

SVIN+: Multi-Modal Framework for Robust Underwater State Estimation using 3D Sonar, Visual-Inertial, Water-Pressure, and DVL

Chinmay Burgul¹, Xiaoming Zhao², Yewei Huang², Michalis Chatzisprou¹,
Marios Xanthidis³, Ioannis Rekleitis¹, Alberto Quattrini Li²

Abstract—As expansion and scalability in underwater operations becomes more important and challenging, autonomous underwater vehicles (AUVs) become more relevant and the necessity for reliable autonomy more critical. A fundamental challenge in any autonomous underwater operation, from control to safe navigation and mapping, is state estimation; a notorious hard problem to address, especially in low visibility conditions. In this paper, we tackle this problem, by presenting SVIn+, a multi-sensor framework for robust state estimation of AUVs equipped with an IMU, cameras, a water-pressure sensor, a Doppler Velocity Log (DVL), and a compact 3D sonar. This framework extends an established open-source state-of-the-art visual-inertial SLAM, SVIn, by integrating water-pressure and DVL measurements into a tightly-coupled nonlinear optimization formulation resulting in improved estimates. To leverage the structural information of recently available 3D acoustic sonars while maintaining computational efficiency, we introduce a hierarchical architecture that incorporates estimates from a decoupled front-end via KISS-ICP, treated as relative pose factors within the factor graph. The proposed framework is validated through field deployments in a shipwreck and coral reef, demonstrating drift mitigation and global consistency. To support further research in underwater multi-modal fusion, the code and the datasets can be found on project website <https://cmburgul.github.io/projects/svin/> <https://dartmouthrobotics.github.io/svin-plus>

I. INTRODUCTION

Reliable autonomy in underwater robotics needs robust Simultaneous Localization and Mapping (SLAM) in perceptually degraded GNSS-denied environments [1], ranging from confined shipwrecks to dynamic coral reefs. These underwater environments are characterized by many perceptual challenges, including light attenuation and limited visibility.

Despite the plethora of SLAM approaches in the literature, robust state estimation in underwater domains remains an open problem. Traditional Visual-Inertial Odometry (VIO) is prone to catastrophic failure due to feature tracking loss in turbid waters or motion-blur artifacts during high-dynamic maneuvers [2]. Integrating water-pressure sensors (e.g., [3]) and Doppler Velocity Logs (DVLs) (e.g., [4]–[6]) can mitigate water-column depth and velocity drift; however, these

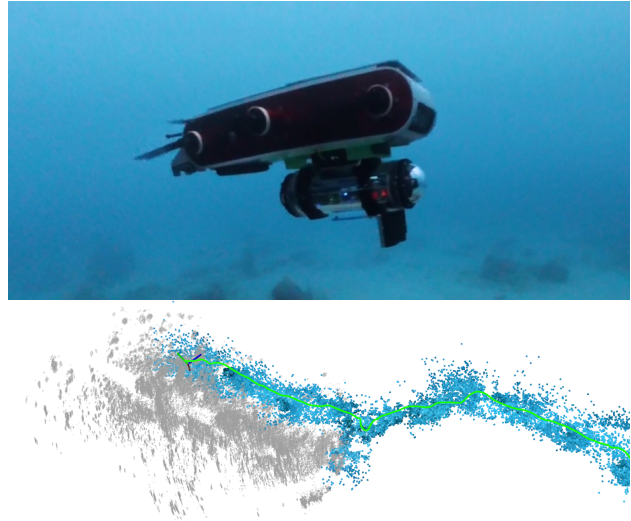


Fig. 1: The Aqua2 robot equipped with multiple sensors during field deployment (top) and proposed SVIn+ multi-sensor odometry pose and trajectory (green), stereo point-cloud (blue) and the current local 3D sonar local map point-cloud (grey) on the Folkestone dataset (bottom).

sensors provide no environmental context, leaving the system unable to perform the loop closures for global consistency. Furthermore, although 2D multibeam sonars can offer some situational awareness, they suffer from geometric ambiguity in complex 3D structures [1].

While a 3D acoustic sonar capable of generating volumetric point clouds has recently become available, and initial studies have begun to explore its potential, its integration into standard multi-sensor SLAM pipelines remains under-explored. Although compact 3D sonars conceptually resemble terrestrial LiDARs, their point clouds are significantly sparser, have a lower update rate (5 Hz), and are heavily corrupted by acoustic multipath noise. The few existing works have primarily addressed sonar-inertial configurations [7] or visual-inertial-acoustic fusion [8], [9], without considering the full suite of sensors commonly used in Autonomous Underwater Vehicles (AUV).

In this paper, we introduce a tightly-coupled multi-modal state estimation framework that achieves robust state estimation in challenging underwater environments. Our framework fuses inertial, camera vision, DVL, and water column depth from water-pressure sensor measurements into the factor graph. To provide structural consistency and mitigate trajectory drift, we integrate relative pose factors derived from a KISS-ICP front-end, which processes the point cloud from a recently released 3D acoustic sonar. This decoupled

¹ Chinmay Burgul, Michalis Chatzisprou, and Ioannis Rekleitis are with the Mechanical Engineering Department, University of Delaware, Newark, DE, USA. [cmburgul, michalis, yiannisr]@udel.edu

² Xiaoming Zhao, Yewei Huang and Alberto Quattrini Li are with Dartmouth College, Hanover, NH, USA {yewei.huang, alberto.quattrini.li}@dartmouth.edu

³ Marios Xanthidis is with the Aquaculture Robotics and Automation Group, SINTEF Ocean, marios.xanthidis@sintef.no

This work was supported by the Research Council of Norway (EchoNav: NO-359447), National Science Foundation (NSF 2545370, 1943205, 2024541, 2144624)

approach abstracts 3D sonar data into relative 6D pose factors, preserving factor graph sparsity and enabling real-time operation on embedded hardware.

In summary, our main contributions are the following:

- 1) Multi-modal state estimation framework, where vision, IMU, DVL, and water-pressure measurements are integrated in a tightly-coupled non-linear optimization for robust state estimation.
- 2) Hierarchical 3D sonar integration, with a decoupled front-end via KISS-ICP for computational efficiency.
- 3) Real-world validation in complex environments including shipwrecks and coral reefs, with the release of the associated code and datasets to the community.

II. RELATED WORK

Robust localization and mapping in underwater environments remains a major challenge due to the absence of GNSS signals forcing underwater vehicles to rely heavily on onboard sensing and dead reckoning, which leads to drift accumulation over time [10]. Furthermore, optical sensing is degraded by weak illumination, backscatter and turbidity [11], while acoustic sensing suffers from multipath effects [12] and limited resolution [13]. As a result, multi-sensor integration is essential for achieving reliable, long-duration underwater SLAM.

Early approaches fuse Inertial Measurement Unit (IMU), Doppler Velocity Logs (DVL), depth, and acoustic positioning measurements within an Extended Kalman Filter (EKF) framework to reduce and manage uncertainties from heterogeneous sensors [10]. While effective at managing short-term uncertainty, these systems lacked the dense environmental context needed to constrain long-term drift.

