IEEE Robotics and Automation Letters (RAL) paper presented at the
2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

Through-water Stereo SLAM with Refraction Correction for AUV Localization

Sudharshan Suresh, Eric Westman, and Michael Kaess

***Abstract*— In this work, we propose a novel method for underwater localization using natural visual landmarks above the water surface. High-accuracy, drift-free pose estimates are necessary for inspection tasks in underwater indoor environments, such as nuclear spent pools. Inaccuracies in robot localization degrade the quality of its obtained map. Our framework uses sparse features obtained via an onboard upward-facing stereo camera to build a global ceiling feature map. However, adopting the pinhole camera model without explicitly modeling light refraction at the water-air interface contributes to a systematic error in observations. Therefore, we use refraction-corrected projection and triangulation functions to obtain true landmark estimates. The SLAM framework jointly optimizes vehicle odometry and point landmarks in a global factor graph using an incremental smoothing and mapping backend. To the best of our knowledge, this is the first method that observes in-air landmarks through water for underwater localization. We evaluate our method via both simulation and real-world experiments in a test-tank environment. The results show accurate localization across various challenging scenarios.**

I. Introduction and Background

Autonomous underwater vehicles (AUVs) can conduct inspection tasks in complex environments inaccessible to humans. They have the potential to create high-fidelity maps of such areas with minimal manual intervention. One such task is imaging and mapping of nuclear waste storage pools, a task critical to the safe operation of the infrastructure.

However, long-term operation of the robot causes drift in its pose estimate if it is solely reliant on dead reckoning. This has a detrimental effect on the resulting map it generates. Thus, there is a need for an accurate robot state estimate.

The problem of underwater localization has received considerable attention over the years. Numerous sensing modalities and algorithms have been explored, as documented by Paull et al. [26]. In this work, we focus on the specific task of AUV localization in nuclear spent pools. These concrete pools are indoor underwater environments with significant clutter in the form of stored nuclear waste.

Recently, [25] proposed an acoustic sensor network to localize a robot swarm in a nuclear storage pond, while [28] used visible light to localize an ROV in a nuclear reactor. Both methods suffer from attenuation in cluttered environments. Vision-based methods are preferable due to excellent visibility and absence of open-sea error sources such as surface disturbances and turbid waters.

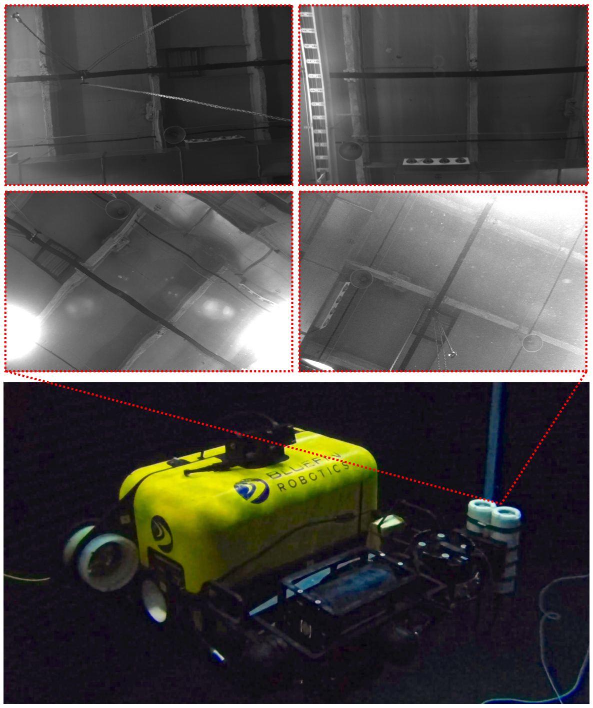


Fig. 1: **(top)** A sampling of ceiling frames taken from the stereo pairs. These highlight challenges for the frontend, including motion blur, light scattering and particulates. **(bottom)** An underwater still of our AUV executing a trajectory in the test environment at a depth of 1m. The upward-facing stereo camera views the ceiling through the water interface.

