Journal of Field Robotics

WILEY

RESEARCH ARTICLE OPEN ACCESS

FLSea: Underwater Visual–Inertial and Stereovision Forward-Looking Data Sets

Yelena Randall[†] | Ori Lifschitz | Tali Treibitz

ViSEAon Marine Imaging Lab, Department of Marine Technologies, University of Haifa, Haifa, Israel

Correspondence: Yelena Randall (y4randall@gmail.com)

Received: 1 August 2025 | **Revised:** 29 April 2026 | **Accepted:** 25 June 2026

Funding: Israel Science Foundation, Grant/Award Number: 680/18; Israeli Ministry of Science and Technology, Grant/Award Number: 3-15621; Israel Data Science Initiative (IDSI); Data Science Research Center at the University of Haifa; European Union's Horizon 2020 research and innovation program, Grant/Award Number: GA 101016958

Keywords: 3D reconstruction | data set | forward-looking | optical | SLAM | stereo | underwater | visual–inertial

ABSTRACT

Visibility underwater is challenging and degrades as the distance between the subject and the camera increases. That is why forward-looking underwater computer vision tasks are difficult. We have collected underwater forward-looking stereovision and visual–inertial image sets using two underwater imaging platforms, a stereo camera rig, and an ROV in the Mediterranean and Red Seas. To our knowledge, there are no other public data sets in the underwater environment with this forward-looking camera-sensor orientation that have published ground-truth depth maps as well as pose. These data sets are critical for the development of several underwater applications, including autonomous obstacle avoidance, visual odometry, 3D tracking, Simultaneous Localization and Mapping and depth estimation through deep learning. The stereo data sets contain synchronized stereo images, and the visual–inertial data sets include monocular images and inertial measurement unit (IMU) measurements with millisecond-level timestamp alignment. All data was collected in dynamic underwater environments with objects of known size. Both sensor configurations allow for scale estimation, with the calibrated baseline in the stereo setup and the IMU in the visual–inertial setup. Ground-truth depth maps were created offline for both data set types using a commercial photogrammetry software (Agisoft Metashape). The ground truth is validated with multiple known measurements placed throughout the imaged environment. There are four stereo and 12 visual–inertial data sets in total, each containing thousands of images, with a range of different underwater visibility and ambient light conditions, natural and man-made structures, and dynamic camera motions. The forward-looking orientation of the camera plus the corresponding ground truth makes these data sets unique and ideal for testing underwater obstacle-avoidance algorithms and for navigation close to the seafloor in dynamic environments. We show results from an experiment with a monocular depth estimation algorithm to demonstrate the applicability of the data sets. With our data sets, we hope to encourage the advancement of autonomous functionality for underwater vehicles in dynamic and/or shallow-water environments.

1 | Introduction

Autonomous underwater vehicles (AUVs) are designed to operate untethered, able to navigate and perform tasks without user input. They use a combination of acoustic, inertial, geomagnetic, and visual sensors to perceive and interact with their environment. Most AUV operations assume a downward-looking sensing configuration from the midwater column

[†]Data sets are available at https://www.kaggle.com/datasets/viseaonlab/flsea-vi and https://www.kaggle.com/datasets/viseaonlab/flsea-stereo.

(Ferrera et al. 2019), enabling safe operation above the seafloor and simplifying obstacle avoidance. In this configuration, the obstacle-avoidance protocol allows the vehicle to maintain sufficient clearance from obstacles below using bathymetry (Carreras et al. 2018). In our work, we consider a less common, alternate scenario, where the AUV senses in the forward-looking field-of-view. This sensor configuration is required for operations in complex, shallow-water environments, where the vehicle must detect and avoid obstacles directly in its path. Such forward-looking autonomy requires scaled three-dimensional (3D) mapping in the direction of motion, which in turn motivates the need for data sets that support benchmarking of perception and navigation algorithms under these conditions. The data sets presented in our work are designed to fill this gap in the existing available data.

There are several reasons that we chose to focus on a forward-looking optical camera in this application: it provides high fidelity data at a close range to the subject; it has low size, weight, and power (SWaP) requirements compared with acoustic sensors; it operates as a passive sensing modality, suiting it to operations where active emissions are undesirable. We provide both monocular and stereo data because on a small system as we imagine, a stereo setup with the required baseline would be cumbersome, and could hinder the vehicle's ability to be agile. Whereas on a larger vehicle, a stereo payload could provide more accurate, scaled depth estimation. Using a visible camera gives us the advantage of high-resolution data at relatively high frame rates. However, there are challenges in using a camera for underwater 3D reconstruction. In water, light is attenuated as a function of range from the subject and scattered by particles suspended in the medium between the camera and the subject (Akkaynak and Treibitz 2018). This often results in an image with low contrast and in which details are veiled by scattered light (backscatter) (Jaffe 1990).

Outside the underwater domain, there has been significant development in monocular 3D scene reconstruction. A few well-known solutions include structure from motion (SFM), simultaneous localization and mapping (SLAM), and deep learning-based depth estimation. SFM is the computer vision solution for reconstructing 3D information from multiple images (Schönberger and Frahm 2016). Performed offline, SFM uses camera motion in an assumed-stationary world to build a 3D sparse depth map of the scene with relative camera poses (Hartley and Zisserman 2004). Several groups have built the foundation for SFM (Longuet-Higgins 1987; Mohr et al. 1995), some have expanded this research to include large-scale and urban image sets (Agarwal et al. 2011; Schaffalitzky and Zisserman 2002), while others have developed well-known open source and commercial SFM software (Schönberger and Frahm 2016; Mapillary 2021; Moulon et al. 2016; Lindenberger et al. 2021; Agisoft 2018). It is important to note that SFM was not conceived in the robotics domain and, therefore, is not designed to run in real time or on embedded platforms. In response, roboticists have applied several key principles of SFM to devise their own solution for navigation, known as SLAM (Mur-Artal et al. 2015).

The goal of visual SLAM is autonomous navigation through unknown environments. Unlike SFM, it operates in real time, tracking platform motion while simultaneously building a map of the scene. Using a practice known as loop closure, SLAM is able to correct for errors accumulated over time by revisiting previously mapped locations. This is particularly useful for global positioning system (GPS)-limited environments, and preferred over purely odometry-based tracking methods due to its ability to correct for localization error over time. SLAM comes in many forms, using different methods (feature-based or direct), producing different map types (sparse or dense), and relying on different sensors, such as monocular or stereo cameras, RGB-D, or light detection and ranging (LiDAR). With our our desired use-case in mind, we will mention several well-known monocular visual methods, such as ORB-SLAM3, LSD-SLAM, DSO, and SVO (Campos et al. 2021; Engel et al. 2014, 2016; Forster et al. 2014). The issue that monocular SLAM, SFM, and deep learning-based depth estimation have in common is that metric scale cannot be derived from monocular images alone. In other words, without a sensor or external cue which provides scaled measurements, scene depth and camera pose are estimated at an arbitrary scale.

Visual–inertial odometry (VIO) is an efficient monocular camera based algorithm compared with computationally intensive SLAM approches that also solves the SLAM problem. Additionally, it solves the unknown scale issue by adding an inertial measurement unit (IMU) (Rosinol et al. 2020; Huai and Huang 2022). An IMU is an affordable small sensor and standard on a robotic system. VIO uses a monocular camera and an IMU to estimate the robot state. With the addition of an inertial sensor, VIO makes it possible to estimate the camera pose in metric scale (Scaramuzza and Zhang 2019). It is becoming increasingly common to combine SLAM and VIO; using them in concert solves the scale issue in monocular SLAM via the relatively minor addition of the IMU, and uses vision-only loop closure to correct for localization error. Some successful examples of this include ORB-SLAM3, ROVIO, maplab, and VINS-Fusion (Campos et al. 2021; Bloesch et al. 2017; Schneider et al. 2017; Qin et al. 2019). Visual–Inertial SLAM outputs a metric-scaled map of the scene, and can correct for error over time using loop closure (Mur-Artal and Tardós 2017).

Outside of robotics, deep learning is increasingly used for depth estimation from monocular images (Li and Snavely 2018; W. Yin et al. 2020; Li et al. 2022; Yuan et al. 2022; Kim et al. 2022; Agarwal and Arora 2022), a designed for real-time or embedded platforms except for Wofk et al. (2019). The output of a monocular deep learning network is a dense depth map with unknown scale.

These algorithms have been successful in a variety of ground domain tasks and there are multiple benchmark data sets used for development, including KITTI, EUROC, and TUM RGB-D (Geiger et al. 2013; Burri et al. 2016; Sturm et al. 2012). While we do see success on land, there are still not many groups utilizing this technology in this particular sensor configuration underwater. There are a few predictable challenges in creating an underwater system that relies on vision for perception and navigation. Aside from the effects of the underwater environment on image quality, the unstructured nature of underwater environments poses a significant issue for algorithms designed

and tested in structured, man-made environments. To move past these challenges, underwater forward-looking visual data is necessary for development and testing, but has proven extremely limited. Collecting data of this type is time-consuming, logistically complex, and often costly. For the underwater autonomous tasks we imagine, one needs forward-looking underwater images at a high frame rate (10 fps or above), scale information for testing and training, and ground truth for validation and supervision. Underwater data is challenging to acquire, and without access to publicly available data, the progress of underwater 3D image reconstruction in the forward-looking view is limited.

