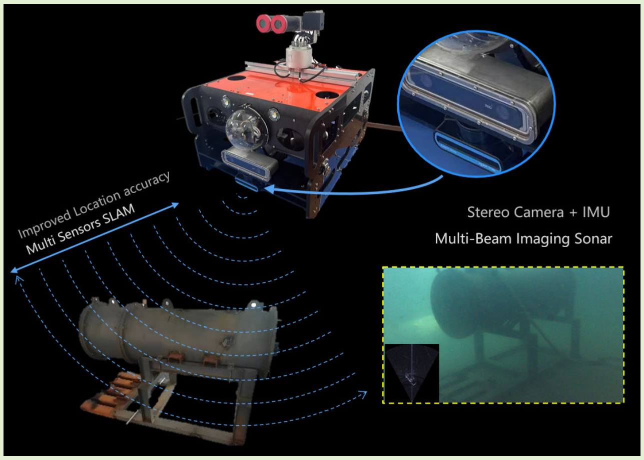
Integration of Sonar and Visual–Inertial Systems for SLAM in Underwater Environments

Jiawei Zhang, Fenglei Han, Duanfeng Han, Jianfeng Yang, Wangyuan Zhao, and Hansheng Li

***Abstract*—Underwater simultaneous localization and mapping (SLAM) encounters challenges in complex environments due to suspended particles, underwater blur, and light and color attenuation. These factors make underwater features less distinct than those in surface images. Moreover, underwater visual features are often unstructured, together with visual failures and low visibility. To address these challenges, a multisensor system is a promising solution. In this article, we introduce a multisensor fusion underwater SLAM method, integrating stereo vision, multibeam imaging sonar, and inertial measurement unit (IMU) data. Our system comprises a visual–inertial subsystem and an acoustic–inertial subsystem. These subsystems collaborate when common features are detected. If one subsystem fails, the other can function independently. The visual–inertial subsystem uses depth information from imaging sonar to optimize error correction during feature tracking. Furthermore, we have optimized the initialization process by matching visual and sonar images and introduced a novel method for depth estimation from sonar images. This dual-sensor strategy improves the system's robustness and adaptability to diverse challenging underwater conditions. Through experiments, we have demonstrated the excellent performance of our algorithm.**



***Index Terms*—Imaging sonar, marine engineering, multisensors' simultaneous localization and mapping (SLAM), optical sensors, stereo vision.**

I. Introduction

In RECENT years, the importance of ocean research and exploration has grown, driven by the global increase in demand for underwater resources, such as oil, gas, minerals, and seafood [1]. At present, underwater mapping often relies on divers with handheld sensors for data gathering, followed by data postprocessing. Remotely operated vehicles (ROVs) offer a reliable method for automating these mapping tasks. Essential in ocean exploration, ROVs find extensive use in areas such as seafloor geology, archeology, and biology. However, the navigation of ROVs remains a particularly challenging task. Most underwater localization and mapping algorithms rely on acoustic sensors, including Doppler velocity log (DVL) [2], ultrashort baseline (USBL) systems [3], and sonars [4]. Nevertheless, the cost of data collection is high due to the highly unstructured nature of underwater environments. In addition, in exploring underwater structures and caves, acoustic positioning devices may be impractical due to obstructions. Recently, numerous vision-based state estimation algorithms have emerged, mainly for use in terrestrial environments [5], [6]. These algorithms typically use single-camera, stereo-camera, or multicamera systems. To improve attitude estimation in complex environments, vision is often combined with inertial measurement units (IMUs), leading to the development of visual and visual–inertial simultaneous localization and mapping (SLAM). However, underwater environments are inherently more complex than terrestrial ones [7]. The scattering and absorption of light by water severely degrade the quality of underwater imagery. As depth increases, red light quickly attenuates, causing images to lean toward blue–green hues, significantly affecting feature extraction accuracy. Moreover, underwater images are often blurred and noisy, further reducing the performance of vision-based SLAM systems. Moreover, the unstructured nature of underwater targets presents significant challenges for vision-based SLAM under these conditions [8].

Furthermore, during ROV missions, there are times when vision is ineffective due to the nature of underwater operations. This results in the camera either failing to capture useful data

Manuscript received 1 February 2024; revised 23 March 2024; accepted 29 March 2024. Date of publication 8 April 2024; date of current version 15 May 2024. This work was supported in part by the Natural Science Foundation of Heilongjiang Province of China, under Grant LH2021E047 and in part by the National Key Research and Development Program of China under Grant 2022YFB3306200. The associate editor coordinating the review of this article and approving it for publication was Dr. Yun-Ju Lee. *(Corresponding author: Fenglei Han.)*

Jiawei Zhang, Fenglei Han, Duanfeng Han, Jianfeng Yang, and Wangyuan Zhao are with the College of Shipbuilding Engineering, Harbin Engineering University, Harbin 150001, China (e-mail: fenglei_han@hrbeu.edu.cn).

Hansheng Li is with Instituto Superior Técnico, University of Lisbon, 1049-001 Lisboa, Portugal.

Digital Object Identifier 10.1109/JSEN.2024.3384301

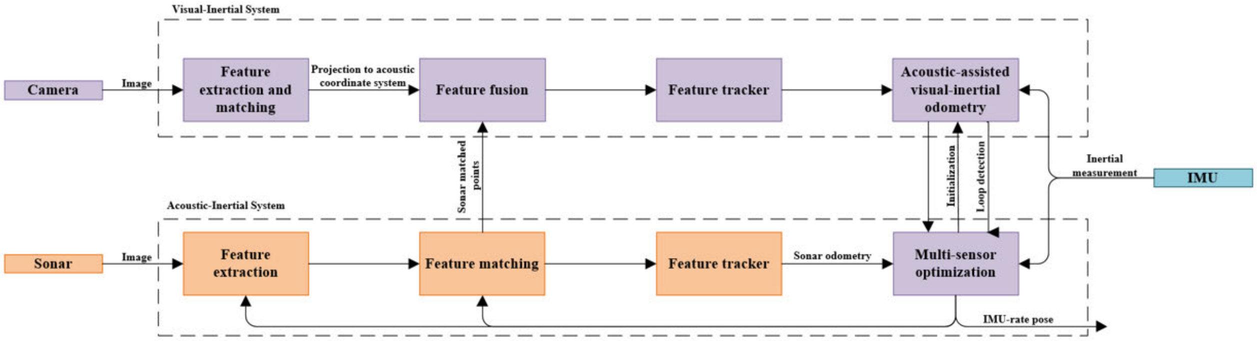


Fig. 1. System structure of our method. This system, which integrates a stereo vision, a multibeam imaging sonar, and an IMU, comprises a visual–inertial subsystem and an acoustic–inertial subsystem. The two subsystems operate independently and can utilize each other's information to enhance the system's accuracy and stability.

or only obtaining visual information at close range because of low visibility. To overcome this challenge, sonar is primarily used [9], [10]. Recently, acoustic SLAM systems have received increased attention in underwater robot localization research. Lee et al. [11] introduced an extended Kalman filter (EKF)-based method for imaging sonar localization, while Wang et al. [12] proposed a belief propagation-based approach for imaging sonar SLAM. Although sonar systems are widely used for underwater localization and mapping, they typically offer limited spatial resolution and are affected by multipath propagation and acoustic reflections. Sonar systems cannot provide sufficient feature points for detailed map construction. Vision systems, offering rich environmental details, perform poorly in low-light and high-turbidity conditions. SLAM systems in underwater environments require highly precise and stable localization capabilities to guide autonomous underwater vehicles (AUVs) or ROVs. Existing solutions often struggle to meet the demands for high accuracy and long-term stability, especially in deep-water environments where GPS signals are unavailable.

In summary, the unique characteristics of underwater environments and operations mean that using a single-sensor type has both limitations and advantages. Therefore, integrating multiple sensors' strengths in SLAM systems significantly enhances performance. Leveraging the complementary information from various sensors, a multisensor fusion approach improves the robustness, accuracy, and reliability of SLAM systems in challenging underwater conditions. This sensor synergy enables more comprehensive perception and mapping, enhancing the effectiveness and reliability of underwater exploration and mapping tasks. Typically, an ROV is outfitted with various sensors, such as cameras, multibeam imaging sonar, IMU, and DVL. However, relying solely on a monocular camera presents initialization challenges due to scale ambiguity, necessitating stable operation of the vehicle. Meeting monocular visual SLAM initialization requirements in complex underwater environments can be challenging [13].

To overcome these challenges and fulfill the practical needs of underwater operations, this work introduces a multisensor fusion underwater SLAM method, integrating stereo vision, multibeam imaging sonar, and IMU data. Our method aims to maintain continuous and stable localization in environments with sparse textures, significant lighting changes, and potential visual failures. Inspired by LVI-SAM [14], as shown in Fig. 1, our system comprises a visual–inertial subsystem and an acoustic–inertial subsystem. These subsystems collaborate when common features are detected. If one subsystem fails, the other can function independently. The visual–inertial subsystem uses depth information from imaging sonar to optimize error correction during feature tracking. In addition, it can utilize sonar data during initialization, serving as the initial input for both subsystems. This dual-sensor strategy improves the system's robustness and adaptability to diverse challenging underwater conditions. The key contributions of this research include the following.