Jung et al. [14, 15] developed visual localization methods for AUVs by installing fiducials underwater. However, spent nuclear pools cannot be modified due to radioactivity and often have nuclear waste and thick sludge deposition at the bottom. Cho et al. [4] performed 3-DOF state estimation of a robot in a reactor vessel through an external camera, by viewing LEDs on the vehicle frame. Later, Lee et al. [18] used a submerged camera and prior map to obtain a full 6-DOF state estimate through fiducial tracking. Both methods, however, are affected by clutter in the line-of-sight between the camera and robot. Consequently, they do not scale to larger environments.

The field-of-view of an upward-facing camera on an AUV is not obstructed while navigating these cluttered environments (Fig. 1). Ceilings in most such environments have robust structural cues for localization, with several notable examples for ground vehicles [7, 13]. In this problem formulation, feature points triangulate to landmarks in air, viewed through the water-air interface. Refraction causes light to

The authors are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. {sudhars1, ewestman, kaess}@andrew.cmu.edu

This work was partially supported by the Department of Energy under contract DE-EM0004383 and by the Office of Naval Research under award N00014-16-1-2103. We thank Bing-Jui Ho for his help in data collection and Chuck Whittaker for his support with survey equipment.

bend at the interface, and generates a systematic error in heights calculated from stereo correspondences. This creates large geometric errors in the global map, and negatively affects the optimized trajectory estimate. To achieve the true SLAM solution, we must explicitly model for the refraction.

Refraction correction was first explored in aerial photogrammetry for shallow water [22, 31]. These works obtain actual water depth from an analytic plotter by applying a correction factor. Fryer et al. later demonstrated that two-camera photogrammetry of submerged objects can only be approximated, as rays from an object have different incident angles with the cameras [8]. This procedure was used to create underwater topographical maps of river beds and reefs [24, 35]. All prior through-water methods (i) are not in the context of SLAM and (ii) consider aerial cameras observing underwater objects.

In this work, we propose a SLAM formulation for AUVs using an onboard upward-facing stereo camera for accurate underwater pose estimation. The method is intended for indoor underwater environments with adequate visibility and lighting. To the best of our knowledge, this is the first through-water visual SLAM technique for underwater vehicles. Concisely, our main contributions are:

(1) An upward-facing stereo SLAM framework for drift-free AUV localization using a ceiling feature map.
(2) A refraction correction module for through-water vision, modeled after prior work in multimedia photogrammetry.
(3) Evaluation in both simulation and real-world settings.

While this work targets nuclear pools, it generalizes to analogous applications such as robotic swimming pool cleaners. This method can also be extended to dense stereo for mapping partially-submerged caves with AUVs [34].

II. Refraction Observation Model

A routine operation in any visual SLAM framework is the projection of 3-D points to image pixel coordinates and the corresponding backprojection of 2-D image points to 3-D points. When operating in a single medium, this is a trivial operation given camera intrinsics and extrinsics. We require compensation factors that enable these operations in a dual-media setting. Thus, we introduce methods for refraction-corrected stereo triangulation (Section II-A) and refraction-corrected projection (Section II-B), both based on previous work. However, the prior algorithms were for aerial photogrammetry through shallow water. We adapt them to the inverse problem of an underwater camera observing points in air. Our SLAM framework (Section III) uses this module at every stage for *true* landmark positions.

A fixed-baseline stereo camera is calibrated underwater and has known, constant camera-robot extrinsics. The single viewpoint pinhole camera model is found to be theoretically inaccurate due to refraction at the camera housing [1, 32]. However, it is a valid approximation if the center of projection and flat port and very close to each other [20], which is the case for underwater housings. Thus, we adopt the pinhole camera model—refraction at the camera's housing is accounted for in the lens distortion parameters.