Despite these challenges, one group has compared visual–inertial-based state estimation algorithms in the underwater domain in the forward-looking perspective (Joshi et al. 2019) and extended OKVIS (Leutenegger et al. 2014) to accommodate data from an underwater acoustic sensor (sonar) (Rahman et al. 2018). In the latter case, sonar is utilized to provide sparse, but robust-scaled information about obstacles in the scene. A second group introduces WaterMBSL, a movable binocular structured-light system with motion compensation and point-cloud registration for high-precision underwater dense reconstruction on a mobile robot (Ou et al. 2024). They then go on to present ROV-Scanner, a 3D dense mapping system using scanning binocular structured light fused with DVL, IMU, and pressure data to enable underwater collision-free navigation in low-light environments (Ou et al. 2025a). Finally, they propose Hybrid-VINS, a tightly coupled hybrid visual–inertial navigation system that fuses active structured-light vision with monocular VINS to improve depth estimation, dense mapping, and loop-closure detection, and validate it with data sets collected in simulated and real-world underwater environments (Ou et al. 2025b). A third group does a comparison of VI-SLAM in the underwater domain using a forward-looking underwater data set, collected on a GoPro, that speaks to the value of inexpensive commercial off-the-shelf (COTS) sensors for underwater navigation and adds forward-looking visual–inertial data to the public domain (Joshi et al. 2022). A fourth group collected a forward-looking stereo data set that allows for 3D object reconstruction using stereo (Lodi Rizzini et al. 2015). A fifth group focused on the task of inspection and maintenance of offshore structures using deep learning-based depth estimation. They collected a stereo image data set called NEREON from the DURIUS buoy for which they generated ground-truth depth maps using a multiple light stripe range and a photometric stereo (Dionísio et al. 2023). This data set is another useful data set for development in underwater depth estimation through deep learning because of the inclusion of ground-truth depth maps, but does not cater to the goal of underwater passive localization and obstacle perception that we focus on in this paper. We found several other multicam inertial and stereo forward-looking data sets collected in test pools which show potential in VI-SLAM and stereo SLAM (Singh et al. 2023; Luczynski et al. 2021). There are also some downward-looking underwater visual data sets available, including AQUALOC: An underwater data set for visual–inertial-pressure localization (Ferrera et al. 2019), the Underwater Caves Sonar Dataset (Mallios et al. 2016), and ACFR Marine (Barrett et al. 2011).

We are pleased to find that more underwater data sets with an optical sensor as the primary sensor have been published in the past five years. One review of SLAM techniques for AUVs cites several of these groups embracing the challenge of underwater vision-based navigation (Zhao et al. 2019). This demonstrates both the need for this type of data as well as the viability of a COTS camera for underwater depth estimation and navigation. Some of the only existing publicly-available forward-looking data to fulfill many of the criteria we describe comes from a single group in South Carolina (Rahman et al. 2018), that focuses on underwater caves with low ambient lighting. To the best of our knowledge, our set, FLSea remains the only underwater forward-looking data set that *publishes ground-truth depth maps* as well as camera pose. A comparison of these data sets can be seen in Table 1.

Access to some of the standard methods for collecting ground truth in air, such as LiDAR scanning, remains an obstacle in collecting this data in underwater environments. The solution we use is a commercial SFM-based software called Agisoft Metashape (Agisoft 2018) to create dense depth maps for each image frame in the data sets using objects of known size to validate the result. In the visual–inertial sets, the objects of known size are used to scale the ground-truth model. In the underwater domain, where access to common "ground-truth" sensors is limited, this process is the closest approximation to the standard methods and what we refer to as ground truth in this paper.

The depth maps provided in this data set are vital for training and testing depth estimation with deep learning. As advances in accuracy and processing speed are made, the use of deep learning-based depth estimation in robotics is becoming more accessible and successful; in addition to using SLAM and VIO underwater, Teixeira et al. (2020) show the potential of using unsupervised methods such as SfMLearner (T. Zhou et al. 2017) and GeoNet (Z. Yin and Shi 2018) for monocular depth estimation underwater. Through the addition of our data set's depth maps, this can be expanded to supervised methods.

Our data set, FLSea, contains two types of forward-looking visual data, collected on a stereo setup and a monocular visual–inertial setup. The stereo data sets consist of synchronized pairs of images, with objects of known size placed in the scene, intrinsic and extrinsic calibrations and ground-truth depth maps for each frame. In the visual–inertial data set, the IMU readings act as a temporal baseline, linking sequential frames by distance traveled as recorded by the IMU. It is important to note here that IMUs are prone to error accumulation over time, also known as drift.

These data sets, together with the ground-truth that we generated, can be used in evaluating and training VIO, VI-SLAM, Stereo SLAM, and monocular depth reconstruction algorithms. Access to public visual data is critical to the development of any visual task, regardless of domain, but particularly in the underwater domain where data of this type is lacking. We hope that FLSea will serve as a benchmark data set for underwater forward-looking tasks; To demonstrate its applicability, we show results in Section 5 from an

TABLE 1 | Review of underwater data sets. A check mark under the illumination column indicates that artificial illumination was used.

Name	Platform	Sensors	Illum.	Ori.	GT	Reference
Natural environment						
FLSea	*ROV, diver*	*Mono. Cam, IMU, and Stereo Cam.*		*FW*	*Pose, depth*	*Ours*
NEREON[a]	Diver	Stereo Cam., LSR, PS	✓	FW	Depth	Dionísio et al. (2023)
GoPro VI-SLAM	Diver	Mono. Cam., IMU	✓	FW	Pose	Joshi et al. (2022)
Aqualoc Dataset	ROV	Mono. Cam., IMU	✓	DW	Pose	Ferrera et al. (2019)
Sonar SLAM	Diver	Stereo Cam., IMU, Sonar	✓	FW	Pose	Rahman et al. (2018)
Underwater cave	AUV	Mono. Cam., IMU, Sonar, DVL	✓	DW	Pose	Mallios et al. (2016)
Maris data	Diver	Stereo Cam.		FW	None	Lodi Rizzini et al. (2015)
ACFR Marine	AUV	Stereo Cam., IMU, Sonar		DW	Pose, labels	Barrett et al. (2011)
Pool						
Multi-Cam. VI	ROV	5 Cam, IMU	✓	FW	Pose	Singh et al. (2023)
Inspection and intervention[a]	ROV	Stereo Cam.	✓	FW	None	Luczynski et al. (2021)

Abbreviations: AUV, autonomous underwater vehicle; DVL, doppler velocity log; DW, downward-looking; FW, forward-looking; GT, ground truth; Illum., illumination; IMU, inertial measurement unit; LSR, laser; Mono., monocular; Ori., orientation; PS, photometric stereo; ROV, remotely operated vehicle; VI, visual; VI-SLAM, visual inertial simultaneous localization and mapping.
[a]These data sets are specifically for the task of inspection, maintenance, and intervention of offshore structures, unlike the others that are more suited to the task discussed in this paper: underwater localization and obstacle perception.

experiment with a monocular depth estimation algorithm using our data sets.

2 | Sensor Setup

The data sets were collected on two different imaging platforms, a diver-held stereo rig and the BlueROV2 (Blue Robotics 2022), the visual–inertial platform.

2.1 | Visual–Inertial

We used the BlueROV2 (Blue Robotics 2022) as our visual–inertial system. It contains open-source electronics and software, and was customized in our lab. It is equipped with an iDS camera and a VectorNav IMU, running on an Nvidia Jetson Nano as the embedded processor in a housing pressure rated to 100 m. The housing is equipped with a dome port (Figure 1). The images are of resolution 968×608, captured at 10 fps (Table 2).

The IMU and camera are not hardware-synchronized; however, both sensors timestamp their measurements using the same system clock with nanosecond resolution. The camera records images at 10 Hz, while the IMU operates at a rate of 100 Hz, providing inertial measurements at approximately 10× the camera frequency. As a result, for each camera frame, there exist multiple IMU samples in close temporal proximity. Given the high IMU sampling rate relative to the camera frame rate, the maximum temporal offset between a camera frame and its nearest IMU measurement is bounded by a few milliseconds. However, for the earlier "Canyon" data sets, the IMU operated at 20 Hz. For these sets, IMU measurements are linearly interpolated to 100 Hz to achieve accurate temporal alignment.

2.2 | Stereo

The stereo rig is comprised of two Nikon D810s, in Hugyfot underwater housings which are mounted on a diver-held rig, with a fixed baseline of 0.317743 m. The housings are pressure rated to 100 m. The housing is equipped with a dome port to minimize distortion that occurs due to refraction at the interface between water and air. The two cameras are hardware-synchronized, using a custom synchronization cable which transmits a signal from one camera to the other when either one of the shutter buttons is pressed (Figure 2). The stereo image resolution is 1280×720 and captured at 10 fps in video mode (Table 3).

Video mode allows for higher frame rate, but only requires one shutter-press at the start and end of each capture. This means that there is a hardware sync for the first stereo image pair, but not for the subsequent images in the video. We performed an experiment to ensure that there was no time-lag between left and right image pairs in this mode. This involved placing a digital stopwatch with millisecond units in view of both cameras and collecting video. We then extracted the stopwatch text from each frame and calculated the time delta between left-right image pairs. This

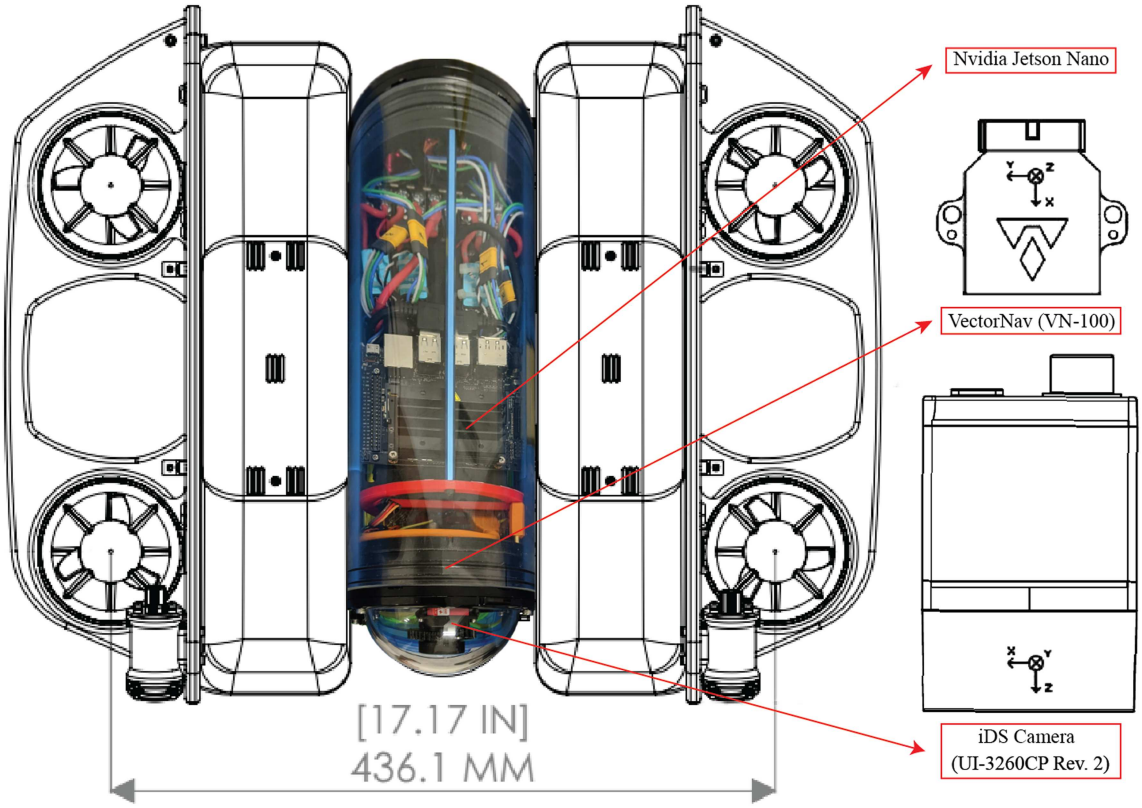

FIGURE 1 | The BlueROV 2 in the four-thruster configuration. The notable sensors here are the VN-100 IMU (VectorNav 2022) and the iDS camera (iDS 2022). Blueprint from (Blue Robotics 2022), center tube is the actual payload tube from our BlueROV2 setup. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2 | Technical specifications of the visual–inertial setup.

