Acoustic-VINS: Tightly Coupled Acoustic-Visual-Inertial Navigation System for Autonomous Underwater Vehicles

Jiangbo Song, Wanqing Li, and Xiangwei Zhu, *Member, IEEE*

***Abstract*—In this work, we present an acoustic-visual-inertial navigation system (Acoustic-VINS) for underwater robot localization. Specifically, we address the problem of the global position of the underwater visual-inertial navigation system being inappreciable by tightly coupling the long baseline (LBL) system into an optimization-based visual-inertial SLAM. In our proposed Acoustic-VINS, the reprojection error, IMU preintegration error, and raw LBL measurement error are jointly minimized within a sliding window factor graph framework. Furthermore, we propose an acoustic-aided initialization method to exhibit an accurate initial state for successful state estimation. Additionally, for wider application, we extend the sensor data of the real-world AQUALOC dataset to obtain the LBL-AQUALOC dataset. Experimental results on the ten sequences of the LBL-AQUALOC dataset in challenging underwater scenes show that our proposed approach outperforms state-of-the-art visual-inertial SLAM.**

***Index Terms*—Underwater SLAM, marine robots, sensor fusion, AUV navigation, localization.**

I. Introduction

UNDERWATER robots such as autonomous underwater vehicles (AUVs) are widely used in ocean exploration, surveying, and mapping. Accurate global positioning and environmental awareness are particularly important for unmanned platforms [1]. Visual-inertial navigation systems (VINS) have shown great promise in providing accurate and robust navigation solutions for a variety of applications, including autonomous driving and drone navigation [2]. However, despite their many advantages, VINS can suffer from several limitations, particularly when applied to AUV navigation [3]. As shown in Fig. 1(b), the complex and challenging underwater environment, characterized by weak textures, low light levels, and high dynamics, can pose significant challenges to visual information, making it difficult for VINS to provide reliable navigation solutions [4]. Therefore, improving the performance of underwater VINS usually requires the fusion of other sensors, such as imaging sonar, Doppler velocity log (DVL), etc [5].

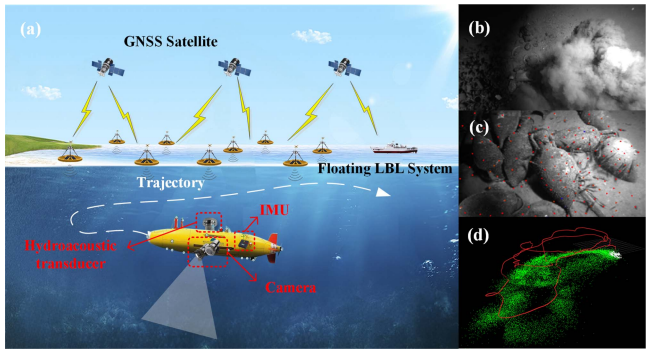

Fig. 1. (a) Application scenarios of Acoustic-VINS and sensors related to positioning and perception. (b) Underwater challenging environment. (c) Underwater visual feature extraction. (d) Acoustic-VINS is employed under the Robot Operation System (ROS) architecture.

There are many impressive visual-inertial algorithms such as filtering-based Open-VINS [6], and optimization-based VINS series [7], [8], [9], ORB-SLAM series [10] and other excellent open source frameworks. In the application of underwater scenes, VINS often needs the assistance of other sensors. SVIn2 [11] enhances the robustness of VINS in underwater visually limited environments at different scales by fusing imaging sonar. To improve the positioning ability, DVL is also used to integrate into VINS [12], which is constrained as the speed measurement, but the improvement of the environment perception ability is limited. Pressure sensor (PS) observations are also often used to constrain VINS, using filtering-based [12] or optimization-based methods [5] to make full use of the advantage of PS elevation non-divergence to improve the performance of VINS. However, current related research on the extension of VINS mainly focuses on self-positioning solutions such as DVL, imaging sonar, and PS [13], [14]. These sensors lack global positioning information or only have partial global information (PS only provides global elevation information), to a

Manuscript received 17 August 2023; accepted 11 November 2023. Date of publication 20 November 2023; date of current version 10 January 2024. This letter was recommended for publication by Associate Editor N. Lawrance and Editor P. Pounds upon evaluation of the reviewers' comments. This work was supported in part by Shenzhen Science and Technology Program under Grants GXWD20201231165807008, 20200830225317001, and ZDSYS20210623091807023, and in part by Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) under Grant SML2021SP408. *(Corresponding authors: Wanqing Li; Xiangwei Zhu.)*

Jiangbo Song is with the School of System and Engineering, Sun Yat-sen University, Guangzhou 510006, China (e-mail: songjb8@mail2.sysu.edu.cn).

Wanqing Li is with the School of Aeronautics and Astronautics, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China (e-mail: liwq223@mail.sysu.edu.cn).

Xiangwei Zhu is with the School of Information and Communication Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China, also with the Shenzhen Key Laboratory of Navigation and Communication Integration, Shenzhen 518107, China, and also with the Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Guangzhou 510006, China (e-mail: zhuxw666@mail.sysu.edu.cn).

The LBL-AQUALOC dataset is available at: https://github.com/SYSU-CPNTLab/LBLAQUALOC-Dataset.

Digital Object Identifier 10.1109/LRA.2023.3334979

certain extent, they cannot accurately position in the geographic coordinate system.

The long baseline (LBL) system provides accurate global position information independent of water depth in large survey areas, which is widely used in underwater high-precision positioning. Previous research has proven that the tight coupling of Strap-down inertial navigation system (SINS) and LBL is an effective solution to high-precision underwater positioning [15], and the reliability of the system can be greatly improved by integrating DVL and PS. Our previous work [16] mainly focused on AUV high-precision positioning based on the fusion of SINS, LBL, DVL, PS, and other sensors. However, previous studies involving LBL did not address the issue of underwater environment perception.

In underwater environments, current research faces two main challenges. First, VINS-based research lacks constraints from global positioning information, making it difficult to eliminate cumulative errors in large-scale, non-closed-loop scenarios. Second, LBL-based research primarily focuses on high-precision positioning while neglecting environmental perception. Therefore, it is essential to investigate the coupling between LBL and VINS to address these issues. This combination cannot only enhance environmental perception capabilities but also achieve high-precision, drift-free global underwater positioning, greatly expanding the operational range of AUVs.