To achieve higher local accuracy, visual-inertial odometry (VIO) emerged as a dominant paradigm. Early vision-based underwater localization and mapping methods [14] are primarily offline and rely on structure-from-motion (SfM) and bundle adjustment. Subsequently, research shifted toward online, loosely coupled SLAM solutions. [15] introduce a loosely coupled SLAM framework for underwater shipwreck inspection using a bag-of-words place recognition module and factor graph optimization. [16] investigated the feasibility of performing underwater SLAM using ORB-SLAM [17]. [5], [18], [19] are loosely coupled systems that integrate DVL with a visual-inertial SLAM framework.

More recently, researchers have moved toward tightly coupled visual-inertial systems for higher localization accuracy and estimator consistency. [4] adopts a Multi-State Constraint Kalman Filter (MSCKF)-based SLAM framework for tightly coupled visual-DVL multi-sensor fusion. [20], [21] use cameras as the primary sensing modality to achieve low-cost and dense reconstruction. Similarly, [22] focuses on mitigating perception degradation through multi-sensor fusion. AquaSLAM [6] proposes an online SLAM framework that jointly estimates the camera-DVL extrinsic transformation within the state estimation process. However, vision-based approaches remain fundamentally limited by restricted field of view, light attenuation, backscatter, and

turbidity, and therefore typically operate only in relatively clear and well-illuminated environments.

To address these challenges, researchers have also adopted sonar as a primary perception modality. Sonar-based, vision-free SLAM methods [23]–[27] typically construct 2D maps under planar motion assumptions and are constrained by relatively low spatial resolution compared to optical sensing. There are also methods using a mechanical sonar [28] or dual sonars [29]–[31] for 3D mapping. SVIn [28] implements a tightly coupled hybrid framework based on OKVIS [32]. While the incorporation of acoustic sensing improves robustness, the relatively low spatial resolution and low update frequency of the sonar measurements limit their ability to support dense reconstruction.

With the recent development of compact 3D sonar sensors, lightweight AUV systems can acquire denser point clouds at relatively higher update rates. There emerges hybrid approaches that combine the 3D sonar with the visual-inertial SLAM pipeline [7]–[9]. [9] presents a loosely coupled fusion of visual and 3D sonar measurements for long-distance cave mapping. [7] adopts a loosely coupled framework that integrates 3D sonar with an inertial navigation system (INS) composed of a DVL, Attitude Heading Reference System (AHRS), and water-pressure sensor for pose estimation and mapping. In contrast, [8] proposes a tightly coupled visual-inertial-sonar SLAM framework without DVL integration, and the system is evaluated only in a relatively small-scale (6m × 7m) environment.

Overall, tightly coupled multi-sensor fusion state-estimation framework for robots equipped with 3D sonar remains an open problem. A reliable and robust long-term multi-sensor state estimation solution that enables dense and accurate reconstruction in underwater environments, particularly in murky water, remains a challenge. In this paper, we address this gap by proposing a partially tightly coupled visual, sonar, DVL, water pressure framework that jointly optimizes measurements from multiple modalities including sonar, camera, water-pressure and DVL to address this gap.

III. PRELIMINARIES

This section introduces the notation and coordinate frames adopted in the paper. The global reference (world) frame is denoted by G . The IMU frame is denoted by I , the camera frame by C , the sonar frame by L , the Doppler Velocity Log (DVL) frame by V , and the water-pressure sensor frame by D . All sensor frames are rigidly attached to the robot body, and their extrinsic transformations are assumed to be known from calibration. The state vector consists of:

$$\mathbf{x}_k = [\mathbf{x}_{I_k}^\top \quad \mathbf{x}_{H_k}^\top \quad \mathbf{x}_{Y_k}^\top]^\top \quad (1)$$

$$\mathbf{x}_{I_k} = [{}^G\bar{\mathbf{q}}^\top \quad {}^G\mathbf{p}_{I_k}^\top \quad {}^G\mathbf{v}_{I_k}^\top \quad \mathbf{b}_g^\top \quad \mathbf{b}_a^\top]^\top \quad (2)$$

$$\mathbf{x}_{H_k} = [\mathbf{x}_{I_{k-1}} \quad \cdots \quad \mathbf{x}_{I_{k-n}}]^\top \quad (3)$$

$$\mathbf{x}_{Y_k} = [{}^G_{I_{n-1}}\bar{\mathbf{q}}^\top \quad {}^G\mathbf{p}_{I_{n-1}}^\top \quad \cdots \quad {}^G_{I_{n-m}}\bar{\mathbf{q}}^\top \quad {}^G\mathbf{p}_{I_{n-m}}^\top]^\top \quad (4)$$

where $\mathbf{x}_{I_k}$ is the current body/IMU state, $\mathbf{x}_{H_k}$ are the n historical non-keyframe states in the local optimization sliding

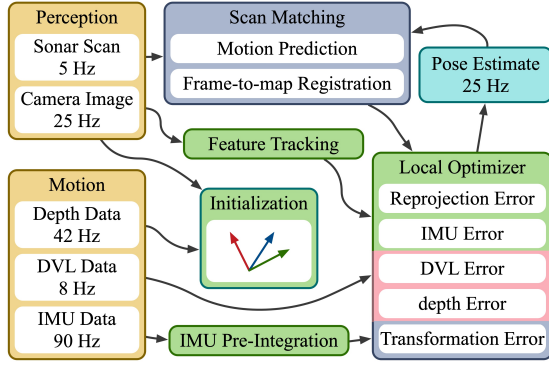


Fig. 2: System pipeline of the proposed framework. Classical camera modules are highlighted in green, the sonar processing pipeline in gray, and marine-specific components in pink. This paper focuses on the integration of the local odometry optimization module.

window, $\mathbf{x}_{Y_k}$ are the m keyframe states in the global frame $\{G\}$ for global pose-graph optimization¹. Let $i \in \{C, L, V, D\}$ denote the sensor modality. The measurement generated by sensor i at time t is denoted by $\mathbf{z}_{i,t}$. When the time index is clear from context, it is abbreviated as $\mathbf{z}_i$ for simplicity.

IV. MULTI SENSOR ODOMETRY

Our main contribution is the design and implementation of a local optimization module that provides high-frequency and reliable state estimation through the tightly coupled fusion of visual, inertial, water-pressure, DVL, and 3D sonar odometry measurements within a nonlinear optimization framework (Fig. 2). The local optimizer treats the state estimation problem as a nonlinear least-squares optimization over a sliding window of size K . Given $\mathbf{X} = \{\mathbf{x}_t, \dots, \mathbf{x}_{t+K}\}$, we formulate the optimization problem as:

$$\mathbf{X} = \arg \min_{\mathbf{X}} J(\mathbf{X}). \quad (5)$$

The cost function $J(\mathbf{X})$ is defined as the sum of squared residuals corresponding to multiple sensing modalities, including the camera reprojection residual $\mathbf{e}_r$, inertial residual $\mathbf{e}_s$, DVL velocity residual $\mathbf{e}_v$, water-depth residual $\mathbf{e}_d$, and 3D sonar residual $\mathbf{e}_t$:

$$J(\mathbf{x}) = \frac{1}{2} \sum_i^n \sum_{k \in \mathcal{K}} \sum_{j \in \mathcal{J}(i,k)} \mathbf{e}_r^{i,j,kT} \boldsymbol{\Omega}_r^k \mathbf{e}_r^{i,j,k} + \frac{1}{2} \sum_{k=1}^{K-1} \mathbf{e}_s^k \boldsymbol{\Omega}_s^k \mathbf{e}_s^k + \frac{1}{2} \sum_k^{K-1} \mathbf{e}_v^k \boldsymbol{\Omega}_v^k \mathbf{e}_v^k + \frac{1}{2} \sum_k^{K-1} \boldsymbol{\Omega}_d^k \left\| e_d^k \right\|^2 + \frac{1}{2} \sum_k^{K-1} \mathbf{e}_L^k \boldsymbol{\Omega}_L^k \mathbf{e}_L^k \quad (6)$$