Camera	
Camera model	iDS camera (UI-3260CP Rev. 2)
Sensor type	Sony IMX249, 1/1.2″ CMOS
Sensor size	2.35 Megapixels
Focal length	4 mm
Opening angle	80°
Resolution	968 × 608
Frequency	10 fps
IMU	
IMU model	VectorNav (VN-100)
Accel. rate	100 Hz[a]
Gyro. rate	100 Hz[a]
Gyro in-run bias	5°/h
Accel. in-run bias	< 0.04 mg
Heading	2.0°
Static pitch and roll	0.5°
Computer	
Processor model	Nvidia Jetson Nano
Underwater housing	
Housing model	Blue robotics 4″ tube
Depth rating	100 m
Port	Blue robotics 4” dome port

Abbreviations: CMOS, complementary metal-oxide-semiconductor; IMU, inertial measurement unit.
[a]20 Hz in “Canyons”, 100 Hz in “Red Sea” data.

experiment showed a consistent stereo synchronization down to millisecond resolution, which is sufficient given our swimming pace of less than 1 m/s.

2.3 | Calibration

Each system has been calibrated to find intrinsic and extrinsic parameters. The intrinsic parameters of a camera are the focal length, optical center, and distortion coefficients. The intrinsic parameters are expressed by a 3 × 3 matrix, K, which holds the focal length f_x, f_y and optical center c_x, c_y:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}. \quad (1)$$

Distortion is classified as either radial or tangential, and the coefficients are solved for during the calibration process. The calibration is performed using a checkerboard of known size and the Matlab camera calibrator package (MATLAB 2022). These parameters are used to correct for lens distortion and project camera points onto the world frame. The extrinsic parameters of the system are the locations of the sensors in the system with respect to a defined origin. We offer the calibration parameters in the form of a yaml configuration file typical to that used in many open-source SLAM algorithms.

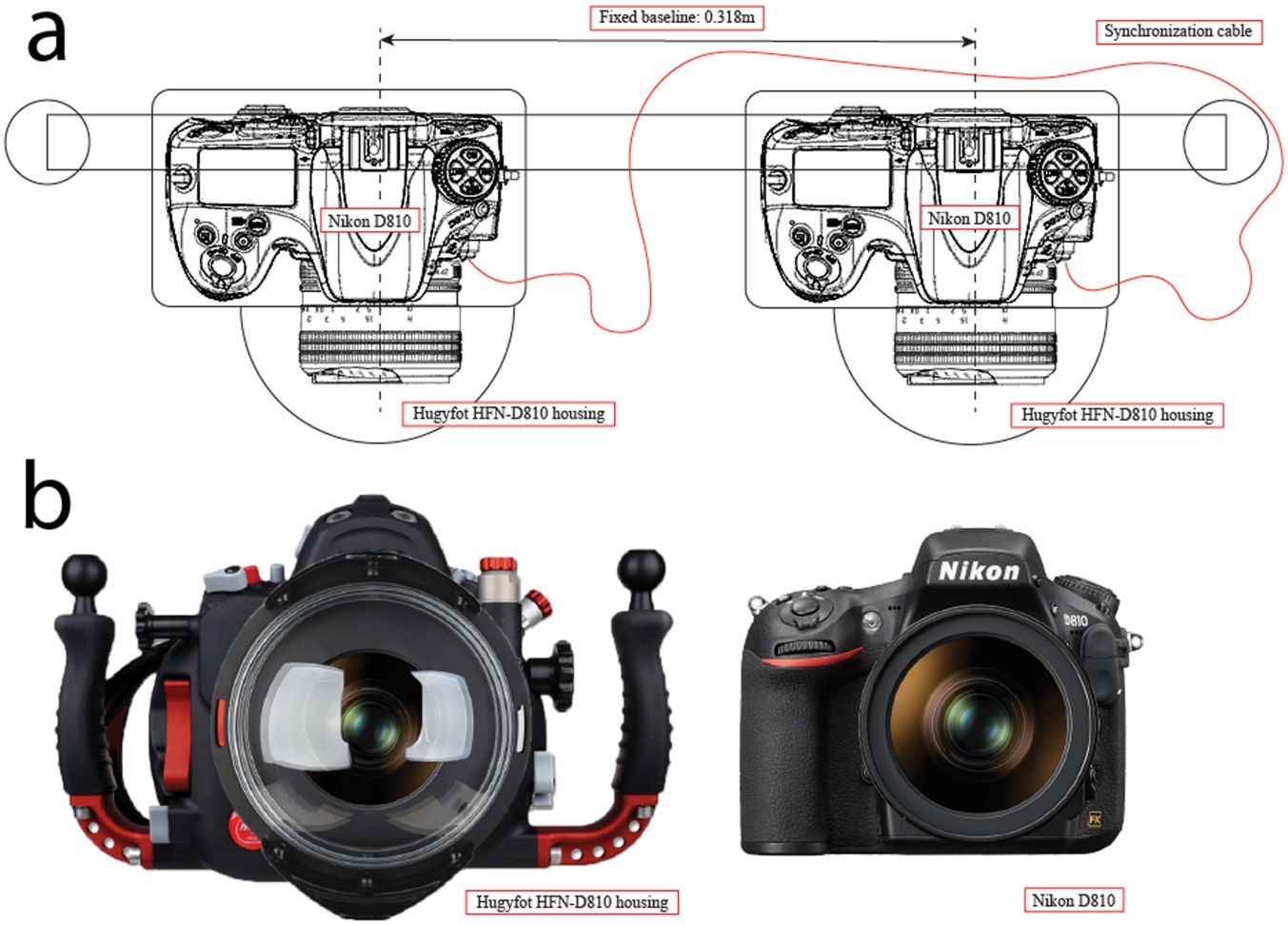

FIGURE 2 | (a) Stereo setup diagram with the synchronization cable and outline of the static jig (fixed baseline) and underwater housings sketched, camera sketches, and (b) Stereo setup components. (Left) Underwater housing (Hugyfot 2022). (Right) A Nikon SLR camera (Nikon 2022). [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 3 | Technical specifications of the stereo setup.

Camera (×2)	
Camera model	Nikon D810
Sensor type	full-frame CMOS sensor
Sensor size	35.9 × 24 mm
Shutter type	Rolling focal-plane
Focal length	35 mm
Opening angle	54.3°
Resolution	1280 × 720
Frequency	10 fps
Underwater housing (×2)	
Housing model	Hugyfot HFN-D810
Depth rating	100 m
Port	Dome port

Abbreviations: CMOS, complementary metal-oxide-semiconductor; IMU, inertial measurement unit.

In the stereo setup the intrinsics must be calculated for each of the cameras as well as the extrinsics of the system. In the case of the stereo setup, the extrinsic parameter is the transformation from one camera (C_1) to the other (C_0). This is expressed by a 4×4 matrix, T_s. A stereo camera setup is often defined by its *baseline*, or the fixed distance between the two cameras, which is the vector length of the translation from T_s.

$$T_s = \left[\begin{array}{ccc|c} & R_{C_1}^{C_0} & & t_{C_1}^{C_0} \\ 0 & 0 & 0 & 1 \end{array}\right] \quad (2)$$

where $R_{C_1}^{C_0}$ is the rotation matrix from C_1 to C_0 and $t_{C_1}^{C_0}$ is the translation from C_1 to C_0. We conducted the extrinsic calibration using a checkerboard and the Matlab stereo camera calibrator package (MATLAB 2022). To find the stereo baseline, the checkerboards are detected in each side of the stereo pair and the pose of one camera in relation to the other is determined, giving us the transformation matrix T_s.

In the visual–inertial setup the extrinsic parameter is the transformation between the IMU and the camera (C). The extrinsic parameters are determined with an AprilTag board and the Kalibr toolbox (Rehder et al. 2016).

$$T_s = \left[\begin{array}{ccc|c} & R^C_{\text{imu}} & & t^C_{\text{imu}} \\ 0 & 0 & 0 & 1 \end{array}\right] \quad (3)$$

where R^C_{imu} is the 3×3 rotation matrix between the IMU and the camera C and t^C_{imu} is the 3×1 translation matrix between the IMU and the camera C.

2.3.1 | *Calibration Method*

A calibration was collected at the start of each data set regardless of imaging setup. The order of actions was to first position the calibration board in a stable location on the seafloor, leaning against a rock or other structure, sometimes using stones to pin the calibration board against the supporting structure. After the board was secured, we began image collection at about 3 m distance from the board, ensuring to capture the calibration board at all locations within the image frame, paying special attention to the corners and edges as to accurately capture the lens distortion. We aimed for around 3–5 min of data collection for each calibration, or 1800–3000 frames. Some variation of distance between the camera setup and calibration board was also incorporated into the calibration collects. For the stereo setup, we made sure to capture plenty of images where the calibration board was fully visible in each camera frame, so that the extrinsic calibration would be possible.

3 | Data Sets

We present two types of data sets, visual–inertial and stereo. There are 12 visual–inertial data sets and four stereo data sets. The data sets are summarized in Table 4. Examples of the image data can be seen in Figures 3 and 4. They were collected in the Mediterranean and Red Sea, over the course of eight dives. The stereo data includes images at 10 fps, with objects of known size placed in the scene. On each dive, a calibration set was collected, used to determine the intrinsic and extrinsic (stereo baseline) parameters. The visual–inertial data includes images at 10 fps, also with objects of known size placed in the scene, and IMU data at 20–100 Hz. A calibration set was also collected for each visual–inertial dive, allowing for calculation of the intrinsic and extrinsic (camera–IMU transformation) parameters. The data sets include the original images that are unenhanced and a second version enhanced with SeaErra software (Treibitz et al. 2022).