To address the aforementioned challenges, this letter proposes an Acoustic-Visual-Inertial Navigation System (Acoustic-VINS) that tightly couples acoustic, visual, and inertial measurements in a sliding window factor graph framework. Additionally, an acoustic-aided initialization method is presented to quickly obtain the initial state for global optimization. The system forms a floating long baseline positioning system (floating LBL) by deploying surface acoustic buoys [17] and submarine transponders, as shown in Fig. 1. Within the working range of the LBL system, global positioning information can be provided to the AUV, and the cumulative error of the VINS can be eliminated when a closed loop cannot be formed. This further improves the AUV's positioning accuracy and stability in challenging underwater environments. Our contributions are concluded as follows:

- The first attempt to tightly couple acoustic LBL observations with an optimization-based visual-inertial navigation system, enabling AUVs to obtain globally drift-free state estimates in large-scale underwater environments without loops.
- The LBL position factor and slant-range factor model are established for initialization and optimization, respectively.
- An acoustic-aided initialization approach is proposed to enable rapid system initialization in visually limited underwater scenes. The initialization experiment verify the effectiveness of the proposed algorithm.
- A LBL-AQUALOC dataset extended from AQUALOC [18]. We open-source the well-synchronized simulated LBL and visual-inertial datasets on GitHub.
- The global localization performance of the proposed Acoustic-VINS is verified to outperform the state-of-the-art (SOTA) algorithms on ten sequences of LBL-AQUALOC.

The remainder of this letter is organized as follows. Section II introduces the system overviews of Acoustic-VINS. Section III introduces the methodology of Acoustic-VINS. The experiments and analyses are presented in Section IV. Finally, Section V summarizes our work.

II. System Overviews

The overall architecture of our proposed Acoustic-VINS is shown in Fig. 2. The entire system is divided into four modules: sensors, pre-processing, initialization, and optimization.

1) Sensors Module: The sensor module includes an acoustic floating LBL system, an IMU, and a camera.

2) Pre-Processing Module: The pre-processing module performs preliminary processing of the raw data from the sensors. The slant-range information of LBL is calculated based on the sound speed and arrival time, and the acoustic position information is further calculated. The IMU measurements between two adjacent LBL observations are pre-integrated [19]. The feature points of image information are detected and tracked.

3) Initialization Module: An acoustic-aided initialization method is proposed to quickly initialize a system in an underwater environment. The position information of LBL is combined with the IMU pre-integration results to perform acoustic-inertial alignment, and the result of visual-only SfM is aligned with inertial information. During the whole acoustic-aided initialization process, accurate initialization information is provided for the back-end optimization module, including pose, speed, gravity, gyro bias,local-global transformation matrix, etc. For the detailed initialization method, please refer to Section III-B.

4) Optimization Module: The optimization module implements the process of optimal state estimation including sliding window factor graph optimization (FGO) and global pose graph optimization (PGO). For the detailed optimization strategy, please refer to Section III-D.

III. Methodology of Acoustic-VINS

This section first introduces the frames and notations used in the Acoustic-VINS; second, it presents the initialization process of Acoustic-VINS; then, it describes the modeling of acoustic, inertial, and visual factors, with a particular focus on acoustic factors related to the LBL; finally, it elaborates the optimization strategy in Acoustic-VINS.

A. Frames and Notations

W stands for the local world frame, and the notation $(\cdot)^w$ represents a quantity in the world frame W. Assuming the direction of gravity is aligned with z^w axis. Similarly, B is the body frame and corresponds to the IMU frame, and the camera frame is denoted by C. The notation $(\cdot)^b$ and $(\cdot)^c$ represent a quantity in the body frame B and camera frame C, respectively. $p^w_{b_k}$ and $q^w_{b_k}$ represent the position and orientation of B with

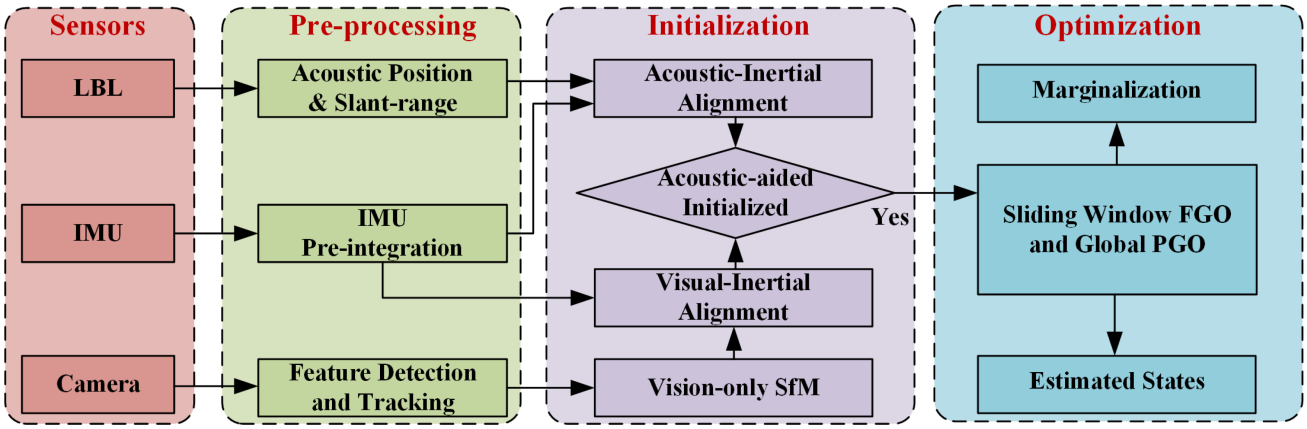

Fig. 2. System overview of the proposed Acoustic-VINS. The system is divided into four modules: sensors, pre-processing, initialization, and optimization. First, the measurements of the sensors are pre-processed. Then, in the initialization module, visual-inertial-initialization is completed by aligning the inertial information and the results of vision-only Structure from Motion (SfM); Acoustic-Inertial-initialization is completed by aligning the inertial information and acoustic information. After the final initialization is successful, it will performs sliding window factor graph optimization (FGO) and global pose graph optimization (PGO).

