sliding window of up to 10 keyframes. When a new keyframe is inserted, the acoustic-inertial BA is performed to fuse the data from DVL, IMU, and camera using the residuals in (9), (17), and (21).

D. Atlas

The Atlas module manages submaps. Different from ORB-SLAM3 which divides submaps into active and inactive ones, the proposed system maintains a consistent map which interlinks all submaps through the acoustic-inertial constraints.

E. Loop Closure

The loop closure thread handles loop fusion within the same map and submap merging. ORB-SLAM3 [5] only corrects the active map and a submap to be merged. Instead, the proposed

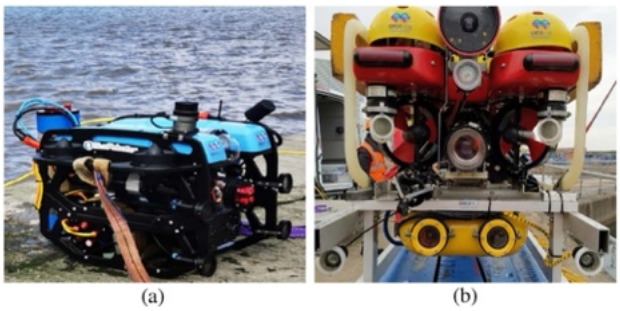


Fig. 6. Underwater vehicles used for experiments. (a) BlueROV2 for WaveTank test. (b) Falcon vehicle for Offshore test.

method corrects all submaps via the acoustic-inertial constraints. Therefore, the overall map and pose errors can be corrected for each time when a submap merging is performed.

F. Sensor Calibration

When the online sensor calibration module is activated, either the extrinsic calibration or the DVL calibration is conducted using the sensor data between 10 keyframes.

VII. Experimental Results

The performances of the proposed SLAM and calibration methods are evaluated in this section.

A. Experiment Settings and Datasets

Due to the lack of a public underwater dataset with the required sensor suite (i.e., stereo camera, IMU, and DVL), two datasets—WaveTank and Offshore datasets—are collected for the experimental evaluation using the two underwater remotely operated vehicles (ROVs) shown in Fig. 6.

1) WaveTank Dataset: The WaveTank dataset is collected in a 9 m × 12 m tank with a structure deployed in the middle, as illustrated in Fig. 7. A BlueROV2 vehicle equipped with a WaterLinked A50 DVL running at 5 Hz, a MICROSTRAIN 3DM-GX5-AHRS IMU at 330 Hz and a custom underwater stereo camera at 20 Hz [34] is used to collect DVL, IMU, and stereo data. In order to have ground truth (GT) for quantitative evaluation, the structure is covered with AprilTag markers [35], as shown in Fig. 7. A fiducial-marker based SLAM is developed to provide GT robot poses. Specifically, the marker poses are initially calibrated using a calibration sequence collected by allowing the vehicle to move slowly and smoothly, generating images where markers are clearly recognizable. In addition, multiple loops are closed to ensure the precision of the landmark poses. This process is crucial to ensure the high accuracy of the marker poses. Following the calibration, the location of each marker is treated as a priori information, and the camera pose is subsequently obtained by being localized in relation to the calibrated marker poses. Therefore, whenever a marker is recognized from an image, an accurate GT camera pose becomes available. Note that the AprilTag markers are only used for generating GT, and they are not used in the SLAM systems to be evaluated. The dataset, along with its GT and evaluation toolset,

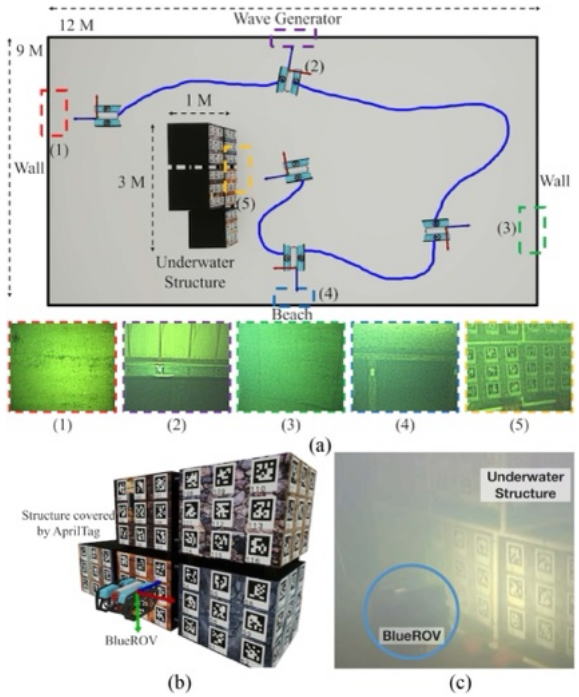


Fig. 7. Experiment settings of the self-collected WaveTank dataset. (a) Overview of the tank. (b) Underwater structure. (c) BlueROV2 moving around the underwater structure.

TABLE I
Statistics of the WaveTank Sequences

Sequence	Velocity Distribution (0 to 0.5 m/s)	Light	GT %
Structure Easy(SE)		on	99.37%
Structure Medium(SM)		off	99.67%
Structure Hard(SH)		off	84.65%
HalfTank Easy(HE)		on	36.96%
HalfTank Medium(HM)		on	44.19%
HalfTank Hard(HH)		off	41.40%
WholeTank Medium(WM)		on	60.73%
WholeTank Hard(WH)		off	18.97%

will be introduced in detail and made publicly accessible in a separate dataset paper.

Eight sequences are employed for the quantitative analysis. The sequences are obtained under varying scenarios and configurations, thereby resulting in assorted levels of difficulty, as summarized in Table I. Factors, such as high speed or the lack of light, significantly augment the difficulty of the sequences. Moreover, the AprilTag coverage rate indicates the proportion of images with AprilTag markers visible to estimate the GT poses, which roughly indicates the visual conditions.

According to the trajectories, three categories exist: structure, half tank, and whole tank. Structure sequences are exclusively collected around the underwater structure and a majority of their images have visible markers. The half-tank sequences are collected along a loop traversing half of the tank. The whole tank sequences are gathered along a loop traversing the entire tank.

TABLE II
COMPETING SLAM METHODS

Method	Sensor	Front-end	Back-end
Ours	Stereo cameras, IMU, DVL	Feature-based	Graph based
Previous Work [12]	Stereo cameras, Gyroscope, DVL	Feature based	EKF, Graph based
SVIN2 [6]	Stereo cameras, IMU	Feature based	Graph based
ORB-SLAM3 [5]	Stereo cameras, IMU	Feature based	Graph based
VINS-Fusion [4]	Stereo cameras, IMU	Optical-Flow	Graph based
Basalt [37]	Stereo cameras, IMU	Optical-Flow	Graph based

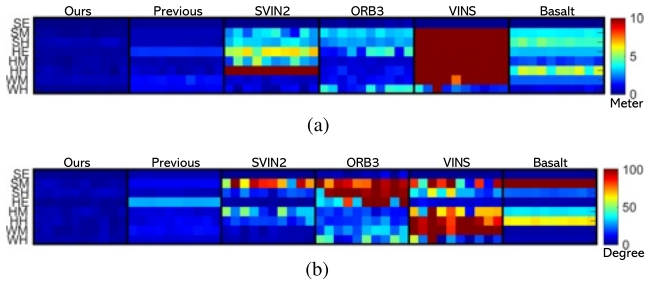


Fig. 8. Odometry results of ten runs on all WaveTank sequences. (a) Errors on odometry translation. (b) Errors on odometry rotation.

Therefore, the half-tank and the whole-tank sequences contain a larger number of images from textureless areas without GT.

2) Offshore Dataset: The Offshore dataset is collected at an offshore wind farm using a Saab Seaeye Falcon ROV as shown in Fig. 6. Since there is no GT in the open sea, our results were compared to those generated by COLMAP [36], a widely used offline structure from motion (SfM) and multiview stereo (MVS) pipeline.

B. Competing Methods

In our study, we conduct a comparative analysis of our results with five state-of-the-art SLAM systems, encompassing two underwater SLAM works, our previous work [12] and SVIN2 [6], and three visual-inertial SLAM works, VINS-Fusion [4], ORB-SLAM3 (ORB3) [5], and Basalt [37]. Our previous work [12] integrates a DVL, a gyroscope, and a stereo camera within a loosely coupled framework. In the absence of a downward-looking sonar system on our ROVs, SVIN2 [6] operates in its stereo-inertial mode. Similarly, VINS-Fusion, ORB-SLAM3, and Basalt are configured to function in their stereo-inertial mode. The distinctive features of these systems are summarized in Table II, providing a comprehensive overview of their respective configurations and capabilities.

To ensure fair comparisons, we used the same parameters, including camera, IMU, and extrinsic parameters, as much as possible for all methods. For unique parameters specific to each method, we used the default settings provided by the authors. Specifically, VINS-Fusion, ORB SLAM3, and Basalt used the EuRoc settings from their original GitHub repositories. SVIN2 used the "svin_stereorig_v2" settings applied to the cave sequence in the original repository. Our previous work used the same parameters as the proposed method.