1) We have developed a tightly coupled imaging sonar-visual–inertial odometry (SVIO) framework based on factor graphs. This framework facilitates the fusion of data from multiple sensors and capitalizes on the advantages offered by each sensor. Thanks to our dual-subsystem design, our framework can sustain stable localization even in the event of visual failures.
2) Establishing correspondence between visual and sonar images and creating joint feature-depth estimation from vision and sonar data improve the accuracy of initialization and positioning. This integration of visual and sonar information leads to enhanced localization precision.
3) We introduced a method to estimate sonar image elevation angles, thereby enhancing the isometric stability and accuracy of the sonar-inertial subsystem by providing the missing elevation angle information.

II. Related Work

A. Filter-Based Underwater SLAM

Filter-based SLAM is a pivotal technique in robotics and autonomous navigation, enabling devices to simultaneously map their environment and determine their location within it [15]. This approach typically employs recursive filters, such as the EKF or the particle filter, to fuse noisy sensor data from various sources in real time. By approximating the system's state and continuously updating it with new measurements, filter-based SLAM can track the robot's position and orientation while constructing or updating a map of the

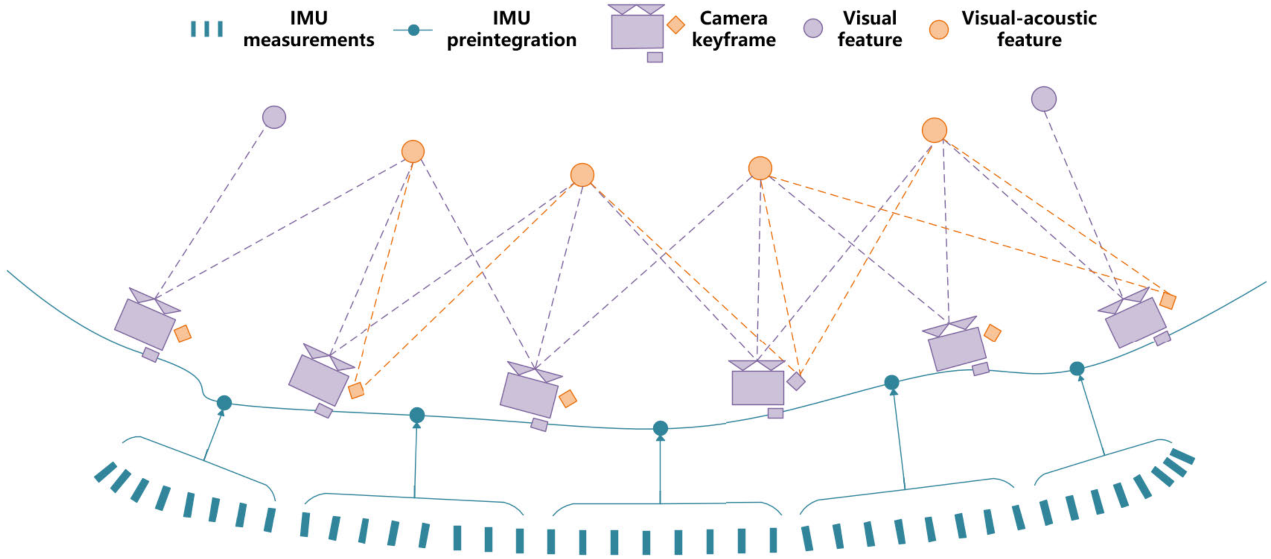


Fig. 2. Visual–inertial–acoustic SLAM framework.

environment. Although challenged by nonlinearities and the need for accurate initial estimates, its efficiency and relatively low computational requirements make filter-based SLAM a popular choice for applications with limited processing power, such as small robots and drones.

For Kalman-based SLAM, Yuan et al. [16] introduced AEKF-SLAM, an innovative underwater SLAM approach using an augmented EKF to enhance navigation accuracy and robustness for underwater robotics. Wang et al. [17] presented an enhanced SLAM algorithm using an adaptive unscented Kalman filter (AUKF) with a noise statistic estimator to address accuracy decline and prevent divergence in traditional UKF-SLAM systems, incorporating a Sage–Husa estimator for real-time noise assessment and a covariance matching technique for stability, thereby significantly improving navigation precision demonstrated through UUV sea trial outcomes. Xing et al. [18] introduced a novel self-localization system for small-sized underwater robots in GPS-denied, structured environments, utilizing a multisensor fusion approach with low-cost sensors and an EKF to integrate data from an IMU, optical flow, pressure sensor, and ArUco markers for precise positioning.

For particle-based SLAM, Chen et al. [19] introduced a Rao–Blackwellized particle filter (RBPF) SLAM algorithm tailored for AUVs using a slow-scanning mechanically scanning imaging sonar (MSIS). He et al. [20] introduced PSO-UFastSLAM, an innovative approach that enhances the unscented-FastSLAM (UFastSLAM) framework with particle swarm optimization (PSO) for underwater mobile robot autonomy.

B. Factor Graph-Based Underwater SLAM

Factor graph-based SLAM surpasses filter-based methods, such as EKF-SLAM in precision, robustness, and scalability. It utilizes optimization for nonlinear estimations, enhancing accuracy in complex environments. Global optimization allows revising past data, reducing errors, and improving overall accuracy. Its scalability supports large-scale mapping, while its flexibility facilitates integrating diverse sensors and constraints. Importantly, it efficiently incorporates loop closure for improved map consistency. This approach offers a superior solution for extensive and accurate environmental mapping, making it ideal for applications demanding high levels of precision and reliability in complex scenarios.

For underwater factor graph-based SLAM, Rahman et al. [21], [22] proposed the integration of sonar sensor and depth sensor into the OKVIS SLAM system, resulting in SVIN. Xu et al. [23] built a new map using the state estimation from DVL within the framework of ORB-SLAM2 and seamlessly connected it to the previous map, thus enhancing the system's robustness. Vargas et al. [24] integrated DVL into the ORB-SLAM3 framework, utilizing motion estimates from both DVL and vision to formulate visual–acoustic residuals, in addition to reprojection errors. They applied visual–acoustic joint optimization in the tracking and local mapping threads of ORB-SLAM3, replacing its original reliance solely on visual data. Lagudi et al. [25] presented a multisensor registration approach for integrating 3-D data from a stereovision system and a 3-D acoustic camera, addressing the challenges of aligning heterogeneous sensor data in close-range acquisition.

III. Methodology

The proposed system is described in this section, as depicted in Fig. 2. The system consists of imaging sonar, stereo camera, IMU, and depth sensor. Owing to the diminished visibility and the presence of dynamic impediments within the aquatic environment, the task of identifying robust features for tracking purposes poses a significant challenge. Beyond the intrinsic constraints of subaqueous vision, such as the attenuation of light and chromatic shifts, vision-based systems are further compromised by the prevalent issue of diminished contrast. To address this, our methodology fortifies the processing pipeline by integrating a sophisticated image preprocessing phase. This phase employs a composite approach of retinex theory and a curvature filter for image enhancement, designed to substantially elevate the discernibility of underwater features [26]. After describing the state variables, we expound upon the multifaceted components of our proposed framework. This includes a detailed introduction of the initialization procedure designed for robust estimation of initial parameters. We then delve into the intricacies of our sensor fusion optimization strategy, which underpins the efficacy of the system in synthesizing data from disparate sensors. Furthermore, we introduce the mechanisms of loop closure and relocalization, which are crucial for maintaining the consistency and accuracy of the system over time. These components collectively contribute to a comprehensive solution aimed at addressing the challenges inherent in dynamic and complex.

A. State Variables

The coordinate systems of the sensor system can be categorized as follows: camera, imaging sonar, IMU, and world that are denoted as C, S, b, and w, respectively. The transformation matrix between two coordinate frames from j to i $\mathbf{T}_{ij} \in \mathrm{SE}(3)$ is defined as $\mathbf{T}_{ij} = [\mathbf{q}_{ij} | \mathbf{t}_{ij}]$, where $R_{ij} \in \mathrm{SO}(3)$ is the rotation matrix that can be represented by quaternion $\mathbf{q}_{ij}$ and $\mathbf{t}_{ij}$ is the translation vector. For the sake of simplifying calculations, we have

$$\mathbf{t}_{CS} = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \quad \mathbf{R}_{CS} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}. \tag{1}$$

The system executes a sliding-window approach, so the system state $\boldsymbol{x}$ can be written as

$$\boldsymbol{x} = [\mathbf{p}_{wb}, \mathbf{q}_{wb}, \mathbf{v}^{w}, \mathbf{b}^{g}, \mathbf{b}^{a}]^{\mathrm{T}} \tag{2}$$

where $\mathbf{v}$ means the linear velocity and $\mathbf{b}_g$ and $\mathbf{b}_a$ mean the gyroscopes and accelerometer bias.