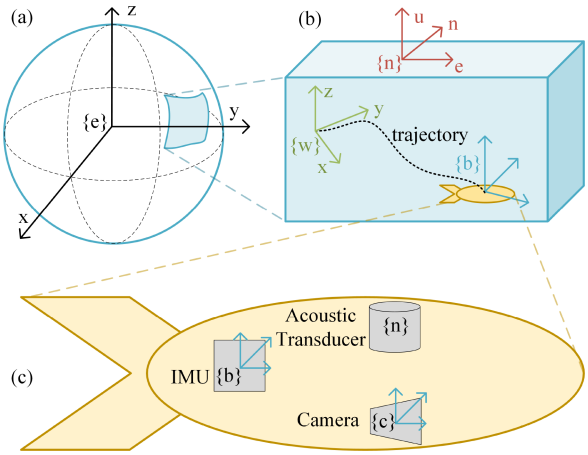

Fig. 3. Illustration of world frames and sensor frames. (a) The origin of the ECEF frame $\{e\}$ is on the center of the earth. The x-axis of the ECEF frame points to the point of intersection between the prime meridian and equator, and the z-axis points to the North Pole. (b) The ENU frame $\{n\}$ and the local world frame for SLAM $\{w\}$ are both located on the ground, with their z-axes pointing upward. (c) $\{b\}$ is the body frame and corresponds to the IMU frame, and the camera frame is denoted by $\{c\}$.

respect to W at time t_k. The velocity of B expressed in W at time t_k is denoted by $v^w_{b_k}$. $R^w_{b_k}$ represents the rotation matrix, which describes the rotation from the moving frame b at time t to the fixed frame w.

As shown in Fig. 3, the Earth-Centered, Earth-Fixed (ECEF) frame $(\cdot)^e$, as a global world frame, is a Cartesian coordinate system that is fixed with respect to the Earth. A semi-global frame, East-North-Up (ENU), is introduced to connect the local world and global world frames. The e-axes, n-axes, and u-axes of the ENU frame $(\cdot)^n$ point to the east, north, and upward direction, respectively. Given a point in the ECEF frame, a unique ENU frame can be determined, with its origin sitting on that point. Note that the u-axis of the ENU frame is also gravity aligned. The system state vector is

Algorithm 1: Acoustic-Aided Initialization.

Require: VINS-based trajectory $\mathbf{p}^w_{VI}$, ILNS-based trajectory $\mathbf{p}^e_{IL}$

Ensure: $\mathbf{R}^n_w$, $\mathbf{p}^n_w$

1: $\mathbf{p}^e_o \leftarrow$ Select the initial position information $\mathbf{p}^e_o$ calculated by ILNS as the reference point

2: $\mathbf{R}^e_n, \mathbf{p}^e_n \leftarrow$ Calculate $\mathbf{R}^e_n$ and $\mathbf{p}^e_n$ according to the reference point $\mathbf{p}^e_o$

3: $\mathbf{p}^n_{IL} \leftarrow$ Convert the ILNS-based trajectory from ECEF to ENU: $\mathbf{p}^n_{IL} = \mathbf{R}^n_e(\mathbf{p}^e_{IL} - \mathbf{p}^e_n)$

4: $\mathbf{p}^n_{IL}, \mathbf{p}^w_{VI} \leftarrow$ Align the time stamps of $\mathbf{p}^n_{IL}$ and $\mathbf{p}^w_{VI}$

5: $s, \mathbf{R}^n_w, \mathbf{p}^n_w \leftarrow$ Obtain $\mathbf{R}^n_w$ and $\mathbf{p}^n_w$ through nonlinear optimization:

$$\min_{s,\mathbf{R}^n_w,\mathbf{p}^n_w} \sum_{l=0}^{L-1} \left\| \mathbf{p}^n_{IL_l} - (s\mathbf{R}^n_w \mathbf{p}^w_{VI_l} + \mathbf{p}^n_w) \right\|^2$$

expressed as

$$\begin{aligned} \mathcal{X} &= \left[\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{T}^b_c, \boldsymbol{T}^n_w, \lambda_0, \lambda_1, \ldots, \lambda_l, \delta\boldsymbol{\rho}\right] \\ \boldsymbol{x}_k &= \left[\boldsymbol{p}^w_{b_k}, \mathbf{q}^w_{b_k}, \boldsymbol{v}^w_{b_k}, \boldsymbol{b}_{g_k}, \boldsymbol{b}_{a_k}\right], k \in [0, n] \\ \boldsymbol{T}^b_c &= \left[\boldsymbol{R}^b_c, \mathbf{p}^b_c\right], \boldsymbol{T}^n_w = \left[\boldsymbol{R}^n_w, \mathbf{p}^n_w\right] \end{aligned} \tag{1}$$

where λ is the inverse depth parameter of the landmark; ρ is the slant-range of the LBL system; T^b_c is the extrinsic parameters between the camera c-frame and the IMU b-frame. In our system, the b-frame is regarded as the coordinate system of the AUV, and the position of the acoustic buoy is in the n-frame.

B. Acoustic-Aided Initialization

The acoustic-aided initialization process of Acoustic-VINS is divided into visual-inertial alignment and acoustic-inertial alignment. When the Acoustic-VINS system starts, the subsystem VINS starts at the same time and performs visual-inertia

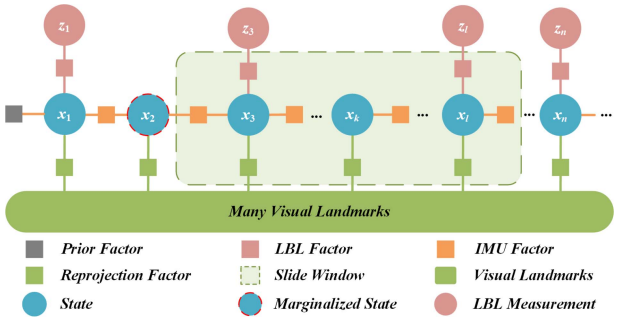

Fig. 4. Acoustic-VINS factor graph framework. The system state is jointly constrained by acoustic, visual, and inertial factors, and is optimized within a sliding window. Note that the frequency of visual factors depends on keyframes and higher than that of LBL factors, the IMU pre-integrates at adjacent keyframes.

initialization which is similar to [7]. At the same time, the subsystem factor graph optimization-based inertial-LBL navigation system [20] (FGO-ILNS) starts to work. Therefore, we can get VINS-based trajectory $\mathbf{p}_{VI}^{w}$ (local frame) and ILNS-based trajectory $\mathbf{p}_{IL}^{e}$ (global frame). The goal of Acoustic-VINS initialization is to obtain the transformation matrix $\boldsymbol{T}_{w}^{n} = [\boldsymbol{R}_{w}^{n}, \mathbf{p}_{w}^{n}]$ from the local to global frame through nonlinear optimization, thereby converting the local trajectory to the global frame trajectory.