C. Evaluation Metrics and Preprocessing

Given that GT is only accessible when the AprilTag markers are visible, i.e., when the robot faces towards the structure, only pose estimates with associated GT are used for quantitative evaluation. For each sequence, we execute each method ten times to compute an average error of the root-mean-square errors (RMSE) and standard deviation (STD) of the absolute error.

1) Preprocessing: The preprocessing of pose estimates from different methods is a crucial step before conducting evaluations, given that these pose estimates are represented in varying reference frames and originate from different timestamps based on the implementations of the methods.

The preprocessing can be divided into the following three parts

1) *Frame Transformation:* Since the implementations of the competing methods use different reference frames (some in camera frames, some in IMU frames) for their poses, we first convert all these poses into the same camera frame.
2) *Time Synchronization:* For a trajectory estimated by a method, its poses are matched with GT based on timestamps.
3) *Trajectory Alignment:* A SE3 transformation is executed to align the first pose of a trajectory with the GT trajectory. This step ensures all trajectory starts from the same origin.

D. Evaluation on Odometry

For the odometry evaluation, the loop-closure detection functions of all methods are disabled.

1) Quantitative Evaluation: The RMSE and STD are detailed in Table III, and the RMSE of the 10 runs are given in Fig. 8 as error maps. These results clearly demonstrate the superiority of our proposed method across all WaveTank sequences in terms of both translational and rotational accuracy. This can be attributed to the tightly coupled integration of acoustic, visual, and inertial sensing in the proposed SLAM methodology.

In comparison, our previous work surpasses other visual-inertial SLAM systems in translational accuracy, largely due to the incorporation of the DVL sensor. However, its rotational performance is compromised. We hypothesize that this is primarily because its loosely-coupled framework does not facilitate real-time correction of gyroscope bias. In addition, the absence of an accelerometer renders the absolute roll and pitch angles unobservable, further impacting its rotational accuracy. To further demonstrate this, a detailed ablation study is presented in Section VII-H1.

Other systems, SVIN2, ORB3, VINS, and Basalt, exhibit significant errors in translation, particularly in sequences characterized by challenging visual conditions. These conditions include scenarios with insufficient features in the images or inadequate lighting, underlining the limitations of these systems in handling the challenges in underwater environments. They also show substantial variability in accuracy across different

TABLE III
ODOMETRY PERFORMANCE IN WAVETANK DATASET AVERAGING 10 RUNS

	Translation Error RMSE (in meter) / STD						Rotation Error RMSE (in degree) / STD					
	Ours	Previous Work	SVIN2	ORB3	VINS	Basalt	Ours	Previous Work	SVIN2	ORB3	VINS	Basalt
Structure Easy	**0.07 / 0.03**	0.21 / 0.12	0.09 / 0.03	0.28 / 0.09	0.13 / 0.04	0.12 / 0.06	**1.62** / 0.68	4.88 / 2.49	1.84 / **0.49**	5.45 / 1.52	2.15 / 0.72	2.57 / 1.02
Structure Medium	**0.18 / 0.08**	0.54 / 0.30	2.94 / 1.64	3.30 / 1.08	NaN / NaN	3.66 / 1.81	**4.17 / 1.57**	11.19 / 5.05	74.56 / 47.51	93.87 / 52.48	66.55 / 22.17	NaN / NaN
Structure Hard	**0.50** / 0.24	0.50 / **0.23**	3.26 / 1.43	2.73 / 1.45	NaN / NaN	4.48 / 2.10	**5.63** / 2.76	5.81 / **2.35**	16.57 / 6.83	98.37 / 60.33	35.61 / 10.83	21.28 / 12.26
HalfTank Easy	**0.28 / 0.17**	1.69 / 1.10	6.01 / 4.33	2.69 / 1.45	29.83 / 20.74	3.12 / 2.14	**2.04 / 0.74**	29.87 / 18.54	3.00 / 1.65	75.35 / 37.99	8.38 / 5.03	4.02 / 1.80
HalfTank Medium	**0.29 / 0.14**	0.44 / 0.22	3.40 / 1.76	0.74 / 0.38	NaN / NaN	1.87 / 0.63	**4.29 / 1.95**	8.49 / 3.75	32.13 / 15.37	16.24 / 7.42	64.33 / 29.31	35.52 / 17.42
HalfTank Hard	**0.36 / 0.22**	0.58 / 0.37	77.60 / 55.07	1.10 / 0.70	NaN / NaN	5.14 / 3.34	**3.84 / 1.99**	10.20 / 5.95	18.96 / 10.37	19.03 / 13.10	97.81 / 44.08	63.50 / 37.69
WholeTank Medium	**0.52 / 0.28**	1.34 / 0.73	0.72 / 0.41	1.18 / 0.71	13.08 / 7.22	2.18 / 1.25	**3.34** / 1.38	12.49 / 5.51	5.29 / 2.12	27.36 / 14.34	NaN / NaN	3.38 / **1.35**
WholeTank Hard	**0.22 / 0.12**	1.11 / 0.83	0.83 / 0.65	2.96 / 2.49	NaN / NaN	0.95 / 0.76	**3.99 / 1.37**	8.14 / 4.17	4.69 / 2.35	30.91 / 24.97	29.24 / 11.20	4.61 / 2.56

The bold entities in these tables indicate the best performance among all the methods.

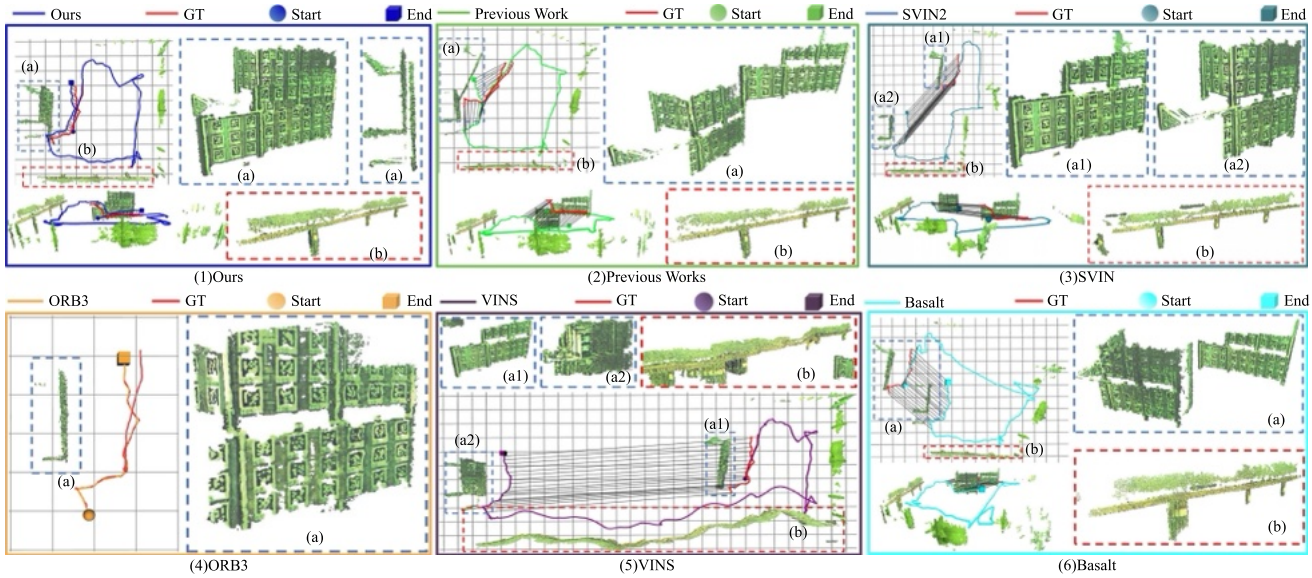


Fig. 9. Estimated trajectories and dense reconstruction of all methods on HE sequence in odometry mode. ◯ and □ mark the start and the end of a trajectory. For each subfigure, the left part (with a grid background) shows the dense map and the estimated trajectory with errors to the GT as −, and its map sections highlighted with dashed lines are shown correspondingly. A grid cell indicates $1m^2$. (a) Ours. (b) Previous works. (c) SVIN. (d) ORB3. (e) VINS. (f) Basalt.

sequences, and this inconsistency is further evident when observing different runs of the same sequence. SVIN2 achieves results comparable to our proposed method on the SE and WH sequences, where the image quality is generally high. In the WH sequence specifically, SVIN2 surpasses other visual-inertial approaches in performance, though it still exhibits a noticeable increase in error compared to our method. However, in more challenging sequences, SVIN2's performance aligns with that of other systems, showing significant odometry drifts. ORB3 performs well on the SE, HM, HH, and WM sequences but fails on other sequences. When ORB3 loses tracking, it resets its pose to the origin and ceases to publish poses. Hence, only a few poses, mainly in the beginning where features are clearly visible, are available. Despite the seemingly acceptable error, in reality, ORB3 mostly fails on these sequences. VINS shows considerable errors in most sequences. Unlike ORB3, when confronted with poor-quality images that cause loss of tracking, VINS drifts rapidly resulting in high errors. Basalt only achieves comparable performance to our proposed method on the SE sequence, drifting on the other sequences with more challenging visual conditions.