In a sonar system, a 3-D point $\mathbf{p}(x, y, z)$ in world can be denoted as $\mathbf{p}_S = (x_S, y_S, z_S)^T$. The Cartesian and spherical coordinates $(\mathcal{R}, \theta, \phi)$ can be transformed as

$$\mathbf{p}_S = \begin{bmatrix} x_S \\ y_S \\ z_S \end{bmatrix} = \mathcal{R} \begin{bmatrix} \cos\phi \sin\theta \\ \cos\phi \cos\theta \\ \sin\phi \end{bmatrix} \tag{3}$$

and

$$\begin{bmatrix} \mathcal{R} \\ \theta \\ \phi \end{bmatrix} = \begin{bmatrix} \sqrt{x_S^2 + y_S^2 + z_S^2} \\ \tan^{-1}(x_S / y_S) \\ \tan^{-1}\left(z_S / \sqrt{(x_S^2 + y_S^2)}\right) \end{bmatrix} \tag{4}$$

where θ and ϕ mean azimuth and elevation angels, respectively. In an imaging sonar, the elevation angle ϕ is lost. Thus, we have sonar image coordinates as

$$\begin{bmatrix} x_s \\ y_s \end{bmatrix} = \begin{bmatrix} \mathcal{R} \sin\theta \\ \mathcal{R} \cos\theta \end{bmatrix}. \tag{5}$$

$(x_s, y_s, 1)^T$ is the sonar image coordinates, where $x_s = \mathcal{R} \sin\theta$ and $y_s = \mathcal{R} \cos\theta$. Thus, (3) can also be written as

$$\mathbf{p}_S = \begin{bmatrix} x_S \\ y_S \\ z_S \end{bmatrix} = \begin{bmatrix} x_s \cos\phi \\ y_s \cos\phi \\ \sqrt{x_s^2 + y_s^2} \sin\phi \end{bmatrix}. \tag{6}$$

In the camera system, the point $\mathbf{p}$ can be denoted as $\mathbf{p}_c = (x_C, y_C, z_C)^T$. $(u, v, 1)^T$ is the image coordinate result of using perspective projection to project a point from the camera coordinate system to the image coordinate system. The mapping from the 3-D camera coordinate system to the 2-D image coordinate system is described by the perspective projection equation

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \mathbf{p}_C = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_C \\ y_C \\ z_C \end{bmatrix} \tag{7}$$

where (f_x, f_y) are the camera's focal lengths, (c_x, c_y) is the optical center of the image, and λ is a scale factor, typically effectively equal to the distance z_C from the object point to the camera center.

Given a calibrated camera and sonar points and knowing the rotation $(\mathbf{R}_{SC})$ and translation $(\mathbf{t}_{SC})$ parameters, we can achieve

$$\begin{aligned} \mathcal{R} &= ||\mathbf{p}_S|| = ||\mathbf{R}_{SC}\mathbf{p}_C + \mathbf{t}_{SC}|| \\ &= \sqrt{||\mathbf{p}_C||^2 + 2\mathbf{t}^T \mathbf{R}_{SC}\mathbf{p}_C + ||\mathbf{t}_{SC}||^2}. \end{aligned} \tag{8}$$

Therefore, there is

$$||\mathbf{p}_C||^2 + 2\mathbf{t}_{SC}^T \mathbf{R}_{SC}\mathbf{p}_C + (||\mathbf{t}_{SC}||^2 - ||\mathbf{p}_S||^2) = 0/ \tag{9}$$

By converting from the camera coordinate system to the image coordinate system, we have

$$||K^{-1}\mathbf{p}_c||^2 z_C^2 + 2\left(\mathbf{t}_{SC}^T \mathbf{R}_{SC} K^{-1}\mathbf{p}_c\right) z_C + ||\mathbf{t}_{SC}||^2 - \mathcal{R}^2 = 0. \tag{10}$$

B. Imaging Sonar and Stereo Camera-Associated Feature Depth

Leveraging the synergistic capabilities of imaging sonar and stereo cameras, our approach effectively resolves the long-standing challenge of feature depth estimation, particularly in complex underwater environments. The integration of these two sensor modalities enables the system to overcome the scale ambiguity typically associated with stereo vision systems. By associating the absolute distance measurements from the imaging sonar with the relative depth information inferred from stereo cameras, we can accurately determine the feature depth with enhanced precision.

The first step involves detecting features in both the imaging sonar data and the stereo camera imagery. In stereo camera data, features are typically identified based on visual markers or distinct environmental points. In contrast, imaging sonar detects features based on the acoustic reflections from objects in the water. Underwater conditions often lead to feature degradation due to factors such as scattering, absorption, and limited visibility. For the above reason, our system utilizes a learning-based feature detector SuperPoint's neural network architecture [27], which is adept at identifying robust features in such degraded images, outperforming traditional feature detectors such as ORB.

Once features are detected, the next critical step is to associate corresponding features across the two sensor modalities. As shown in Algorithm 1, to associate the feature in two

Algorithm 1 Imaging Sonar and Stereo Camera Feature Correlation Algorithm

```
Input: Visual landmark set $\mathcal{L}_v$ at time k
       Sonar feature set $\mathcal{L}_s$ at time k
       Distance threshold D
Output: Visual sonar joint landmark set $\mathcal{L}_s$ at time k
    /*Create list for visual sonar joint landmark set*/
 1: $\mathcal{L}_{joint} = \emptyset$
    /*Compute distance from visual landmark to sonar landmark*/
 2: for $l_i$ in $\mathcal{L}_v$ do
    /*visual landmark in sonar coordinate*/
 3:   $l'_i = \mathbf{T}_{SC}\mathbf{T}_{Cw}l_i = [x, y, z]^T$
 4:   $l''_i = \mathbf{T}_{sS}l'_i = [x_s, y_s]^T$
 5:   for $l_j$ in $\mathcal{L}_s$ do
 6:     $d = ||l_j - l''_i||$
 7:     if $d < D_{min}$ then
 8:       $D_{min} = d$
 9:       $l'_j = l_j$
10:     end if
11:   end for
12:   if $D_{min} < D$ then
13:     $\mathcal{L}_{joint} = \mathcal{L}_{joint} \cup l'_j$
14:   end if
15: end for
16: return $\mathcal{L}_{joint}$
```

sensors, we project the visual landmarks onto the sonar image coordinate system. Assuming that the projected coordinate is $(x, y, z)^T$ in the sonar coordinate system, the coordinates of this point in the sonar image would be

$$x'_s = \frac{x\sqrt{x^2+y^2}}{\sqrt{x^2+y^2+z^2}}$$
$$y'_s = \frac{y\sqrt{x^2+y^2}}{\sqrt{x^2+y^2+z^2}}. \tag{11}$$

For consistency in writing, it is represented by $\mathbf{T}_{sS}$ in line 4 in Algorithm 1.

For every visual landmark, the algorithm computes the Euclidean distance to each sonar landmark, retaining the closest match within a predefined minimum distance threshold. If this minimum distance is less than the set threshold, the corresponding sonar landmark is added to the joint landmark set. The final output is this joint set of correlated visual and sonar landmarks, which is critical for enhancing the accuracy of underwater navigation and mapping.

C. Feature Depth Estimation Using Sonar Image

Fig. 3 illustrates two sonar image frames in which the same target point is detected. R and t denote the rotation matrix and translation vector, respectively, that describe the transformation between two frames in the world coordinate system. We observe the acoustic projection methodology utilized to translate a point's position from the physical environment onto the sonar imaging plane. The solid line delineates the true location of the observed point, while the dashed line represents its acoustic projection within the constraints of sonar imaging, specifically where the elevation angle ϕ equals zero. This results in the point's apparent position being depicted on the imaging plane along the dashed arc trajectory. This indicates that the matched point is the intersection of two arcs. By solving the equations of these arcs simultaneously, we can determine the coordinates of the intersection point.

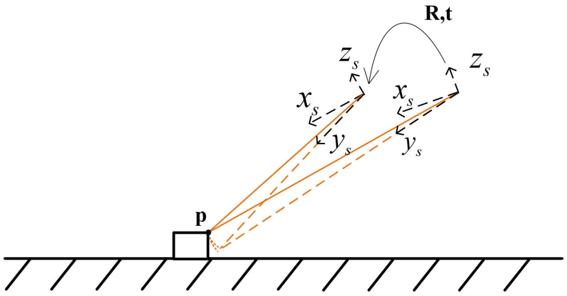


Fig. 3. Two sonar image frames in which the same target point p is detected.