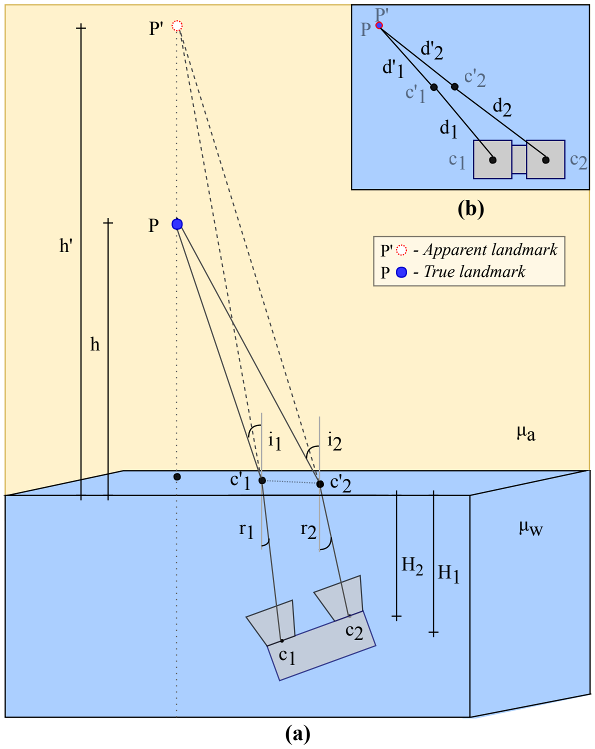


Fig. 2: **(a)** Geometry of refraction-corrected stereo triangulation for a single landmark in air. While the stereo pair incorrectly triangulates a measurement to *apparent* position P', we perform correction to obtain the *true* position P. **(b) (inset)** top view of the geometry, showing directly observable quantities in the XY plane.

We mathematically model the water surface as a plane, an approximation that is commonly made in related works [21, 31]. The camera viewing direction is not required to be perpendicular to the water surface. Lacking this assumption, we cannot model refraction at the water interface as a radial distortion [29]. We establish a sign-convention for the Z direction: the water surface is the XY plane, points in air are negative and points underwater are positive. The *apparent* landmark is that triangulated without considering refraction at the interface. The *true* landmark is that obtained from explicitly modeling this refraction.

A. Refraction-Corrected Stereo Triangulation

Given pixel correspondences in an image pair, we wish to calculate the *true* position of a landmark. Fig. 2 (a) illustrates the geometry for a single point landmark observed by a stereo pair. We assume known (i) pose of the cameras and (ii) refractive indices of the media (we consider $\mu_w = 1.33$ for water and $\mu_a = 1$ for air, but we can modify the indices to represent any general pair of media). The cameras are at positions c_1 and c_2, having depths H_1 and H_2 below the water surface respectively. The *apparent* landmark P' has a height h', while the *true* landmark P has a depressed height h. c'_1 and c'_2 are the interface intercept points obtained by tracing the rays from P' to c_1 and c_2 respectively. For rays from P to c_1 and c_2, the incidence angles with the interface are i_1, i_2 and refracted angles are r_1, r_2.

Fig. 2 (b) is the top view of Fig. 2 (a) showing the distances in the XY dimensions:

$$\begin{aligned} d_1 &= \|c_{1\mathbf{xy}} - c'_{1\mathbf{xy}}\| & d_2 &= \|c_{2\mathbf{xy}} - c'_{2\mathbf{xy}}\| \\ d'_1 &= \|c'_{1\mathbf{xy}} - P'_{\mathbf{xy}}\| & d'_2 &= \|c'_{2\mathbf{xy}} - P'_{\mathbf{xy}}\| \end{aligned} \tag{1}$$

From Fig. 2, r_1 and r_2 are:

$$r_1 = \tan^{\text{-}1}\left(\frac{d_1}{H_1}\right) \qquad r_2 = \tan^{\text{-}1}\left(\frac{d_2}{H_2}\right) \tag{2}$$

Snell's law relates the refractive indices of media with the direction of light propagation. Further, i_1 and i_2 are:

$$\frac{\sin i_1}{\sin r_1} = \frac{\sin i_2}{\sin r_2} = \frac{\mu_w}{\mu_a} \tag{3}$$

