Check for updates

Data paper

The International Journal of Robotics Research
2026, Vol. 45(4) 541–551
© The Author(s) 2025

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02783649251364904
journals.sagepub.com/home/ijr

Tank dataset: An underwater multi-sensor dataset for SLAM evaluation

Shida Xu[1], Jonatan Scharff Willners[2], Joshua Roe[2], Sean Katagiri[2], Tomasz Luczynski[2], Yvan Petillot[2] and Sen Wang[1]

Abstract

Underwater visual Simultaneous Localization and Mapping (SLAM) is essential for autonomous underwater navigation and close-range underwater inspection. However, the turbid and low-light conditions common underwater severely limit visibility and cause motion blurring, posing significant open challenges for visual SLAM approaches deployed underwater. On the other hand, the scarcity of public underwater multi-sensor datasets, coupled with the lack of 6-Degree-of-Freedom (6-DoF) ground truth data for SLAM evaluation, hinders the advancement of underwater visual SLAM research. To address these problems, this paper introduces an underwater dataset encompassing multi-sensor data from a stereo camera, an Inertial Measurement Unit, a Doppler Velocity Log, and a pressure sensor. To cover various difficulty levels for underwater SLAM evaluation, it provides eight sequences collected under different speed and illumination conditions. Extrinsic and intrinsic calibration parameters are also provided for multi-sensor fusion. Additionally, we present TankGT, a fiducial-marker-based SLAM system designed to provide highly accurate 6-DoF ground truth poses in underwater environments, enabling rigorous quantitative and qualitative benchmarking for underwater SLAM algorithms. We demonstrate the effectiveness of the proposed Tank dataset with four SLAM algorithms. The dataset is released to facilitate underwater SLAM research in the community at https://senseroboticslab.github.io/underwater-tank-dataset.

Keywords

underwater dataset, underwater robotics, visual SLAM

Received: 10 December 2024; Revised 11 May 2025; Accepted 4 June 2025
Senior Editor: James Forbes
Associate Editor: Ayoung Kim

1. Introduction

Underwater Simultaneous Localization and Mapping (SLAM) is vital for enabling autonomous navigation and task execution in underwater environments. Two primary approaches in underwater SLAM are sonar-based SLAM and visual SLAM. Sonar systems offer a significantly longer operational range compared to cameras, making them effective in low-visibility conditions. However, their low resolution, high noise, and limited observability make it difficult to recover full 6-Degree-of-Freedom (6-DoF) poses accurately. Additionally, the high cost of sonar systems limits their widespread adoption.

In contrast, visual SLAM is attractive due to the widespread availability, low cost, and rich perceptual capabilities of camera systems commonly used on underwater robots for underwater inspection tasks. Nevertheless, underwater environments pose several unique challenges for visual SLAM. Turbid water reduces visibility, producing noisy, low-contrast, and textureless images that hinder feature extraction and matching. Motion blur, caused by poor lighting and external disturbances such as currents and waves, further degrades tracking performance. Featureless open areas provide little visual structure for SLAM algorithms to exploit. Additionally, suspended particles scatter light and introduce artifacts such as marine snow, which severely affect image quality and disrupt feature matching (Guth et al., 2014; Zhang et al., 2022). These challenges make vision-only SLAM systems unreliable or even non-functional in many real-world underwater scenarios.

[1]Department of Electrical and Electronic Engineering, Imperial College London, London, UK
[2]School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, UK

Corresponding author:
Sen Wang, Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK.
Email: sen.wang@imperial.ac.uk

Consequently, multi-sensor fusion, using cameras with modalities such as Inertial Measurement Unit (IMU), Doppler Velocity Log (DVL), and depth sensor, has emerged as a promising strategy to enhance the robustness and accuracy of underwater visual SLAM systems.

In recent years, research in underwater visual SLAM has increasingly adopted a multi-modal approach (Vargas et al., 2021; Xu et al., 2021; Thoms et al., 2023; Zhao et al., 2023; Xu et al., 2025), integrating sensors such as camera, IMU, DVL, and pressure sensors, to address the challenges of the underwater environment. However, the lack of publicly available underwater datasets with accurate 6-DoF ground truth (GT) poses has significantly limited the benchmarking of SLAM methods and slowed progress in the field. Obtaining reliable GT poses underwater remains a major challenge. Motion capture systems, commonly used for aerial SLAM, suffer from limited operational range and coverage underwater because their performance degrades sharply in water. Acoustic positioning systems, such as Ultra-Short Baseline (USBL), offer an alternative but are expensive, difficult to deploy, and often unreliable in shallow water (Guth et al., 2014). Structure-from-Motion (SfM) techniques (Schönberger and Frahm, 2016) are frequently used to generate GT poses offline from visual data. For instance, COLMAP, a widely used SfM toolbox, has been employed in datasets such as AQUALOC (Ferrera et al., 2019) and Eiffel Tower (Boittiaux et al., 2023). However, SfM methods are effective only under ideal visual conditions, requiring good visibility, sufficient environmental features, and slow platform motion to avoid motion blur. As such, they are poorly suited for evaluating SLAM performance in challenging, real-world underwater scenarios. To the best of our knowledge, generating accurate and comprehensive 6-DoF ground truth for underwater SLAM remains an open and unsolved problem, presenting a key barrier to benchmarking and advancing state-of-the-art methods.

Furthermore, as summarized in Table 1, most existing public underwater datasets lack comprehensive multi-sensor data and the necessary calibration parameters. These parameters, including camera intrinsics and the extrinsic transformations between sensors, are crucial for multi-sensor SLAM systems. Since each sensor measures data in its own local reference frame, accurate extrinsic calibration is required to transform all sensor data into a common frame before sensor fusion. Inaccurate extrinsics can cause spatial misalignments, severely degrading both localization and mapping performance. This limitation significantly hinders the benchmarking and development of robust multi-sensor SLAM algorithms.

In this paper, we propose the Tank dataset, an underwater dataset that includes multi-sensor data from a stereo camera, an IMU, a DVL, and a depth sensor. To the best of our knowledge, it is the first public underwater dataset that incorporates this multi-sensor configuration and provides accurate extrinsic sensor calibration parameters. Moreover, a physical underwater structure and a fiducial-marker-based SLAM system, termed TankGT, are developed to generate accurate 6-DoF GT poses for underwater visual SLAM evaluation. Eight sequences are captured in different routes, velocities and lighting settings to reflect varying difficulty levels. A convenient evaluation tool set is also developed to generate visually appealing and publication-ready tables and figures for SLAM evaluation. We demonstrate the effectiveness of this Tank dataset on four state-of-the-art SLAM algorithms.

2. Related work

In this section, we provide a review of existing underwater SLAM datasets.