Assuming the existence of two arcs, $\mathcal{A}_1$ and $\mathcal{A}_2$, which represent observations of a matched point from two consecutive sonar frames, then these arcs can be described as follows:

$$\mathcal{A}_1 = \begin{bmatrix} x_{s_1}\cos\phi_1 \\ y_{s_1}\cos\phi_1 \\ \mathcal{R}_1\sin\phi_1 \end{bmatrix}, \quad \mathcal{A}_2 = \begin{bmatrix} x_{s_2}\cos\phi_2 \\ y_{s_2}\cos\phi_2 \\ \mathcal{R}_2\sin\phi_2 \end{bmatrix}. \tag{12}$$

The arcs $\mathcal{A}_1$ and $\mathcal{A}_2$, representing observations from consecutive sonar frames, can be transmitted to the same coordinate system by the equation $\tilde{\mathcal{A}}_2 = R\mathcal{A}_2 + t$, where R is the rotation matrix and t is the translation vector between the two frames. By minimizing the discrepancy between the arc $\mathcal{A}_1$ and $\tilde{\mathcal{A}}_2$ which is subjected to rotational constraints, the coordinates of an intersection point p can be accurately deduced

$$\min_{\phi} ||\mathcal{A}_1 - \tilde{\mathcal{A}}_2||^2, \quad \textbf{s.t. } \phi \in [-6^\circ, 6^\circ]. \tag{13}$$

Given the sonar's measurement radius ranging from 1 to 10 m, the task of calculating the minimum distance between two arcs can indeed be simplified to finding the minimum distance between their corresponding chords, where the endpoints of the chords are identical to the endpoints of the arcs

$$\mathcal{A} = \begin{bmatrix} x_s\cos\phi \\ y_s\cos\phi \\ \mathcal{R}\sin\phi \end{bmatrix} \rightarrow \mathbf{L} = \mathbf{a} + t\mathbf{b} \tag{14}$$

where $\mathbf{a} = [x_s\cos(-6^\circ), y_s\cos(-6^\circ), \mathcal{R}_1\sin(-6^\circ)]^T$ and $\mathbf{b} = [0, 0, 2\mathcal{R}\sin(6^\circ)]^T$ This approximation is valid within this context as the curvature of the arcs is relatively slight due to the short radius, allowing the chords to serve as a close proxy for the actual arcs, thus simplifying the computational process while retaining a high level of accuracy for most practical applications in sonar-based systems. Equation (14) can be modified as follows:

$$\min_{t} ||\mathbf{L}_1 - \tilde{\mathbf{L}}_2||^2, \quad \textbf{s.t. } t \in [0, 1]. \tag{15}$$

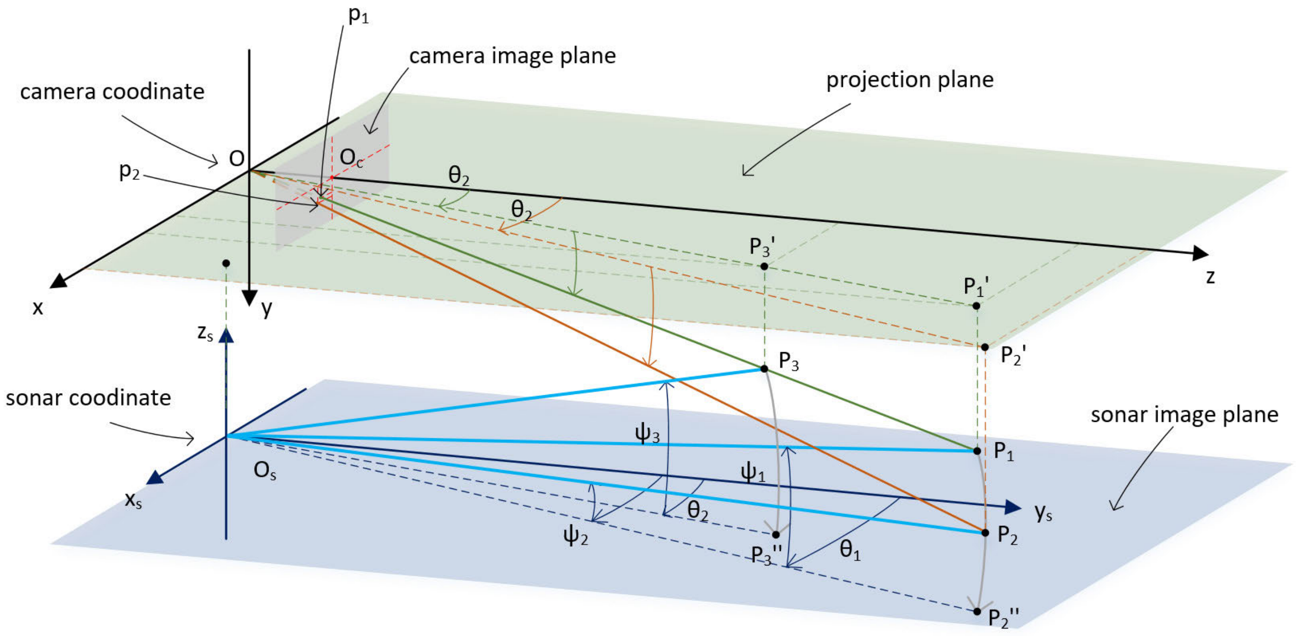


Fig. 4. Joint coordinate system for left camera and sonar.

D. Multisensor Initialization

The goal of system initialization is to get the initial values such as velocity, gravity direction, and IMU bias. For complex, tightly coupled nonlinear systems used in underwater applications, having a strong and accurate initialization is crucial for the system's success. Incorrect initialization can lead to errors and divergence in the system's operation. Underwater conditions often present obstacles for visual systems, such as occlusion (blocked vision) and a lack of distinct features to track. Meanwhile, the underwater currents make it difficult for the robot to maintain a stable posture during the initialization process. These make initialization more challenging compared to land or air environments.

To address these challenges, a new method of initialization is proposed. This method utilizes data from three different types of sensors: stereo cameras, IMUs, and imaging sonar. Due to the robot's slow movement velocity and a detection distance of 10 m, the delay in sonar information caused by the speed of sound can be disregarded. Since imaging sonar is a type of forward-looking sonar, it always maintains a covisibility relationship with visual information. Meanwhile, the sonar data, which provide absolute distance measurements, are used to calibrate and refine the relative depth information obtained from the stereo camera. This process helps resolve the scale ambiguity in the stereo vision and provides a more accurate depth estimation for the detected features. The combination of these sensors helps to introduce scale constraints that are crucial for accurate initial estimations of the system's state.

For initial tracking, the number of visual feature points should be greater than the minimum value of 15, which is set in the system. As shown in Fig. 4, for the keyframe of vision and the paired sonar image, taking the optical center of the camera as the starting point, the ray passing through the projection point of the matching point on the image plane intersects at the matching point with the arc on the imaging sonar passing through the matching point and the projection point on the sonar image. Thus, by constructing the intersection equation, we can solve for the coordinates of the matching point. First, convert the matched points in the camera coordinate system and the sonar coordinate system to the world coordinate system

$$\mathbf{L}_S = T_{wC}T_{CS}\mathbf{p}_S, \quad \mathbf{L}'_C = \lambda T_{wC}K^{-1}\mathbf{p}_c. \tag{16}$$

Once the points from the camera and sonar systems are transformed into the world coordinate system, they can be used together to determine the location of the matched points in the world frame

$$\mathbf{L}'_C - \mathbf{L}_S = 0. \tag{17}$$

However, due to the existence of noise and errors, this equation is not necessarily strictly 0. Therefore, the least squares form is used instead of the zero solution

$$\begin{aligned}&\min_{\lambda\phi} \left\|\mathbf{L}'_C - \mathbf{L}_S\right\|^2\\ &\textbf{s.t. } \phi \in [-6^\circ, 6^\circ], \quad \lambda > 0.\end{aligned} \tag{18}$$

After computing the values of λ and ϕ, we can determine the coordinates of the matched points. After that, we have

$$\mathbf{p}_w = \omega\mathbf{L}_C + (1-\omega)\mathbf{L}'_C \tag{19}$$

where ω is the weighting factor that balances the contribution of each set of coordinates. By adjusting ω, we can control how much influence each coordinate set has on the final position of $\mathbf{p}_w$

For the calibration of the extrinsic parameters between the camera and the IMU, as well as the estimation of IMU biases, gravity direction, and initial velocity, the methodology presented in the VINS-Mono [13] can be referenced.

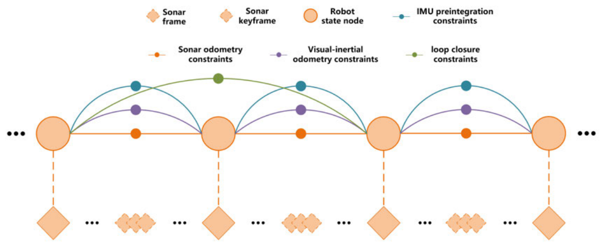

Sonar frame
Sonar keyframe
Robot state node
IMU preintegration constraints
Sonar odometry constraints
Visual-inertial odometry constraints
loop closure constraints
...
...
...
...
...
...
...
...

Fig. 5. Illustration of relative sonar and visual–inertial factor graph.

E. Tightly Coupled SLAM With Visual, Inertial, and Imaging Sonar

In the context of tightly coupled SLAM integrating visual, inertial, and imaging sonar data, we employ a refined approach using factor graphs for nonlinear optimization, allowing for incremental backend optimization. As shown in Fig. 5, the full state vector of the optimization problem is given the following:

$$\min_{\mathcal{X}}\left\{\sum_{i\in B}\|\mathbf{r}_b\|^2_{\Sigma_{b_ib_{i+1}}}+\sum_{(i,j)\in F}\rho\left\|\mathbf{r}_f\right\|^2_{\Sigma^{c_i}_{\mathbf{f}_j}}+\sum_{(i,j)\in A}\rho\|\mathbf{r}_a\|^2_{\Sigma^{c_i}_{\mathbf{f}_j}}\right.$$
$$\left.+\sum_{(i,j)\in S}\rho\|\mathbf{r}_s\|^2_{\Sigma^{s_i}_{\mathbf{f}_j}}\right\} \tag{20}$$

where $\mathbf{r}_b$, $\mathbf{r}_f$, $\mathbf{r}_a$, and $\mathbf{r}_s$ are residuals for IMU, visual measurement, associated sonar-visual measurement, and sonar measurement respectively. B is the IMU measurements set, F is the visual feature set, A is the associated feature set, and S is the sonar feature set. ρ is the Huber kernel core, defined as