In contrast, our method incorporates the use of a DVL to enhance the accuracy of translational estimation. This approach significantly improves both performance and robustness, offering a notable advantage over the visual-inertial systems, especially in underwater environments where image quality is usually compromised.

2) Qualitative Evaluation on HalfTank Easy (HE) Sequence: Fig. 9 shows the trajectories estimated on the HE sequence by all the methods. This sequence encompasses challenging scenarios in which images lack distinct features. As a consequence, this sequence presents a decent level of difficulty. In this sequence, the vehicle initially traverses the front side of the structure where GT is available. Subsequently, it moves through half of the tank along its boundary where GT is unavailable. The vehicle finally returns to the front side of the structure where GT becomes accessible again.

The results indicate that our proposed method achieves the lowest level of drift. Our previous work exhibits the second-lowest drift. SVIN2, VINS, and Basalt exhibit significantly higher drift than the proposed method, primarily due to the rapid accumulation of translation errors under degraded visual

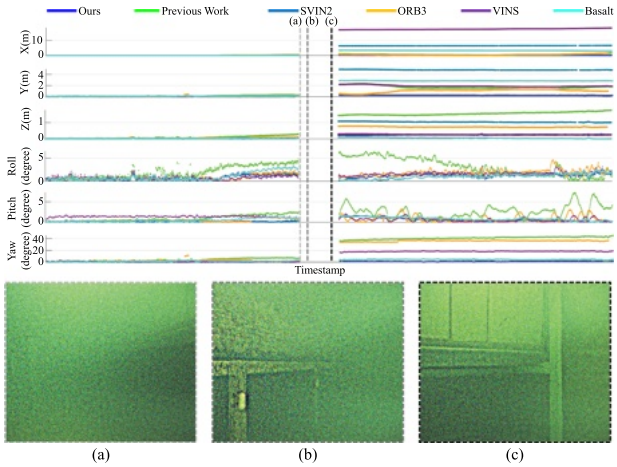


Fig. 10. Odometry errors on HE sequence and visually degraded cases. The timestamp gaps indicate the times without GT poses. (a)–(c) Challenging visual conditions in visually degraded cases.

conditions. In contrast, ORB3 loses tracking until the robot traverses a significant portion of the trajectory and returns to the structure where it successfully initializes again.

Fig. 9(a) and (b) provides the detailed views of the reconstructed map sections. The presence of misalignment or deformation in these reconstructions is indicative of the errors that have accumulated over the course of traversing the entire trajectory. Our proposed method achieves the most accurate and comprehensive dense reconstruction of the structure and the tank's boundary.

Fig. 10 shows the evolution of errors over time across the 6-DoF poses. The proposed method consistently achieves high accuracy across all axes. Visual-inertial methods, including SVIN2, ORB-SLAM3, VINS, and Basalt, demonstrate comparable performance in roll and pitch but exhibit noticeable drift on other axes due to the challenging visual conditions. Our previous work [12] exhibited significant drift on orientation, contributing to overall trajectory drift. We hypothesize that this was primarily due to the exclusion of accelerometer integration and the assumption of zero gyroscope bias. To further demonstrate the effects of bias correction and accelerometer integration, a detailed ablation study is presented in Section VII-H1.

3) Qualitative Evaluation on WholeTank Hard (WH) Sequence: Fig. 11 presents the trajectories on the WH sequence. This sequence involves a longer traversal distance and more aggressive vehicular motion, apart from challenging visual conditions. Initially, the vehicle circumnavigates the structure where GT is accessible, and it subsequently traverses the entire tank by following its boundary. The vehicle finally returns to the structure where GT becomes available again.

The trajectories depicted in Fig. 11 demonstrate that our proposed method exhibits the least drift after navigating the entire tank. All other methods show noticeable drifts. The drift in our previous work is primarily along the z-axis, while SVIN2 and Basalt experience mainly translational drift. ORB3 and VINS demonstrate substantially higher overall drift compared to the other methods.

E. Evaluation on SLAM

For the SLAM evaluation, the loop closure functionality of each method is activated.

1) Quantitative Evaluation: Similar to the odometry evaluation, for each sequence, a method is run 10 times to compute the average RMSE and STD in Table IV. The results of the 10 runs are also given in Fig. 12. We can see that our method surpasses others in most scenarios. Although the SLAM performances of SVIN2, ORB3 and VINS show improvement compared with their odometry, they still exhibit significant errors in translation, particularly in sequences characterized by challenging visual conditions. The loop closure mechanisms somewhat mitigate the drifts, but the lack of sufficient features or lighting significantly degrades the performance and reliability of visual-inertial systems in underwater settings. Basalt only achieves slightly better performance to our proposed method on the SE sequence, drifting on sequences with more challenging visual conditions. We hypothesize that the superior performance of Basalt on the SE sequence is due to its use of an offline loop closure approach, which allows for extensive frame matching and larger-scale optimization. In contrast, our approach employs an online loop correction in real-time and achieves comparable performance. In addition, Basalt frequently encounters numerical faults, leading to system crashes when running on sequences with challenging visual conditions.

The temporal evolution of errors across the 6-DoF poses is given in Fig. 13. Most methods, except SVIN2 and Basalt, exhibit a significant reduction in error following loop detection for enhanced trajectory accuracy. Our proposed method consistently achieves superior accuracy both before and after loop closure. In comparison, our previous work realizes comparable accuracy in translation and yaw errors post loop closure. However, its errors in roll and pitch remain uncorrected. SVIN2 does not detect the loop, and its loop-closure module, implemented as a separate ROS node from the tracking node, publishes poses at a low frequency. This results in only a few data points being represented. After the loop closures, ORB3 approaches the accuracy of the proposed method in the translation and yaw axes. However, it exhibits significant errors in roll and pitch at the start of the second trajectory segment. As discussed before, ORB3 cannot localize the robot when traversing along the tank wall. VINS exhibits a noticeable reduction in x-translation error and gradually corrects for drift, resulting in a stepwise decrease in error. Basalt exhibits significant errors in trajectory due to false positive loop closure detections under challenging visual conditions.

2) Qualitative Evaluation on HE Sequence: Fig. 14 illustrates the trajectories and dense reconstructions produced by the methods in SLAM mode for the HE sequence. Our proposed method, alongside our previous work, estimates trajectories well aligned with GT, producing good-quality 3-D maps. The map misalignments observed in the odometry mode in Fig. 9(a) and (b) have also been corrected in the SLAM mode. Meanwhile, ORB3 achieves significantly improved results compared to its odometry mode, yet it fails to localise around the tank wall. The results of SVIN2 and VINS still show clear errors in the

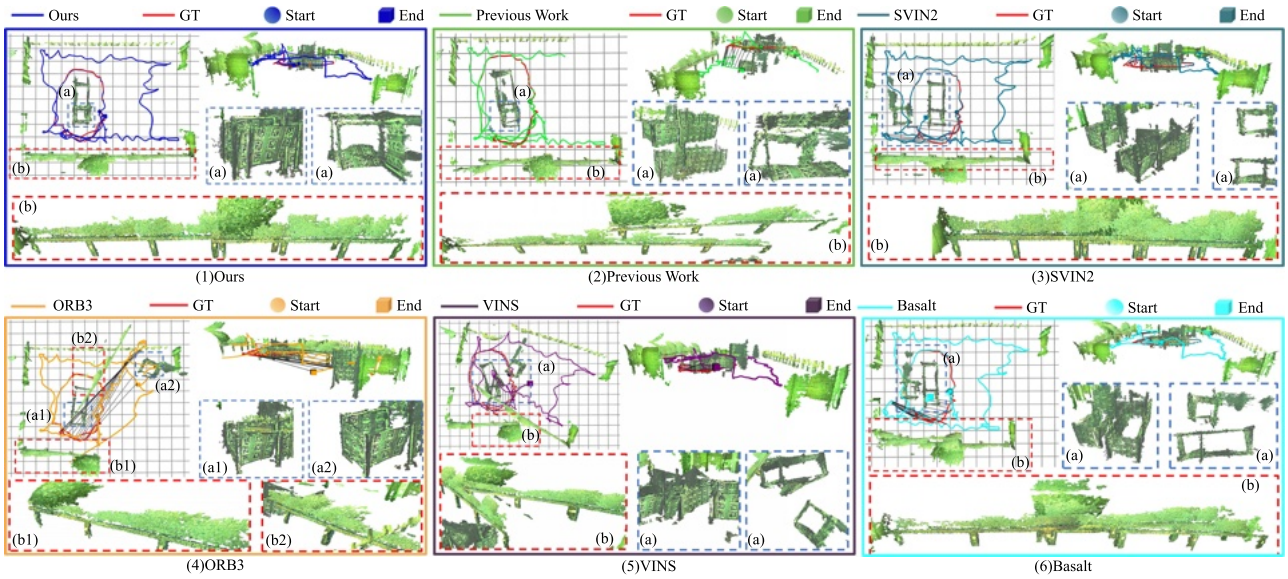


Fig. 11. Estimated trajectories and dense maps on WH sequence in odometry mode. The figure legends and grid size are the same as Fig. 9. (a) Ours. (b) Previous work. (c) SVIN2. (d) ORB3. (e) VINS. (f) Basalt.