3.1 | Visual–Inertial Data

The visual–inertial data sets were captured in a few different locations, and are organized according to their environment type. The Canyons data sets were captured in Nachsholim, the Mediterranean, Israel, in three different canyons, totaling four data sets. The water depth for these four data sets range from 4 to 7 m. The Red Sea data sets were captured in Eilat, Israel, in a few different areas, totaling eight data sets. The water depth for these eight data sets range from 3 to 8 m. The visual–inertial data includes RGB images at 10 fps and IMU data taken at 20 Hz (Canyons) and 100 Hz (Red Sea). Each of the data sets include a measuring tape laid on the seafloor, with visual targets of known size spaced a meter apart along the measuring tape as well as the calibration checkerboard or AprilTag board placed somewhere in the scene. Some data sets include loop-closure events, and some of the data sets include the same objects or similar trajectory as other data sets. The canyon data sets were collected in an area with substantial natural 3D structure, which serves as a challenging test for navigation and obstacle avoidance in complex environments. There is no man-made structure in the Canyon data sets. The only objects of known size are the ones placed in the scene as mentioned earlier. More information about these objects can be found in the Supplementary material. The Red Sea data sets do have some man-made structure in the scene (a pier, cement blocks, and coral tables) and much less natural structure than the Canyon data sets. In addition to the objects of known size placed in the scene, some of the man-made structures were measured as detailed in the Appendix.

There are a few image artifacts present in these data sets that are representative of the common challenges faced in underwater imaging. For a few of these sequences, the exposure was set to a static value, meaning that when there is a transition from low ambient light (in the canyon) to higher ambient light (out of the canyon, more shallow), some of the image frames are over exposed. While not ideal, this is an issue that is encountered when imaging in complex underwater environments with a system that does not have artificial lighting and uses a fixed exposure. Underwater caustics are present in many of the data sets. Underwater caustics can be observed in shallow water, and are the result of a wavy sea-surface, which reflects or refracts light rays (Agrafiotis et al. 2018), causing ever-changing light patterns on the seafloor. All of the data sets also exhibit attenuation and turbidity, affecting the range at which scene objects can be seen and the clarity at which we see them. Examples of all of these phenomena can be seen in Figure 5.

3.1.1 | *Visual–Inertial Data Collection Method*

The visual–inertial data sets were collected with the BlueROV2. To ensure a smooth route, the thrusters were shut off and the diver manually maneuvered the ROV throughout the environment. The diver maintained a slow pace of less than 1 m/s, and strived to limit movement on the vertical axis. The goal in this methodology was to keep a slow pace as to collect a large amount of frames per object in the scene. The camera and IMU were controlled by an operator on topside. The topside operator monitored the video feed and IMU measurements and adjusted camera exposure and gain if a large change in ambient light occurred. At the beginning of each dive a calibration set was collected, where the calibration board was positioned and intrinsic and extrinsic sets were collected. After the calibration set, the measuring tape and visual targets spaced a meter apart were laid on the seafloor and data collection commenced. Each data set begins at the measuring tape. The diver then swims the ROV around the environment at a slow pace which mimics that of a small underwater vehicle.

TABLE 4 | FLSea data set summary.

Visual–inertial							
Location	**Name**	**Images**	**Length**	**Average speed (m/s)**	**Distance (m)**	**Desc.**[a]	**Sync**
Canyons	U Canyon	2895	0:04:49	0.16	46.9	TDC	Camera (10 Hz) and IMU (20 Hz, interpolated to 100 Hz) with millisecond timestamps from shared clock
	Flatiron	2474	0:04:07	0.19	47.4	TDC	
	Horse Canyon	1520	0:02:35	0.19	43.8	TDCL	
	Tiny Canyon	958	0:01:49	0.12	16.3	TDCL	
	Calibration	535	0:00:53	N/A	N/A	N/A	
Red Sea	Northeast Path	1759	0:02:59	0.23	59.8	TDC	Camera (10 Hz) and IMU (100 Hz) with millisecond timestamps from shared clock
	Landward Path	1204	0:02:02	0.18	21.6	TDC	
	Dice Path	1428	0:02:24	0.21	31.5	TDC	
	Pier Path	1336	0:02:15	0.25	42.3	TDC	
	Coral Table Loop	1017	0:01:43	0.16	16.2	TD	
	Cross Pyramid Loop	1652	0:02:48	0.19	31.5	TDC	
	Big Dice Loop	3159	0:05:20	0.27	86.2	TDC	
	Sub Pier	1091	0:01:55	0.16	18.6	TDC	
	Calibration	1849	0:03:07	N/A	N/A	N/A	
Total	*Not including cal.*	**20,493**	**0:34:46**				
Stereo							
Location	**Data set name**	**Images**	**Length**	**Average speed (m/s)**	**Distance (m)**	**Average dispersion (pix)**	**Sync**
Canyons	Canyon 1	3803 L/R pairs	0:06:20	0.29	90.2	75	Hardware at video start
	Canyon 2	2362 L/R pairs	0:03:56	0.24	62.5	71	
Shallow	Rock Garden 1	867 L/R pairs	0:01:26	0.16	79.9	77	
	Rock Garden 2	305 L/R pairs	0:00:30	0.43	10.5	86	
	Calibration	413 L/R pairs	0:00:41	N/A	N/A	N/A	
Total	*Not including cal.*	**7337**	**0:12:12**				

[a]C, caustics; D, dynamic objects; Desc., visual descriptor key; IMU, inertial measurement unit; L, low light; T, turbidity.

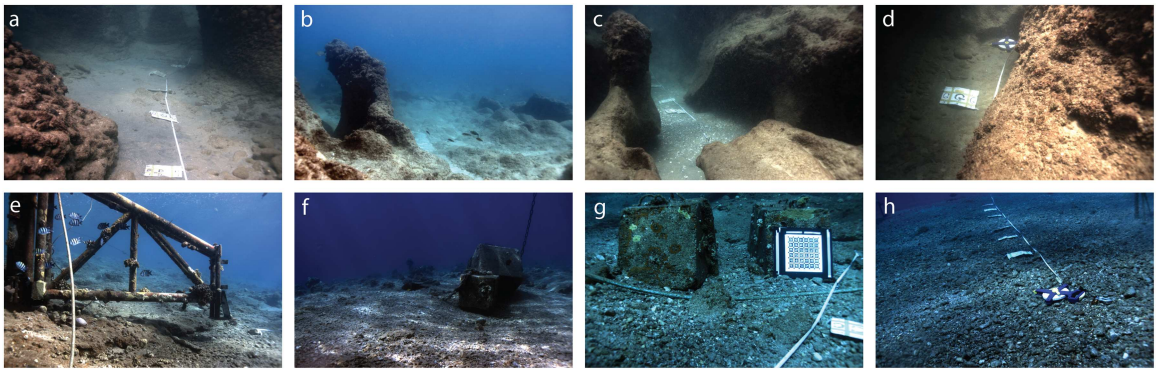

FIGURE 3 | Examples from the visual–inertial data sets. (a) U Canyon, (b) Flatiron, (c) Horse Canyon, (d) Tiny Canyon, (e) Northeast Path, (f) Big Dice Loop, (g) Landward Path, and (h) Cross Pyramid Loop. [Color figure can be viewed at wileyonlinelibrary.com]

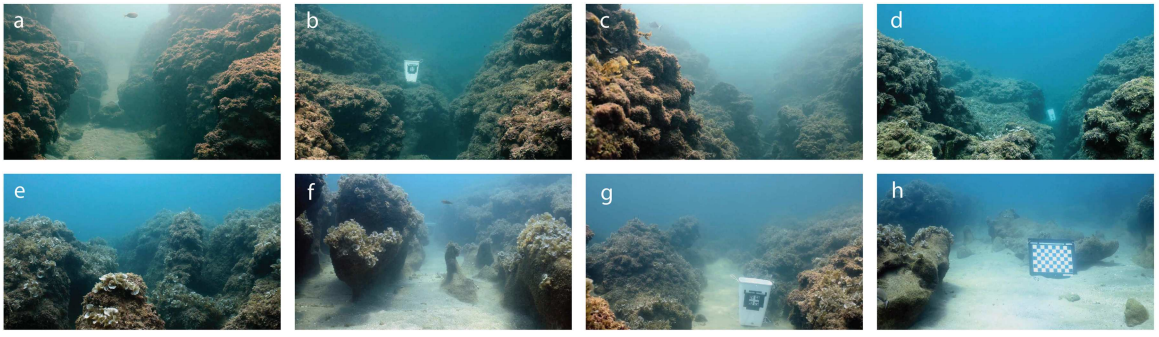

FIGURE 4 | Examples from the stereo data sets: (a–d) Canyon 1, (e–g) Rock Garden 1, and (h) Rock Garden 2. [Color figure can be viewed at wileyonlinelibrary.com]