$$\rho(s)=\begin{cases}1, & s\geq 1\\ 2\sqrt{s}-1, & s<1.\end{cases} \tag{21}$$

For IMU measurement between frame b_i and b_{i+1}, the residual with preintergated IMU data is shown as follows:

$$\mathbf{r}_b\left(\hat{\mathbf{z}}_{b_{k+1}b_k},\mathcal{X}\right)=\begin{bmatrix}\mathbf{r}_p\\ \mathbf{r}_q\\ \mathbf{r}_v\\ \mathbf{r}_{ba}\\ \mathbf{r}_{bg}\end{bmatrix}$$
$$=\begin{bmatrix}\mathbf{q}_{b_iw}\left(\mathbf{p}_{wb_j}-\mathbf{p}_{wb_i}-\mathbf{v}^w_i\Delta t+\frac{1}{2}\mathbf{g}^w\Delta t^2\right)-\boldsymbol{\alpha}_{b_ib_j}\\ 2\left[\mathbf{q}_{b_jb_i}\otimes\left(\mathbf{q}_{b_iw}\otimes\mathbf{q}_{wb_j}\right)\right]_{xyz}\\ \mathbf{q}_{b_iw}\left(\mathbf{v}^w_j-\mathbf{v}^w_i+\mathbf{g}^w\Delta t\right)-\boldsymbol{\beta}_{b_ib_j}\\ \mathbf{b}^a_j-\mathbf{b}^a_i\\ \mathbf{b}^g_j-\mathbf{b}^g_i\end{bmatrix} \tag{22}$$

where $\boldsymbol{\alpha}_{b_ib_j}$, $\boldsymbol{\beta}_{b_ib_j}$, and $\mathbf{q}_{b_ib_j}$ are preintergated terms between frame i and $i+1$ and are defined as follows:

$$\boldsymbol{\alpha}_{b_ib_j}=\iint_{t\in[i,j]}\left(\mathbf{q}_{b_ib_t}\mathbf{a}^{b_t}\right)\delta t^2$$
$$\boldsymbol{\beta}_{b_ib_j}=\int_{t\in[i,j]}\left(\mathbf{q}_{b_ib_t}\mathbf{a}^{b_t}\right)\delta t$$
$$\mathbf{q}_{b_ib_j}=\int_{t\in[i,j]}\mathbf{q}_{b_ib_t}\otimes\begin{bmatrix}0\\ \frac{1}{2}\boldsymbol{\omega}^{b_t}\end{bmatrix}\delta t. \tag{23}$$

In the error terms, the displacement $\mathbf{r}_p$, velocity $\mathbf{r}_v$, and bias $\mathbf{r}_b$ are obtained by direct subtraction. The second term is related to the rotational error about the quaternion, where $[\cdot]xyz$ denotes the 3-D vector formed by taking only the imaginary part (x, y, z) of the quaternion.

In the context of visual optimization for SLAM, bundle adjustment (BA) is utilized to refine the 3-D coordinates of points as observed in multiple camera frames. The optimization function here minimizes the reprojection error, which is the discrepancy between the observed keypoint positions in the camera frames and the projected positions of the corresponding 3-D points, i.e., written as

$$\left\|\mathbf{r}_f\right\|^2_{\Sigma^{c_i}_{\mathbf{f}_j}}=\left\|\mathbf{z}^{c_i}_{f_j}-\pi\left(\mathbf{q}_{wc_i},\mathbf{p}_{wc_i},\mathbf{f}_j\right)\right\|^2_{\Sigma} \tag{24}$$

where Σ is the covariance matrix associated with the scale of the keypoint. The projection function $\pi_{(\cdot)}$ is a projection model that maps 3-D points onto the image plane of a rectified stereo camera, considering camera intrinsics such as focal lengths, the principal point, and the baseline of the stereo setup, which are defined as follows:

$$\pi\left(\begin{bmatrix}X\\ Y\\ Z\end{bmatrix}\right)=\begin{bmatrix}f_x\dfrac{X}{Z}+c_x\\ f_y\dfrac{Y}{Z}+c_y\\ f_x\dfrac{X-b}{Z}+c_x\end{bmatrix} \tag{25}$$

where b is the baseline for the stereo camera.

The sonar visual associated optimization term involves integrating sonar measurements with visual data to improve the estimation of the environment and the sensor's pose. This term would account for the differences between the observed sonar data and the expected measurements derived from the visual map and the current state estimate. Referring to (10) and (11), we can extrapolate the method of re-projecting matched points from the word system into the sonar image frame. This process involves transforming the coordinates using the camera's and sonar's extrinsic and intrinsic parameters. The detail can be written as

$$\mathbf{p}'_s=\begin{bmatrix}x'_s\\ y'_s\end{bmatrix}=\begin{bmatrix}\dfrac{z_C\mathbf{r}_1\mathbf{p}_c+t_x}{\sqrt{1-(z_C\mathbf{r}_3\mathbf{p}_c/\mathcal{R}+t_z/\mathcal{R})^2}}\\ \dfrac{z_C\mathbf{r}_2\mathbf{p}_c+t_y}{\sqrt{1-(z_C\mathbf{r}_3\mathbf{p}_c/\mathcal{R}+t_z/\mathcal{R})^2}}\end{bmatrix} \tag{26}$$

where $\mathbf{r}_i(i=1,2,3)$ represents the row vectors of the rotation matrix $\mathbf{R}_{SC}$. Based on the equations referenced, we can articulate the associated term by minimizing the reprojection error. The error is defined as

$$\|\mathbf{r}_a\|^2_{\Sigma^{c_i}_{\mathbf{f}_j}}=\left\|\mathbf{p}_s-\mathbf{p}'_s\right\|^2_{\Sigma} \tag{27}$$

where Σ is the covariance matrix.

For optimizing the sonar term, we adopt an acoustic BA approach, projecting the sonar landmarks back onto the sonar image plane. The optimization function is formulated as follows:

$$\|\mathbf{r}_s\|^2_{\Sigma^{s_i}_{\mathbf{f}_j}} = \left\| \mathbf{z}^{s_i}_{f_j} - \pi\left(\mathbf{q}_{wc_i}, \mathbf{p}_{wc_i}, \mathbf{f}_j\right)\right\|^2_{\Sigma} \tag{28}$$

where Σ is the covariance matrix associated with the scale of the keypoint. The projection function $\pi_{(\cdot)}$ is a projection model that maps 3-D points onto the image plane of an imaging sonar, which is the same as (26). This approach allows us to refine the position of sonar landmarks by minimizing the reprojection error on the sonar image plane, enhancing the overall accuracy of the sonar-based mapping and localization within the SLAM framework.

IV. Experiment

The algorithm's performance underwent a thorough evaluation through a series of controlled tank experiments. During these experiments, the target was intentionally placed at the submerged base of the tank. For in-depth information about the test tank's specifications, the characteristics of the target, and a comprehensive description of the experimental setup, please refer to Section IV-B.

A. Experimental on Public Dataset

In this experiment, our proposed multisensor fusion SLAM algorithm was validated using three publicly available terrestrial standard datasets from EuRoC. We evaluated the algorithm's performance in traditional environments and compared it with the results of other visual–inertial SLAM algorithms. Specifically, we compared our algorithm with ORB SLAM3 [28], a representative feature-based SLAM method, and VINS Fusion [29], a strongly representative direct-based SLAM method. When assessing the performance of an SLAM algorithm, various aspects such as complexity, computation time, and accuracy are considered. Among these factors, accuracy is the most critical indicator. In the context of the TUM dataset construction work, an accuracy metric was introduced: absolute trajectory error (ATE). This metric effectively evaluates the accuracy of SLAM algorithms and is widely applied. ATE can be computed using metrics such as median or mean error, but the root mean square error (RMSE) is typically used to represent these metrics. RMSE provides a comprehensive measure of accuracy, summarizing the overall performance of the SLAM algorithm in terms of trajectory estimation. Sequences in EuRoC are categorized into different levels of difficulty, namely, easy (MH01), medium (MH02 and MH03), or hard (MH04 and MH05), based on factors such as lighting conditions, scene texture, and the movement of the vehicle. We selected one sequence from each difficulty level to evaluate the three algorithms. The evaluation results are presented in Table I.

From the above results, it is evident that our proposed multisensor fusion underwater SLAM algorithm can successfully operate on standard datasets and is on par with other vision-based methods. It consistently provides accurate trajectory estimation.

TABLE I
RMSE Results for EuRoC Sequence in Meter

	MH01	MH03	MH04
vins fusion	0.060	0.330	0.780
Orbslam3	0.022	0.027	0.089
Ours	0.018	0.028	0.119

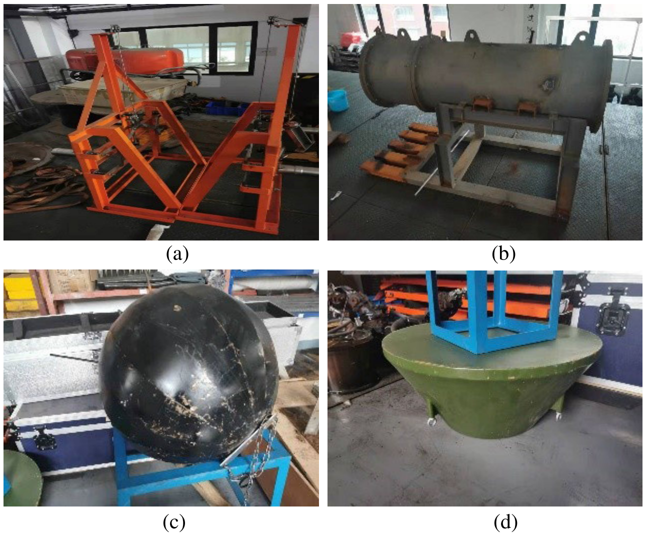

(a)
(b)
(c)
(d)

Fig. 6. Structures used in the experiment. (a) Shelf. (b) Tank. (c) Ball. (d) Platform.