Here, k denotes the frame index within the set of frames $\mathcal{K}$ contained in the sliding window, i represents the camera index (allowing straightforward extension to an n camera system), and $j \in \mathcal{J}$ indexes the visual landmarks observed in the k -th frame. For a detailed derivation of the camera reprojection residual within a sliding-window nonlinear optimization framework, readers are referred to OKVIS [32]. The

¹ ${}^G \tilde{q}$ is the Hamilton unit quaternion corresponding to the rotation ${}^G \mathbf{R}$ from $\{I\}$ to $\{G\}$, ${}^G \mathbf{p}_k$ and ${}^G \mathbf{v}_k$ are the position and velocity of $\{I\}$ in $\{G\}$, and $\mathbf{b}_g$ and $\mathbf{b}_a$ are the biases of the gyroscope and accelerometer, respectively.

inertial residual $\mathbf{e}_s$ is defined based on the IMU preintegration model, as described in [33].

The matrices $\boldsymbol{\Omega}_r^k, \boldsymbol{\Omega}_s^k, \boldsymbol{\Omega}_v^k$, and $\boldsymbol{\Omega}_t^k$ are the information matrices (inverse covariance matrices) for the visual, IMU, DVL, and 3D Sonar measurements at frame k , respectively. Note that while the visual, inertial, DVL, and sonar residuals are vectors, the depth error e_d is a scalar, and is therefore weighted by the scalar $\boldsymbol{\Omega}_d^k$. The remainder of this section introduces the derivation of the DVL velocity residual, the depth residual, and the 3D sonar residual in detail.

A. Water-Pressure Sensor Fusion

The water-pressure sensor measures the water-pressure, which can be converted into water column depth z_d relative to the water surface. The depth residual e_d^k for a measurement $z_{d,k}$ at time step k can be formulated similarly to [4] as:

$$e_d^k = D_0 - \mathbf{e}_3^\top ({}^I_k \mathbf{R}^\top {}^I \mathbf{p}_D + {}^G \mathbf{p}_k) \quad (7)$$

where $\mathbf{e}_3^\top = [0 \ 0 \ 1]$ extracts the Z-axis position in global frame, ${}^I \mathbf{p}_D$ is the translation from IMU to Water-Pressure Sensor. D_0 is the scalar depth of global frame $\{G\}$, computed using the initial depth measurement $z_{d,0}$ as $D_0 = z_{d,0} - \mathbf{e}_3^\top ({}^G \mathbf{p}_0 + {}^G \mathbf{R}^\top {}^I_0 \mathbf{p}_D)$. Jacobians with respect to state can be defined as $\frac{\partial e_d}{\partial {}^G \mathbf{p}_I} = -\mathbf{e}_3^\top$, $\frac{\partial e_d}{\partial {}^G \mathbf{R}} = \mathbf{e}_3^\top {}^I \mathbf{R} [{}^I \mathbf{p}_D \times]$. Where $[\cdot \times]$ denotes the skew-symmetric matrix operator mapping $\mathbf{v} \in \mathbb{R}^3$ to its corresponding skew-symmetric matrix in Lie algebra $\mathfrak{so}(3)$. Furthermore, given the short duration of the experiments compared to the tidal timescale, we assume a constant water surface level, neglecting the effects of tidal cycles and variations in ambient pressure.

B. Doppler Velocity Log (DVL) Fusion

The DVL provides linear velocity measurements $\mathbf{z}_v$ in the sensor frame. We relate these velocity to the system state velocity, following the approach in [4], [6] to define the DVL residual $\mathbf{e}_{v,k}$ at time step k as:

$$\mathbf{e}_{v,k} = \mathbf{z}_{v,k} - {}^V \mathbf{R}^I {}^I_k \mathbf{R}^G \mathbf{v}_{I_k} - [{}^I \omega_{k-1:k} \times] {}^I \mathbf{p}_V \quad (8)$$

where ${}^V \mathbf{R}$ and ${}^I \mathbf{p}_V$ are the extrinsic rotation from the IMU frame to DVL frame and position of DVL in IMU frame respectively. $\omega_{k-1:k}$ is the gyroscope angular velocity from $k-1$ to k time steps. The corresponding Jacobians with respect to state variables can be analytically derived as $\frac{\partial \mathbf{e}_v}{\partial {}^I \mathbf{R}} = -{}^V \mathbf{R} [{}^I_k \mathbf{R}^G \mathbf{v} \times]$, $\frac{\partial \mathbf{e}_v}{\partial {}^G \mathbf{v}} = -{}^V \mathbf{R} {}^I \mathbf{R}$. We construct the information matrix $\boldsymbol{\Omega}_v$ using the velocity covariance $\boldsymbol{\Sigma}_v$ provided by the DVL sensor, allowing us to accurately weigh the measurement within the optimization framework.

C. 3D Sonar Fusion

To process 3D Sonar measurements, we perform scan registration using KISS-ICP [34]. Unlike the original implementation, which assumes a constant-velocity kinematic model for motion prediction, we use the pose estimate of previous step from SVIn+ as motion prior to register the current scan against a local map, yielding the sonar odometry ${}^W \mathbf{T}_L \in SE(3)$. Here, W denotes the sonar odometry world

frame, which is initialized to align with the first sonar frame S_0 . This frame is distinct from the SVIn+ world frame G .