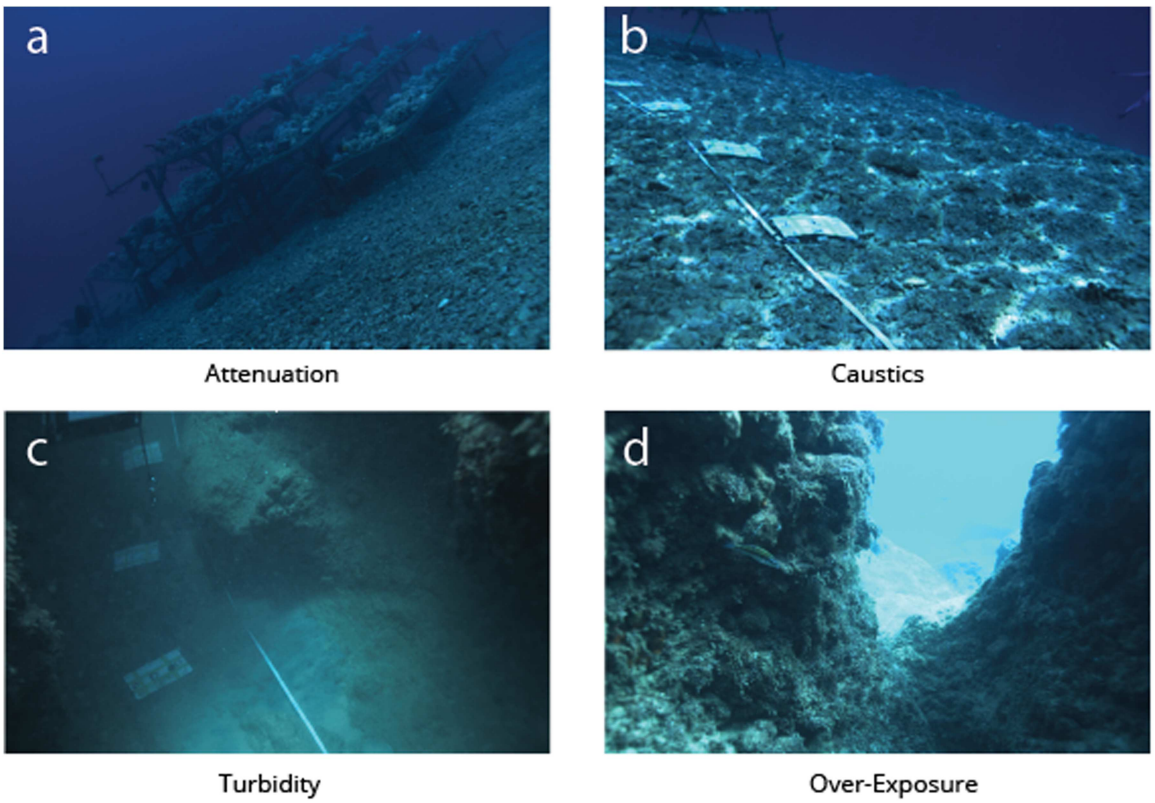

FIGURE 5 | Underwater image phenomena. (a) Attenuation, (b) caustics, (c) turbidity, and (d) overexposure. [Color figure can be viewed at wileyonlinelibrary.com]

3.1.2 | *Visual–Inertial Data Set Format*

The visual–inertial data sets are organized as follows.

```
vi
└── location_name
    ├── calibration
    └── dataset_name
        ├── depth
        │   ├── timestamp_SeaErra_abs_depth.tif
        │   └── ...
        ├── imgs
        │   ├── timestamp.tiff
        │   └── ...
        ├── metashape
        │   ├── R_t.txt
        │   ├── point_cloud_RGB.txt
        │   └── dataset_name_unenhanced.ply
        ├── seaErra
        │   ├── timestamp_SeaErra.tiff
        │   └── ...
        ├── imu.txt
        └── notes.txt
```

Each data set includes the original images, enhanced images, an "imu.txt," file, depth maps for each image frames, model exports which contain camera poses and XYZ points, video files displaying different elements of the data set and a "notes.txt," file which contains metadata details about the data set.

3.2 | Stereo Data

The stereo data sets were captured on a single day, over the course of two dives in Nachsholim, Israel. The water depth in these data sets ranges from 3 to 8 m. There are two data sets captured within a canyon, at a maximum water depth of 8 m and three data sets captured on a flat shallower area at a maximum water depth of 5 m. The data sets include RGB stereo images collected at 10 fps. Each data set includes objects of known size, spaced evenly throughout the scene. Most of the data sets include loop-closure events, meaning that locations were revisited throughout the data set. There is also some challenging camera motion in these data sets, mainly being fast rotation. Other than the objects that were placed in the scene, there is no man-made structure in these scenes. The canyon data sets include plenty of natural structure in every frame. The flat data sets, however, have limited 3D structure and lots of homogeneous sandy areas. There is more turbidity than in the canyon data sets.

3.2.1 | *Stereo Data Set Collection Method*

The stereo data sets were collected on the stereo rig in video mode. At the beginning of each dive, the exposure and focus were set, a calibration was performed and 3D objects of known size were placed throughout the scene. After the setup, data collection was performed by swimming through the environment with the stereo rig. During data collection, a point was made to revisit known locations, providing opportunity for "loop-closure" events.

3.2.2 | *Stereo Data Set Format*

The stereo data sets are organized as follows.

```
stereo
├── calibration
└── dataset_name
    ├── depth
    │   ├── LFT
    │   │   ├── LFT_image_number_abs_depth.tif
    │   │   └── ...
    │   └── RGT
    │       ├── RGT_image_number_abs_depth.tif
    │       └── ...
    ├── imgs
    │   ├── LFT
    │   │   ├── LFT_image_number.tiff
    │   │   └── ...
    │   └── RGT
    │       ├── RGT_image_number.tiff
    │       └── ...
    └── notes.txt
```

4 | Ground Truth

Obtaining ground-truth underwater is a notorious challenge. GPS, a method for determining absolute position, is not

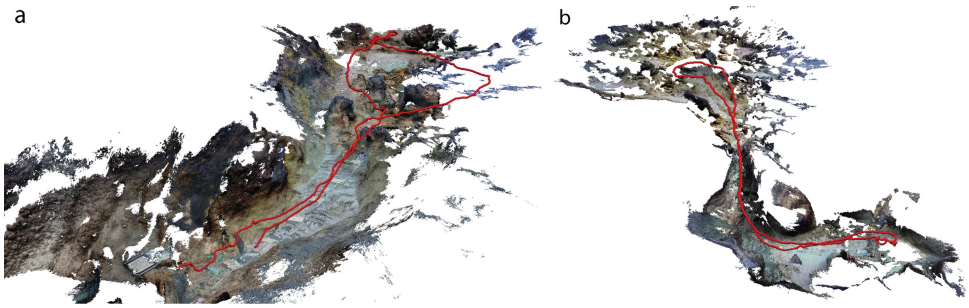

FIGURE 6 | Examples of Agisoft Meshes from (a) Flatiron and (b) U Canyon with pose trajectory in red. [Color figure can be viewed at wileyonlinelibrary.com]

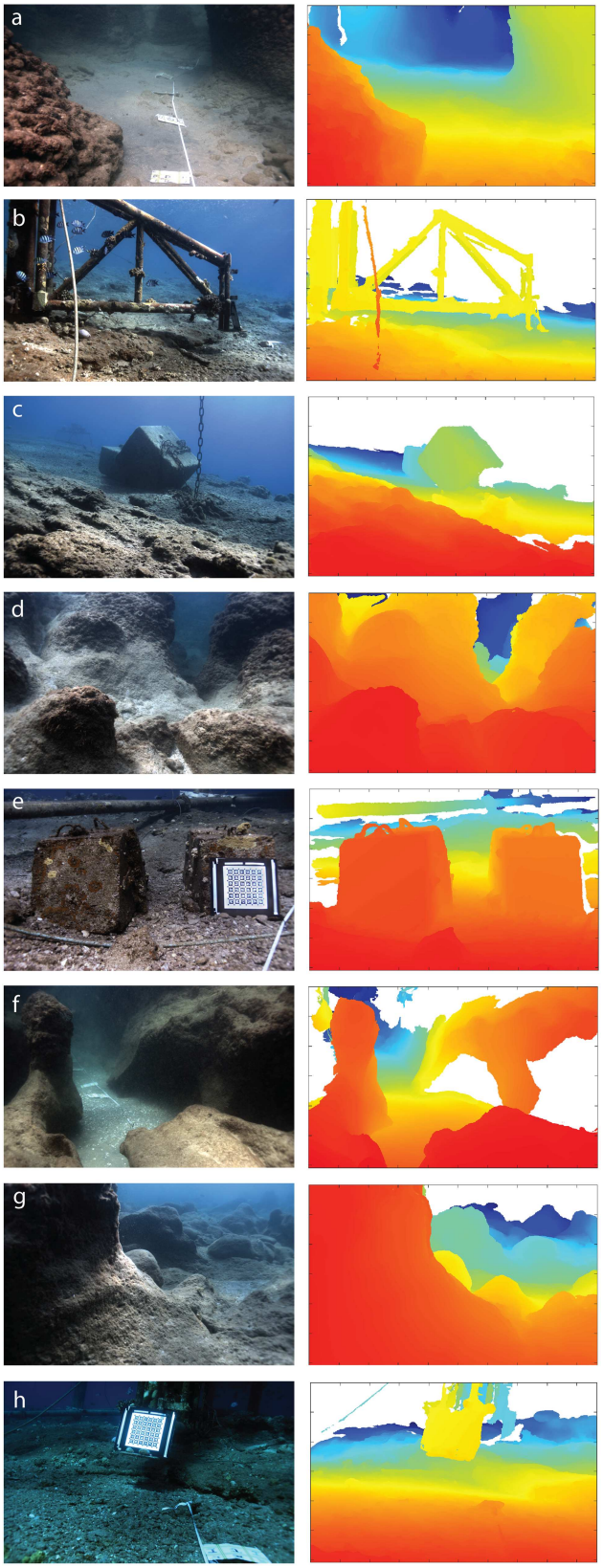

FIGURE 7 | Ground-truth examples from (a) U Canyon, (b) Northeast Path, (c) Big Dice Loop, (d) Horse Canyon, (e) Landward Path, (f) Horse Canyon, (g) Flatiron, and (h) Northeast Path. The white areas are where depth was not resolved. In this set of examples, depth ranges from 0 m at its closest to 12 m at its furthest. [Color figure can be viewed at wileyonlinelibrary.com]

available, and traditional methods for collecting 3D data, such as LiDAR, are limited underwater due to scattering and attenuation. So, we use Agisoft Metashape (Agisoft 2018) to generate per frame-scaled depth maps offline from images and scale cues placed in the scene. Metashape uses SFM to estimate the camera poses and 3D structure of the scene given a sequence of multiple images. The software works by finding feature points in each image and matching them across images to create "tie points." The output of this step is a tie point cloud and a camera pose for each image. At this stage, the scale references are identified and set, scaling the model to the actual scene scale. Sparse depth maps are also generated at this time. The next step is to generate a dense point cloud from the camera poses and input images. The final step is to create the mesh model from the dense point cloud as seen in Figure 6, from which we extract dense depth maps for each image frame as seen in Figure 7. The depth maps are then validated using a few methods. One method is to check known measurements of known distances in the scene. Another method we used for validation was by visual inspection. In other words, the depth maps are inspected by overlaying them on the input images, to check that objects represented in the depth maps align with the actual objects in the scene. The final validation is to compute the absolute error for objects of known size with AprilTags of known dimensions on them in the depth maps. The error found in this analysis was consistently less than 0.5 cm in the x- and y-directions. It must be noted here that unfortunately these objects not present in all images, therefore we could only quantitatively check a small subset of the ground truth. The ground-truth portion of the data sets includes dense depth maps, camera poses, and measurements of some objects in the scene (Figure S10a–f).