TABLE IV
SLAM PERFORMANCE IN WAVETANK DATASET AVERAGING 10 RUNS

	Translation Error RMSE (in meter) / STD						Rotation Error RMSE (in degree) / STD					
	Ours	Previous Work	SVIN2	ORB3	VINS	Basalt	Ours	Previous Work	SVIN2	ORB3	VINS	Basalt
Structure Easy	0.06 / 0.03	0.18 / 0.11	0.09 / 0.04	0.20 / 0.09	0.22 / 0.07	**0.04 / 0.01**	1.51 / 0.67	4.11 / 2.23	1.76 / 0.45	4.18 / 1.82	2.32 / 0.67	**0.89 / 0.38**
Structure Medium	**0.16 / 0.10**	0.49 / 0.29	2.11 / 1.36	3.49 / 1.17	NaN / NaN	NaN / NaN	**3.39 / 1.48**	10.98 / 5.41	57.51 / 39.42	90.01 / 49.18	65.96 / 23.40	NaN / NaN
Structure Hard	**0.31 / 0.17**	0.43 / 0.23	3.59 / 1.54	2.93 / 1.30	NaN / NaN	4.17 / 2.15	**4.10 / 2.06**	5.92 / 2.76	23.44 / 7.56	NaN / NaN	38.24 / 14.15	14.42 / 7.86
HalfTank Easy	**0.18 / 0.13**	1.12 / 0.86	4.51 / 3.44	2.21 / 1.76	26.93 / 19.67	17.46 / 15.20	**1.84 / 0.65**	19.83 / 13.66	3.46 / 1.65	48.61 / 38.08	8.43 / 5.12	66.63 / 19.90
HalfTank Medium	**0.24 / 0.16**	0.26 / 0.17	2.90 / 2.07	0.71 / 0.42	NaN / NaN	NaN / NaN	**3.17 / 1.64**	5.93 / 3.51	24.15 / 13.89	15.16 / 7.77	45.99 / 21.16	NaN / NaN
HalfTank Hard	**0.24 / 0.16**	0.37 / 0.26	68.70 / 47.50	1.15 / 0.75	NaN / NaN	NaN / NaN	**3.45 / 1.94**	9.85 / 6.36	15.65 / 8.41	22.24 / 15.57	92.07 / 48.77	65.16 / 37.67
WholeTank Medium	**0.14 / 0.11**	0.27 / 0.13	0.41 / 0.24	0.68 / 0.22	4.11 / 2.24	2.17 / 1.25	**1.64 / 0.87**	9.20 / 4.88	6.58 / 3.25	8.19 / 3.22	91.50 / 45.44	3.36 / 1.38
WholeTank Hard	**0.12 / 0.07**	0.30 / 0.17	0.27 / 0.21	2.37 / 1.95	NaN / NaN	0.95 / 0.76	**2.83 / 1.21**	10.70 / 6.87	3.80 / 1.90	30.50 / 23.99	28.58 / 11.13	4.54 / 2.57

The bold entities in these tables indicate the best performance among all the methods.

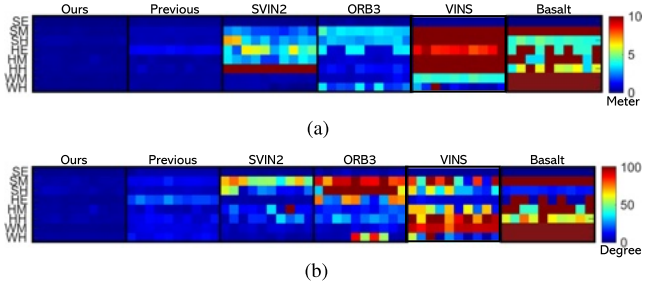


Fig. 12. SLAM results of 10 runs on all WaveTank sequences. (a) SLAM translation error. (b) SLAM rotation error.

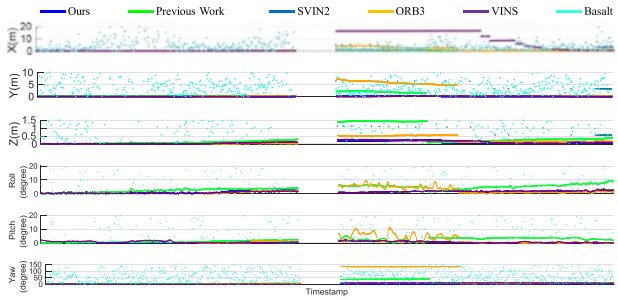


Fig. 13. SLAM errors of six axis on HE sequence. The gaps indicate the times without GT poses.

GT. Basalt exhibits significant errors in the trajectory and dense reconstruction on the HE sequence. Furthermore, it consistently crashes on the WH sequence, preventing result acquisition. Therefore, the maximum error is manually set for the WH sequence. We hypothesize that this is caused by false positive loop closure detections due to poor image quality.

3) Qualitative Evaluation on WH Sequence: The quantitative SLAM results on WH sequence are also shown in Fig. 15, which provides similar insights into the SLAM performance. The absence of Basalt's results is due to its consistent crashes on this sequence.

F. Validation in Real Offshore Environments

The offshore experiments were conducted near an offshore wind turbine in the North Sea. The Falcon ROV, localized using the proposed SLAM method, was deployed to operate around the turbine foundation. Due to the lack of GT data in open-sea

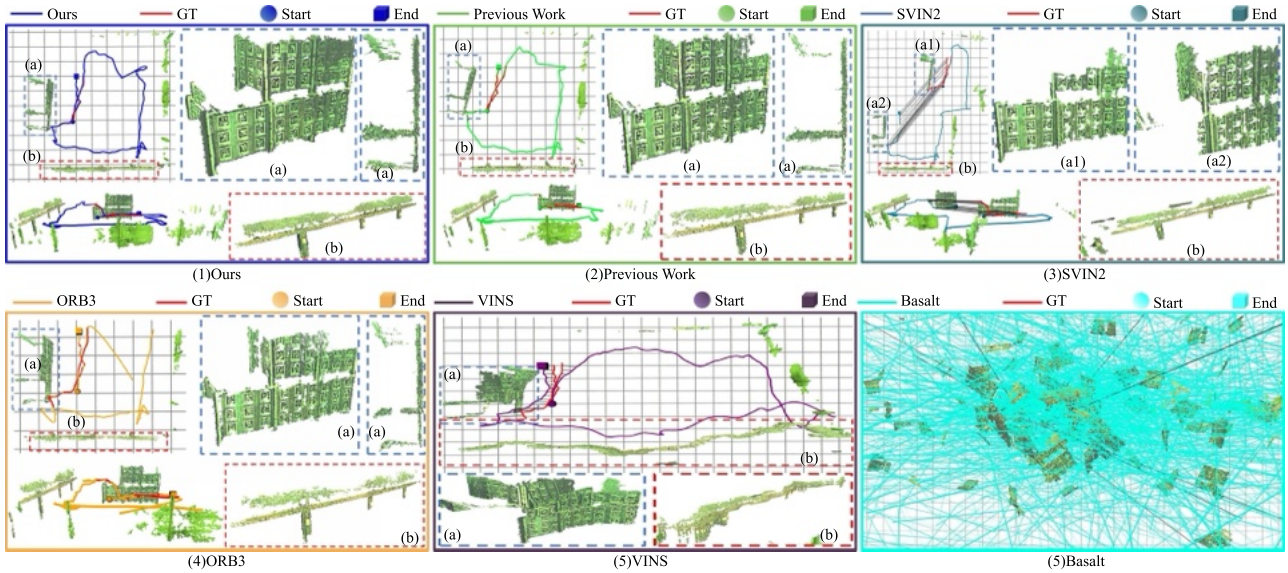


Fig. 14. Estimated trajectories and dense maps on HE sequence in SLAM mode. The figure legends and grid size are the same as Fig. 9. (a) Ours. (b) Previous work. (c) SVIN2. (d) ORB3. (e) VINS. (f) Basalt.