To account for the time offsets between the asynchronous sonar odometry and SVIn+ states, we perform $SE(3)$ interpolation on the bounding sonar odometry to derive the poses at the SVIn+ time steps $k-1$ and k . By extracting these time-aligned odometry ${}^W_{L_{k-1}}\mathbf{T}$ and ${}^W_{L_k}\mathbf{T}$, we compute the relative transformation ${}^{L_{k-1}}_{L_k}\mathbf{T}_m$. Then, we construct the relative pose measurement defined by the rotation vector ${}^{L_{k-1}}_{L_k}\boldsymbol{\theta}_{L_k} \in \mathbb{R}^3$ and translation ${}^{L_{k-1}}_{L_k}\mathbf{p}_{L_k} \in \mathbb{R}^3$:

$$\begin{aligned} {}^{L_{k-1}}_{L_k}\mathbf{T}_m &= {}^W_{L_{k-1}}\mathbf{T}^{-1} {}^W_{L_k}\mathbf{T} \\ &= \begin{bmatrix} {}^W_{L_{k-1}}\mathbf{R}^\top {}^W_{L_k}\mathbf{R} & {}^W_{L_{k-1}}\mathbf{R}^\top ({}^W\mathbf{p}_{L_k} - {}^W\mathbf{p}_{L_{k-1}}) \\ \mathbf{0}^\top & 1 \end{bmatrix} \end{aligned} \quad (9)$$

$${}^{L_{k-1}}_{L_k}\mathbf{p}_{L_k} = {}^W_{L_{k-1}}\mathbf{R}^\top ({}^W\mathbf{p}_{L_k} - {}^W\mathbf{p}_{L_{k-1}}) \quad (10)$$

$$\boldsymbol{\theta}_{L_{k-1:k}} = \log({}^W_{L_{k-1}}\mathbf{R}^\top {}^W_{L_k}\mathbf{R})^\vee \quad (11)$$

where $\log(\cdot)^\vee$ extracts the rotation vector $\boldsymbol{\theta} \in \mathbb{R}^3$ directly from the $SO(3)$ manifold. These components form the relative sonar measurement vector $\mathbf{z}_{L,k-1:k}$, subject to additive noise $\mathbf{n}_L \in \mathbb{R}^6$:

$$\mathbf{z}_{L,k-1:k} = \begin{bmatrix} {}^{L_{k-1}}_{L_k}\mathbf{p}_{L_k} \\ {}^{L_{k-1}}_{L_k}\boldsymbol{\theta}_{L_k} \end{bmatrix} + \mathbf{n}_L \quad (12)$$

To formulate the residual, we predict this relative transformation, denoted as ${}^{L_{k-1}}_{L_k}\hat{\mathbf{T}}$, using corresponding SVIn+ pose estimates ${}^G_{I_{k-1}}\mathbf{T}$ and ${}^G_{I_k}\mathbf{T}$ along with the IMU-Sonar extrinsic transformation:

$${}^{L_{k-1}}_{L_k}\hat{\mathbf{T}} = {}^L_{I_k}\mathbf{T} {}^G_{I_{k-1}}\mathbf{T}^{-1} {}^G_{I_k}\mathbf{T}^\top {}^L_{I_k}\mathbf{T} \quad (13)$$

$${}^{L_{k-1}}_{L_k}\hat{\mathbf{p}}_{L_k} = {}^L_{I_k}\mathbf{R}^\top [{}^G_{I_{k-1}}\mathbf{R}^\top ({}^G_{I_k}\mathbf{R}^\top \mathbf{p}_{L_k} + {}^G\mathbf{p}_{I_k} - {}^G\mathbf{p}_{I_{k-1}}) - {}^I\mathbf{p}_L] \quad (14)$$

$${}^{L_{k-1}}_{L_k}\hat{\boldsymbol{\theta}}_{L_k} = \log({}^{L_{k-1}}_{L_k}\hat{\mathbf{T}})^\vee = \log({}^L_{I_k}\mathbf{R}^\top {}^G_{I_{k-1}}\mathbf{R}^\top {}^G_{I_k}\mathbf{R}^\top {}^L_{I_k}\mathbf{R})^\vee \quad (15)$$

Finally, the sonar relative pose residual $\mathbf{e}_{L,k-1,k}$ is computed as the difference between the interpolated measurement and the state prediction:

$$\mathbf{e}_{L,k-1,k} = \begin{bmatrix} {}^{L_{k-1}}_{L_k}\tilde{\mathbf{p}}_{L_k} \\ {}^{L_{k-1}}_{L_k}\tilde{\boldsymbol{\theta}}_{L_k} \end{bmatrix} = \begin{bmatrix} {}^{L_{k-1}}_{L_k}\mathbf{p}_{L_k} - {}^{L_{k-1}}_{L_k}\hat{\mathbf{p}}_{L_k} \\ \log({}^{L_{k-1}}_{L_k}\mathbf{R}^\top {}^L_{I_k}\mathbf{R})^\vee \end{bmatrix} \quad (16)$$

The corresponding Jacobians with respect to state variables can be analytically derived as $\frac{\partial {}^{L_{k-1}}_{L_k}\tilde{\mathbf{p}}_{L_k}}{\partial {}^G\mathbf{p}_{I_k}} = {}^L_{I_k}\mathbf{R}^\top {}^G_{I_{k-1}}\mathbf{R}^\top$, $\frac{\partial {}^{L_{k-1}}_{L_k}\tilde{\boldsymbol{\theta}}_{L_k}}{\partial {}^G_{I_k}\mathbf{R}} = -{}^L_{I_k}\mathbf{R}^\top {}^G_{I_{k-1}}\mathbf{R}^\top {}^G_{I_k}\mathbf{R}^\top [{}^I\mathbf{p}_L]$.

V. EXPERIMENTAL VALIDATION

To evaluate the performance and robustness of the proposed SVIn+, we collected a series of datasets in diverse and challenging underwater environments off the coast of Barbados. Our experiments aim to demonstrate that the integration of 3D sonar constraints into a tightly-coupled visual-inertial-DVL estimator mitigates trajectory drift by leveraging acoustic structural context. In the following subsections, we detail our hardware configuration and extrinsic calibration procedures, describe the characteristics of our shipwreck and coral reef datasets, and provide a quantitative comparison against state-of-the-art baselines.

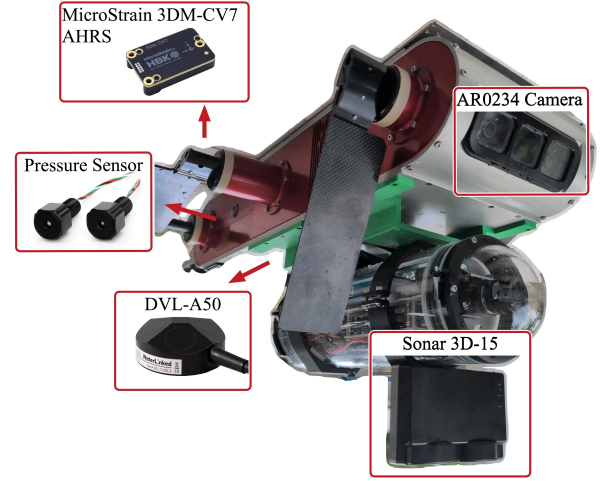


Fig. 3: A detailed view of the onboard sensor configuration.

A. Experimental Setup

We utilize the Aqua2 AUV from independent robotics equipped with a MicroStrain 3DM-CV7 AHRS, two AR0234 global shutter cameras in stereo configuration with hardware synchronization, a Water Linked DVL-A50 Doppler Velocity Log, a Bar30 pressure sensor, and a Water Linked 3D-15 sonar (Fig. 3).