The ACFR dataset (Steinberg et al., 2010) includes 22 AUV dives off Tasmania in 2008, with annotations on every 100th image from over 100,000 stereo pairs. Each annotated image has 50 labeled points covering biological, abiotic, and ambiguous classes. Nevertheless, it lacks raw IMU readings and does not provide the original DVL or pressure sensor data, restricting its value for systems that perform their own fusion or sensor calibration. Furthermore, the dataset does not include camera intrinsic parameters or extrinsic transformations between the sensors. This is a significant limitation for benchmarking multi-sensor SLAM algorithms. In addition, the ACFR dataset employs an extended information filter-based SLAM algorithm (Mahon et al., 2008) to generate GT poses. However, this filter-based SLAM is no longer state-of-the-art, and may not be suitable to provide GT data for evaluating modern SLAM algorithms.

Table 1. Comparison of underwater datasets suitable for underwater SLAM.

					Parameter		GT		
Dataset	Camera	IMU	Pressure	DVL	Intrinsic	Extrinsic	Method	DoF	Scale
ACFR Steinberg et al. (2010)	Stereo	×	✗	✗	×	×	SLAM	6	✓
Caves Mallios et al. (2017)	Monocular	✓	✓	✓	✓	✓	Landmarks	1	✓
AQUALOC Ferrera et al. (2019)	Monocular	✓	✓	×	✓	✓	COLMAP	6	✓
AURORA Bernardi et al. (2022)	Monocular	✗	✗	✗	×	Translation	×	×	×
Eiffel Tower Boittiaux et al. (2023)	Monocular	×	×	×	×	×	COLMAP	6	×
Ours	Stereo	✓	✓	✓	✓	✓	AprilTag	6	✓

The Caves dataset (Mallios et al., 2017), collected in in an underwater cave complex using a diver-guided AUV, includes sonar, DVL, IMUs, depth, and downward-facing camera data. It provides both ROS bag files and processed text formats for ease of use. However, it only provides roughly measured GT data using 1-D distances between cone pairs (as landmarks). Moreover, it only employs a monocular camera.

The AQUALOC dataset (Ferrera et al., 2019) provides 17 sequences recorded at depths up to 380 m using ROVs equipped with a monocular camera, IMU, and pressure sensor. Data is provided as ROS bags and raw files, with offline SfM-based trajectories for benchmarking. However, the dataset lacks DVL data, which is crucial for underwater SLAM. The SfM-based GT poses are generated using COLMAP (Schönberger and Frahm, 2016), which may not be suitable for evaluating SLAM algorithms in challenging underwater conditions.

The AURORA dataset (Bernardi et al., 2022) offers a more diverse sensor suite, including sidescan sonar, multibeam echosounder, and visual data, collected during surveys in the Greater Haig Fras Marine Conservation Zone. However, it only provides fused navigation outputs rather than time-synchronized raw measurements from the DVL, IMU, and depth sensors. This omission limits its utility for evaluating or developing tightly-coupled SLAM pipelines that require direct access to raw measurements for accurate state estimation and uncertainty propagation.

The Eiffel Tower dataset (Boittiaux et al., 2023) provides only monocular camera data, which significantly limits its applicability to multi-sensor SLAM research. Without data from inertial sensors, depth sensors, or sonar, it cannot support the development or benchmarking of sensor fusion frameworks that are increasingly central to robust SLAM in challenging environments such as underwater domains.

As far as we know, the existing datasets do not provide a comprehensive multi-sensor configuration, calibration parameters with accurate 6-DoF GT poses for underwater SLAM evaluation. The Tank dataset aims to fill this gap by providing a multi-sensor dataset with accurate GT poses, enabling the development and benchmarking of robust underwater SLAM algorithms.

3. Dataset collection and formation

3.1. Environment setup

The data is collected in a 9 × 12 m water tank with a dedicated underwater structure covered by AprilTag (Ferrera et al., 2019) markers in the middle. As indicated in Figure 1, the environment has five different areas: (1) textureless wall, (2) wave generator, (3) textureless wall, (4) beach, and (5) underwater structure. These areas provide a variety of scenarios: the beach area provides some visible features when the vehicle traverses closely, being a relatively easy region; the wall and generator areas do not have much salient texture to track, hence challenging for visual SLAM; The underwater structure provides clear features as well AprilTag markers for generating GT poses.

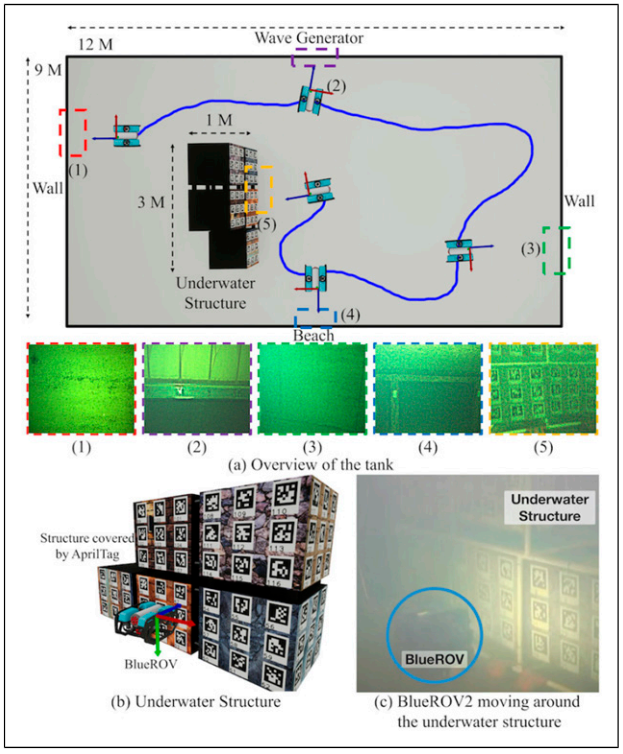


Figure 1. Experiment setting for the data collection.

3.2. Sensor configuration

To collect the data, a BlueROV2 vehicle equipped with a WaterLinked A50 DVL, a MICROSTRAIN 3DM-GX5-AHRS IMU, a Blue Robotics Bar30 Pressure/Depth sensor, and a custom underwater stereo camera was manually piloted through the water tank. The detailed specifications of the sensors are summarized in Table 2.

Camera intrinsic parameters. The intrinsic parameters of the stereo cameras are calibrated using the Pinax Model (Łuczyński et al., 2017) which corrects the refraction and distortion caused by the camera's waterproof housing. An underwater dehazing algorithm (Łuczyński and Birk, 2017) is also enabled to enhance the camera visibility in the water.