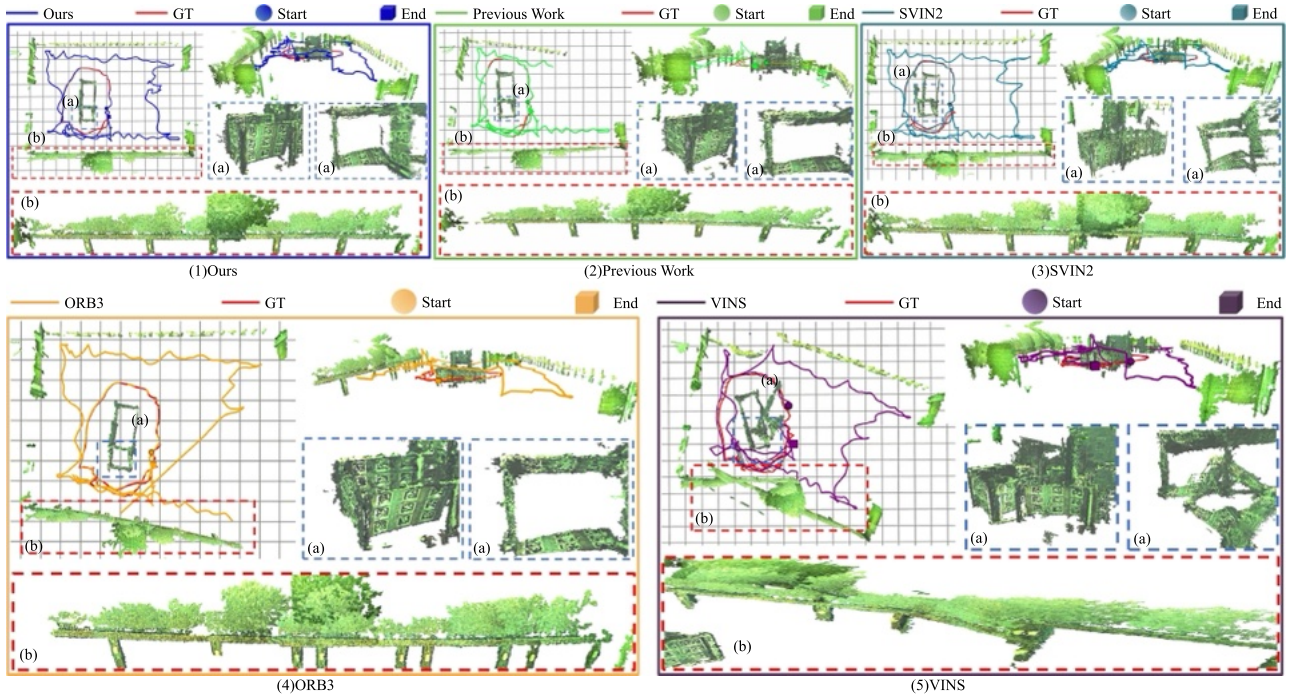


Fig. 15. Estimated trajectories and dense maps on WH sequence in SLAM mode. The figure legends and grid size are the same as Fig. 9. (a) Ours. (b) Previous work. (c) SVIN2. (d) ORB3. (e) VINS.

environments, we use trajectory and reconstruction generated by COLMAP [36] as a reference to evaluate our performance. However, the challenging nature of the Offshore image sequences compromised the performance of COLMAP's MVS component, resulting in noisy dense reconstructions. Consequently, only the SfM generated trajectory from COLMAP was used as a reference. To assess reconstruction quality, point clouds obtained through stereo matching were incrementally fused using both the trajectory produced by our method and that of COLMAP's SfM.

Fig. 1 shows the dense point cloud covering an expansive region of about 50-m depth from the top to the bottom of a turbine base. Despite poor image quality in certain areas, the dense reconstruction remains consistent, verifying the accuracy of the poses estimated by the proposed SLAM technique and its viability in real offshore scenarios. In addition, we compare the reconstructions generated by our method and COLMAP in Fig. 16. The COLMAP trajectory exhibits misalignment with our trajectory in both the top and bottom areas. The top area, covered in algae, led to noisy results from COLMAP, evident

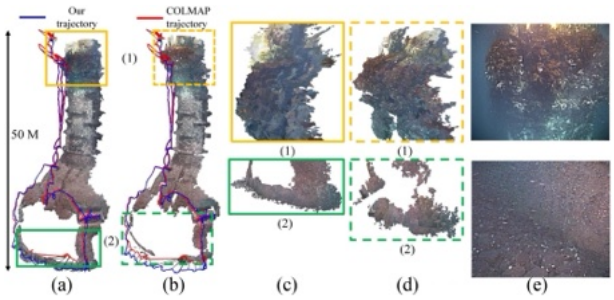


Fig. 16. Offshore 3-D reconstruction and estimated trajectory comparison with COLMAP. (a) Our reconstruction. (b) COLMAP reconstruction. (c) Our details. (d) COLMAP details. (e) Image input.

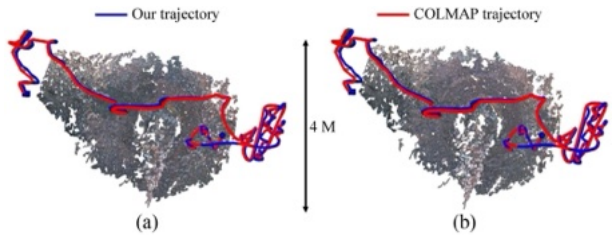


Fig. 17. Offshore 3-D reconstruction and estimated trajectory comparison with COLMAP. (a) Our reconstruction. (a) COLMAP reconstruction.

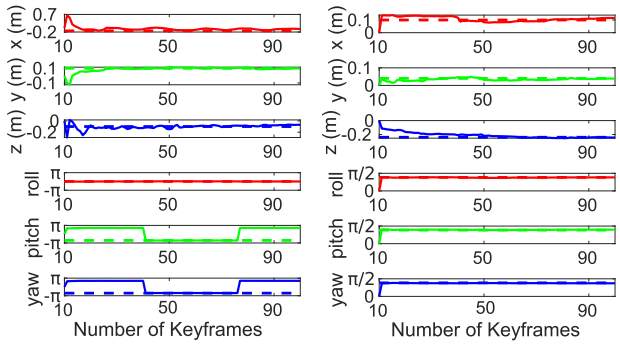
in the inconsistent reconstruction shown in Fig. 16(a). The bottom area, containing motion blur and textureless images, also challenged COLMAP, resulting in a suboptimal reconstruction shown in Fig. 16(b). In contrast, our method's dense reconstruction is more consistent and accurate, even in these areas with poor image quality.

Furthermore, we compared our results with COLMAP in the region surrounding a cable socket on the wind turbine base, as shown in Fig. 17. Both trajectories align well, and the dense reconstructions are consistent. The offshore experiments demonstrate the robustness of the proposed method in challenging real-world ocean environments.

G. Evaluation on Sensor Calibration

1) Extrinsic Calibration of DVL, IMU, and Camera: A segment of the Easy Structure sequence which has relatively clear and richly featured images is chosen for the extrinsic calibration among the DVL, IMU, and camera sensors. An identity matrix is set as the initial extrinsic transformation for the calibration optimization, assuming no prior information about the calibration parameters. Then the $\mathbf{T}_{\mathtt{ID}}$ (the extrinsic parameters between the IMU and the DVL) and $\mathbf{T}_{\mathtt{DC}}$ (the extrinsic parameters between the camera and the DVL) are optimized. The calibration process commences after the insertion of 10 keyframes and ends when 100 keyframes have been received.

The extrinsic calibration result is shown in Fig. 18. We can see the estimated extrinsic parameters gradually converge to the measured values. Note the slight offsets between the calibrated parameters and the manual measures are likely caused by the inevitable measurement errors, particularly for the orientations. Nonetheless, the results demonstrate that the proposed method can achieve decent calibration even without prior information.



(1) Extrinsic parameters $\mathbf{T}_{\mathtt{ID}}$ (2) Extrinsic parameters $\mathbf{T}_{\mathtt{DC}}$

Fig. 18. Results of extrinsic calibration. Solid lines: Estimation. Dashed lines: Manual measure. (a) Extrinsic parameters $\mathbf{T}_{\mathtt{ID}}$. (b) Extrinsic parameters $\mathbf{T}_{\mathtt{DC}}$.

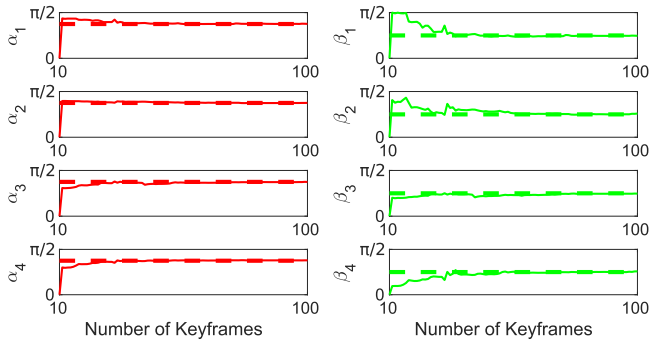


Fig. 19. DVL misalignment calibration. Solid lines: estimation. dashed lines: manual measure.

2) DVL Misalignment Calibration: The DVL misalignment calibration is also performed on the Easy Structure sequence. The results are given in Fig. 19. Over time, we can observe the calibrated transducer alignment parameters converging to the default setting. Since our DVL is rather new and well-maintained with no substantial misalignment, the calibration results closely match the default setting. These results demonstrate the capability of the proposed method to calibrate the transducer alignment parameters accurately.

H. Ablation Study

1) Effect of Bias Correction and Accelerometer Residual: In previous sections, we hypothesized that the drift observed in our prior work was primarily due to the lack of accelerometer integration and the assumption of zero gyroscope bias. To further substantiate this hypothesis and analyze the factors contributing to the performance improvement of our proposed method over the previous work, we present a detailed ablation study in this section.