Accurately knowing the extrinsic transformation across this multi-sensor suite is critical for robust state estimation. The transformation between Camera $\{C\}$ and IMU $\{I\}$ reference frames is calibrated using Kalibr [35]. The rigid extrinsic transformations from IMU to DVL $\{V\}$ and Water-Pressure Sensor $\{D\}$ frames are provided by the manufacturer of the Aqua2 robot. Finally, the relative rotation and translation between 3D Sonar $\{L\}$ and Camera $\{C\}$ are determined following the approach in [9]. We solve the Perspective-n-Point (PnP) problem using 3D-2D correspondences extracted from four cinder block targets placed underwater and visible in both the sonar point clouds and visual images. All experiments are conducted on a workstation running Ubuntu 24.04, equipped with an Intel Core Ultra 7 155H CPU, 64 GB of RAM, and an NVIDIA RTX 5000 Ada Generation GPU. The pipeline was also successfully evaluated in real-time on a low-power laptop (Intel Core i7-10510U CPU, 16 GB of RAM, NVIDIA GTX 1050), demonstrating its suitability for deployment on resource-constrained AUV hardware.

B. Dataset Description

We collected three datasets in two distinct locations off the coast of Barbados using the Aqua2 platform. To the best of our knowledge, there are currently no publicly available underwater datasets that provide a synchronized suite comprising stereo vision, DVL, and 3D sonar data in such complex environments. These sequences encompass a range of perceptual challenges, from confined structural interiors to expansive open-water reefs:

- *BajanQueen-01* (Shipwreck Interior): This sequence consists of a moderate-distance trajectory ~ 75 meters over about 8 minutes, within the confined, structured

environment of a shipwreck. Both missions start and end at a fixed ArUco marker, visited a total of 3 times, providing a reliable reference to evaluate start-to-end translation and rotation drift.

- *BajanQueen-02* (Shipwreck): A longer, ~ 600 -meter trajectory over around 55 minutes, which covers the entire shipwreck, testing robustness against significant depth variations and sparse visual features.
- *Folkestone-01* (Shallow Reef): conducted in a natural unstructured environment, the trajectory spans ~ 280 meters over 1252 seconds, challenging the vision front-end with dynamic lighting despite a relatively flat trajectory.

C. Evaluation of Return-to-Origin Trajectory

We evaluate the accumulated trajectory drift using *BajanQueen-01* sequence. Since global ground truth (e.g., GPS or long-baseline acoustic positioning) is unavailable in these unstructured underwater environments, we rely on a “return-to-origin” metric. We compare our proposed *SVIn+* framework against two baselines: *SVIn2* visual-inertial framework and KISS-ICP as (*Sonar Odom*). Global loop closures were disabled to isolate the performance of the underlying continuous odometry, ensuring that comparisons were not confounded by global correction mechanisms. In addition to the online baselines, we utilize an offline Structure-from-Motion (SfM) pipeline [36]. We performed an extensive parameter tuning to get the best results from each method for fair comparison.

Fig. 4 shows the trajectories over a complete start-to-end loop, with the initial and final positions denoted by a circle and a star, respectively. While *Sonar Odom* provides geometric context, it exhibited significantly higher drift due to the low-frequency nature of acoustic updates and noise (e.g., coming from multipath effects) in confined shipwreck compartments. The offline SfM reference (Colmap) shows noticeable discontinuities where visual tracking failed due to environmental conditions.

SVIn+ overall maintains a consistent path throughout the dataset, where the start and end are close to each other, compared to *SVIn2*. As shown qualitatively in Fig. 9, the *SVIn+* trajectory results in the ArUco marker board being closer compared to *SVIn2* baselines. Quantifying the accumulated drift via the ArUco marker board reveals that the proposed *SVIn+* achieves a return-to-origin error of just 0.52 m, representing a substantial improvement over the 0.81 m error yielded by *SVIn2*. This lower drift is further evidenced by the resulting point cloud from *SVIn+*, which clearly shows three corridors on the left side of the shipwreck and the room on the right side (Fig. 5). These results highlight the effectiveness of the proposed multi-sensor fusion in bounding global drift and providing a coherent structural reconstruction.

D. Evaluation on Long-Duration Field Sequences

To evaluate the long-term stability of *SVIn+*, we further validate the system on two extended sequences characterized

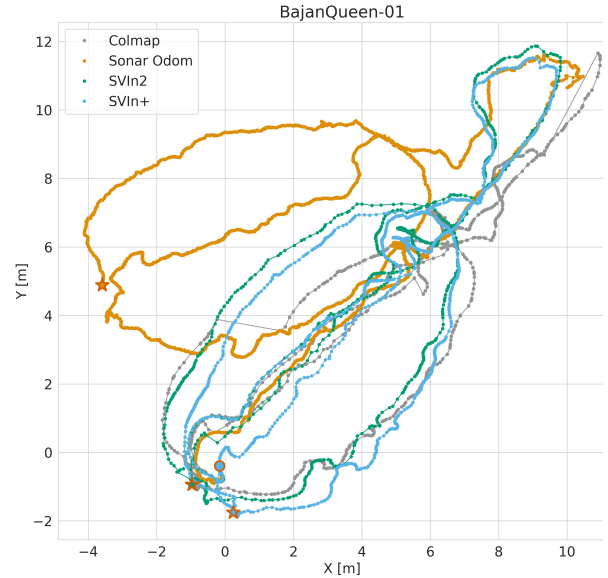


Fig. 4: Trajectory comparison between the proposed *SVIn+* and the baseline *SVIn2*, *Sonar Odom* on the *BajanQueen-01* dataset. Evaluation of the start-to-end loop demonstrates the reduced accumulated drift achieved by the proposed multi-sensor fusion.

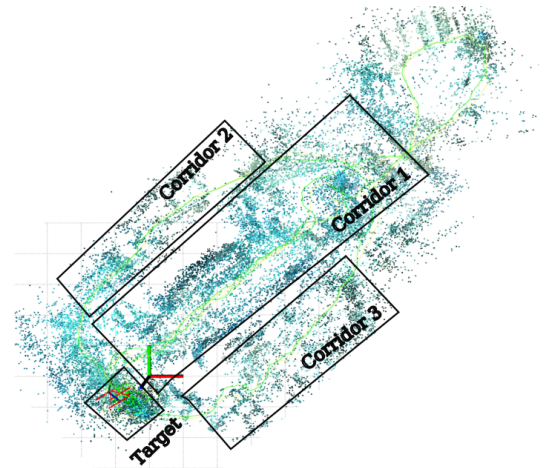


Fig. 5: Visualization of the point-cloud from the *SVIn+* on the *BajanQueen-01* dataset.

by significantly greater distances and durations.