Knowing the angles of incidence, we obtain the corrected height of the landmark for each camera (h_{c_1}, h_{c_2}). From Fig. 2, in a similar fashion to Equation 2, we have:

$$h_{c_1} = d'_1/\tan(i_1) \qquad h_{c_2} = d'_2/\tan(i_2) \tag{4}$$

They are found to be slightly different, as no unique solution exists when rays from the *true* 3-D point landmark have different incident angles with the cameras [8]. However, an approximate solution suffices for landmark initialization. We take the average, giving the final corrected height:

$$h = (h_{c_1} + h_{c_2})/2 \tag{5}$$

Thus, refractive triangulation gives us the *true* position of landmarks. This ensures consistent triangulation regardless of robot location and assures geometrically accurate maps.

B. Refraction-Corrected Projection

Given the *true* position of a landmark—P—we wish to project it to image coordinates. First, we calculate the *shifted* position P^* the camera views the landmark at. This is done by radially shifting it parallel to the water surface.

P^* lies on the ray joining the *apparent* landmark position P' with the camera center (Fig. 2). Due to bending of light at the interface, *true* landmark position and camera center are no longer collinear. We radially shift the landmark with respect to the camera center before projection [21].

Fig. 3 shows the problem geometry when viewed perpendicular to the direction of the ray. The *true* landmark to be imaged is P, at a height H_p. The projection center of the camera viewing the landmark is C, at depth H_c. The incident and refracted rays make angles i, r with the interface. There is no closed form solution as C' is unknown—we follow an iterative method. This process is formalized in Algorithm 1. We initialize the *shifted* radial distance R^* to the *true* radial distance R itself. Knowing its position and applying Snell's law, we can compute the angles i and r. The radial shift, ΔR, is computed as $\Delta R = R^* - R$. From Fig. 3:

$$\begin{aligned} R &= H_p \tan i + H_c \tan r \\ R^* &= (H_p + H_c) \tan r \end{aligned} \tag{6}$$

These are directly obvious from the right triangles that i and r are part of. Using Equation 6, we compute ΔR at every iteration and radially shift the point until convergence (i.e. $|\Delta R| < \epsilon$). We convert the expression to Cartesian coordinates to get the *shifted* landmark that we can project trivially, as in the single-media case. In initial tests, we get convergence to within a few *mm* from ground truth in a 100 iterations.

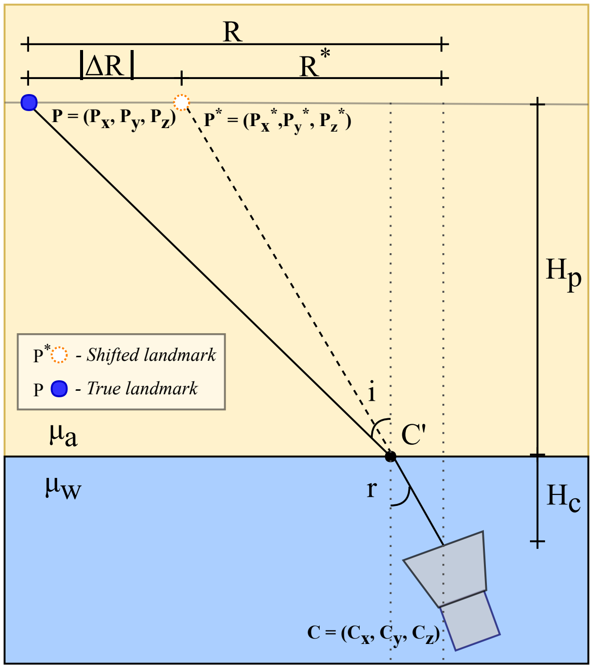


Fig. 3: Radial shift geometry for the estimated *true* position of a landmark P with respect to camera center C. An iterative procedure converges to P^*, which is the *shifted* position of the landmark. This allows us to trivially project the 3-D landmark into the camera similar to a single-medium setting.