5 | Experiments

To show applicability of our data set we performed an experiment with GLPDepth (Kim et al. 2022), a state-of-the-art monocular depth estimation CNN. GLPDepth proposes a novel global–local path architecture for monocular depth estimation as well as an improved depth-specific data augmentation method for training. GLPDepth performs very well on the benchmark NYU Depth Dataset V2 (Silberman et al. 2012). We performed two experiments. The first was simply to use the model trained by GLPDepth on the NYU V2 data set for inference on FLSea underwater data. The second experiment was to train the model with our data and compare the results. We evaluated the results using a scale-invariant metric proposed by Eigen et al. (2014), which we will be referring to as scale-invariant log loss or SI-log. The equation for SI-log is as follows.

$$L(\hat{d}, d) = \frac{1}{n}\sum_{p}(\ln \hat{d}_p - \ln d_p)^2 - \frac{\lambda}{n^2}\left(\sum_{p}(\ln \hat{d}_p - \ln d_p)\right)^2, \tag{4}$$

where d_p is the ground-truth depth per pixel, $\hat{d}_p$ is estimated depth per pixel, n is the number of pixels in the image, and λ is

the level of scale-invariance, where $\lambda = 0$ means the equation is not scale-invariant at all and $\lambda = 1$ means the equation is fully scale-invariant. A lower SI-log is better.

The train/test split was 80/20, and curated such that there was a fair distribution of depth, underwater structure and visual artifacts between the sets. The results can be seen in Table 5, and Figures 8 and 9 where training with our data set improves results.

The results indicate that training with FLSea VI improves GLPDepth's monocular depth estimation on underwater images compared with the model trained on NYU V2, an indoor data set. We can also see qualitatively that the model performs better on objects at greater distance in the background of the image in comparison to the NYU V2 model.

TABLE 5 | GLPDepth results.

Training	SI-log
NYU V2	0.2276
FLSea VI	0.1012

6 | Known Issues

There are a few known issues in the data sets that we would like to draw attention to for data users. Firstly, in the visual–inertial sets, a few of the images are over exposed. The overexposure is present when there were drastic changes in light, for example, between inside a canyon and out (U Canyon). This was resolved in later data sets. Another issue in the canyon data sets is that the IMU was set to collect inertial measurements at 20 Hz (U Canyon, Flatiron, Horse Canyon, and Tiny Canyon). Because there is no hardware synchronization between the IMU and the camera, and the camera collects images at 10 Hz, this is not ideal for visual–inertial applications. To solve this, the inertial measurements were linearly interpolated to 100 Hz and in the red sea data sets the IMU was set to collect measurements at 100 Hz.

Another known issue is in the ground-truth section of the data set. Being that we create the ground truth with SFM, a photogrammetry method, it is still prone to the negative effects of imaging underwater. This causes slight imperfections in the depth maps especially with objects in the far background. This is best visualized in the <dataset_name>_overlay.avi video included in the data sets.

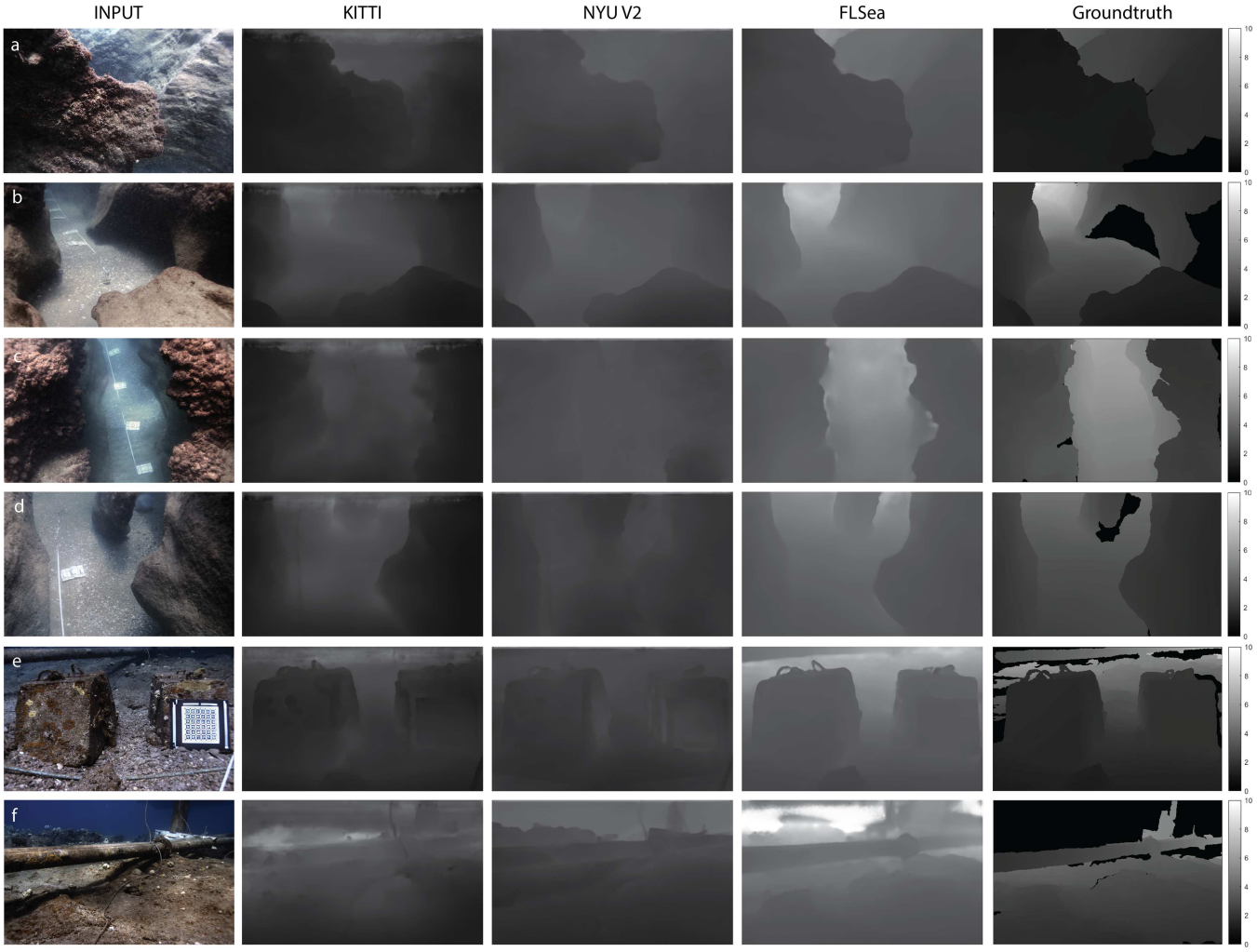

FIGURE 8 | GLPDepth inference on FLSea test images trained on KITTI, NYU V2, and FLSea, plus the FLSea ground truth. [Color figure can be viewed at wileyonlinelibrary.com]

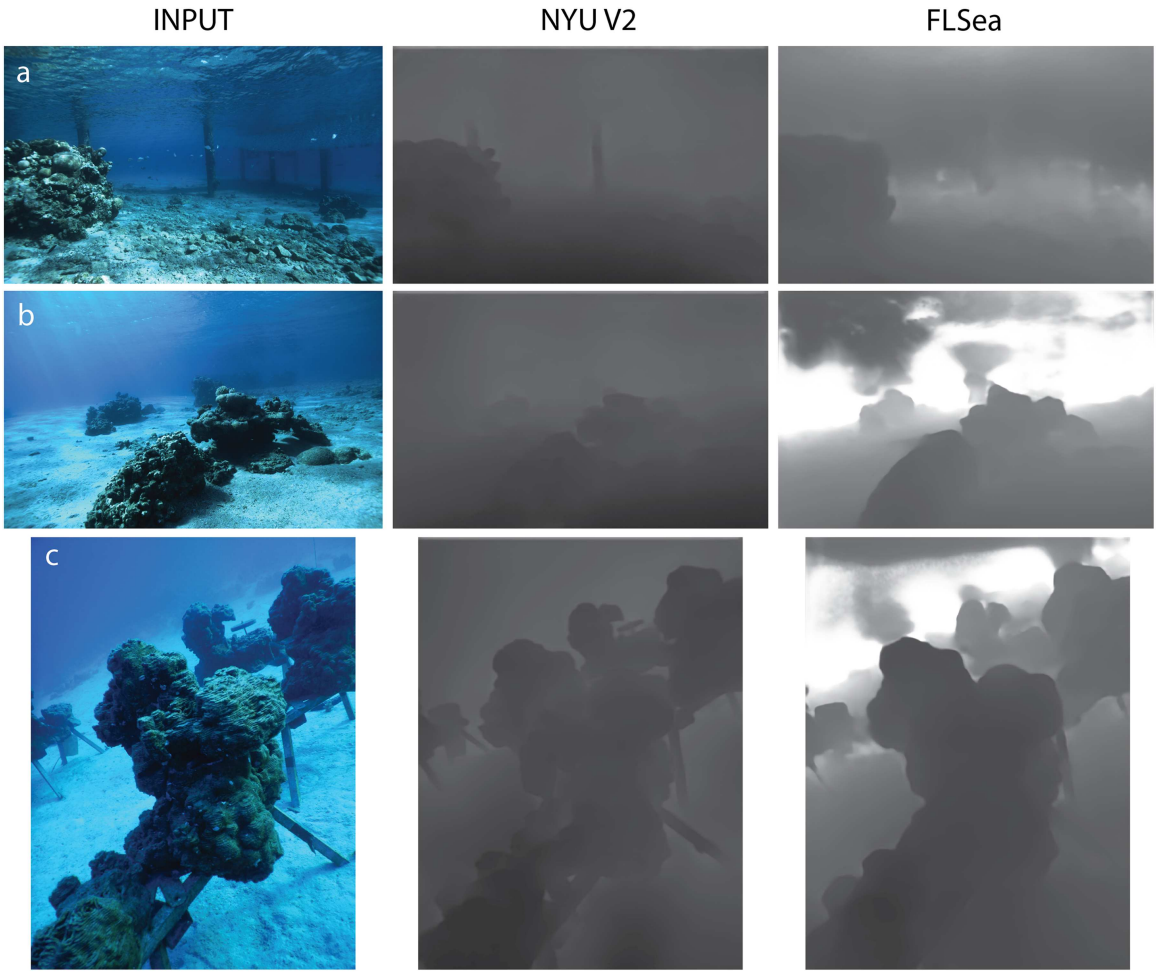

FIGURE 9 | GLPDepth inference on underwater images trained on NYU V2 and FLSea. [Color figure can be viewed at wileyonlinelibrary.com]