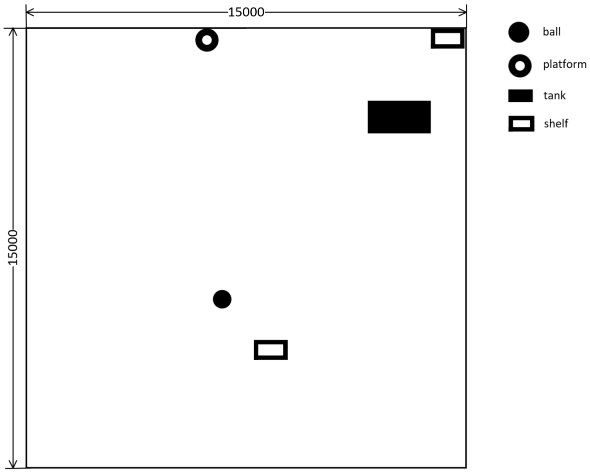


Fig. 7. Arrangement of the structures.

B. Experimental Environment and Equipment

To evaluate our proposed method, we collected datasets from a water tank.

The experimental site for the underwater tank experiments was situated at the Ocean Underwater Engineering Science Research Institute of Shanghai Jiao Tong University. The tank boasted a substantial depth of 6 m and spacious dimensions of 15 m in both length and width, rendering it ideally suited to accommodate the demands of algorithmic simulation. To simulate scenarios where the ROV operates within a submerged structure environment, four different types of structures as shown in Fig. 6 were strategically positioned underwater. The specific layout and placement locations are depicted in Fig. 7

TABLE II
LOCATION OF THE STRUCTURES

	platform	ball	shelf1	tank	shelf2
Location(m)	(6.1,14.6)	(6.6,5.7)	(8.3,4.2)	(12.7,11.9)	(14.3,14.6)

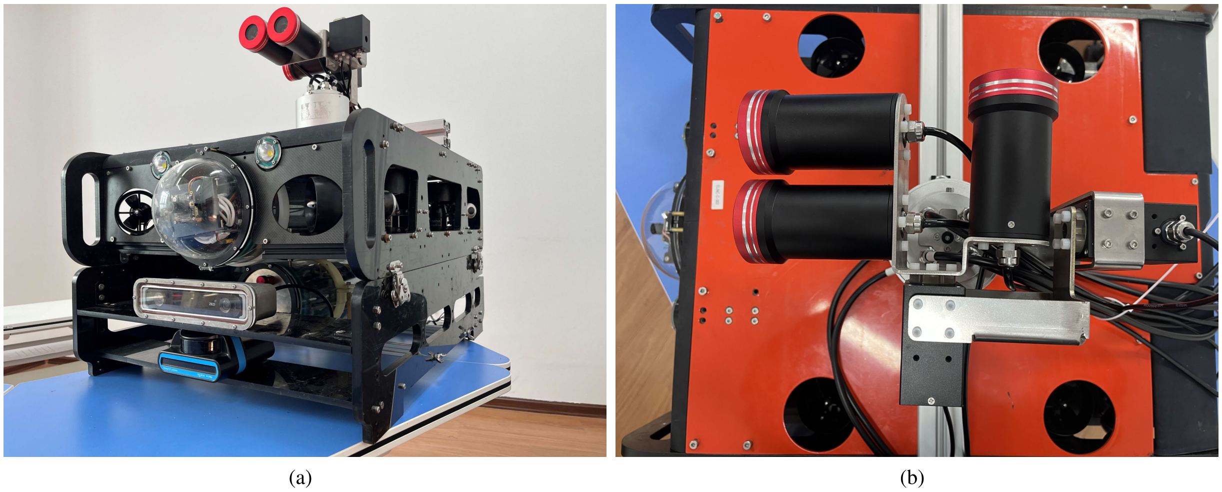
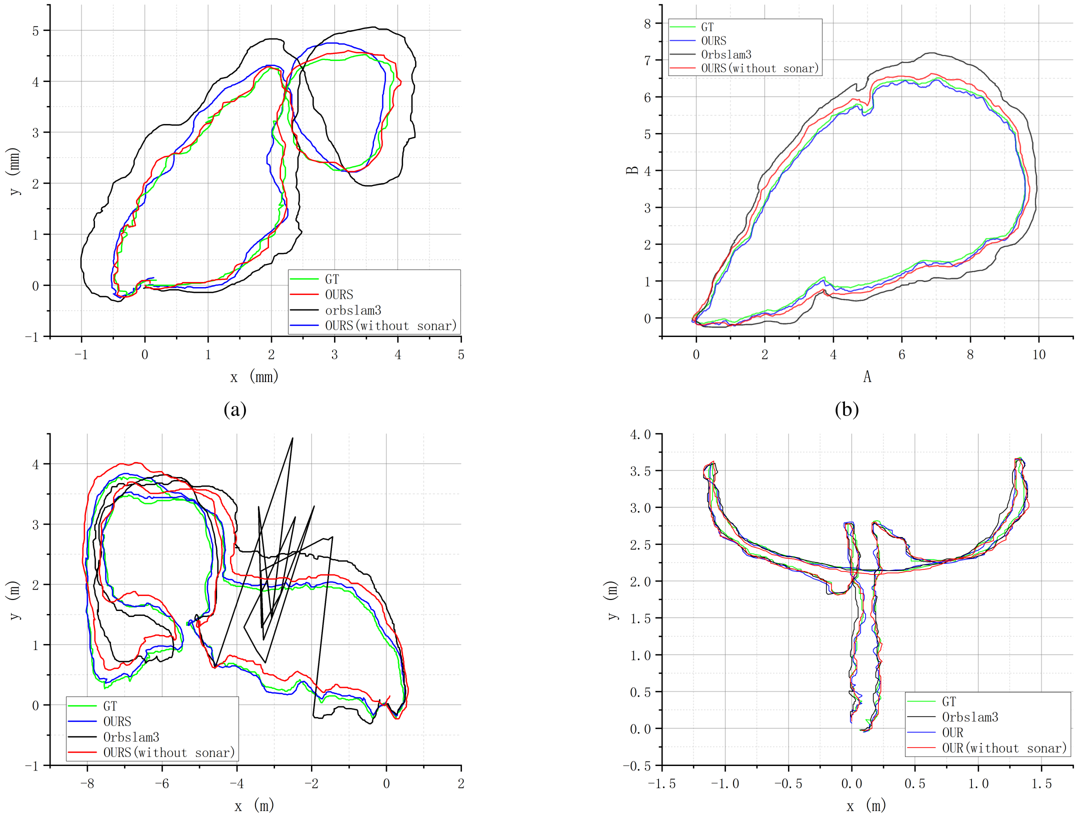

(a) (b)

Fig. 8. Equipment used in the experiment (a) ROV platform with stereo camera, imaging sonar, IMU, and positioning equipment. (b) Positioning equipment.

and detailed in Table II (The origin position is located in the bottom-left corner.). It is worth noting that these structures were divided into two operational zones, intentionally positioned out of each other's visual sensor range. This setup aimed to emulate the ROV's capability for cross-zone operations and simulate operational scenarios where camera failure occurs.

The ROV system used for the experiments in Fig. 8(a) was equipped with a comprehensive array of hardware and software components to ensure precise navigation and data collection. In addition to the eight thrusters for propulsion, it incorporated several essential sensors and devices: stereo cameras, imaging sonar, IMU, depth sensor, and laser rangefinders. Three laser rangefinders were employed for precise positioning and distance measurement, contributing to obstacle avoidance and navigation accuracy. Among these, the ROV was equipped with a ZED2i stereo camera and an IMU, while it utilized the Blueprint Oculus Multibeam Sonar M1200D for multibeam sonar operations. For efficient data collection and processing, the robot system was configured with the robot operating system (ROS), a versatile and widely used framework for robotics applications. The onboard processing unit was NVIDIA Xavier, a powerful processor, responsible for executing SLAM operations, which allowed the ROV to create maps of its surroundings and determine its position in real time.