Algorithm 1 Iterative radial-shift for refraction correction.

```
$R^* = R = \sqrt{(P_x - C_x)^2 + (P_y - C_y)^2}$
repeat
    $r = \tan^{\text{-}1} \frac{R^*}{C_z + P_z}$
    $i = \sin^{\text{-}1}\left(\frac{\mu_w}{\mu_a} \sin r\right)$
    $\Delta R = R^* - (H_p \tan i + H_c \tan r)$
    $R^* \leftarrow R^* + \Delta R$
until $(|\Delta R| < \epsilon)$
$P^*_x = C_x + \frac{R^*}{R}(P_x - C_x)$
$P^*_y = C_y + \frac{R^*}{R}(P_y - C_y)$
$P^*_z = P_z$
```

III. Proposed SLAM Formulation

A. Factor Graph Representation

We represent the problem as a factor graph optimization, as commonly done in the SLAM literature. A factor graph is a bipartite graph comprised of *variables* to be optimized and *measurements* that constrain the system. In a landmark-based SLAM formulation, all poses in the trajectory and their accompanying observed landmarks make up the *variables*. This is graphically represented in Fig. 4. Typically, underwater vehicles have a pressure sensor that directly observes depth (Z). Detecting the direction of gravity allows the inertial measurement unit (IMU) to provide absolute pitch and roll

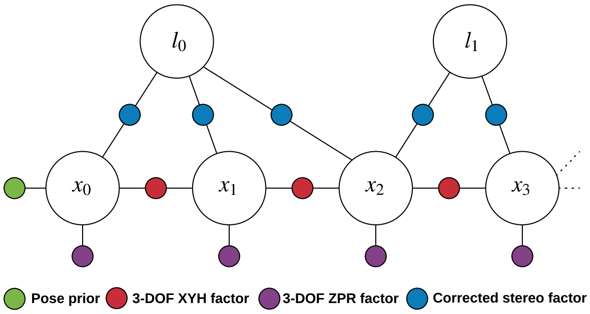


Fig. 4: Factor graph representing our SLAM formulation. Variable nodes are the large circles that represent either poses (x_i) or landmarks (l_i). Measurement factors are denoted by smaller, colored circles. As opposed to conventional landmark-based stereo SLAM, our method incorporates a refraction-corrected stereo factor between poses and landmarks.

measurements. The remaining degrees of freedom—X, Y and yaw—are obtained via dead reckoning and are subject to drift. This gives accurate measurements locally, but the pose estimate drifts over long dives. We represent the vehicle pose as:

$$x_i = [\underbrace{t_{i,x}, t_{i,y}, t_{i,z}}_{\substack{\text{translational}\\\text{components}}},\ \underbrace{\phi_i, \theta_i, \psi_i}_{\substack{\text{yaw, pitch and}\\\text{roll angles}}}]^T \tag{7}$$

Thus, we can split the vehicle odometry into two independent constraints, similar to [36]: **(i)** a 3-DOF pose-to-pose relative odometry constraint on XYH (X, Y, yaw) and **(ii)** a 3-DOF unary constraint on ZPR (Z, pitch, roll). At a given timestep i: an XYH factor u_{i-1} is added between x_{i-1} and x_i, and a ZPR factor v_i is added to vehicle pose x_i. A corrected stereo measurement factor m_k joins any observed 3-D point landmark l_j with pose x_i. This factor m_k is the corrected stereo landmark pixel observations, which is a four-vector for unrectified stereo. We attach a pose prior measurement p_0 to x_0 to bind the entire trajectory to a global coordinate frame. The state and measurement vectors are:

$$\begin{aligned} \mathcal{X} &= \{x_0, \ldots, l_0, \ldots\} \\ \mathcal{Z} &= \{p_0, u_0, \ldots, v_0, \ldots, m_0, \ldots\} \end{aligned} \tag{8}$$