7 | Summary

In summary, we present the public with forward-looking stereo and visual–inertial underwater data sets. They are complete with either stereo image pairs or monocular images and inertial measurements, enhanced images, intrinsic and extrinsic calibrations, ground truth–scaled depth maps and camera poses and measurements of known objects in the scene. We give our thanks to those who have already used our data. This includes experiments in photorealistic 3D mesh mapping for AUVs (Lee and Cho 2024), underwater image dehazing (Huang et al. 2023), lightweight underwater depth estimation using deep learning (K. Zhou et al. 2024), for sonar data simulation in a forward-looking sonar denoising model (Yang et al. 2024), as a benchmark for testing a large underwater synthetic data set (Lv et al. 2024), for training and benchmarking models for underwater depth estimation (Liu et al. 2024; Ding et al. 2024), for testing monocular visual SLAM (X. Wang et al. 2025), in monocular depth estimation (C. Wang et al. 2024; Ebner et al. 2024), and for testing underwater odometry and VIO (Singh et al. 2024; Singh and Alexis 2024). We hope these data sets will continue to be used to further the development of underwater autonomous systems, with the goal of working in complex environments close to obstacles, where a forward-looking view is necessary. These data sets can be useful for testing SLAM and VIO algorithms and deep learning networks for depth estimation, to name a few. With the addition of depth maps in our data sets, these methods can be evaluated and compared, which is useful for the underwater robotics community.

Acknowledgments

A big thank you goes to all members of the VISEAON Marine Imaging Lab. Thank you to Ohad Inbar for assistance with the BlueROV2, Deborah Levy for her help with image enhancement, ground-truth generation, and data collection, and to Naama Pearl for her help with experiments with monocular depth estimation neural networks, to Judith Fischer, Aviad Avni, and Matan Yuval for their help in data collection, and to Lucyana Randall for her editorial insight. The research was funded by Israel Science Foundation grant #680/18, the Israeli Ministry of Science and Technology grant #3-15621, the Israel Data Science Initiative (IDSI) of the Council for Higher Education in Israel, the Data Science Research Center at the University of Haifa, and the European Union's Horizon 2020 research and innovation program under grant agreement no. GA 101016958.

Data Availability Statement

The data that support the findings of this study are openly available in FLSea VI at https://www.kaggle.com/datasets/viseaonlab/flsea-vi, reference number 10450291 and FLSea Stereo at https://www.kaggle.com/datasets/viseaonlab/flsea-stereo.

References

Agarwal, A., and C. Arora. 2022. *Attention Attention Everywhere: Monocular Depth Prediction With Skip Attention*. https://doi.org/10.48550/ARXIV.2210.09071.

Agarwal, S., Y. Furukawa, N. Snavely, et al. 2011. "Building Rome in a Day." *Communications of the ACM* 54, no. 10: 105–112. https://doi.org/10.1145/2001269.2001293.

Agisoft. 2018. *Metashape PhotoScan Professional (Software). Version 1.4.5*. http://www.agisoft.com/downloads/installer/.

Agrafiotis, P., D. Skarlatos, T. Forbes, C. Poullis, M. Skamantzari, and A. Georgopoulos. 2018. "Underwater Photogrammetry in Very Shallow Waters: Main Challenges and Caustics Effect Removal." *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2: 15–22.

Akkaynak, D., and T. Treibitz. 2018. "A Revised Underwater Image Formation Model." *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*: 6723–6732.

Barrett, N., L. Meyer, N. Hill, and P. Walsh. 2011. *Methods for the Processing and Scoring of AUV Digital Imagery From South Eastern Tasmania*. University of Tasmania. https://figshare.utas.edu.au/articles/report/Methods_for_the_processing_and_scoring_of_AUV_digital_imagery_from_South_Eastern_Tasmania/23996787.

Bloesch, M., M. Burri, S. Omari, M. Hutter, and R. Siegwart. 2017. "Iterated Extended Kalman Filter Based Visual–Inertial Odometry Using Direct Photometric Feedback." *International Journal of Robotics Research* 36, no. 10: 1053–1072. https://doi.org/10.1177/0278364917728574.

Blue Robotics. 2022. *BlueROV2—Affordable and Capable Underwater ROV*. https://bluerobotics.com/store/rov/bluerov2/.

Burri, M., J. Nikolic, P. Gohl, et al. 2016. "The EuRoC Micro Aerial Vehicle Datasets." *International Journal of Robotics Research* 35, no. 10: 1157–1163. https://doi.org/10.1177/0278364915620033.

Campos, C., R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós. 2021. "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM." *IEEE Transactions on Robotics* 37, no. 6: 1874–1890.

Carreras, M., J. D. Hernández, E. Vidal, N. Palomeras, D. Ribas, and P. Ridao. 2018. "Sparus II AUV—A Hovering Vehicle for Seabed Inspection." *IEEE Journal of Oceanic Engineering* 43, no. 2: 344–355. https://doi.org/10.1109/JOE.2018.2792278.

Ding, Y., K. Li, H. Mei, S. Liu, and G. Hou. 2024. *WaterMono: Teacher-Guided Anomaly Masking and Enhancement Boosting for Robust Underwater Self-Supervised Monocular Depth Estimation*. https://arxiv.org/abs/2406.13344.

Dionísio, J. M. M., P. N. A. A. S. Pereira, P. N. Leite, F. S. Neves, J. M. R. S. Tavares, and A. M. Pinto. 2023. "NEREON—An Underwater Dataset for Monocular Depth Estimation." In *OCEANS 2023—Limerick*, 1–7. https://doi.org/10.1109/OCEANSLimerick52467.2023.10244675.

Ebner, L., G. Billings, and S. Williams. 2024. "Metrically Scaled Monocular Depth Estimation Through Sparse Priors for Underwater Robots." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 3751–3757. https://doi.org/10.1109/ICRA57147.2024.10611007.

Eigen, D., C. Puhrsch, and R. Fergus. 2014. "Depth Map Prediction From a Single Image Using a Multi-Scale Deep Network." *CoRR, abs/1406.2283*. http://arxiv.org/abs/1406.2283.

Engel, J., V. Koltun, and D. Cremers. 2016. "Direct Sparse Odometry." *CoRR, abs/1607.02565*. http://arxiv.org/abs/1607.02565.

Engel, J., T. Schöps, and D. Cremers. 2014. "LSD-SLAM: Large-Scale Direct Monocular SLAM." In *Computer Vision—ECCV 2014*, edited by D. Fleet, T. Pajdla, B. Schiele and T. Tuytelaars, 834–849. Springer International Publishing.

Ferrera, M., V. Creuze, J. Moras, and P. Trouvé-Peloux. 2019. "AQUALOC: An Underwater Dataset for Visual–Inertial–Pressure Localization." *International Journal of Robotics Research* 38, no. 14: 1549–1559.

Forster, C., M. Pizzoli, and D. Scaramuzza. 2014. "SVO: Fast Semi-Direct Monocular Visual Odometry." In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 15–22. https://doi.org/10.1109/ICRA.2014.6906584.

Geiger, A., P. Lenz, C. Stiller, and R. Urtasun. 2013. "Vision Meets Robotics: The KITTI Dataset." *International Journal of Robotics Research (IJRR)* 32, no. 11: 1231–1237.

Hartley, R. I., and A. Zisserman. 2004. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd ed. ISBN: 0521540518.

Huai, Z., and G. Huang. 2022. "Robocentric Visual–Inertial Odometry." *International Journal of Robotics Research* 41, no. 7: 667–689.

Huang, H., Y. Jin, S. Sun, and M. Zhang. 2023. "A Novel Multiscale Fusion Algorithm for Underwater Image Dehazing and Edge Preserving." In *2023 IEEE International Conference on Unmanned Systems (ICUS)*, 329–334. https://doi.org/10.1109/ICUS58632.2023.10318329.

Hugyfot. 2022. *Hfn-d810: Nikon d810 Underwater Housing*. https://www.hugyfot.com/l.

iDS. 2022. *Ui-3260cp Rev. 2*. https://www.ids-imaging.us/.

Jaffe, J. 1990. "Computer Modeling and the Design of Optimal Underwater Imaging Systems." *IEEE Journal of Oceanic Engineering* 15, no. 2: 101–111. https://doi.org/10.1109/48.50695.

Joshi, B., N. Vitzilaios, I. Rekleitis, et al. 2019. "Experimental Comparison of Open Source Visual–Inertial-Based State Estimation Algorithms in the Underwater Domain." *IROS*: 7227–7233. https://doi.org/10.1109/IROS40897.2019.8968049.

Joshi, B., M. Xanthidis, S. Rahman, and I. Rekleitis. 2022. "High Definition, Inexpensive, Underwater Mapping." *IEEE International Conference on Robotics and Automation (ICRA)*: 1113–1121. https://doi.org/10.1109/ICRA46639.2022.9811695.

Kim, D., W. Ga, P. Ahn, D. Joo, S. Chun, and J. Kim. 2022. "Global–Local Path Networks for Monocular Depth Estimation With Vertical Cutdepth." *CoRR, abs/2201.07436*. https://arxiv.org/abs/2201.07436.

Lee, J., and Y. Cho. 2024. *Mesh-Based Photorealistic and Real-Time 3D Mapping for Robust Visual Perception of Autonomous Underwater Vehicle*. https://arxiv.org/abs/2404.18395.

Leutenegger, S., P. Furgale, V. Rabaud, M. Chli, K. Konolige, and R. Siegwart. 2014. "Keyframe-Based Visual–Inertial SLAM Using Nonlinear Optimization." *Proceedings of the Robotics*.