Extrinsic parameters. Extrinsic parameters between sensor A and B can be defined as a transformation matrix as:

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

where $\mathbf{T}_{\mathrm{AB}}$ standing for the pose of sensor B with respect to sensor A, $\mathbf{R}_{\mathrm{AB}} \in \mathrm{SO}(3)$ describes its rotation, and ${}_{\mathrm{A}}\mathbf{p}_{\mathrm{AB}} \in \mathbb{R}^3$ is its translation expressed in frame A.

Table 2. Sensor specifications.

Underwater Stereo Camera	
Resolution	612 × 512×2
Frame Rate	20 Hz
Field of View	90 degrees
Baseline	0.12 m
MICROSTRAIN 3DM-GX5-AHRS IMU	
Ping Rate	333 Hz
Gyroscope	
Range	300°/sec
Resolution	0.003°/sec
Bias Instability	8°/hr
Initial Bias Error	±0.04°/sec
Sampling Rate	4 kHz
Accelerometer	
Range	±8 g
Resolution	0.02 mg
Bias Instability	±0.04 mg
Initial Bias Error	±0.002 g
Sampling Rate	1 kHz
WaterLinked A50 DVL	
Transducer Frequency	1 MHz
Transducer Setup	4-beam convex Janus array
Transducer Beam Angle	22.5 degrees
Ping Rate	5 Hz
Min Altitude	5 cm
Max Altitude	50 m
Max Velocity	3.75 m/s
Velocity Resolution	0.1 mm/s
Long Term Accuracy	±1.01%
Blue Robotics Bar 30 Pressure/Depth	
Depth Range	±200 mbar (204 cm in fresh water)
Resolution	0.2 mbar (2 mm in fresh water)
Blue Robotics Lumen Subsea Light	
Maximum Brightness	1500 lumens
Color Temperature	5700 kelvin
Beam Angle	135 degrees in water

The sensor coordinate frames are shown in Figure 2. We provide $\mathbf{T}_{\mathrm{IC}}$, the extrinsic parameters between the IMU and the left camera, and $\mathbf{T}_{\mathrm{DC}}$, the extrinsic parameters between the DVL and the left camera. Two different sets of sensor extrinsic parameters are supplied because the extrinsic configurations changed slightly during the period of data collection (but maintained within an individual sequence). The extrinsic parameters are provided as transformation matrices in a YAML file.

Extrinsic calibration algorithm. The extrinsic parameters between the DVL and the camera are obtained using our extrinsic calibration algorithm proposed in Xu et al. (2021), being initialized with manual measures for better convergence. Specifically, trajectories estimated separately from DVL based dead reckoning and camera-based visual SLAM are obtained. The DVL and camera trajectories are defined as a set of relative transformations:

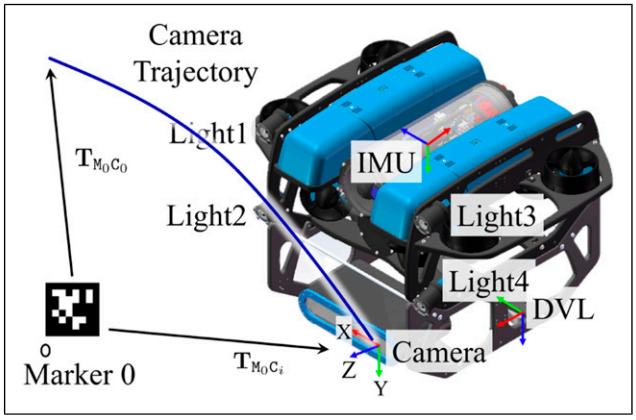


Figure 2. Coordinate frames of the sensors and the GT poses.

$$\mathcal{S}_d \doteq \{\mathbf{T}_{\mathrm{D}_i\mathrm{D}_j}\}_{i,j\in\mathcal{K}} \quad \text{and} \quad \mathcal{S}_c \doteq \{\mathbf{T}_{\mathrm{C}_i\mathrm{C}_j}\}_{i,j\in\mathcal{K}} \tag{2}$$

where i and j are consecutive timestamps, and $\mathcal{K}$ is the set of all timestamps where DVL and camera data are available. Given the extrinsic parameters $\mathbf{T}_{\mathrm{DC}}$, a pair of DVL and camera poses at timestamp i, we can transform the camera trajectory to the DVL frame as:

$$\mathbf{T}_{\mathrm{D}_i\mathrm{D}_j} = \mathbf{T}_{\mathrm{DC}}\mathbf{T}_{\mathrm{C}_i\mathrm{C}_j}\mathbf{T}_{\mathrm{DC}}^{-1} \tag{3}$$

Splitting (3) into rotation and translation, we have:

$$\begin{aligned} \mathbf{R}_{\mathrm{D}_i\mathrm{D}_j} &= \mathbf{R}_{\mathrm{DC}}\mathbf{R}_{\mathrm{C}_i\mathrm{C}_j}\mathbf{R}_{\mathrm{DC}}^{-1} \\ {}_{\mathrm{D}_i}\mathbf{p}_{\mathrm{D}_i\mathrm{D}_j} &= {}_{\mathrm{D}}\mathbf{p}_{\mathrm{DC}} + \mathbf{R}_{\mathrm{DC}\mathrm{C}_i}\mathbf{p}_{\mathrm{C}_i\mathrm{C}_j} - \mathbf{R}_{\mathrm{DC}}\mathbf{R}_{\mathrm{C}_i\mathrm{C}_j}\mathbf{R}_{\mathrm{DC}}^{-1}{}_{\mathrm{C}}\mathbf{p}_{\mathrm{CD}} \end{aligned} \tag{4}$$

Therefore, rotational and translational errors can be calculated as:

$$\begin{aligned} \mathbf{r}_{\mathbf{R}_{ij}} &= \mathtt{log}\left(\mathbf{R}_{\mathrm{D}_i\mathrm{D}_j}^{-1}\mathbf{R}_{\mathrm{DC}}\mathbf{R}_{\mathrm{C}_i\mathrm{C}_j}\mathbf{R}_{\mathrm{DC}}^{-1}\right) \\ \mathbf{r}_{\mathbf{p}_{ij}} &= {}_{\mathrm{D}}\mathbf{p}_{\mathrm{DC}} + \mathbf{R}_{\mathrm{DC}\mathrm{C}_i}\mathbf{p}_{\mathrm{C}_i\mathrm{C}_j} - \mathbf{R}_{\mathrm{DC}}\mathbf{R}_{\mathrm{C}_i\mathrm{C}_j}\mathbf{R}_{\mathrm{DC}}^{-1}{}_{\mathrm{C}}\mathbf{p}_{\mathrm{CD}} - {}_{\mathrm{D}_i}\mathbf{p}_{\mathrm{D}_i\mathrm{D}_j} \end{aligned} \tag{5}$$

where $\log(\cdot)$ is the logarithmic function to map a quantity from SO(3) to vectorized Lie algebra $\mathfrak{so}(3)$. The extrinsic parameter $\mathbf{T}_{\mathrm{DC}}$ is computed by minimizing the following errors using Gauss–Newton optimization:

$$\min_{\mathbf{T}_{\mathrm{DC}}} \sum_{i,j\in\mathcal{K}} \left\|\mathbf{r}_{\mathbf{R}_{ij}}\right\|^2 + \left\|\mathbf{r}_{\mathbf{p}_{ij}}\right\|^2 \tag{6}$$

Extrinsic calibration between the IMU and the camera is performed using the Kalibr toolbox (Furgale et al., 2013).