Figure 6 illustrates the *SVIn+* performance on the *BajanQueen-02* shipwreck sequence, while Figure 7 demonstrates its capabilities on the *Folkestone-01* coral reef sequence. Throughout these deployments, the estimator maintains a highly consistent global depth profile. The shipwreck side-view (Fig. 6 bottom) demonstrates the ability to capture the shipwreck shape. The y-z profile over the coral reef (Fig. 7, bottom) shows that the trajectory accurately captures the mission profile: the Aqua2 initially descends, executes a wide exploratory loop, and gradually ascends as it returns toward the shallows. Qualitatively, the visual coherence of the reconstructed shipwreck and reef features over their respective 215 m and 285 m transits demonstrates that the fusion of 3D sonar, water-pressure and DVL measurements provides robust constraints. This integration maintains a stable state estimate even in the presence of dynamic lighting

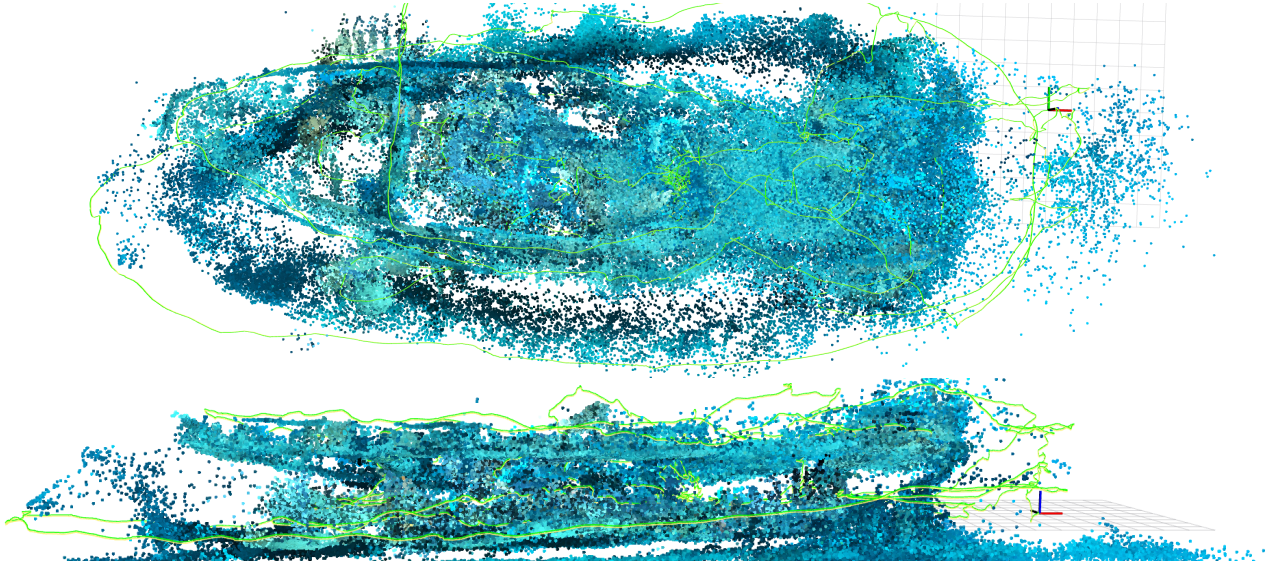


Fig. 6: Visualization of the estimated trajectory and corresponding 3D point-cloud generated by *SVIn+* Odometry on the *BajanQueen-02* sequence. *Top*: Top-down perspective of the mapped environment. *Bottom*: Side view illustrating the depth profile as the Aqua2 robot navigates through the interior and around the exterior of the shipwreck structure.

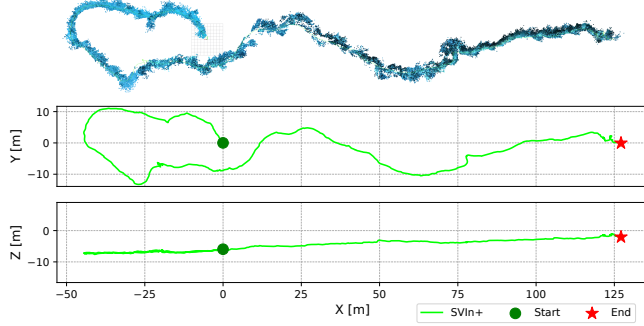


Fig. 7: Top-down visualization of the trajectory with point-cloud from *SVIn+* on the *Folkestone* dataset (circle and star indicate start and end, respectively).

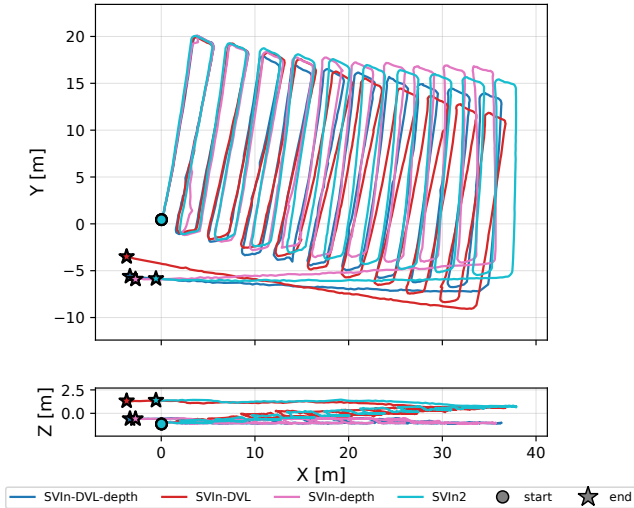


Fig. 8: Comparison of estimated boustrophedon trajectories using the baseline *SVIn2* framework against the proposed multi-sensor integrations (*SVIn*+Depth, *SVIn*+DVL, and *SVIn*+Depth+DVL).

and repetitive textural characteristics typical of shallow reefs.

TABLE I: Comparison of tag revisitation error, quantified by the average (and std) of the spatial deviations between reprojections of revisits of a fixed ArUco marker (as visualized in Fig. 9).

Configuration	Reprojection Distance (m)
SVIn2	0.53 (0.11)
SVIn2 (loop-closure)	0.22 (0.11)
SVIn-depth	0.29 (0.11)
SVIn-dvl	0.52 (0.23)
SVIn-Sonar	0.29 (0.15)
SVIn+	0.21 (0.08)

E. Ablation study

To evaluate the individual contribution of each sensor modality, we performed an incremental ablation study. First, we performed an evaluation with the Aqua2 going over a shallow area at fixed depth and following a lawnmower pattern for qualitative assessment. Fig. 8 compares the estimated trajectories of baseline *SVIn2* with DVL and Depth from water pressure sensor integration. The incorporation of DVL measurements effectively constrains drift in the $x-y$ plane, while depth integration bounds vertical z -axis error. The tightly-coupled fusion of both the sensors (*SVIn*+Depth+DVL) achieves the lowest overall positional drift from start to end.

To systematically isolate the contribution of each sensor modality, we evaluate five degraded configurations against our full system on the *BajanQueen-01* dataset (Table I, and Figs. 9, 10), using the fixed marker in the environment. Starting from the visual-inertial baseline (*SVIn2*, mean error: 0.53 m), the integration of the depth sensor alone (*SVIn*-depth) constrains vertical drift, reducing the mean deviation to 0.29 m (45.3% improvement). Fusing only the DVL (*SVIn*-dvl) provides bounded velocity constraints, yielding a mean error of 0.52 m with higher lateral consistency. The joint tightly-coupled integration of all modalities in *SVIn+* yields the most accurate and geometrically consistent estimate (0.21 m, a 60.4% overall error reduction). Notably, the open-loop state estimation of *SVIn+* (0.21 m