In typical scenarios, obtaining ground-truth (GT) data in a water tank can be challenging due to environmental constraints. To address this issue, we employed the experimental setup depicted in Fig. 8(b). The system is equipped with a trio of laser rangefinders: two are positioned to face forward and one is directed to the right. These rangefinders are strategically arranged in two groups, with one group aligned vertically with the other. This configuration is crucial for the dual purpose of distance measurement and maintaining a stable orientation relative to a reference, such as a wall or underwater structure. To ensure precise control of the device's orientation, the system is integrated with two servo motors. These servos are responsible for adjusting the pitch and roll angles, thereby compensating for any tilts or rotations that could affect the accuracy of the measurements. Moreover, a brushless direct current (BLDC) motor is employed to preserve a constant heading, countering the effects of underwater currents and drift that can lead to rotational discrepancies. The innovative aspect of this system lies in its utilization of the pair of forward-facing laser rangefinders to ascertain and stabilize the device's heading. By continuously comparing the distance readings from both lasers, the system can detect any angular deviation from its intended direction. If one laser reports a greater distance than the other, it indicates that the system is beginning to veer off course. To rectify this, the system actively adjusts its orientation to equalize the readings from both lasers, thereby ensuring that the heading remains unchanged. Through meticulous experimental calibrations, the refractive index of water has been factored into the system with a fine-tuned constant of 0.7375. This calibration is critical for the lasers to yield precise measurements within a 1% accuracy margin, which is of paramount importance for assessing the feasibility of using lasers underwater. The refraction correction allows for reliable distance readings up to a maximum range of 8 m. However, underwater scattering—caused by particles in the water—can severely affect the range and accuracy of laser measurements. This scattering phenomenon is highly variable and can be influenced by the turbidity and particulate matter present in the water. To circumvent the limitations imposed by scattering and to extend the operational range of the device, a strategic mechanism is in place. When the rangefinders approach their maximum effective range, the entire measurement apparatus



(a)

(b)

(c)

(d)

Fig. 9. Underwater experiment without visual failure. (a) Seq. 01. (b) Seq. 02. (c) Seq. 03. (d) Seq. 04.

is designed to rotate by 90°. This pivotal maneuver enables the system to continue acquiring accurate distance data even beyond the 8-m mark, up to 15 m.

C. Experimental on Underwater Dataset

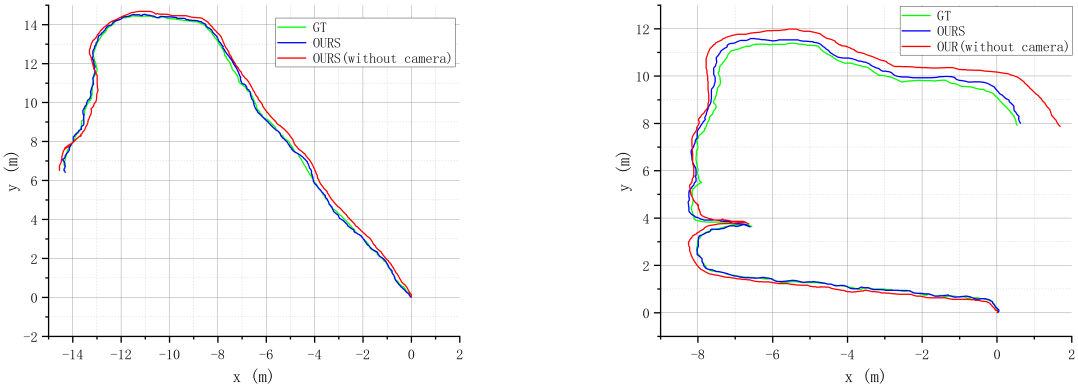
The testing of the underwater dataset is primarily divided into two parts: one involves scenarios where there is no visual failure during the operation, while the other simulates situations with visual failures during the operation. These two parts aim to assess the performance of the SLAM system in real operational environments under different conditions. The underwater dataset comprises a total of six sequences, with sequences 1–4 representing datasets without visual failures, while sequences 5 and 6 simulate underwater scenarios with visual.

1) Underwater Experiment Without Visual Failure: In these four sequences, it is important to note that laser sensors were unable to capture angular information. As a result, we relied on two key evaluation metrics: RMSE for ATE and mean ATE. Regarding these two parameters, we conducted a comparative analysis between our method, ORB SLAM3, and our method without sonar integration. This analysis allowed us to assess and contrast the performance of these approaches comprehensively.

As shown in Fig. 9, the trajectories of these three method and GT are plotted. In sequences 01 and 02, where lighting conditions were relatively poor, it can be observed that ORB SLAM3 exhibited larger errors compared to our method. This discrepancy can be attributed to the characteristics of the structures in these sequences. In particular, sequence 01 featured relatively few structural features, and the shelf structures were complex, offering abundant ORB features. In contrast, the ball structures were simpler, primarily composed of smooth surfaces with fewer ORB features. As a result, within the visual range, the features of the shelf structures dominated. However, since the shelf structures were relatively distant from the ROV, this led to a relatively large error in ORB SLAM3's results. Consequently, ORB SLAM3 exhibited larger errors compared to our method, which relies on the robustness of feature extract and is better equipped to handle such scenarios. Sequence 03 was captured under relatively good lighting conditions. It can be observed that both ORB SLAM3 and our method (without sonar integration) exhibited reduced errors in such conditions. This reduction in errors can be attributed to the influence of lighting on visual-based approaches. During the return journey, there was a significant amount of suspended particles in the pool due to the turbulence caused by the thrusters. ORB SLAM3 experienced trajectory loss in this scenario. However, due to its multimap functionality, it was able to recover the trajectory during loop-closure detection. This demonstrates that traditional feature-based methods face significant challenges when dealing with complex underwater environments. In contrast, our method, aided by sonar information, exhibited relatively consistent error behavior across various environments. Furthermore, as indicated in Table III, enabling sonar even in conditions without visual failures significantly improved the accuracy of localization. Sequence

TABLE III
EVALUATION RESULT FOR UNDERWATER EXPERIMENT WITHOUT VISUAL FAILURE

		Ours		Orbslam3		Our(without sonar)	
seq	Durations(s)	RMSE(m)	MEAN(m)	RMSE(m)	MEAN(m)	RMSE(m)	MEAN(m)
01	127.64	0.213	0.167	0.943	0.892	0.475	0.443
02	242.35	0.175	0.153	1.234	1.193	0.533	0.517
03	320.78	0.156	0.138	0.721	0.646	0.323	0.289
04	190.26	0.093	0.088	0.123	0.096	0.105	0.091



(a) (b)

Fig. 10. Underwater experiment with visual failure. (a) Seq. 05. (b) Seq. 06.

TABLE IV
EVALUATION RESULT FOR UNDERWATER EXPERIMENT WITH VISUAL FAILURE

		Ours		Our(without camera)	
seq	Durations(s)	RMSE(m)	MEAN(m)	RMSE(m)	MEAN(m)
05	327.64	0.256	0.243	0.57	0.45
06	342.35	0.452	0.423	0.89	0.62

TABLE V
RMSE AMONG SONAR-BASED METHODS

	Ours	Method1	Method2
RMSE(m)	0.1467	0.2917	0.1675

04 involved close-range movement around the platform structure under well-lit conditions. It can be observed that the three methods exhibited relatively small differences in errors under these conditions. This is primarily due to the close proximity of the structures, the presence of numerous target features, and the favorable lighting conditions. This indicates that the three algorithms perform similarly in environments with good lighting and ample features, showcasing that their precision is not significantly different in favorable conditions.

2) Underwater Experiment With Visual Failure: In the two sequences with visual failures, we also utilized RMSE and mean APE as evaluation metrics to gauge accuracy. Since visual SLAM methods such as ORB SLAM3 cannot operate properly in the presence of visual failures, we solely compared our method and our method without camera integration. Furthermore, due to the challenges associated with initializing systems that incorporate sonar with IMU, our method without camera integration was initialized using the camera, after which the camera information was subsequently disabled.

As shown in Fig. 10, the trajectories of these three method and GT are plotted. It can be observed that without the constraint of visual data to reduce errors, disabling the camera led to a significant accumulation of errors. While the system continued to operate correctly after camera deactivation, the absence of camera-based reprojection resulted in increased cumulative errors. The detailed error can be found in Table IV.

To provide a more comprehensive comparison of the algorithm's advancement, we also included two additional sonar-based SLAM methods for comparison. These methods are referred to as "Method 1" and "Method 2."

1) *Method 1 [30]:* This is a purely acoustic-based method. It optimizes the elevation information lost from the imaging sonar as part of the optimization process.
2) *Method 2 [31]:* This is a fusion approach that combines both sonar and optical data. It utilizes stereo vision along with imaging sonar fusion to enhance its performance.

The results from these two methods are derived from their respective research articles. When we compare our method to these two methods, we calculate the average performance of our method after excluding Sequence 04 from the evaluation. As shown in Table V, the results indicate that our method achieves higher accuracy compared to these two methods. It is worth noting that both our method and Method 2 utilize visual sensors, and their errors are significantly lower than Method 1, which relies solely on sonar for SLAM. This observation underscores the effectiveness of multisensor fusion in underwater SLAM, as it enhances the robustness of underwater localization.