Sequence configuration

Eight sequences under various underwater conditions were collected to evaluate SLAM performance. Table 3 outlines the configuration of each sequence, including robot velocity, lighting conditions, and GT percentage. The GT percentage metric represents the proportion of images in which AprilTags are visible within a sequence. A lower GT percentage

indicates that the vehicle moves further from the underwater structure, likely encountering more textureless regions, as illustrated in Figure 1 (1) (2) (3) (4), thereby increasing the difficulty. The velocity, lighting conditions, and GT percentage collectively characterize the difficulty of each sequence.

Three difficulty levels are defined for evaluation:

- The easy level features low velocity, good lighting conditions, and appropriate distances between the camera and objects to minimize motion blur and ensure good visual input for SLAM systems.
- The medium level introduces challenges such as poor lighting, occasional aggressive motion, and a lack of structural features, making it more difficult for SLAM systems.
- The hard level combines aggressive motion, poor or varying lighting, and long-term structureless or textureless scenes, posing significant challenges for SLAM systems.

Examples of these challenging visual conditions, including lack of structure, motion blur, and lighting variations, are shown in Figure 3.

We designed three types of sequences: Structure (SE, SM, and SH), HalfTank (HE, HM, and HH), and WholeTank (WM and WH).

- The Structure sequences (SE, SM, and SH) are collected around the structure, with the vehicle following a looped trajectory, as illustrated by the SE sequence in Figure 4(a).
- The HalfTank sequences (HE, HM, and HH) cover a larger area, starting from the structure and traversing half of the tank, including the wall and beach areas. The trajectory and reconstruction of the HE sequence are shown in Figure 4(b).
- The WholeTank sequences (WM and WH) span an even larger area, where the vehicle begins near the structure, traverses the entire tank and returns to the structure for loop closure, as demonstrated by the WM sequence in Figure 4(c).

Table 3. Statistics of the eight sequences. Velocity histograms show velocity distribution between 0 and 0.5 m/s.

Sequence	Velocity	Light	GT %
Structure Easy (SE)		✓	99.37%
Structure Medium (SM)		×	99.67%
Structure Hard (SH)		×	84.65%
HalfTank Easy (HE)		✓	36.96%
HalfTank Medium (HM)		✓	44.19%
HalfTank Hard (HH)		× and ✓	41.40%
WholeTank Medium (WM)		✓	60.73%
WholeTank Hard (WH)		×	18.97%

For the Structure sequences, 6-DoF GT poses are mostly available. For both the HalfTank and WholeTank sequences, the vehicle starts in the structure area with GT, moves through challenging textureless regions without GT, and then returns to the structure area where GT is available again. This setup allows us to evaluate visual odometry drift or visual SLAM errors by comparing GT poses before and after traversing challenging regions.

Data format

ROS bag. The data is provided as ROS bags and associated parameter YAML files. The ROS bag files contain the sensor data. Each ROS bag file has six topics:

(1) **`/camera/left/image_dehazed/compressed`:sensor_msgs/CompressedImage** Left image from the stereo camera.
(2) **`/camera/right/image_dehazed/compressed`:sensor_msgs/CompressedImage** Right image from the stereo camera.
(3) **`/imu/data`:sensor_msgs/Imu** IMU data.
(4) **`/DVL/data`:waterlinked_a50_ros_driver/DVL** DVL data in its raw format.
(5) **`/depth/data`:nav_msgs/Odometry** Depth data from pressure sensor.
(6) **`/apriltag_slam/GT`:nav_msgs/Odometry** GT poses provided by the TankGT.
(7) **`/apriltag_slam/GT_full`:nav_msgs/Odometry** GT poses fused by the TankGT and AQUA SLAM (Xu et al., 2025).

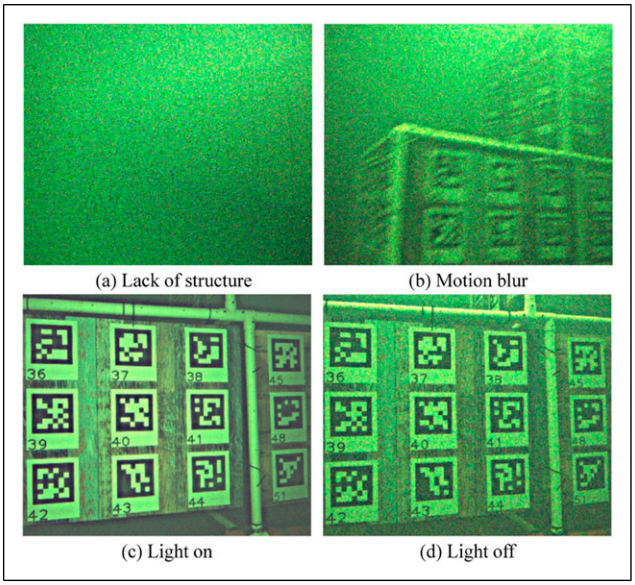


Figure 3. Some visually challenging scenarios in the dataset.

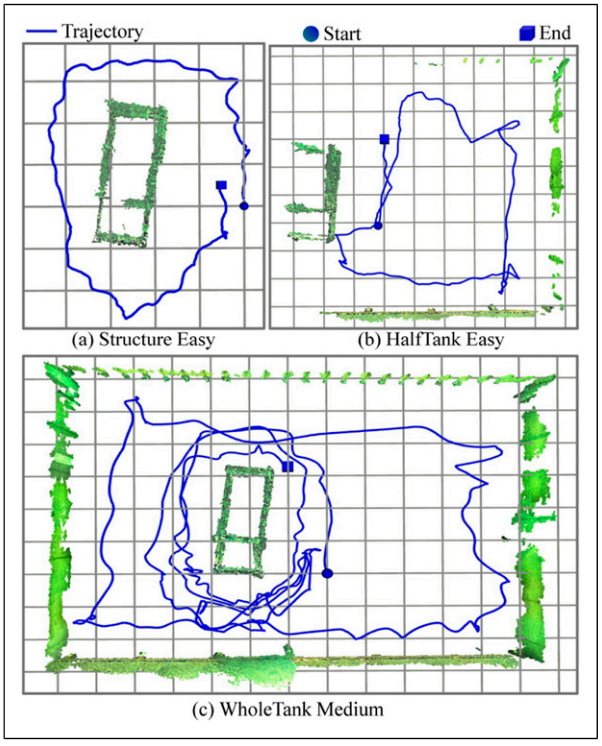


Figure 4. Trajectories and 3D maps of three sequences. A grid is with size 1*m* × 1*m*.