To evaluate their impact, we disabled the accelerometer-related residuals [defined in (19) and (20)] and bias correction in the optimization process. The results, illustrated in Fig. 20, reveal increased drift in roll, pitch, and z-axes when either accelerometer residuals or bias correction is disabled individually. The highest error occurs when both are disabled simultaneously. This is because the IMU residuals constrain roll and pitch through gravity, and gyroscope bias correction further reduces

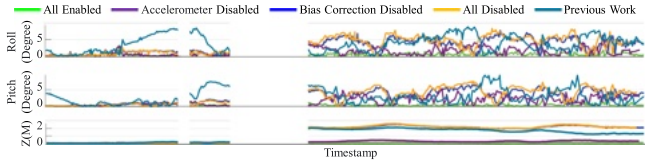


Fig. 20. Abalation study of the effect of bias correction and accelerometer integration. The gaps indicate the times without GT poses.

TABLE V
ABLATION STUDY OF THE EFFECT OF SENSOR CALIBRATION

	Translation Error (in meter)			Rotation Error (in degree)		
	All Calibrated	Extrinsics Uncalibrated	DVL Uncalibrated	All Calibrated	Extrinsics Uncalibrated	DVL Uncalibrated
Seq 1	**0.308**	1.421	1.082	**1.438**	4.030	4.065
Seq 2	**0.039**	0.167	0.300	**0.544**	2.136	0.998
Seq 3	**0.070**	1.118	2.187	**0.427**	11.675	13.511

The bold entities in these tables indicate the best performance among all the methods.

noise in gyroscope measurements. Disabling these components introduces errors in roll and pitch estimation, and since the BlueROV2 moves primarily in the y direction, roll drift directly translates into z-axis drift. Furthermore, we observe similar performance to our previous work when disabling both the accelerometer residuals and bias correction. This validates that the performance improvements compared to our previous work stem from: 1) the newly introduced tightly coupled formulation, which enables bias correction, and 2) the newly added accelerometer residuals.

2) Effect of Sensor Calibration: To further investigate the impact of sensor calibration, an ablation study is performed in the UUV simulation environment [38]. Noises to the extrinsic parameters and the DVL transducer orientation are manually added. Then, using the proposed calibration method, these parameters are recalibrated automatically. Subsequently, we compared the SLAM performance using both the calibrated and uncalibrated parameters, on three sequences. The results, presented in Table V, demonstrate that the calibrated parameters significantly reduce translation and rotation errors compared to the uncalibrated ones. These findings underscore the importance and effectiveness of the proposed sensor calibration methods.

3) Challenging Cases: Thanks to the integration of DVL, our SLAM system can perform robustly in visually challenging environments over extended periods. During testing on the datasets, no failure cases were observed. However, for improved precision, we recommend that the SLAM system is initialized in visually favorable conditions and undergo sufficient motion to complete IMU initialization, which corrects bias and estimates the gravity direction. Insufficiently excited IMU initialization may lead to suboptimal precision, especially when visual conditions are poor and the system relies primarily on IMU and DVL for pose estimation.

Fig. 21(a) shows a challenging scenario where the SLAM system was initialized under poor visual conditions, resulting in a distorted sparse reconstruction. Three factors contribute to this distortion: 1) The image quality during initialization was poor, as illustrated in Fig. 21(b) and (c); 2) initialization occurred during an aggressive motion, causing severe motion blur in the images [see Fig. 21(c)], which led to IMU initialization failure and default zero bias; 3) aggressive motion, low image quality,

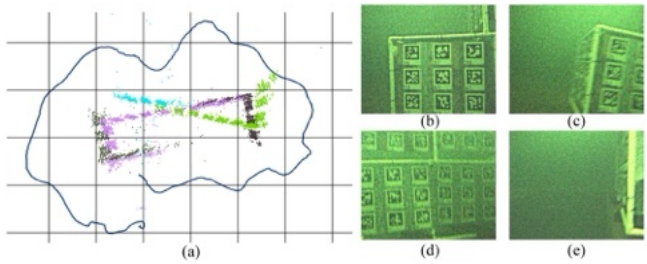


Fig. 21. Trajectory and sparse reconstruction under challenging visual conditions but without proper initialization.

and persistent motion blur throughout the sequence forced the SLAM system to mainly rely on IMU and DVL for pose estimation. Since the IMU bias was not properly initialized, drift, especially in orientation, was introduced. These factors together caused pose drift and misalignment in the sparse map, as shown in Fig. 21(a).

Therefore, we recommend initializing the SLAM system in visually optimal conditions with sufficient motion to ensure proper IMU initialization, which corrects bias and estimates gravity direction. Without proper IMU initialization, precision may degrade, particularly in challenging visual environments.

VIII. CONCLUSION

This article proposes a novel underwater acoustic-visual-inertial SLAM which tightly fuses DVL, camera, and IMU sensors in a graph optimization framework. DVL measurement is rigorously investigated, and its DVL preintegration model is derived in detail. The proposed SLAM method leverages DVL sensing to enhance its accuracy and robustness in challenging underwater environments. Meanwhile, novel techniques for DVL-camera-IMU extrinsic calibration and DVL transducer misalignment calibration are presented, further expedited with a rapid linear approximation method. These methods enable precise calibration of the extrinsic parameters among the DVL, camera, and IMU, as well as the orientation of the DVL transducer, even without the need for a dedicated underwater facility. The proposed methods are evaluated qualitatively and qualitatively in a tank and at an offshore site. The results demonstrate that the proposed method not only achieves superior accuracy and robustness compared to state-of-the-art underwater SLAM methods and visual-inertial SLAM methods, but also has the capability to enable real-world underwater applications. These applications include precise navigation and mapping for deep-sea exploration, accurate localization for the maintenance and inspection of under-water pipelines and cables, detailed monitoring of coral reef health and other marine ecosystems, and efficient search and recovery operations in underwater archaeology and wreck exploration. In future, we will investigate temporal calibration for acoustic-visual-inertial systems.

APPENDIX

A. DVL Translation Measurement Model Derivation

Given the definitions of ${}_{C_0}\mathbf{p}_{C_0D_j}$, ${}_{C_0}\mathbf{p}_{C_0D_i}$ and ${}_{C_0}\mathbf{p}_{D_iD_j}$ in (7), we can reformulate (7) as follows:

$$
\begin{aligned}
&\mathbf{p}_j - \mathbf{R}_j\mathbf{R}_{\mathtt{DC}\mathtt{D}}^T\mathbf{p}_{\mathtt{DC}} = \mathbf{p}_i - \mathbf{R}_i\mathbf{R}_{\mathtt{DC}\mathtt{D}}^T\mathbf{p}_{\mathtt{DC}} \\
&\quad + \sum_{k=i}^{j-1}\mathbf{R}_i\mathbf{R}_{\mathtt{IC}}^T\Delta\mathbf{R}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}}\,({}_{\mathtt{D}_i}\tilde{\mathbf{v}} - \boldsymbol{\eta}^D)\Delta t \\
&\overset{(a)}{\Longrightarrow} \mathbf{p}_j - \mathbf{R}_j\mathbf{R}_{\mathtt{DC}\mathtt{D}}^T\mathbf{p}_{\mathtt{DC}} - \mathbf{p}_i + \mathbf{R}_i\mathbf{R}_{\mathtt{DC}\mathtt{D}}^T\mathbf{p}_{\mathtt{DC}} \\
&= \sum_{k=i}^{j-1}\mathbf{R}_i\mathbf{R}_{\mathtt{IC}}^T\Delta\mathbf{R}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}}\,({}_{\mathtt{D}_i}\tilde{\mathbf{v}} - \boldsymbol{\eta}^D)\Delta t \\
&\overset{(b)}{\Longrightarrow} \mathbf{R}_{\mathtt{IC}}{\mathbf{R}_i}^T\mathbf{p}_j - \mathbf{R}_{\mathtt{IC}}{\mathbf{R}_i}^T\mathbf{R}_j\mathbf{R}_{\mathtt{DC}\mathtt{D}}^T\mathbf{p}_{\mathtt{DC}} - \mathbf{R}_{\mathtt{IC}}{\mathbf{R}_i}^T\mathbf{p}_i \\
&+ \mathbf{R}_{\mathtt{IC}}{\mathbf{R}_i}^T\mathbf{R}_i\mathbf{R}_{\mathtt{DC}\mathtt{D}}^T\mathbf{p}_{\mathtt{DC}} = \underbrace{\sum_{k=i}^{j-1}\Delta\mathbf{R}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}}\,({}_{\mathtt{D}_i}\tilde{\mathbf{v}} - \boldsymbol{\eta}^D)\Delta t}_{\Delta_{\mathtt{D}_i}\mathbf{p}_{\mathtt{D}_i\mathtt{D}_j}} \\
&\overset{(c)}{\Longrightarrow} \mathbf{R}_{\mathtt{ID}}\mathbf{R}_{\mathtt{DC}}{\mathbf{R}_i}^T\mathbf{p}_j - \mathbf{R}_{\mathtt{ID}}\mathbf{R}_{\mathtt{DC}}\mathbf{R}_i{}^T\mathbf{R}_j\mathbf{R}_{\mathtt{DC}\mathtt{D}}^T\mathbf{p}_{\mathtt{DC}} \\
&- \mathbf{R}_{\mathtt{ID}}\mathbf{R}_{\mathtt{DC}}{\mathbf{R}_i}^T\mathbf{p}_i + \mathbf{R}_{\mathtt{ID}\mathtt{D}}\mathbf{p}_{\mathtt{DC}} = \Delta_{\mathtt{D}_i}\mathbf{p}_{\mathtt{D}_i\mathtt{D}_j} \\
&\overset{(d)}{\Longrightarrow} \underbrace{\mathbf{R}_{\mathtt{ID}}\left({}_{\mathtt{D}}\mathbf{p}_{\mathtt{DC}} - \mathbf{R}_{\mathtt{DC}}\mathbf{R}_i^T\mathbf{R}_j\mathbf{R}_{\mathtt{DC}\mathtt{D}}^T\mathbf{p}_{\mathtt{DC}} + \mathbf{R}_{\mathtt{DC}}(\mathbf{R}_i^T\mathbf{p}_j - \mathbf{R}_i^T\mathbf{p}_i)\right)}_{h_{D_t}(\mathcal{X}_i,\mathcal{X}_j)} \\
&= \Delta_{\mathtt{D}_i}\mathbf{p}_{\mathtt{D}_i\mathtt{D}_j}.
\end{aligned}
$$