V. Conclusion

This article introduces an innovative multisensor underwater SLAM system that synergizes the capabilities of stereo vision, imaging sonar, and IMU to tackle the complex problem of ROV localization. Designed to enhance the precision and reliability of underwater navigation, our system is structured around two core subsystems: the Vision-IMU subsystem and the Acoustic-IMU subsystem. A key feature of this architecture is its shared feature depth estimation, which ensures that the system remains operational and robust even when one of the subsystems encounters difficulties. This intersubsystem cooperation not only boosts the system's resilience against individual sensor failures but also significantly improves localization accuracy by leveraging shared features for optimized initialization and employing sonar image-based feature depth estimation for precise positioning. The system's effectiveness and adaptability to challenging underwater conditions, such as fluctuating light conditions and instances of visual occlusion, have been rigorously validated through a series of experiments. Unlike existing approaches that predominantly rely on a single type of sensor, our method integrates diverse sensory inputs, significantly enhancing the robustness and accuracy of underwater SLAM in complex environments. This comprehensive integration addresses the limitations associated with single-sensor systems, such as poor visibility conditions for visual sensors and the low spatial resolution of acoustic sensors. These tests confirm the algorithm's robust performance, marking a significant advancement in the field of underwater robotic navigation and opening new avenues for the exploration and operational use of ROVs in complex aquatic environments. Similarly, there are areas for improvement in our method. First, our current system does not incorporate semantic information into the SLAM system, which could significantly enhance its reliability. In addition, we have not fully leveraged the long-range detection capabilities of acoustic devices to use acoustic features as prior information for visual features. Therefore, our future research will focus on developing a multisensor SLAM system enriched with semantic information and integrating sonar prior information for visual system loop closure detection. This advancement aims to improve recall rates and further reduce cumulative errors, paving the way for more accurate and robust underwater navigation solutions.

References

[1] J. Henderson, O. Pizarro, M. Johnson-Roberson, and I. Mahon, "Mapping submerged archaeological sites using stereo-vision photogrammetry," *Int. J. Nautical Archaeol.*, vol. 42, no. 2, pp. 243–256, Sep. 2013.

[2] C. M. Lee, P. M. Lee, S. W. Hong, S. M. Kim, and W. Seong, "Underwater navigation system based on inertial sensor and Doppler velocity log using indirect feedback Kalman filter," *Int. J. Offshore Polar Eng.*, vol. 15, no. 2, Jun. 2005.

[3] P. Rigby, O. Pizarro, and S. Williams, "Towards geo-referenced AUV navigation through fusion of USBL and DVL measurements," in *Proc. IEEE OCEANS*, Sep. 2006, pp. 1–6.

[4] J. Snyder, "Doppler velocity log (DVL) navigation for observation-class ROVs," in *Proc. OCEANS MTS/IEEE SEATTLE*, Sep. 2010, pp. 1–9.

[5] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual-inertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, 2015.

[6] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint Kalman filter for vision-aided inertial navigation," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2007, p. 6.

[7] S. Ding, T. Ma, Y. Li, S. Xu, and Z. Yang, "RD-VIO: Relative-depth-aided visual-inertial odometry for autonomous underwater vehicles," *Appl. Ocean Res.*, vol. 134, May 2023, Art. no. 103532.

[8] S. Xu et al., "An effective stereo SLAM with high-level primitives in underwater environment," *Meas. Sci. Technol.*, vol. 34, no. 10, Oct. 2023, Art. no. 105405.

[9] E. Westman, A. Hinduja, and M. Kaess, "Feature-based SLAM for imaging sonar with under-constrained landmarks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 3629–3636.

[10] M. D. Aykin and S. Negahdaripour, "Forward-look 2-D sonar image formation and 3-D reconstruction," in *Proc. OCEANS*, Sep. 2013, pp. 1–10.

[11] Y. Lee, J. Choi, and H.-T. Choi, "Experimental results on EKF-based underwater localization algorithm using artificial landmark and imaging sonar," in *Proc. OCEANS*, Sep. 2014, pp. 1–3.

[12] J. Wang, F. Chen, Y. Huang, J. McConnell, T. Shan, and B. Englot, "Virtual maps for autonomous exploration of cluttered underwater environments," *IEEE J. Ocean. Eng.*, vol. 47, no. 4, pp. 916–935, Oct. 2022.

[13] 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.

[14] T. Shan, B. Englot, C. Ratti, and D. Rus, "LVI-SAM: Tightly-coupled LiDAR-visual-inertial odometry via smoothing and mapping," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 5692–5698.

[15] F. Hidalgo and T. Bräunl, "Review of underwater SLAM techniques," in *Proc. 6th Int. Conf. Autom., Robot. Appl. (ICARA)*, Feb. 2015, pp. 306–311.

[16] X. Yuan, J.-F. Martínez-Ortega, J. Fernández, and M. Eckert, "AEKF-SLAM: A new algorithm for robotic underwater navigation," *Sensors*, vol. 17, no. 5, p. 1174, May 2017.

[17] H. Wang, G. Fu, J. Li, Z. Yan, and X. Bian, "An adaptive UKF based SLAM method for unmanned underwater vehicle," *Math. Problems Eng.*, vol. 2013, pp. 1–12, Sep. 2013.

[18] H. Xing et al., "A multi-sensor fusion self-localization system of a miniature underwater robot in structured and GPS-denied environments," *IEEE Sensors J.*, vol. 21, no. 23, pp. 27136–27146, Dec. 2021.

[19] L. Chen, A. Yang, H. Hu, and W. Naeem, "RBPF-MSIS: Toward rao-blackwellized particle filter SLAM for autonomous underwater vehicle with slow mechanical scanning imaging sonar," *IEEE Syst. J.*, vol. 14, no. 3, pp. 3301–3312, Sep. 2020.

[20] B. He et al., "Autonomous navigation based on unscented-FastSLAM using particle swarm optimization for autonomous underwater vehicles," *Measurement*, vol. 71, pp. 89–101, Jul. 2015.

[21] 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. (IROS)*, Nov. 2019, pp. 1861–1868.

[22] S. Rahman, A. Quattrini Li, and I. Rekleitis, "SVIn2: A multi-sensor fusion-based underwater SLAM system," *Int. J. Robot. Res.*, vol. 41, nos. 11–12, pp. 1022–1042, Sep. 2022.

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

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

[25] A. Lagudi, G. Bianco, M. Muzzupappa, and F. Bruno, "An alignment method for the integration of underwater 3D data captured by a stereovision system and an acoustic camera," *Sensors*, vol. 16, no. 4, p. 536, Apr. 2016.

[26] J. Zhang et al., "Object measurement in real underwater environments using improved stereo matching with semantic segmentation," *Measurement*, vol. 218, Aug. 2023, Art. no. 113147.

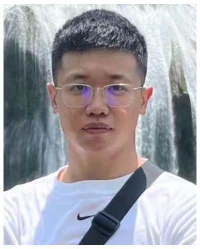
[27] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 224–236.

[28] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. 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.

[29] T. Qin, J. Pan, S. Cao, and S. Shen, “A general optimization-based framework for local odometry estimation with multiple sensors,” 2019, *arXiv:1901.03638*.

[30] J. Li, M. Kaess, R. M. Eustice, and M. Johnson-Roberson, “Pose-graph SLAM using forward-looking sonar,” *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2330–2337, Jul. 2018.

[31] H. Jang, S. Yoon, and A. Kim, “Multi-session underwater pose-graph SLAM using inter-session opti-acoustic two-view factor,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 11668–11674.



Jiawei Zhang received the B.S. degree in engineering from Dalian University of Technology, Dalian, China, in 2017. He is currently pursuing the degree in ship and marine structure design and manufacturing with Harbin Engineering University, Harbin, China, focusing on multisensor fusion for underwater simultaneous localization and mapping (SLAM).

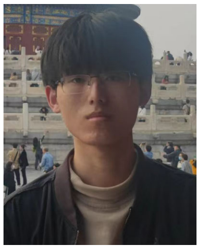
His research interests include underwater robot control and the application of multisensors in SLAM.

Fenglei Han received the Ph.D. degree in engineering from Harbin Engineering University, Harbin, China, in 2016.

He is currently an Associate Professor with the College of Shipbuilding Engineering, Harbin Engineering University. His research interests include underwater robot vision technology, intelligent navigation, path planning, and smart ship design. He conducts fundamental research using various approaches, including theoretical analysis, mathematical modeling, numerical simulation, experiments, and virtual simulation, within these research areas.

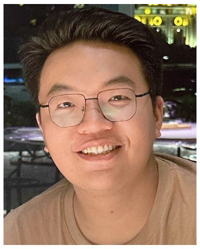
Duanfeng Han received the Ph.D. degree in engineering from Harbin Engineering University, Harbin, China, in 2002.

He is currently a Professor with the College of Shipbuilding Engineering, Harbin Engineering University. His research interests include polar ship and offshore engineering design and manufacturing technology, ice load computation and analysis, research on new icebreaking methods and mechanisms, theoretical prediction methods for ice-covered ship performance, and ice tank experimental techniques.



Jianfeng Yang received the B.S. degree in engineering and the master’s degree from Harbin Engineering University, Harbin, China, in 2020. He is currently pursuing the degree in design and manufacturing of ships and marine structures, focusing on the application of computer vision.

His research interests include underwater image enhancement and the application of multisensors in SLAM.



Wangyuan Zhao received the B.S. degree in engineering from Harbin Engineering University, Harbin, China, in 2018, where he is currently pursuing the Ph.D. degree, with a specialization in the intelligent application of optical sensors in underwater and surface environments.

His research interests include the utilization of deep learning techniques for optical image analysis in underwater robotics and perception and parameter measurement in highly complex underwater environments.

Hansheng Li received the B.S. degree in engineering from Harbin Engineering University, Harbin, China, in 2017. He is currently pursuing the degree in design and manufacturing of ships and marine structures, focusing on the application of computer vision.

He is engaging in hydrodynamic modeling, fluid–structure interaction, and control of bioinspired marine robotics, by applicating STAR-CCM+ and OpenFOAM.