The specific implementation of Acoustic-VINS initialization is shown in Algorithm 1. Specifically, selecting the initial position information $\mathbf{p}_{o}^{e}$ calculated by FGO-ILNS as the reference point, which can be used as the origin of the n-frame. According to the reference point $\mathbf{p}_{o}^{e}$, $\mathbf{R}_{n}^{e}$ and $\mathbf{p}_{n}^{e}$ can be calculated. Therefore, the ILNS-based trajectory can be converted from $\mathbf{p}_{IL}^{e}$ to $\mathbf{p}_{IL}^{n}$, namely

$$\mathbf{p}_{IL}^{n} = \mathbf{R}_{e}^{n}\left(\mathbf{p}_{IL}^{e} - \mathbf{p}_{n}^{e}\right). \tag{2}$$

At this time, two sets of location information can be obtained, namely $\mathbf{p}_{IL}^{n}$ and $\mathbf{p}_{VI}^{w}$. Then, the two trajectories are time-synchronized by aligning the timestamps. The local-global transformation parameters $\mathbf{R}_{w}^{n}$ and $\mathbf{p}_{w}^{n}$ can be obtained through a nonlinear optimization method, namely

$$\min_{s,\mathbf{R}_{w}^{n},\mathbf{p}_{w}^{n}} \sum_{l=0}^{L-1} \left\| \mathbf{p}_{IL_l}^{n} - \left(s\mathbf{R}_{w}^{n}\mathbf{p}_{VI_l}^{w} + \mathbf{p}_{w}^{n}\right)\right\|^{2}. \tag{3}$$

Finally, we can convert the local trajectory to the global trajectory.

C. Factor Graph Optimization

This subsection introduces the Maximum A Posteriori (MAP) estimation problem in Acoustic-VINS and the modeling of LBL position factor and slant-range factor, IMU pre-integration, and camera projection residuals. The factor graph framework see Fig. 4.

1) Maximum a Posteriori Estimation: The MAP estimation problem in Acoustic-VINS can be described as minimizing the sum of the prior and the Mahalanobis norm of all observations:

$$\min_{\mathcal{X}} \left\{ \begin{array}{l} \|\mathbf{r}_{p} - \mathbf{H}_{p}\mathcal{X}\|^{2} + \sum_{k\in\mathcal{B}} \left\|\mathbf{r}_{\mathcal{B}}\left(\tilde{\boldsymbol{z}}_{k,k+1}^{\mathcal{B}}, \mathcal{X}\right)\right\|_{\boldsymbol{\Sigma}_{k,k+1}^{\mathcal{B}}}^{2} \\ + \sum_{l\in\mathcal{L}} \left\|\mathbf{r}_{\mathcal{C}}\left(\tilde{\boldsymbol{z}}_{l}^{\mathcal{C}_{i,j}}, \mathcal{X}\right)\right\|_{\boldsymbol{\Sigma}_{l}^{\mathcal{V}_{i,j}}}^{2} \\ + \sum_{h\in\mathcal{A}} \left\|\mathbf{r}_{\mathcal{A}}\left(\tilde{\boldsymbol{z}}_{h}^{\mathcal{A}}, \mathcal{X}\right)\right\|_{\boldsymbol{\Sigma}_{h}^{\mathcal{A}}}^{2} \end{array} \right\} \tag{4}$$

where $\mathbf{r}_{\mathcal{B}}$ is the residuals of the IMU preintegration measurements; $\mathbf{r}_{\mathcal{C}}$ is the residuals of the visual measurements; $\mathbf{r}_{\mathcal{A}}$ is the residuals of the LBL measurements; Σ is the covariance for each measurement; $\mathbf{r}_{p}$, $\mathbf{H}_{p}$ represent the prior information from marginalization; $\mathcal{A}$ is the number of the LBL measurements in the sliding window; $\mathcal{L}$ is the landmark map in the sliding window; l is the landmark on the map; i denotes the reference keyframe of the landmark l; j is another keyframe.

2) LBL Position Factor and Slant-Range Factor: This section introduces two factor models of the LBL system, one for the position factor during the initialization phase and one for the slant-range factor for back-end optimization.

During the initialization phase, the acquisition of the ILNS-based trajectory $\mathbf{p}_{IL}^{e}$ requires factor modeling of LBL's position. The LBL position information in the geographic coordinate system is converted to the local world coordinate system W, taking into account the external parameters l_{LBL}^{b} of the LBL acoustic transducer and the carrier coordinate system B, and its residual model can be expressed as

$$\mathbf{r}_{\mathcal{A}}(\tilde{z}_{h}^{\mathcal{A}}, X) = p_{wb_h}^{w} + R_{b_h}^{w} l_{LBL}^{b} - \hat{p}_{\mathcal{A},h}^{w}. \tag{5}$$

Next introducing the slant-range factor. r stands for acoustic hydrophone, which is installed in a buoy to receive and emit acoustic signals; s represents the acoustic source, acoustic transducer equipped on AUV; LBL slant-range measurements can be modeled as

$$\tilde{\rho}_{s,r}^{i} = \|\mathbf{p}_{s}^{e} - \mathbf{p}_{r}^{e}\| + c\boldsymbol{\delta}\mathbf{t} + v_{\rho} \tag{6}$$

where $\mathbf{p}_{s}^{e}$ is the position of the acoustic transducer; $\mathbf{p}_{r}^{e}$ is the position of the acoustic hydrophone; c is the acoustic average speed, and $\boldsymbol{\delta}\mathbf{t}$ is the arrival time bias between the sound source and the transducer, therefore, $c\boldsymbol{\delta}\mathbf{t}$ represents the propagation delay error; v_{ρ} is measurement noise. The rotation matrix from the ENU frame to the ECEF frame is

$$\mathbf{R}_{n}^{e} = \begin{bmatrix} -\sin\lambda & -\sin\phi\cos\lambda & \cos\phi\cos\lambda \\ \cos\lambda & -\sin\phi\sin\lambda & \cos\phi\sin\lambda \\ 0 & \cos\phi & \sin\phi \end{bmatrix}. \tag{7}$$

The relationship between the ECEF and local world coordinates of the AUV's acoustic transducer can be expressed as

$$\mathbf{p}_{r}^{e} = \mathbf{R}_{n}^{e}\mathbf{R}_{w}^{n}\left(\mathbf{p}_{r}^{w} - \mathbf{p}_{\text{ori}}^{w}\right) + \mathbf{p}_{\text{ori}}^{e}. \tag{8}$$

The position of the AUV's acoustic transducer in the local world frame can be associated with the system states by

$$\mathbf{p}_{r}^{w} = \mathbf{p}_{b}^{w} + \mathbf{R}_{b}^{w}\mathbf{p}_{r}^{b}. \tag{9}$$

Consequently, the residual of LBL slant-range measurement in t_k can be formulated as

$$\mathbf{r}_{\mathcal{A}}\left(\tilde{\boldsymbol{z}}_{h_k}^{\mathcal{A}}, \mathcal{X}\right) = \left\|\mathbf{p}_{s_k}^{e} - \mathbf{p}_{r_k}^{e}\right\| + c\boldsymbol{\delta}\mathbf{t} + v_{\rho} - \tilde{\rho}_{s,r}^{i}. \tag{10}$$