Li, Z., and N. Snavely. 2018. "MegaDepth: Learning Single-View Depth Prediction From Internet Photos." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, Z., X. Wang, X. Liu, and J. Jiang. 2022. *BinsFormer: Revisiting Adaptive Bins for Monocular Depth Estimation*. https://doi.org/10.48550/ARXIV.2204.00987.

Lindenberger, P., P. Sarlin, V. Larsson, and M. Pollefeys. 2021. "Pixel-Perfect Structure-From-Motion With Featuremetric Refinement." *CoRR, abs/2108.08291*. https://arxiv.org/abs/2108.08291.

Liu, T., S. Zhang, and Z. Yu. 2024. "Redefining Accuracy: Underwater Depth Estimation for Irregular Illumination Scenes." *Sensors* 24, no. 13: 4353. https://doi.org/10.3390/s24134353.

Lodi Rizzini, D., F. Kallasi, F. Oleari, and S. Caselli. 2015. "Investigation of Vision-Based Underwater Object Detection With Multiple Datasets." *International Journal of Advanced Robotic Systems (IJARS)* 12, no. 77: 1–13. https://doi.org/10.5772/60526.

Longuet-Higgins, H. 1987. "A Computer Algorithm for Reconstructing a Scene From Two Projections." In *Readings in Computer Vision*, edited

by M. A. Fischler and O. Firschein, 61–62. Morgan Kaufmann. https://www.sciencedirect.com/science/article/pii/B978008051581650012X.

Luczynski, T., J. S. Willners, E. Vargas, et al. 2021. "Underwater Inspection and Intervention Dataset." *CoRR, abs/2107.13628*.

Lv, Q., J. Dong, Y. Li, et al. 2024. *UWStereo: A Large Synthetic Dataset for Underwater Stereo Matching*. https://arxiv.org/abs/2409.01782.

Mallios, A., P. Ridao, D. Ribas, M. Carreras, and R. Camilli. 2016. "Toward Autonomous Exploration in Confined Underwater Environments." *Journal of Field Robotics* 33, no. 7: 994–1012. https://doi.org/10.1002/rob.21640.

Mapillary. 2021. *OpenSFM, Version 0.5.1*. https://opensfm.org/.

MATLAB. 2022. *R2022a*. MathWorks Inc.

Mohr, R., L. Quan, and F. Veillon. 1995. "Relative 3D Reconstruction Using Multiple Uncalibrated Images." *International Journal of Robotics Research* 14, no. 6: 619–632. https://doi.org/10.1177/027836499501400607.

Moulon, P., P. Monasse, R. Perrot, and R. Marlet. 2016. "OpenMVG: Open Multiple View Geometry." In *International Workshop on Reproducible Research in Pattern Recognition*, 60–74. Springer.

Mur-Artal, R., J. M. M. Montiel, and J. D. Tardós. 2015. "ORB-SLAM: A Versatile and Accurate Monocular SLAM System." *IEEE Transactions on Robotics* 31, no. 5: 1147–1163. https://doi.org/10.1109/TRO.2015.2463671.

Mur-Artal, R., and J. D. Tardós. 2017. "Visual–Inertial Monocular SLAM With Map Reuse." *IEEE Robotics and Automation Letters* 2, no. 2: 796–803. https://doi.org/10.1109/LRA.2017.2653359.

Nikon. 2022. *Nikon d810: Full-Frame DSLR*. https://www.nikonusa.com/l.

Ou, Y., J. Fan, C. Zhou, L. Cheng, and M. Tan. 2024. "Water-MBSL: Underwater Movable Binocular Structured Light-Based High-Precision Dense Reconstruction Framework." *IEEE Transactions on Industrial Informatics* 20, no. 4: 6142–6154. https://doi.org/10.1109/TII.2023.3342899.

Ou, Y., J. Fan, C. Zhou, et al. 2025a. "Structured Light-Based Underwater Collision-Free Navigation and Dense Mapping System for Refined Exploration in Unknown Dark Environments." *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 55, no. 1: 110–123. https://doi.org/10.1109/TSMC.2024.3370917.

Ou, Y., J. Fan, C. Zhou, P. Zhang, and Zeng-Guang. 2025b. "Hybrid-VINS: Underwater Tightly Coupled Hybrid Visual Inertial Dense SLAM for AUV." *IEEE Transactions on Industrial Electronics* 72, no. 3: 2821–2831. https://doi.org/10.1109/TIE.2024.3440515.

Qin, T., J. Pan, S. Cao, and S. Shen. 2019. *A General Optimization-Based Framework for Local Odometry Estimation With Multiple Sensors*.

Rahman, S., A. Quattrini Li, and I. Rekleitis. 2018. "Sonar Visual Inertial SLAM of Underwater Structures." In *2018 IEEE International Conference on Robotics and Automation (ICRA)*.

Rehder, J., J. Nikolic, T. Schneider, T. Hinzmann, and R. Siegwart. 2016. "Extending Kalibr: Calibrating the Extrinsics of Multiple IMUs and of Individual Axes." In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 4304–4311. https://doi.org/10.1109/ICRA.2016.7487628.

Rosinol, A., M. Abate, Y. Chang, and L. Carlone. 2020. "Kimera: An Open-Source Library for Real-Time Metric-Semantic Localization and Mapping." In *IEEE International Conference on Robotics and Automation (ICRA)*. https://github.com/MIT-SPARK/Kimera.

Scaramuzza, D., and Z. Zhang. 2019. "Visual–Inertial Odometry of Aerial Robots." *CoRR, abs/1906.03289*. http://arxiv.org/abs/1906.03289.

Schaffalitzky, F., and A. Zisserman. 2002. "Multi-View Matching for Unordered Image Sets, or "How DOI Organize My Holiday Snaps?" In *Computer Vision—ECCV 2002*, edited by A. Heyden, G. Sparr, M. Nielsen and P. Johansen, 414–431. Springer.

Schneider, T., M. Dymczyk, M. Fehr, et al. 2017. "maplab: An Open Framework for Research in Visual–Inertial Mapping and Localization." *CoRR, abs/1711.10250*. http://arxiv.org/abs/1711.10250.

Schönberger, J. L., and J.-M. Frahm. 2016. "Structure-From-Motion Revisited." In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Silberman, N., D. Hoiem, P. Kohli, and R. Fergus. 2012. "Indoor Segmentation and Support Inference From RGBD Images." In *ECCV*.

Singh, M., and K. Alexis. 2024. "Online Refractive Camera Model Calibration in Visual Inertial Odometry." In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 12609–12616. https://doi.org/10.1109/IROS58592.2024.10802302.

Singh, M., M. Dharmadhikari, and K. Alexis. 2023. "An Online Self-Calibrating Refractive Camera Model With Application to Underwater Odometry." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*.

Singh, M., M. Dharmadhikari, and K. Alexis. 2024. "An Online Self-Calibrating Refractive Camera Model With Application to Underwater Odometry." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 10005–10011. https://doi.org/10.1109/ICRA57147.2024.10610643.

Sturm, J., N. Engelhard, F. Endres, W. Burgard, and D. Cremers. 2012. "A Benchmark for the Evaluation of RGB-D SLAM Systems." In *Proceedings of the International Conference on Intelligent Robot Systems (IROS)*.

Teixeira, B., H. Silva, A. Matos, and E. Silva. 2020. "Deep Learning for Underwater Visual Odometry Estimation." *IEEE Access* 8: 44687–44701. https://doi.org/10.1109/ACCESS.2020.2978406.

Treibitz, T., D. Levy, Y. Goldfracht, D. Akkaynak, and Y. Bekerman. 2022. *Seaerra*. SeaErra.

VectorNav. 2022. *Vn-100 IMU*. https://www.vectornav.com/.

Wang, C., H. Xu, G. Jiang, M. Yu, T. Luo, and Y. Chen. 2024. "Underwater Monocular Depth Estimation Based on Physical-Guided Transformer." *IEEE Transactions on Geoscience and Remote Sensing* 62: 1–16. https://doi.org/10.1109/TGRS.2024.3373904.

Wang, X., X. Fan, Y. Liu, Y. Xin, and P. Shi. 2025. "EUM-SLAM: An Enhancing Underwater Monocular Visual SLAM With Deep Learning-Based Optical Flow Estimation." *IEEE Transactions on Instrumentation and Measurement*: 1. https://doi.org/10.1109/TIM.2025.3541785.

Wofk, D., F. Ma, T.-J. Yang, S. Karaman, and V. Sze. 2019. "FastDepth: Fast Monocular Depth Estimation on Embedded Systems." In *2019 International Conference on Robotics and Automation (ICRA)*, 6101–6108. https://doi.org/10.1109/ICRA.2019.8794182.

Yang, T., T. Zhang, and Y. Yao. 2024. "Simnfnd: A Forward-Looking Sonar Denoising Model Trained on Simulated Noise-Free and Noisy Data." *Remote Sensing* 16, no. 15: 2815. https://doi.org/10.3390/rs16152815.

Yin, W., J. Zhang, O. Wang, et al. 2020. "Learning to Recover 3D Scene Shape From a Single Image." *CoRR, abs/2012.09365*. https://arxiv.org/abs/2012.09365.

Yin, Z., and J. Shi. 2018. "GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose." *CoRR, abs/1803.02276*. http://arxiv.org/abs/1803.02276.

Yuan, W., X. Gu, Z. Dai, S. Zhu, and P. Tan. 2022. *NeW CRFs: Neural Window Fully-Connected CRFs for Monocular Depth Estimation*. https://arxiv.org/abs/2203.01502.

Zhao, W., T. He, A. Y. M. Sani, and T. Yao. 2019. "Review of SLAM Techniques for Autonomous Underwater Vehicles." In *Proceedings of the 2019 International Conference on Robotics, Intelligent Control and*

Artificial Intelligence, RICAI'19, 384–389. Association for Computing Machinery. https://doi.org/10.1145/3366194.3366262.

Zhou, K., J. Chen, S. Gui, and Z. Wang. 2024. "Towards Lightweight Underwater Depth Estimation." In *2024 IEEE Conference on Artificial Intelligence (CAI)*, 1442–1445. https://doi.org/10.1109/CAI59869.2024.00258.

Zhou, T., M. Brown, N. Snavely, and D. G. Lowe. 2017. "Unsupervised Learning of Depth and Ego-Motion From Video." *CoRR, abs/1704.07813*. http://arxiv.org/abs/1704.07813.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.
Figure S1: Known Measurements.