Step (a) moves $\mathbf{p}_i - \mathbf{R}_i\mathbf{R}_{\mathtt{DC}\mathtt{D}}^T\mathbf{p}_{\mathtt{DC}}$ to the left-hand side of the equation. Step (b) multiplies $\mathbf{R}_{\mathtt{IC}}{\mathbf{R}_i}^T$ at the both sides of the equation. Step (c) substitutes $\mathbf{R}_{\mathtt{IC}}$ with $\mathbf{R}_{\mathtt{ID}}\mathbf{R}_{\mathtt{DC}}$ as defined at (1). Step (d) reformulates the equation as the DVL translation measurement model defined at (8).

B. DVL Preintegration Derivation

The relative DVL translation incremental $\Delta_{\mathtt{D}_i}\mathbf{p}_{\mathtt{D}_i\mathtt{D}_j}$ can be further reformulated as

$$
\begin{aligned}
\Delta_{\mathtt{D}_i}\mathbf{p}_{\mathtt{D}_i\mathtt{D}_j} &= \sum_{k=i}^{j-1}\Delta\mathbf{R}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}}\,({}_{\mathtt{D}_i}\tilde{\mathbf{v}} - \boldsymbol{\eta}^D)\Delta t \\
&\overset{(a)}{\approx} \sum_{k=i}^{j-1}\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}(\mathbf{I} - \delta\hat{\boldsymbol{\phi}}_{\mathtt{I}_i\mathtt{I}_k}^{\wedge})\mathbf{R}_{\mathtt{ID}\,\mathtt{D}_i}\tilde{\mathbf{v}}\Delta t - \Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}}\,\boldsymbol{\eta}^D\Delta t \\
&\overset{(b)}{=} \sum_{k=i}^{j-1}\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}\,\mathtt{D}_i}\tilde{\mathbf{v}}\Delta t \\
&\quad - \Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\delta\hat{\boldsymbol{\phi}}_{\mathtt{I}_i\mathtt{I}_k}^{\wedge}\mathbf{R}_{\mathtt{ID}\,\mathtt{D}_i}\tilde{\mathbf{v}}\Delta t - \Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}}\,\boldsymbol{\eta}^D\Delta t \\
&\overset{(c)}{=} \underbrace{\sum_{k=i}^{j-1}\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}\,\mathtt{D}_i}\tilde{\mathbf{v}}\Delta t}_{\Delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j}} - \\
&\quad \underbrace{\sum_{k=i}^{j-1}[-\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}(\mathbf{R}_{\mathtt{ID}\,\mathtt{D}_i}\tilde{\mathbf{v}})^{\wedge}\cdot\delta\hat{\boldsymbol{\phi}}_{\mathtt{I}_i\mathtt{I}_k}\Delta t + \Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}}\,\boldsymbol{\eta}^D\Delta t]}_{\delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j}}.
\end{aligned}
$$

Step (a) first replaces $\Delta\mathbf{R}_{\mathtt{I}_i\mathtt{I}_k}$ with $\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\mathrm{Exp}(-\delta\hat{\boldsymbol{\phi}}_{\mathtt{I}_i\mathtt{I}_k})$ as in (14) then applies the first order approximation of $\mathrm{Exp}(-\delta\hat{\boldsymbol{\phi}}_{\mathtt{I}_i\mathtt{I}_k})$ as $\mathrm{Exp}(\boldsymbol{\theta}) \approx (\mathbf{I} + \boldsymbol{\theta}^{\wedge})$ [31]. Step (b) expands the brace and reformulates the equation. Step (c) applies the property of $\boldsymbol{u}^{\wedge}\boldsymbol{v} = -\boldsymbol{v}^{\wedge}\boldsymbol{u}$ and reformulates to $\Delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j}$ and $\delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j}$ defined in (8).

C. Jacobian of Extrinsic Calibration Approximation Iteration Derivation

Following the definition of derivative, we have

$$
\begin{aligned}
\frac{\partial\Delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j}}{\partial\boldsymbol{\phi}_{\mathtt{ID}}} &= \lim_{\Delta\boldsymbol{\phi}_{\mathtt{ID}}\to 0}\frac{\begin{bmatrix}\sum_{k=i}^{j-1}\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\mathrm{Exp}(\Delta\boldsymbol{\phi}_{\mathtt{ID}})\mathbf{R}_{\mathtt{ID}\,\mathtt{D}_i}\tilde{\mathbf{v}}\Delta t \\ -\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}\,\mathtt{D}_i}\tilde{\mathbf{v}}\Delta t\end{bmatrix}}{\Delta\boldsymbol{\phi}_{\mathtt{ID}}} \\
&\overset{(a)}{\approx} \lim_{\Delta\boldsymbol{\phi}_{\mathtt{ID}}\to 0}\frac{\begin{bmatrix}\sum_{k=i}^{j-1}\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}(\mathbf{I} + \Delta\boldsymbol{\phi}_{\mathtt{ID}}^{\wedge})\mathbf{R}_{\mathtt{ID}\,\mathtt{D}_i}\tilde{\mathbf{v}}\Delta t \\ -\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}\,\mathtt{D}_i}\tilde{\mathbf{v}}\Delta t\end{bmatrix}}{\Delta\boldsymbol{\phi}_{\mathtt{ID}}} \\
&\overset{(b)}{=} \lim_{\Delta\boldsymbol{\phi}_{\mathtt{ID}}\to 0}\frac{\sum_{k=i}^{j-1}\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\Delta\boldsymbol{\phi}_{\mathtt{ID}}^{\wedge}\mathbf{R}_{\mathtt{ID}\,\mathtt{D}_i}\tilde{\mathbf{v}}\Delta t}{\Delta\boldsymbol{\phi}_{\mathtt{ID}}} \\
&\overset{(c)}{=} \lim_{\Delta\boldsymbol{\phi}_{\mathtt{ID}}\to 0}\frac{\sum_{k=i}^{j-1}-\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}(\mathbf{R}_{\mathtt{ID}\,\mathtt{D}_i}\tilde{\mathbf{v}})^{\wedge}\Delta\boldsymbol{\phi}_{\mathtt{ID}}\Delta t}{\Delta\boldsymbol{\phi}_{\mathtt{ID}}} \\
&\overset{(d)}{=} \sum_{k=i}^{j-1}-\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}(\mathbf{R}_{\mathtt{ID}\,\mathtt{D}_i}\tilde{\mathbf{v}})^{\wedge}\Delta t.
\end{aligned}
$$

We assume left multiplication update is adopted during iteration and $\mathrm{Exp}(\Delta\boldsymbol{\phi}_{\mathtt{ID}})$ stand for the incremental update. Step (a) applies the first order approximation of $\mathrm{Exp}(\Delta\boldsymbol{\phi}_{\mathtt{ID}})$. Step (b) applies $\boldsymbol{u}^{\wedge}\boldsymbol{v} = -\boldsymbol{v}^{\wedge}\boldsymbol{u}$. Steps (c) and (d) reformulate the equation as the form in (25).

D. Jacobian of Misalignment Calibration Approximation Iteration Derivation

We have the below derivation for (26)

$$
\begin{aligned}
\frac{\partial\Delta_{\mathtt{D}_i}\bar{\mathbf{p}}_{\mathtt{D}_i\mathtt{D}_j}}{\partial_{\mathtt{D}_i}\tilde{\mathbf{v}}} &= \lim_{\Delta\tilde{\mathbf{v}}\to 0}\frac{\begin{bmatrix}\sum_{k=i}^{j-1}\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}}\,({}_{\mathtt{D}_i}\tilde{\mathbf{v}} + \Delta\tilde{\mathbf{v}})\Delta t \\ -\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}\,\mathtt{D}_i}\tilde{\mathbf{v}}\Delta t\end{bmatrix}}{\Delta\tilde{\mathbf{v}}} \\
&\overset{(a)}{=} \lim_{\Delta\tilde{\mathbf{v}}\to 0}\frac{\sum_{k=i}^{j-1}\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}}\,\Delta\tilde{\mathbf{v}}\Delta t}{\Delta\tilde{\mathbf{v}}} \\
&\overset{(b)}{=} \sum_{k=i}^{j-1}\Delta\hat{\mathbf{R}}_{\mathtt{I}_i\mathtt{I}_k}\mathbf{R}_{\mathtt{ID}}\,\Delta t.
\end{aligned}
$$

ACKNOWLEDGMENT

The authors would like to thank Dr Jonatan Scharff Willners, Joshua Roe, and Sean Katagiri for their support on the data collection and hardware.