3) IMU Preintegration Factor: The residual that relates the Acoustic-VINS states and pre-integrated IMU measurements can be formulated as

$$\mathbf{r}_{\mathcal{B}}\left(\tilde{\boldsymbol{z}}_{k-1,k}^{\mathcal{B}}, \mathcal{X}\right) = \left[\delta\boldsymbol{\alpha}_{b_{k+1}}^{b_k}, \delta\boldsymbol{\beta}_{b_{k+1}}^{b_k}, \delta\boldsymbol{\theta}_{b_{k+1}}^{b_k}, \delta\mathbf{b}_a, \delta\mathbf{b}_g\right] \quad (11)$$

where $\delta\boldsymbol{\alpha}_{b_{k+1}}^{b_k}, \delta\boldsymbol{\beta}_{b_{k+1}}^{b_k}, \delta\boldsymbol{\theta}_{b_{k+1}}^{b_k}$ encapsulate relative position, speed, and rotation information. Details about IMU pre-integration can be found in [9].

4) Visual Reprojection Factor: The visual reprojection residual can be formulated as

$$\mathbf{r}_{\mathcal{C}}\left(\hat{\mathbf{z}}_l^{c_j}, \mathcal{X}\right) = \begin{bmatrix}\mathbf{b}_1 & \mathbf{b}_2\end{bmatrix}^T \cdot \left(\hat{\overline{\mathcal{P}}}_l^{c_j} - \frac{\mathcal{P}_l^{c_j}}{\left\|\mathcal{P}_l^{c_j}\right\|}\right) \quad (12)$$

Details of the visual reprojection factor can be found in [7].

D. Optimization Strategy of Acoustic-VINS

In Acoustic-VINS, state estimation is modeled as a MAP problem. To obtain more accurate state estimation results under large-scale conditions without loop closure, we adopted a coarse-to-fine optimization strategy. Initially, the coarse optimization is achieved by the sliding window optimization method, which provides a preliminary fusion result of acoustic-visual-inertial observations. This coarse trajectory optimization result is referred to as Acoustic-VIO. Subsequently, to obtain a globally consistent trajectory, we perform global pose graph optimization while integrating LBL measurement information, fixing the map points, and only optimizing the keyframe poses. The trajectory after global pose graph optimization is referred to as Acoustic-VINS, which is the final refined optimization result. Meanwhile, to ensure system real-time performance, the size of the sliding window and the frequency of global PGO can be adjusted according to computer performance.

IV. Experimental Results

To evaluate the initialization and global localization performance of the proposed Acoustic-VINS, we tested our system on the underwater dataset LBL-AQUALOC which is derived from AQUALOC [18]. The proposed system was implemented under the ROS framework. Unless otherwise stated, all experiences were performed on the same desktop with an Intel Core i7-11700K at 3.6 GHz and 32 GB RAM.

A. LBL-AQUALOC Dataset Description

The public underwater dataset AQUALOC collected ten sequence data in the scene of underwater archaeological sites. To address the lack of LBL data, we constructed a new dataset, LBL-AQUALOC, using a semi-physical simulation. This new dataset includes measurements from LBL, a mono camera, a PS, and a low-cost MEMS-IMU, packaged in Rosbag format. The detailed parameters of the data acquisition system and various sensors are shown in Table I. In original AQUALOC dataset, given the difficulty of acquiring ground truth in natural underwater environments, SOTA Structure-from-Motion (SfM) software, Colmap [21], was employed to compute 3D reconstructions of each sequence offline and extract reliable trajectories. Next, we will introduce the underwater visual-inertial data and the detailed simulation process of LBL data.

TABLE I
Technical Details of the LBL-AQUALOC Dataset and Acquisition System

Dataset	**Specification**
Number of Sequence	Total 10 Sequences
Duration	Total 60'39"
Length	Total 571.5 m
Depth	Seq01-03: About 270 m Seq04-10: About 380 m
Acoustic sensor	**Floating LBL System**
Output frequency	1Hz
Ranging accuracy	0.05 m
Relative positioning accuracy	0.2 m
Buoy coordinates (Seq01-03)	buoy01: (-150m, 75m, 270m) buoy02: (75m, 150m, 270m) buoy03: (150m, -75m, 270m) buoy04: (-75m, -150m, 270m)
Buoy coordinates (Seq04-10)	buoy01: (-150m, 75m, 380m) buoy02: (75m, 150m, 380m) buoy03: (150m, -75m, 380m) buoy04: (-75m, -150m, 380m)
*__Camera sensor__	**UEye - UI-3260CP**
Resolution	968×608 px
Sensor	Monochromatic
Frames per second	20 fps
*__Lens__	**Kowa LM6NCH C-Mount**
Focal length	6 mm
*__Inertial Measurement Unit__	**MEMS - MPU-9250**
Gyroscope frequency	200 Hz
Accelerometer frequency	200 Hz
Magnetometer frequency	200 Hz
*__Housing__	**3" Blue Robotics Enclosure**
Enclosure	25.8 x 8.9 cm
Enclosure Material	Aluminium
Dome	3" Blue Robotics Dome End Cap

* The corresponding parameters are derived from the AQUALOC [18].

1) Underwater Visual-Inertial Data: The sequences of the archaeological sites were captured in the Mediterranean Sea near Corsica. In total, 10 sequences were documented, with 3 from the first site and 7 from the second. The first site was approximately 270 meters deep and had remnants of an old shipwreck. The second site was around 380 meters deep. As shown in Fig. 5, the underwater archaeological environment has characteristics such as low light, weak texture, repetitive texture, dynamic object interference, as well as turbidity caused by seabed sediment, and interference caused by the movement of mechanical arms. In such an environment, visual information faces significant challenges.