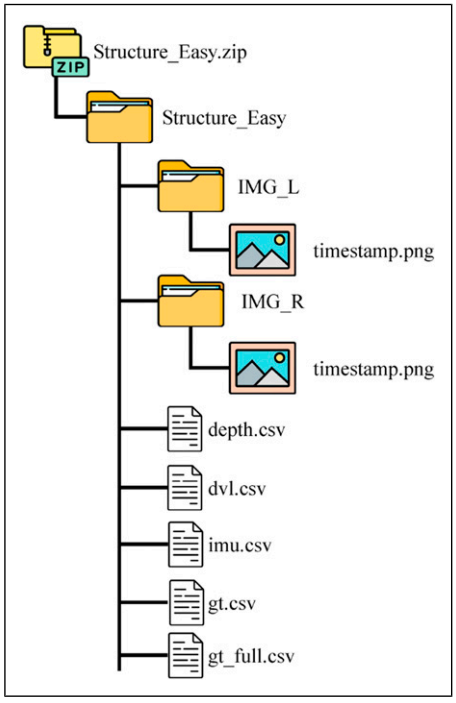


Figure 5. The folder structure of the raw data.

Raw data. In addition to the ROS bag files, the raw data is also provided in a folder structure. The raw data includes the images from the stereo camera, the depth data from the pressure sensor, the DVL data, the IMU data, and the GT data. An example of the folder structure of Structure_Easy sequence is shown in Figure 5.The name of the image files is the timestamp of the image in nanosecond. The depth data is stored in a CSV file with the timestamp and the depth value. The DVL data is stored in a CSV file with the timestamp and the 3-DoF velocity in DVL body frame and four radial velocities along each transducer. The IMU data is stored in a CSV file with the timestamp and the acceleration and angular velocity values. The GT data is stored in a CSV file with the timestamp and the GT pose values.

Parameter YAML. The parameter files provide the camera intrinsic parameters and extrinsic parameters between the sensors in the YAML format as shown in Figure 6.

Ground truth generation

Underwater structure with AprilTag markers

The underwater structure is built with waterproof panels which are covered with AprilTag markers, as shown in Figure 1(b) and (c). As illustrated in Figure 7, a panel that measures 1 × 0.84 m has nine markers arranged with a horizontal interval of 0.333 m and a vertical interval of 0.28 m. This configuration allows for multiple markers to be observed simultaneously as the robot moves around the structure, enhancing the robustness and accuracy of the GT generation.

TankGT algorithm

To obtain accurate and reliable GT pose data, we developed TankGT, a fiducial marker-based SLAM system. It utilizes the relative pose information from AprilTags to simultaneously estimate the camera poses and the poses of the AprilTag markers. It includes two modes: calibration mode and localization mode.

Calibration mode. The calibration mode is designed to calibrate the poses of all AprilTag markers under ideal conditions. Challenging underwater conditions can compromise the accuracy of estimated marker poses. Therefore, we estimate the marker poses in a rather optimal scenario where the robot moves slowly around the underwater structure for good image quality with multiple loop closures to reduce errors. This calibration mode is formulated as a SLAM problem with a factor graph optimization. Specifically, considering all camera frames $\mathcal{K}_n$ up to n and all markers $\mathcal{K}_m$ up to m, the state $\mathcal{X}$ is defined as a set of the camera poses $\mathcal{C}$ and the marker poses $\mathcal{M}$:

```
#Camera Intrinsics
Camera.fx: 655.0
Camera.fy: 655.0
Camera.cx: 306.0
Camera.cy: 256.0

# stereo baseline times fx
Camera.bf: 78.89165891925023

# Sensor Extrinsic between IMU and Camera
T_imu_camera: !!opencv-matrix
  rows: 4
  cols: 4
  dt: f
  data: [ -0.0165,   -0.0175,   -0.9997,   -0.211687435,
          -0.0106,    0.9998,   -0.0173,    -0.097259169,
          0.9998,    0.0103,   -0.0167,    -0.056238089,
          0,  0,   0,    1 ]

# Sensor Extrinsic between DVL and Camera
T_dvl_camera: !!opencv-matrix
  rows: 4
  cols: 4
  dt: f
  data: [ -0.0030,    0.0292,    0.9996,   0.132344740,
          0.9995,    0.0328,    0.0021,    0.032085643,
          -0.0327,    0.9990,   -0.0293,    -0.240100175,
          0,  0,   0,    1 ]
```

Figure 6. The parameter YAML file.

$$\begin{aligned} \mathcal{C} &\doteq \{\mathbf{T}_{\mathrm{M}_0\mathrm{C}_i}\}, \quad i \in \mathcal{K}_n \\ \mathcal{M} &\doteq \{\mathbf{T}_{\mathrm{M}_0\mathrm{M}_j}\}, \quad j \in \mathcal{K}_m \\ \mathcal{X} &\doteq \{\mathcal{C}, \mathcal{M}\} \end{aligned} \tag{7}$$

where the camera poses and marker poses are all defined with respect to the marker 0 M_0. When the marker j is visible in the ith camera frame, it can provide a relative pose measurement $\widehat{\mathbf{T}}_{\mathrm{M}_j\mathrm{C}_i}$ based on the AprilTag pose estimation algorithm (Wang and Olson, 2016). Therefore, the residual $\mathbf{r}_{i,j}$ is defined as

$$\mathbf{r}_{i,j} \doteq \mathrm{Log}\left(\mathbf{T}^{-1}_{\mathrm{M}_0\mathrm{M}_j}\mathbf{T}_{\mathrm{M}_0\mathrm{C}_i}\widehat{\mathbf{T}}^{-1}_{\mathrm{M}_j\mathrm{C}_i}\right) \tag{8}$$

where $\mathrm{Log}(\cdot): \mathrm{SE}(3) \to \mathbb{R}^6$ stands for the logarithm map from SE(3) to a vectorized $\mathfrak{se}(3)$. Finally, the objective function is

$$\mathcal{X}^* = \arg\min_{\mathcal{X}} \sum_{i\in\mathcal{K}_n}\sum_{j\in\mathcal{K}_m} \left\|\mathbf{r}_{i,j}\right\|^2 \tag{9}$$

The optimization problem is solved by g2o (Kümmerle et al., 2011) using the Levenberg-Marquardt algorithm. After this calibration, the poses of all AprilTag markers are fixed for the localization mode.