REFERENCES

[1] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.

[2] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, Mar. 2018.

[3] D. Luo, Y. Zhuang, and S. Wang, "Hybrid sparse monocular visual odometry with online photometric calibration," *Int. J. Robot. Res.*, vol. 41, no. 11/12, pp. 993–1021, 2022.

[4] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, Aug. 2018.

[5] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.

[6] S. Rahman, A. Q. Li, and I. Rekleitis, "SVIn2: A multi-sensor fusion-based underwater SLAM system," *Int. J. Robot. Res.*, vol. 41, no. 11/12, pp. 1022–1042, 2022.

[7] D. Rudolph and T. A. Wilson, "Doppler velocity log theory and preliminary considerations for design and construction," in *Proc. 2012 IEEE Southeastcon*, 2012, pp. 1–7.

[8] Y. Huang et al., "Tightly-coupled visual-DVL fusion for accurate localization of underwater robots," in *Proc. 2023 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 8090–8095.

[9] L. Zhao, M. Zhou, and B. Loose, "Tightly-coupled visual-DVL-inertial odometry for robot-based ice-water boundary exploration," in *Proc. 2023 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 7127–7134.

[10] A. Thoms, G. Earle, N. Charron, and S. Narasimhan, "Tightly coupled, graph-based DVL/IMU fusion and decoupled mapping for SLAM-centric maritime infrastructure inspection," *IEEE J. Ocean. Eng.*, vol. 48, no. 3, pp. 663–676, Jul. 2023.

[11] E. Vargas et al., "Robust underwater visual SLAM fusing acoustic sensing," in *Proc. 2021 IEEE Int. Conf. Robot. Automat.*, 2021, pp. 2140–2146.

[12] S. Xu et al., "Underwater visual acoustic SLAM with extrinsic calibration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 7647–7652.

[13] K. H. Strobl and G. Hirzinger, "Optimal hand-eye calibration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2006, pp. 4647–4653.

[14] B. Xu and Y. Guo, "A novel DVL calibration method based on robust invariant extended Kalman filter," *IEEE Trans. Veh. Technol.*, vol. 71, no. 9, pp. 9422–9434, Sep. 2022.

[15] Q. Fu, Q. Shen, D. Wei, F. Wu, and G. Yan, "Multiposition alignment for rotational INS based on real-time estimation of inner lever arms," *IEEE Trans. Instrum. Meas.*, vol. 71, Jun. 2022, Art. no. 8503208.

[16] R. Eustice, O. Pizarro, and H. Singh, "Visually augmented navigation in an unstructured environment using a delayed state history," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2004, pp. 25–32.

[17] P. Ozog and R. M. Eustice, "Real-time SLAM with piecewise-planar surface models and sparse 3D point clouds," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2013, pp. 1042–1049.

[18] A. Kim and R. M. Eustice, "Real-time visual SLAM for autonomous underwater hull inspection using visual saliency," *IEEE Trans. Robot.*, vol. 29, no. 3, pp. 719–733, Jun. 2013.

[19] S. Hong and J. Kim, "Three-dimensional visual mapping of underwater ship hull surface using piecewise-planar SLAM," *Int. J. Control, Automat. Syst.*, vol. 18, pp. 564–574, 2020.

[20] E. Westman and M. Kaess, "Underwater AprilTag SLAM and calibration for high precision robot localization," Tech. Rep. CMU-RI-TR-18-43, 2018. [Online]. Available: https://www.ri.cmu.edu/publications/underwater-apriltag-slam-and-calibration-for-high-precision-robot-localization/

[21] S. Rahman, A. Q. Li, and I. Rekleitis, "Sonar visual inertial SLAM of underwater structures," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 5190–5196.

[22] S. Leutenegger, P. Furgale, V. Rabaud, M. Chli, K. Konolige, and R. Siegwart, "Keyframe-based visual-inertial SLAM using nonlinear optimization," in *Proc. Robot. Sci. Syst.*, 2013, doi: 10.15607/RSS.2013.IX.037.

[23] S. Rahman, A. Q. Li, and I. Rekleitis, "SVIn2: An underwater SLAM system using sonar, visual, inertial, and depth sensor," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* 2019, pp. 1861–1868.

[24] B. Joshi, H. Damron, S. Rahman, and I. Rekleitis, "SM/VIO: Robust underwater state estimation switching between model-based and visual inertial odometry," in *Proc. 2023 IEEE Int. Conf. Robot. Automat.*, 2023, pp. 5192–5199.

[25] C. Gu, Y. Cong, and G. Sun, "Environment driven underwater camera-IMU calibration for monocular visual-inertial SLAM," in *Proc. Int. Conf. Robot. Automat.*, 2019, pp. 2405–2411.

[26] D. Yang, B. He, M. Zhu, and J. Liu, "An extrinsic calibration method with closed-form solution for underwater opti-acoustic imaging system," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 9, pp. 6828–6842, Sep. 2020.

[27] D. Li, J. Xu, B. Zhu, and H. He, "A calibration method of DVL in integrated navigation system based on particle swarm optimization," *Measurement*, vol. 187, 2022, Art. no. 110325.

[28] L. Luo, Y. Huang, G. Wang, Y. Zhang, and L. Tang, "An on-line full-parameters calibration method for SINS/DVL integrated navigation system," *IEEE Sensors J.*, vol. 23, no. 24, pp. 30927–30939, Dec. 2023.

[29] P. Furgale, "Representing robot pose: The good, the bad, and the ugly," (n.d.). [Online]. Available: https://paulfurgale.info/news/2014/6/9/representing-robot-pose-the-good-the-bad-and-the-ugly

[30] N. Brokloff, "Matrix algorithm for doppler sonar navigation," in *Proc. OCEANS'94*, 1994, pp. III/378–III/383.

[31] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual–inertial odometry," *IEEE Trans. Robot.*, vol. 33, no. 1, pp. 1–21, Feb. 2017.

[32] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—A modern synthesis," in *Proc. Int. Workshop Vis. Algorithms Corfu*, Greece, 2000, pp. 298–372.

[33] S. Xu, J. S. Willners, Z. Hong, K. Zhang, Y. R. Petillot, and S. Wang, "Observability-aware active extrinsic calibration of multiple sensors," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 2091–2097.

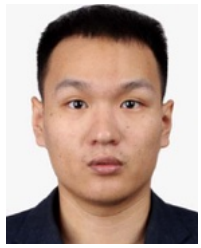
[34] T. Łuczyński, P. Łuczyński, L. Pehle, M. Wirsum, and A. Birk, "Model based design of a stereo vision system for intelligent deep-sea operations," *Measurement*, vol. 144, pp. 298–310, 2019.

[35] J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2016, pp. 4193–4198.

[36] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4104–4113.

[37] V. Usenko, N. Demmel, D. Schubert, J. Stückler, and D. Cremers, "Visual-inertial mapping with non-linear factor recovery," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 422–429, Apr. 2020.

[38] M. M. M. Manhães, S. A. Scherer, M. Voss, L. R. Douat, and T. Rauschenbach, "UUV simulator: A Gazebo-based package for underwater intervention and multi-robot simulation," in *Proc. OCEANS 2016 MTS/IEEE Monterey*, Sep. 2016, pp. 1–8, doi: 10.1109/OCEANS.2016.7761080.



Shida Xu received the M.Sc. degree in advanced computer science from the University of Sheffield, Sheffield, U.K., in 2020, and the Ph.D. degree in electrical engineering from Heriot-Watt University, Edinburgh, United Kingdom in 2025.

He is currently a Research Associate with the Sense Robotics Lab, Department of Electrical and Electronic Engineering, Imperial College London, London, U.K. His research interests include underwater SLAM, sensor fusion, and autonomous navigation.

Kaicheng Zhang received the B.Sc. degree in telecommunication engineering from Xidian University, Xi'an, China, in 2020, and the Ph.D. degree in electrical engineering from Heriot-Watt University, Edinburgh, U.K., in January 2025.

His research interests include point cloud processing and simultaneous localization and mapping.

Sen Wang received the M.Eng. degree in control theory and engineering from the Harbin Institute of Technology, Harbin, China, in 2011, and the Ph.D. degree in computer science from the University of Essex, Colchester, U.K., in 2015.

He is currently a Senior Lecturer (an Associate Professor) in robotics and autonomous systems with Imperial College London, London, U.K. His research interests include robot perception and autonomy using probabilistic and learning approaches, especially autonomous navigation, robotic vision, SLAM, and robot learning.

Dr. Wang was an Associate Editors for IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, IEEE ROBOTICS AND AUTOMATION LETTERS, ICRA, and IROS.