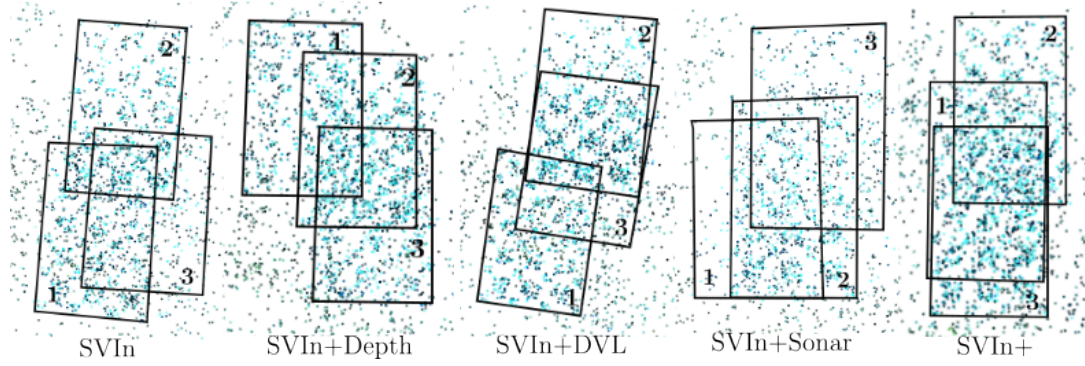


Fig. 9: Visualization of accumulated drift utilizing an ArUco marker board. The return-to-origin error is quantified by evaluating the spatial deviation between the initial marker projection (1) and the projection recorded during the third revisit (3). Note that apparent variations in marker size are solely a result of perspective scaling due to varying distances from the camera.

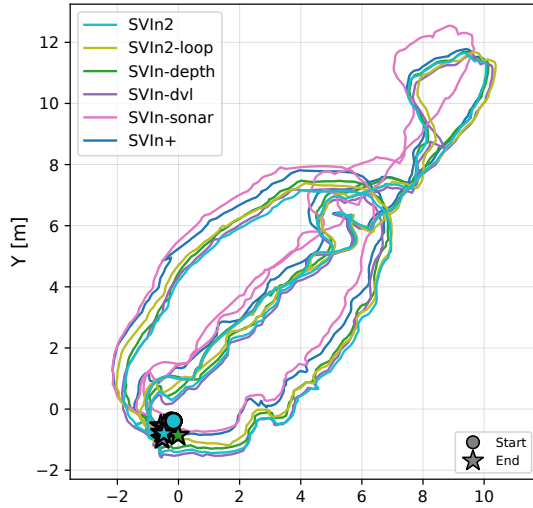


Fig. 10: Trajectory comparison of baseline SVIn with adding different sensor modalities.

deviation) outperforms the baseline visual-inertial SLAM running with active loop closures (SVIn2-loop, 0.22 m). This highlights that multi-modal fusion can match the global accuracy of loop-closed systems without the risk of false loop acceptance. As shown in Figure 9, the reprojections from the three successive passes are concentrated within a noticeably tighter error radius, demonstrating enhanced geometric consistency.

Furthermore, Fig. 11 illustrates the dense map constructed from 3D Sonar scans compared to the 3D points extracted from visual measurements. This highlights the advantage of 3D Sonar: beyond enhancing localization, it generates rich environmental maps essential for subsequent path planning and execution of autonomous tasks.

VI. LESSONS LEARNED AND FUTURE WORK

The stability and drift reduction achieved by SVIn+ establish a robust foundation for several future research directions: *Robustness to sensor degeneracy*: While the current factor graph manages uncertainty, extreme conditions can render specific modalities completely unreliable (e.g., sparse 3D sonar registration failures, as depicted in Figure 12). We plan to develop an adaptive switching mechanism for proactive failure detection and system resilience.

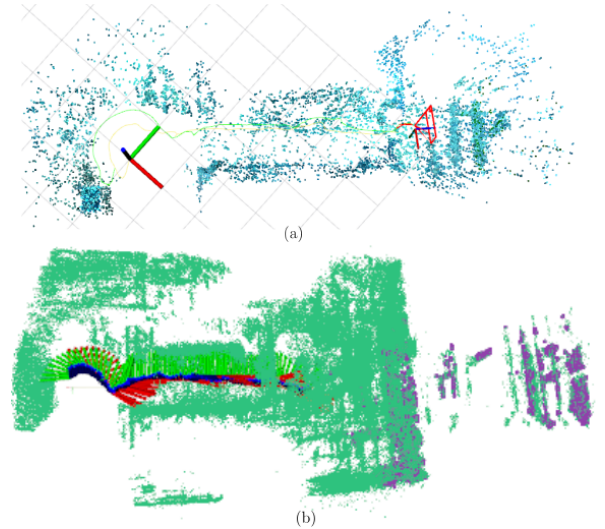


Fig. 11: Comparison of (a) the 3D feature point cloud generated by the visual sensor and (b) the local map point cloud (green) constructed by registering 3D sonar point scans (purple) from the BajanQueen-01 dataset.

Global mapping and loop closure: To transition from continuous odometry to a complete SLAM system for long-endurance missions, we will integrate multi-modal global place recognition. Leveraging both visual and acoustic signatures will ensure reliable re-localization and correct accumulated error.

Online spatial calibration: Because unmodeled dynamic alignment variations (such as DVL misalignment) introduce systematic errors that offline calibration cannot fully resolve, we plan to implement an online extrinsic calibration backend to dynamically maximize trajectory accuracy.

VII. CONCLUSIONS

In this paper, we presented SVIn+, a robust multi-modal framework for underwater state estimation that integrates measurements from 3D acoustic sonar, visual-inertial sensors, a water-pressure sensor and a Doppler Velocity Log (DVL). To overcome the perceptual degradation common in underwater environments, such as light attenuation and turbidity, our tightly-coupled nonlinear optimization framework jointly optimizes these heterogeneous measurements within a local sliding window. To efficiently leverage the structural context of 3D sonars without compromising real-

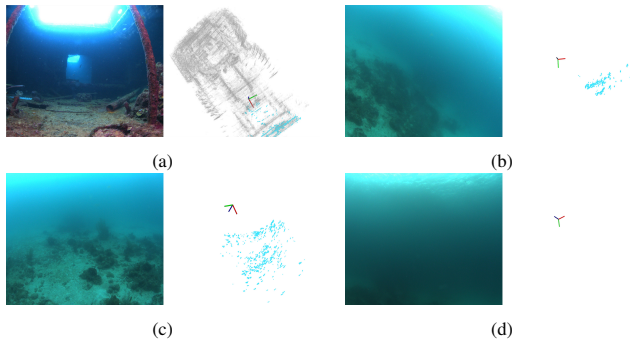


Fig. 12: Comparison of visual images and sonar point cloud scan (blue) for a successful registration (a) and representative failure cases, where points are very sparse or non-present (b) (c) (d).

time performance, we introduced a hierarchical architecture that incorporates 6D relative pose factors from a decoupled KISS-ICP front-end. Extensive field experiments conducted with Aqua2 platform in complex environments, including a shipwreck and a coral reef, validate the effectiveness of our approach. Our evaluation demonstrates that the proposed multi-sensor fusion limits accumulated drift, outperforming the baseline SVIn2 framework and providing a robust foundation for future integration of global loop-closure and SLAM capabilities.