2) Simulation of LBL Measurements: Assuming that a floating LBL system was deployed on the water surface as shown in Fig. 1. The trajectory obtained from Colmap is used as the ground truth of the AUV motion trajectory. We further simulate the position of the acoustic buoy, the sound speed profile file (assuming that the sound speed changes uniformly with depth), and the time of arrival (TOA) at each moment (adding measurement deviation to TOA). The final simulated LBL measurements include the buoy number, buoy position information, arrival time, and average sound speed obtained by the AUV at each moment. These acoustic measurements are time-synchronized with visual-inertial measurements. Through

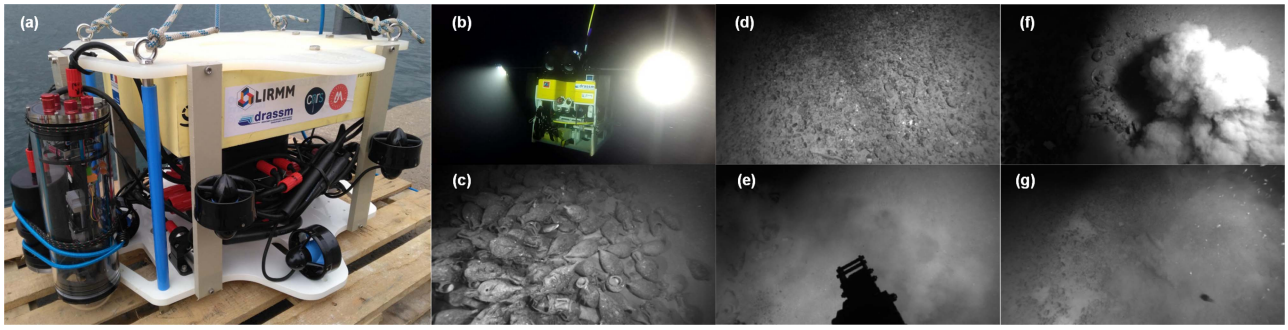

Fig. 5. (a) Data acquisition equipment ROV. And underwater challenging environment, (b) low light; (c) archaeological sites; (d) texture repetitive; (e) robotic arm; (f) sandy cloud; (g) dynamic object. Reproduced with permission of [18], Copyright 2019, SAGE Publications.

TABLE II
ATE OF ACOUSTIC-AIDED INITIALIZATION UNDER DIFFERENT DISTANCE

ATE (m)	Distance (m)				
	2	4	6	8	10
Seq #1	0.43	0.22	0.11	0.13	0.25
Seq #3	0.38	0.21	0.09	0.16	0.20
Seq #5	0.54	0.23	0.13	0.19	0.31
Seq #7	0.44	0.22	0.12	0.11	0.26
Seq #10	0.33	0.25	0.11	0.18	0.27

LBL measurements, we can calculate the acoustic positioning accuracy results at the centimeter level, as well as the corresponding slant-range information.

B. Results and Analysis

1) Initialization Experiment: Through acoustic-aided initialization, we can obtain the transformation matrix from the local frame to the global frame. With this transformation matrix, we can convert the trajectory in the local frame to the global coordinates, and then calculate the absolute trajectory error to evaluate the performance of the initialization algorithm. We selected several sequences from the LBL-AQUALOC dataset for verification. The ground truth trajectory is the trajectory processed by Colmap provided by the original AQUALOC dataset.

To verify how the acoustic-aided initialization process is affected by the trajectory distance, we tested its performance at initialization distances of 2 m, 4 m, 6 m, 8 m, and 10 m. This specific trajectory is divided into separate, non-overlapping parts for each distance criterion to avoid skewing the findings in favor of the first half of the trajectory. The initialization technique was carried out individually on each segment, and the statistics obtained about the precision of the initialized system to ENU transform are displayed in Table II .

From Table II, it can be seen that as the initialization distance increases, the performance of initialization also gets better. However, when the distance exceeds 6 m, the error will increase due to the accumulation of errors in local pose estimation. Therefore, we take 6 m as the initialization distance.

2) Global Localization Experiment: The proposed Acoustic-VINS is compared with Acoustic-VIO, FGO-ILNS, and VINS-Fusion. All four algorithms are multi-sensor combination algorithms based on sliding window graph optimization. Among them, Acoustic-VIO integrates LBL, visual, and inertial measurement, without performing global FGO, as described in Section III-D; FGO-ILNS is a subsystem that fuses LBL-inertial; VINS-Fusion is a SOTA visual-inertial system. The proposed system is suitable for AUVs in large-scale environments without loop closure, therefore, the comparison algorithms do not consider loop closure correction.

We tested on ten sequences of the LBL-AQUALOC dataset, and the results are depicted in Fig. 6. In such challenging datasets, Acoustic-VINS shows excellent accuracy. As can be seen from Fig. 6, Acoustic-VINS has very little drift in all ten sequences, while VINS-Fusion exhibits significant drift, and the trajectory of Acoustic-VIO and FGO-ILNS are rough. Compared to Acoustic-VIO, the proposed Acoustic-VINS performs global pose graph optimization and better optimizes the odometry trajectory. The unsmooth trajectory of Acoustic-VIO is due to its ability to provide accurate relative motion estimation only in a short time, and as time goes on, errors accumulate, while LBL observations constrain VIO, so Acoustic-VIO does not show large drift but estimates an unsmooth trajectory. The positioning performance of FGO-ILNS mainly depends on the LBL, and it is affected by LBL noise, exhibiting characteristics of no drift but high noise. In contrast, the globally optimized Acoustic-VINS can provide more accurate and smoother global trajectory estimation. We also tested the monocular-inertial fusion of VINS-Fusion and found that it would produce large accumulated errors in the LBL-AQUALOC dataset due to the lack of global information constraints. Moreover, in actual tests, we found that VINS-Fusion had difficulty initializing in some sequences and even diverged in estimation results. Benefiting from the proposed acoustic-aided initialization method, Acoustic-VINS can achieve fast initialization even when initial visual information is weakly textured or highly dynamic.

We calculated the absolute trajectory error (ATE) of ten sequences in the LBL-AQUALOC dataset with Colmap's calculation results as the reference ground truth, as shown in Fig. 7. Acoustic-VINS produces the highest accuracy in all ten sequences and significantly improves accuracy compared to Acoustic-VIO and FGO-ILNS. Due to the lack of acoustic information assistance and the impact of challenging underwater environments, the visual-inertial system VINS-Fusion exhibits the worst accuracy. Compared to VINS-Fusion, Acoustic-VIO and FGO-ILNS also yield a significant improvement in accuracy due to their constraint by acoustic information and because

TABLE III
AVERAGE ATE OF DIFFERENT ALGORITHMS ON TEN SEQUENCES OF THE LBL-AQUALOC DATASET

[1]Method	[2]Sensors	Mean(m)	Seq #1	Seq #2	Seq #3	Seq #4	Seq #5	Seq #6	Seq #7	Seq #8	Seq #9	Seq #10
VINS-Fusion	*V+I*	2.42	2.26	1.60	3.20	0.85	1.20	1.80	4.01	3.05	3.46	2.72
FGO-ILNS	*L+I*	0.27	0.26	0.24	0.27	0.27	0.28	0.28	0.25	0.28	0.27	0.28
Acoustic-VIO	*L+V+I*	0.59	0.58	0.19	0.85	0.51	0.59	1.40	0.26	0.32	0.34	0.50
Acoustic-VINS	*L+V+I*	**0.13**	**0.10**	**0.13**	**0.13**	**0.12**	**0.12**	**0.11**	**0.11**	**0.12**	**0.20**	**0.11**

[1] All algorithms set the same sliding window size and process the Rosbag data in real-time to output the results.
[2] *L* means acoustic LBL measurements; *V* represents visual measurements; *I* means IMU measurements.