Localization mode. The localization mode estimates the camera poses directly using the known AprilTag poses obtained in the calibration mode. Hence, the objective function to solve is reformulated as

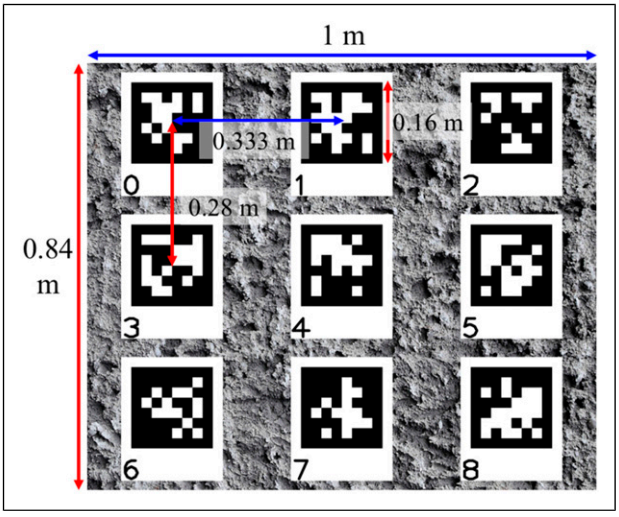


Figure 7. The design of the AprilTag markers on each panel.

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \sum_{i\in\mathcal{K}_n}\sum_{j\in\mathcal{K}_m} \left\|\mathbf{r}_{i,j}\right\|^2 \tag{10}$$

The design of this TankGT algorithm with two modes ensures efficient computation, good reliability and decent localization precision while preventing the localized poses from drifting over time. However, it can only be used in the underwater structure area where the AprilTag markers are visible.

To provide a complete GT trajectory, we include optimized camera poses by fusing the TankGT poses with the AQUA-SLAM algorithm (Xu et al., 2025). AQUA-SLAM is an underwater SLAM framework that fuses stereo camera, IMU, and DVL data in a tightly-coupled framework to achieve reliable pose estimation in challenging underwater conditions. For our purposes, we disable the loop closure module in AQUA-SLAM to obtain smooth odometry constraints between consecutive camera frames. To correct drift, the TankGT poses are fused as landmarks measurements, serving as external references. The overall problem is then formulated and solved as a factor graph optimization, as described below:

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \sum_{i\in\mathcal{K}_n}\sum_{j\in\mathcal{K}_m} \left\|\mathbf{r}_{i,j}\right\|^2 + \sum_{i\in\mathcal{K}_n} \left\|{}_{\mathrm{odom}}\mathbf{r}_{i,i+1}\right\|^2 \tag{11}$$

where

$${}_{\mathrm{odom}}\mathbf{r}_{i,i+1} \doteq \mathrm{Log}\left(\mathbf{T}^{-1}_{\mathrm{M}_0\mathrm{C}_i}\mathbf{T}_{\mathrm{M}_0\mathrm{C}_{i+1}}\widehat{\mathbf{T}}_{\mathrm{C}_i\mathrm{C}_{i+1}}\right) \tag{12}$$

and $\widehat{\mathbf{T}}_{\mathrm{C}_i\mathrm{C}_{i+1}}$ is the odometry constraints from the AQUA-SLAM algorithm (Xu et al., 2025). Since each sequence begins and ends near the underwater structure with AprilTag markers, pose estimates in regions where the AprilTag markers are not visible can be corrected using the TankGT poses, resulting in a complete and drift-free GT trajectory.

Accuracy of the generated ground truth

Since the AprilTag design on each panel is known (Figure 7), we have the theoretical relative pose between any pair of markers on the same panel. This allows us to compute an error against the calibrated poses of the AprilTag markers, to some extent, validating the accuracy of the generated ground truth camera poses. Figure 8 shows the cumulative density of the translation errors between the theoretical and the estimated AprilTag poses. We can see that over 90% of the estimated marker poses are within 3 cm, which is considered sufficient for the GT generation using measurement constraints from multiple AprilTag markers. Note that the actual errors are likely lower than this value given the fact that the panels may experience slight deformation in water.

SLAM evaluation using the dataset

In this section, we introduce our proposal to evaluate SLAM performance using the dataset and its tool set, before benchmarking four SLAM algorithms on the dataset.

Trajectory alignment

To evaluate a SLAM algorithm, its estimated poses are first paired with the GT poses using the timestamp. SE(3) transformation is then performed to align the estimated trajectory with the GT trajectory by minimizing the distances between the estimated poses and the GT poses. Specifically, given a set of estimated poses $\mathcal{S}_{\mathtt{est}} \doteq \{\mathbf{T}_{\mathtt{E}\mathtt{C}_i}\}, i \in \mathcal{K}_n$ in the SLAM frame E and its time-synchronized GT poses $\mathcal{S}_{\mathtt{gt}} \doteq \{\mathbf{T}_{\mathtt{M}_0\mathtt{C}_i}\}, i \in \mathcal{K}_n$ in the GT frame M_0, the optimal transformation $\mathbf{T}_{\mathtt{E}\mathtt{M}_0}$ that minimizes the sum of the squared distances between the estimated trajectory and the GT trajectory can be computed by solving

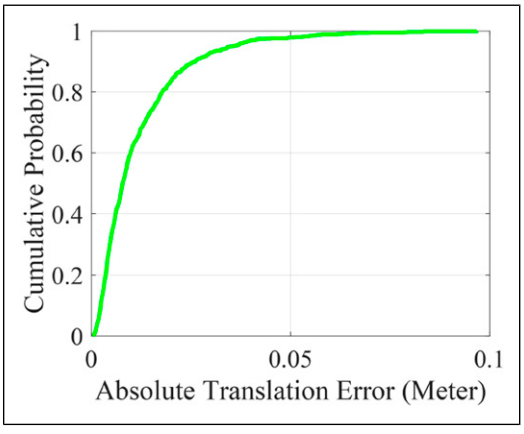


Figure 8. Cumulative translation error of the calibrated AprilTag poses.

$$\mathbf{T}^*_{\mathtt{E}\mathtt{M}_0} = \arg\min_{\mathbf{T}_{\mathtt{E}\mathtt{M}_0}} \sum_{i \in \mathcal{K}_n} \left\| \log\left(\mathbf{R}_{\mathtt{E}\mathtt{C}_i}\left(\mathbf{R}_{\mathtt{E}\mathtt{M}_0}\mathbf{R}_{\mathtt{M}_0\mathtt{C}_i}\right)^{\mathrm{T}}\right) \right\|^2 + \left\| {}_{\mathtt{E}}\mathbf{p}_{\mathtt{E}\mathtt{C}_i} - \mathbf{R}_{\mathtt{E}\mathtt{C}_i}\left(\mathbf{R}_{\mathtt{E}\mathtt{M}_0}\mathbf{R}_{\mathtt{M}_0\mathtt{C}_i}\right)^{\mathrm{T}}\left(\mathbf{R}_{\mathtt{E}\mathtt{M}_0\mathtt{M}_0}\mathbf{p}_{\mathtt{M}_0\mathtt{C}_i} + {}_{\mathtt{E}}\mathbf{p}_{\mathtt{E}\mathtt{M}_0}\right) \right\|^2 \tag{13}$$

where $\log(\cdot) : \mathrm{SO}(3) \to \mathbb{R}^3$ stands for the logarithm map from Lie Group SO(3) to a vectorized Lie algebra $\mathfrak{so}(3)$. Now the GT trajectory can be aligned with the estimated trajectory using $\mathbf{T}_{\widehat{\mathtt{E}\mathtt{C}_i}} = \mathbf{T}^*_{\mathtt{E}\mathtt{M}_0}\mathbf{T}_{\mathtt{M}_0\mathtt{C}_i}$.