REFERENCES

- [1] J. McConnell, I. Collado-Gonzalez, and B. Englot, "Perception for underwater robots," *Current Robotics Reports*, 2022.
- [2] B. Joshi, S. Rahman, M. Kalaitzakis, B. Cain, J. Johnson, M. Xanthidis, N. Karapetyan, A. Hernandez, A. Quattrini Li, N. Vitzilaos, and I. Rekleitis, "Experimental Comparison of Open Source Visual-Inertial-Based State Estimation Algorithms in the Underwater Domain," in *Proc. IROS*, Macau, 2019, pp. 7221–7227.
- [3] C. Hu, S. Zhu, Y. Liang, Z. Mu, and W. Song, "Visual-pressure fusion for underwater robot localization with online initialization," *IEEE Robotics and Automation Letters*, 2021.
- [4] L. Zhao, M. Zhou, and B. Loose, "Tightly-coupled visual- DVL-inertial odometry for robot-based ice-water boundary exploration," in *Proc. IROS*, 2023.
- [5] E. Vargas, R. Scona, J. S. Willners, T. Luczynski, Y. Cao, S. Wang, and Y. R. Petillot, "Robust underwater visual SLAM fusing acoustic sensing," in *Proc. ICRA*, 2021.
- [6] S. Xu, K. Zhang, and S. Wang, "AQUA-SLAM: Tightly-coupled underwater acoustic-visual-inertial slam with sensor calibration," *IEEE Trans. Robot.*, 2025.
- [7] S. Archieri, A. Cinar, S. Pan, J. S. Willners, M. Grimaldi, I. Carlucho, and Y. Petillot, "InsSo3D: Inertial navigation system and 3D Sonar SLAM for turbid environment inspection," in *Proc. ICRA*, 2026.
- [8] S. Pan, S. Archieri, A. Cinar, J. S. Willners, I. Carlucho, and Y. Petillot, "VISO: Robust underwater visual-inertial-sonar SLAM with photometric rendering for dense 3D reconstruction," in *Proc. ICRA*, 2026.
- [9] C. Burgul, Y. Huang, M. Chatzisprou, I. Rekleitis, A. Quattrini Li, and M. Xanthidis, "Underwater dense mapping with the first compact 3d sonar," in *Proc. ICRA*, 2026.
- [10] J. C. Kinsey, R. M. Eustice, and L. L. Whitcomb, "A survey of underwater vehicle navigation: Recent advances and new challenges," in *IFAC conference of manoeuvring and control of marine craft*, 2006.
- [11] Y. Yang and G. Huang, "Acoustic-inertial underwater navigation," in *Proc. ICRA*, 2017.
- [12] P. Blondel and B. J. Murton, *Handbook of seafloor sonar imagery*. Wiley Chichester, 1997.
- [13] S. Negahdaripour, "On 3-d motion estimation from feature tracks in 2-d fs sonar video," *IEEE Transactions on Robotics*, 2013.
- [14] O. Pizarro, R. Eustice, and H. Singh, "Large area 3d reconstructions from underwater surveys," in *Oceans' 04 MTS/IEEE Techno-Ocean'04 (IEEE Cat. No. 04CH37600)*, 2004.
- [15] A. Kim and R. M. Eustice, "Real-time visual SLAM for autonomous underwater hull inspection using visual saliency," *IEEE Transactions on Robotics*, 2013.
- [16] F. Hidalgo, C. Kahlefeldt, and T. Bräunl, "Monocular ORB-SLAM application in underwater scenarios," in *2018 OCEANS-MTS/IEEE Kobe Techno-Oceans (OTO)*, 2018.
- [17] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, 2015.
- [18] A. G. Chavez, Q. Xu, C. A. Mueller, S. Schwertfeger, and A. Birk, "Adaptive navigation scheme for optimal deep-sea localization using multimodal perception cues," in *Proc. IROS*, 2019.
- [19] S. Xu, T. Luczynski, J. S. Willners, Z. Hong, K. Zhang, Y. R. Petillot, and S. Wang, "Underwater visual acoustic SLAM with extrinsic calibration," in *Proc. IROS*, 2021.
- [20] M. Leonardi, A. Stahl, E. F. Brekke, and M. Ludvigsen, "UVS: underwater visual SLAM—a robust monocular visual SLAM system for lifelong underwater operations," *Autonomous Robots*, 2023.
- [21] J. Song, O. Bagoren, R. Andigani, A. Sethuraman, and K. A. Skinner, "Turtlmap: Real-time localization and dense mapping of low-texture underwater environments with a low-cost unmanned underwater vehicle," in *Proc. IROS*, 2024.
- [22] S. Ding, T. Zhang, D. Jiang, and M. Lei, "Underwater visual-inertial-acoustic-depth SLAM with DVL preintegration for degraded environments," *arXiv preprint arXiv:2510.21215*, 2025.
- [23] M. F. Fallon, J. Folkesson, H. McClelland, and J. J. Leonard, "Relocating underwater features autonomously using sonar-based SLAM," *IEEE Journal of Oceanic Engineering*, 2013.
- [24] J. Li, M. Kaess, R. M. Eustice, and M. Johnson-Roberson, "Pose-graph SLAM using forward-looking sonar," *IEEE Robotics and Automation Letters*, 2018.
- [25] J. Wang, S. Bai, and B. Englot, "Underwater localization and 3d mapping of submerged structures with a single-beam scanning sonar," in *Proc. ICRA*, 2017.
- [26] J. Wang, F. Chen, Y. Huang, J. McConnell, T. Shan, and B. Englot, "Virtual maps for autonomous exploration of cluttered underwater environments," *IEEE Journal of Oceanic Engineering*, 2022.
- [27] S. Archieri, J. Drupt, A. Cinar, M. Grimaldi, I. Carlucho, J. S. Willners, and Y. R. Petillot, "3dssdf: Underwater 3d sonar reconstruction using signed distance functions," in *Proc. ICRA*, 2025.
- [28] S. Rahman, A. Quattrini Li, and I. Rekleitis, "SVIn2: A Multi-sensor Fusion-based Underwater SLAM System," *International Journal of Robotics Research*, July 2022.
- [29] L. Chen, S. Wang, H. Hu, Y.-s. Ryu, and G.-H. Yang, "3d mapping by a robotic fish with two mechanical scanning sonars," in *Intelligent Autonomous Systems (IAS)*, 2015.
- [30] A. Mallios, E. Vidal, R. Campos, and M. Carreras, "Underwater caves sonar data set," *International Journal of Robotics Research*, 2017.
- [31] J. McConnell, I. Collado-Gonzalez, P. Szenher, and B. Englot, "Large-scale dense 3-d mapping using submaps derived from orthogonal imaging sonars," *IEEE Journal of Oceanic Engineering*, 2024.
- [32] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual-inertial odometry using nonlinear optimization," *International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.
- [33] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual-inertial odometry," *IEEE Transactions on Robotics*, 2016.
- [34] I. Vizzo, T. Guadagnino, B. Mersch, L. Wiesmann, J. Behley, and C. Stachniss, "KISS-ICP: In Defense of Point-to-Point ICP – Simple, Accurate, and Robust Registration If Done the Right Way," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 1029–1036, 2023.
- [35] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *Proc. IROS*, 2013.
- [36] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. CVPR*, 2016.