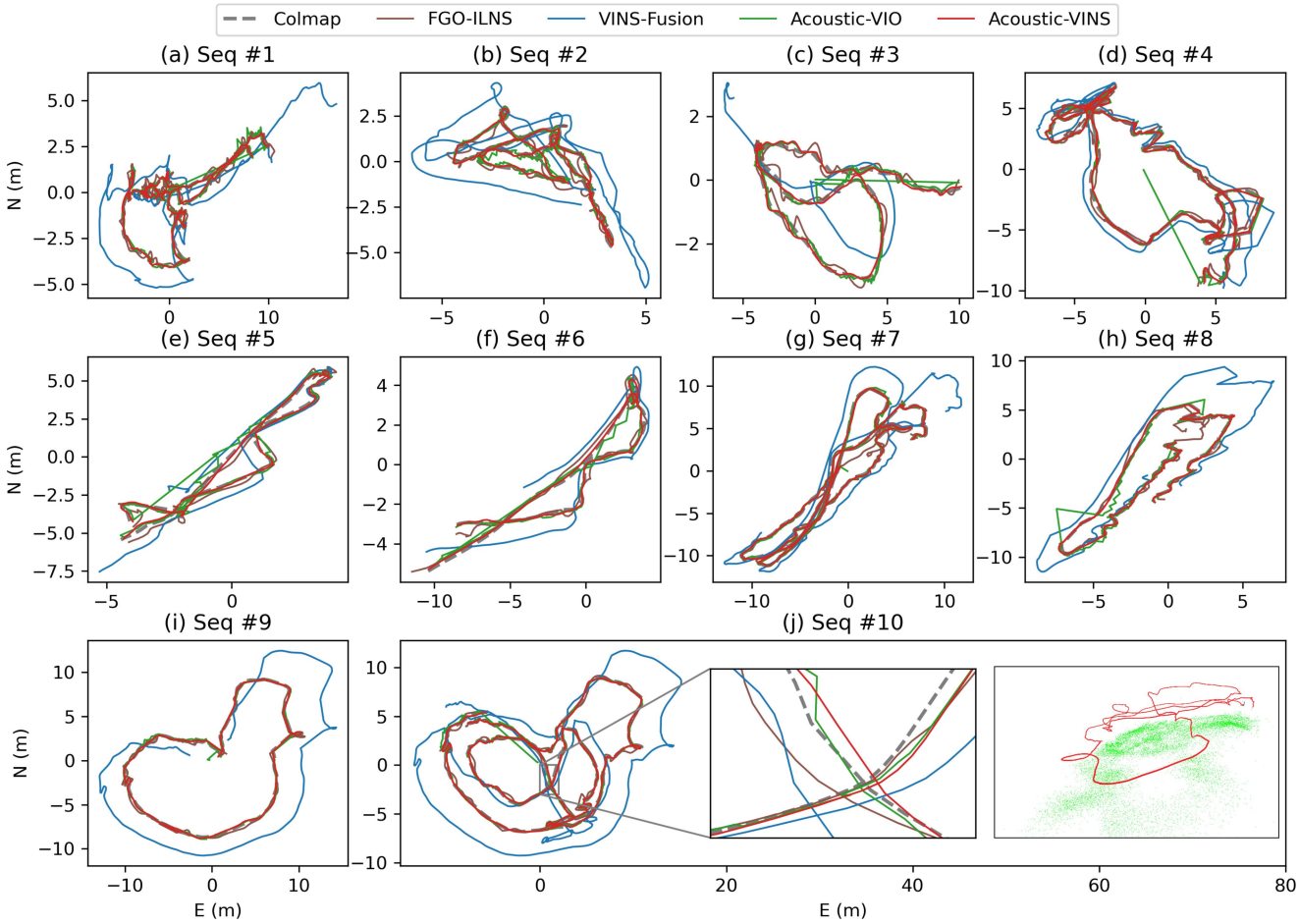

Fig. 6. (a)–(j) Trajectory of the proposed Acoustic-VINS, which calculates in ten sequences of the LBL-AQUALOC dataset and compares it with the trajectories calculated by Acoustic-VIO (without global optimization), the inertial-LBL algorithm FGO-ILNS, the state-of-the-art algorithm VINS-Fusion, and Colmap (as ground truth). (j) In the results of the tenth sequence, the center subgraph shows a local magnified view, while the right subgraph shows the visualization results in Rviz.

LBL's positioning error does not accumulate over time. In Table III, we calculated the average ATE of each algorithm on ten sequences. The proposed Acoustic-VINS improved accuracy by 77.97%, 51.85%, and 94.63% compared to Acoustic-VIO, FGO-ILNS, and VINS-Fusion respectively.

Fig. 8 shows the errors of three methods in three directions on ten sequences. From the box plot, it can be seen that the accuracy performance is similar in all three directions. Moreover, the median values of Acoustic-VINS, FGO-ILNS, and Acoustic-VIO are close and much smaller than that of VINS-Fusion, indicating that Acoustic-VINS, FGO-ILNS, and Acoustic-VIO have high positioning accuracy. Furthermore, the volatility of Acoustic-VINS is smaller than that of Acoustic-VIO and FGO-ILNS, demonstrating that Acoustic-VINS has smaller error fluctuations and estimates a smoother trajectory.

The above experimental results show that the proposed Acoustic-VINS can be applied to complex underwater scenarios and can accurately estimate AUV's pose in real-time under interference such as low light, weak texture, high dynamics, etc. Moreover, thanks to the acoustic-aided initialization method, the system can achieve fast initialization and obtain better state estimation results even when initial visual information is weakly

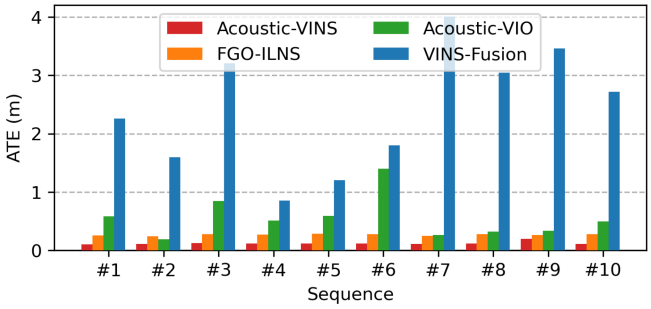

Fig. 7. Absolute Translational Error (ATE) values in ten sequences of the LBL-AQUALOC dataset. The result between the trajectories calculated by VINS-Fusion, FGO-ILNS, Acoustic-VIO, and Acoustic-VINS and the ground truth trajectory of Colmap.