Error calculation

After alignment, the absolute errors between the estimated trajectory and the transformed GT trajectory can be calculated as the RMSE rotation error

$$\mathbf{E}_{\mathrm{R}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\log\left(\mathbf{R}_{\mathtt{E}\mathtt{C}_i}\left(\mathbf{R}_{\widehat{\mathtt{E}\mathtt{C}_i}}\right)^{\mathrm{T}}\right)\right\|^2} \tag{14}$$

and the RMSE translation error

$$\mathbf{E}_{\mathrm{T}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|{}_{\mathtt{E}}\mathbf{p}_{\mathtt{E}\mathtt{C}_i} - {}_{\mathtt{E}}\mathbf{p}_{\widehat{\mathtt{E}\mathtt{C}_i}}\right\|^2} \tag{15}$$

where n is the number of paired poses in the trajectory.

Notably, the GT data is only available for the underwater structure area. For most sequences, such as the HalfTank and WholeTank sequences, the GT data is only available for part of the sequences (see the GT percentage in Table 3). This makes the relative error imprecisely reflect the performance. Therefore, the absolute error, instead of the relative error, is proposed for evaluation.

Evaluation tool set

To facilitate the evaluation, we also provide evaluation tools to generate visually appealing and publication-ready tables and figures for SLAM evaluation. The evaluation tools include:

(1) **Error Table**: A table of RMSE translation and rotation errors.
(2) **Error Distribution Heat Map**: A distribution plot of the RMSE translation and rotation errors of multiple runs to evaluate the robustness of a SLAM system.
(3) **Error with Time**: A figure shows the pose errors in six axes evolving with time.
(4) **Stereo Reconstruction**: A 3D reconstruction using the estimated camera trajectory and the stereo depth.
(5) **Trajectory Plot**: A 3D plot of the estimated trajectory and GT trajectory.

Table 4. SLAM performance (average of 10 runs) using the proposed Tank dataset.

	Translation error (in meter)				Rotation error (in degree)			
	UVA	SVIN2	ORB3	VINS	UVA	SVIN2	ORB3	VINS
Structure Easy	0.178	**0.090**	0.199	0.219	4.109	**1.756**	4.183	2.321
Structure Medium	**0.489**	2.111	3.494	42,359.413	**10.982**	57.510	90.014	65.957
Structure Hard	**0.432**	3.589	2.933	5165.797	**5.923**	23.439	102.615	38.244
HalfTank Easy	**1.121**	4.508	2.213	26.930	19.827	**3.456**	48.613	8.435
HalfTank Medium	**0.256**	2.902	0.708	15,772.612	**5.935**	24.153	15.159	45.991
HalfTank Hard	**0.367**	68.703	1.147	31,722.982	**9.850**	15.647	22.236	92.070
WholeTank Medium	**0.267**	0.406	0.682	4.106	9.202	**6.578**	8.195	91.497
WholeTank Hard	0.297	**0.274**	2.370	35,999.266	10.697	**3.796**	30.504	28.585

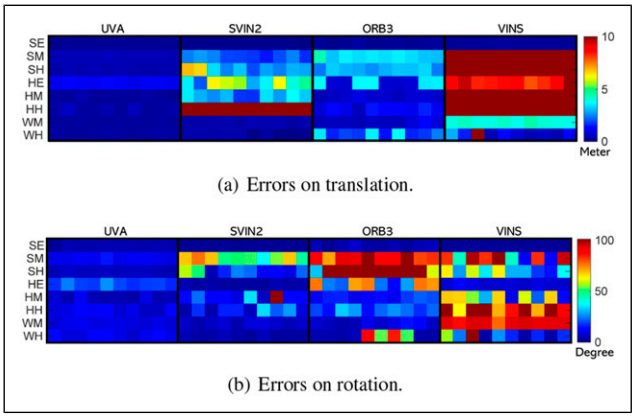


Figure 9. Results of 10 runs on all Tank sequences.

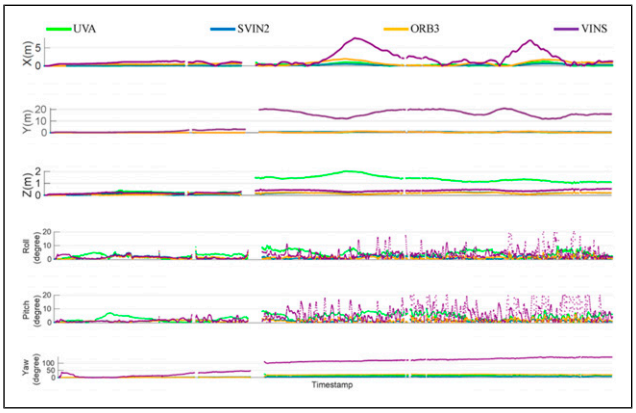


Figure 10. Errors on WholeTank Medium sequence. The timestamp gaps indicate the times without GT poses.

Evaluation results

We evaluate the performance of four SLAM systems, that is, Underwater Visual Acoustic SLAM (UVA) (Xu et al., 2021), SVIN2 (Rahman et al., 2022), ORB SLAM3 (Campos et al., 2020), and VINS-Fusion (Lin et al., 2018), on the proposed Tank dataset by reporting the generated results using the evaluation tool set. UVA is an underwater SLAM system fusing stereo cameras, a DVL, and a gyroscope in a loosely-coupled framework. SVIN2 is an underwater SLAM system fusing a stereo camera, an IMU, and a sonar. We only use the stereo camera and IMU data since sonar data is unavailable in the dataset. ORB SLAM3 and VINS-Fusion are two state-of-the-art visual-inertial SLAM systems. ORB SLAM3 is a feature-based SLAM system, while VINS-Fusion is an optical flow based system.

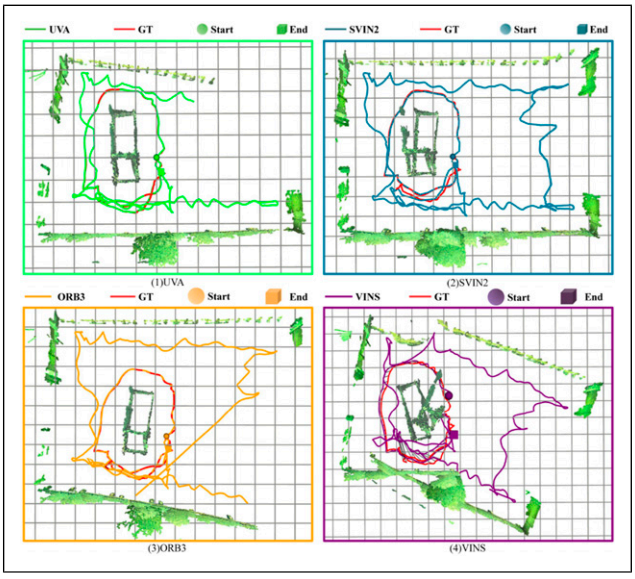


Figure 11. 3D reconstruction and SLAM trajectories on the WholeTank Hard sequence.

Error table. The RMSE absolute errors are presented in Table 4. UVA demonstrates the best performance in translation error across most sequences, attributed to the integration of the DVL. Regarding the rotation error, SVIN2 outperforms on less challenging sequences, while UVA excels in more challenging ones. ORB-SLAM3 performs well on the SE sequence but loses track in more challenging scenarios. VINS-Fusion also performs adequately on the SE sequence, but drifts rapidly in other sequences.

Error distribution heat map. The error distribution heat map, presented in Figure 9, illustrates the error deviations across 10 runs for each method. The UVA method demonstrates the highest robustness across the majority of the sequences.

Error with time. The pose errors on the WholeTank Medium sequence are shown in Figure 10.

Stereo reconstruction. We provide a convenient script to generate 3D reconstructions using the estimated camera trajectory generated by SLAM and the stereo depth generated by Block-Matching stereo matching algorithm. An example of the 3D reconstruction, SLAM trajectories and associated GT trajectories are shown in Figure 11.

Conclusion

In this paper, we introduce the Tank dataset, which includes multi-sensor data from a stereo camera, an IMU, a DVL, and a depth pressure sensor. Accurate 6-DoF ground truth poses are also provided by using the proposed TankGT algorithm along with AprilTag markers on an underwater structure. We validate the effectiveness of the dataset by using it to benchmark four SLAM systems. The dataset and the evaluation tools are publicly available at https://senseroboticslab.github.io/underwater-tank-dataset.

ORCID iDs

Tomasz Luczynski https://orcid.org/0000-0002-0970-8051
Sen Wang https://orcid.org/0000-0003-1537-8834

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplemental Material

Supplemental material for this article is available online.

References

Bernardi M, Hosking B, Petrioli C, et al. (2022) Aurora, a multi-sensor dataset for robotic ocean exploration. *The International Journal of Robotics Research* 41(5): 461–469.

Boittiaux C, Dune C, Ferrera M, et al. (2023) Eiffel tower: a deep-sea underwater dataset for long-term visual localization. *The International Journal of Robotics Research* 42(9): 689–699.

Campos C, Elvira R, Rodríguez JJG, et al. (2020) Orb-slam3: an accurate open-source library for visual, visual-inertial and multi-map slam. arXiv preprint arXiv:2007.11898.

Ferrera M, Creuze V, Moras J, et al. (2019) Aqualoc: an underwater dataset for visual–inertial–pressure localization. *The International Journal of Robotics Research* 38(14): 1549–1559.

Furgale P, Rehder J and Siegwart R (2013) Unified temporal and spatial calibration for multi-sensor systems. In: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1280–1286.

Guth F, Silveira L, Botelho S, et al. (2014) Underwater slam: challenges, state of the art, algorithms and a new biologically-inspired approach. In: *5th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics*, 981–986. DOI: 10.1109/BIOROB.2014.6913908.

Kümmerle R, Grisetti G, Strasdat H, et al. (2011) G2o: a general framework for graph optimization. In: *2011 IEEE International Conference on Robotics and Automation*, 3607–3613. DOI: 10.1109/ICRA.2011.5979949.

Lin Y, Gao F, Qin T, et al. (2018) Autonomous aerial navigation using monocular visual-inertial fusion. *Journal of Field Robotics* 35(1): 23–51.

Łuczyński T and Birk A (2017) Underwater image haze removal with an underwater-ready dark channel prior. In: *OCEANS 2017-Anchorage*. IEEE, 1–6.

Łuczyński T, Pfingsthorn M and Birk A (2017) The pinax-model for accurate and efficient refraction correction of underwater cameras in flat-pane housings. *Ocean Engineering* 133: 9–22.

Mahon I, Williams SB, Pizarro O, et al. (2008) Efficient view-based slam using visual loop closures. *IEEE Transactions on Robotics* 24(5): 1002–1014.

Mallios A, Vidal E, Campos R, et al. (2017) Underwater caves sonar data set. *The International Journal of Robotics Research* 36(12): 1247–1251.

Rahman S, Quattrini Li A and Rekleitis I (2022) Svin2: a multi-sensor fusion-based underwater SLAM system. *The International Journal of Robotics Research* 41(11-12): 1022–1042.

Schönberger JL and Frahm JM (2016) Structure-from-motion revisited. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Steinberg DM, Williams SB, Pizarro O, et al. (2010) Towards autonomous habitat classification using Gaussian mixture models. In: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 4424–4431. DOI: 10.1109/IROS.2010.5652480.

Thoms A, Earle G, Charron N, et al. (2023) Tightly coupled, graph-based dvl/imu fusion and decoupled mapping for slam-centric maritime infrastructure inspection. *IEEE Journal of Oceanic Engineering* 48(3): 663–676.

Vargas E, Scona R, Willners JS, et al. (2021) Robust underwater visual slam fusing acoustic sensing. In: *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2140–2146.

Wang J and Olson E (2016) AprilTag 2: efficient and robust fiducial detection. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Xu S, Luczynski T, Willners JS, et al. (2021) Underwater visual acoustic slam with extrinsic calibration. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7647–7652.

Xu S, Zhang K and Wang S (2025) Aqua-slam: tightly coupled underwater acoustic-visual-inertial slam with sensor calibration. *IEEE Transactions on Robotics* 41: 2785–2803.

Zhang S, Zhao S, An D, et al. (2022) Visual slam for underwater vehicles: a survey. *Computer Science Review* 46: 100510.

Zhao L, Zhou M and Loose B (2023) Tightly-coupled visual-dvl-inertial odometry for robot-based ice-water boundary exploration. In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7127–7134.