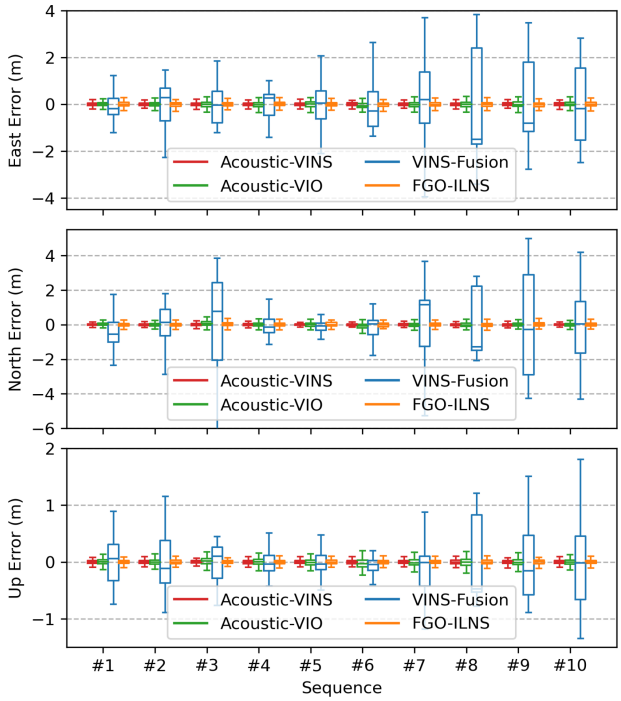

Fig. 8. Positioning error results in the (a) East, (b) North, and (c) Up directions of ten sequences of the LBL-AQUALOC dataset. The positioning errors of VINS-Fusion, FGO-ILNS, Acoustic-VIO, and Acoustic-VINS were compared with Colmap.

textured or highly dynamic. In summary, the proposed Acoustic-VINS can effectively eliminate the impact of challenging visual scenarios and achieve high robustness and high-precision positioning performance.

V. Conclusion

A robust, real-time, acoustic-visual-inertial navigation system for underwater scenarios was presented in this letter. Due to the great challenges posed by low light, weak texture, and high dynamics in underwater environments to visual-inertial systems, we proposed a visual-inertial system that integrates an acoustic LBL system. Moreover, with the acoustic-aided initialization method, the impact of harsh underwater environments on system initialization was eliminated, allowing for tight coupling of acoustics, vision, and inertia in a factor graph optimization framework. In addition, we extended the sensor data of the public AQUALOC dataset by adding LBL data and conducted experiments on ten sequences of the LBL-AQUALOC dataset. The experimental results demonstrated that Acoustic-VINS exhibited superior robustness and accuracy compared to SOTA algorithms in degraded and challenging underwater scenarios.

References

[1] C. Cadena et al., "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, Dec. 2016.

[2] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, "Simultaneous localization and mapping: A survey of current trends in autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 2, no. 3, pp. 194–220, Sep. 2017.

[3] A. Manzanilla, S. Reyes, M. García, D. Mercado, and R. Lozano, "Autonomous navigation for unmanned underwater vehicles: Real-time experiments using computer vision," *IEEE Robot. Automat. Lett.*, vol. 4, pp. 1351–1356, Apr. 2019.

[4] G. Billings, R. Camilli, and M. Johnson-Roberson, "Hybrid visual SLAM for underwater vehicle manipulator systems," *IEEE Robot. Automat. Lett.*, vol. 7, no. 3, pp. 6798–6805, Jul. 2022.

[5] C. Hu, S. Zhu, Y. Liang, and W. Song, "Tightly-coupled visual-inertial-pressure fusion using forward and backward IMU preintegration," *IEEE Robot. Automat. Lett.*, vol. 7, no. 3, pp. 6790–6797, Jul. 2022.

[6] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "Openvins: A research platform for visual-inertial estimation," in *Proc. IEEE Int. Conf. Robot. Automat., Conf. Proc.*, 2020, pp. 4666–4672.

[7] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, pp. 1004–1020, 2018.

[8] T. Qin, S. Cao, J. Pan, and S. Shen, "A general optimization-based framework for global pose estimation with multiple sensors," 2019, *arXiv:1901.03642*.

[9] S. Cao, X. Lu, and S. Shen, "GVINS: Tightly coupled GNSS–visual–inertial fusion for smooth and consistent state estimation," *IEEE Trans. Robot.*, vol. 38, no. 4, pp. 2004–2021, Aug. 2022.

[10] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. 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.

[11] S. Rahman, A. Quattrini 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.

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

[13] E. R. Potokar, K. Norman, and J. G. Mangelson, "Invariant extended Kalman filtering for underwater navigation," *IEEE Robot. Automat. Lett.*, vol. 6, no. 3, pp. 5792–5799, Jul. 2021.

[14] M. M. D. Santos, G. G. D. Giacomo, P. L. J. Drews, and S. S. C. Botelho, "Matching color aerial images and underwater sonar images using deep learning for underwater localization," *IEEE Robot. Automat. Lett.*, vol. 5, no. 4, pp. 6365–6370, Oct. 2020.

[15] T. Zhang, L. Chen, and Y. Yan, "Underwater positioning algorithm based on SINS/LBL integrated system," *IEEE Access*, vol. 6, pp. 7157–7163, 2018.

[16] J. Song, W. Li, X. Zhu, Z. Dai, and C. Ran, "Underwater adaptive height-constraint algorithm based on SINS/LBL tightly coupled," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, 2022.

[17] K. Watanabe, K. Utsunomiya, K. Harada, and Q. Shen, "Development of a floating LBL system and a lightweight ROV for sky to water system," in *Proc. IEEE/MTS SEATTLE*, pp. 1–6, 2019.

[18] M. Ferrera, V. Creuze, J. Moras, and P. Trouvé-Peloux, "AQUALOC: An underwater dataset for visual-inertial-pressure localization," *Int. J. Robot. Res.*, vol. 38, no. 14, pp. 1549–1559, 2019.

[19] 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.

[20] J. Song, W. Li, R. Liu, and X. Zhu, "FGO-ILNS: Tightly coupled multi-sensor integrated navigation system based on factor graph optimization for autonomous underwater vehicle," Oct. 2023, *arXiv:2310.14163*.

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