We compute the *maximum a posteriori* (MAP) estimate, which predicts variable values that maximally agree with the given measurements:

$$\begin{aligned} \mathcal{X}^* &= \underset{\mathcal{X}}{\operatorname{argmax}}\ p(\mathcal{X}|\mathcal{Z}) \\ &= \underset{\mathcal{X}}{\operatorname{argmax}}\ p(\mathcal{X})p(\mathcal{Z}|\mathcal{X}) \\ &= \underset{\mathcal{X}}{\operatorname{argmax}}\ \underbrace{p(x_0)}_{\text{prior}} \prod_{i=1}^{n} \underbrace{p(u_i|x_{i-1}, x_i)}_{\text{XYH}}\ \underbrace{p(v_i|x_i)}_{\text{ZPR}} \\ &\qquad \prod_{k=1}^{m} \underbrace{p(m_k|x_i, l_j)}_{\text{corr. stereo factor}} \end{aligned} \tag{9}$$

We consider all four measurements as normally distributed random variables with covariances Σ_0, Ψ_i, Φ_i, Γ_k:

$$\begin{aligned} p(x_0) &= \mathcal{N}(p_0, \Sigma_0) \\ p(u_i|x_{i-1}, x_i) &= \mathcal{N}(\mathcal{U}(x_{i-1}, x_i), \Psi_i) \\ p(v_i|x_i) &= \mathcal{N}(\mathcal{V}(x_i), \Phi_i) \\ p(m_k|x_i, l_j) &= \mathcal{N}(\mathcal{M}(x_i, l_j), \Gamma_k) \end{aligned} \tag{10}$$

In Equation 10:

(i) p_0 represents the pose prior.
(ii) $\mathcal{U}(x_{i-1}, x_i)$ represents the relative transform between consecutive poses in $[t_{i,x}, t_{i,y}, \phi_i]$.
(iii) $\mathcal{V}(x_i)$ is the direct measurement of $[t_{i,z}, \theta_i, \psi_i]$.
(iv) $\mathcal{M}(x_{i_k}, l_{j_k})$ is the refraction-corrected stereo measurement function. It projects l_j into the stereo cameras at vehicle pose x_i while accounting for refraction. The output is a four-vector of stereo pixel measurements.

Assuming Gaussian noise models reduces the inference to a nonlinear least squares optimization [6]:

$$\begin{aligned} \mathcal{X}^* &= \underset{\mathcal{X}}{\operatorname{argmin}} - \log \Bigg(p(x_0) \prod_{i=1}^{n} p(u_i|x_{i-1}, x_i) p(v_i|x_i) \\ &\qquad \prod_{k=1}^{m} p(m_k|x_i, l_j) \Bigg) \\ &= \underset{\mathcal{X}}{\operatorname{argmin}} \|p_0 \ominus x_0\|^2_{\Sigma_0} + \sum_{k=1}^{m} \|m_k - \mathcal{M}(x_i, l_j)\|^2_{\Gamma_k} \\ &\quad + \sum_{i=1}^{n} \Big(\|u_i - \mathcal{U}(x_{i-1}, x_i)\|^2_{\Psi_i} + \|v_i - \mathcal{V}(x_i)\|^2_{\Phi_i} \Big) \end{aligned} \tag{11}$$

The 6-DOF pose prior is in the $SE(3)$ Lie group, and $\ominus$ represents the logarithm map of the relative transformation between the elements [2]. The notation of the form $\|w\|^2{}_{\Lambda} = w^T \Lambda^{-1} w$ is the Mahalanobis distance of w.

We use incremental methods to obtain optimized vehicle pose and landmark estimates at every timestep [16, 17]. Instead of re-calculating the entire system each time, it updates the previous matrix factorization with the new measurements. The sparse nature of the system (i.e. pose-landmark connectivity) assures computational efficiency.

B. *Feature Extraction*

Our technique uses sparse stereo feature points. Existing benchmarks for feature detectors underwater focus on repeatability in turbid environments [10], which is not required in our clear conditions. Our preliminary investigation demonstrated no discernible upside to using other feature detectors such as SIFT, SURF, or MSER. Moreover, we value the efficiency of ORB features for near real-time implementation of stereo visual SLAM [27]. We detect a large number of ORB feature points and prune them through adaptive non-maximal suppression [3], selectively choosing keypoints based on corner strength and spatial location. This prevents clustering, degeneracy, and speeds up computation. We establish matches between the stereo pairs based on the Hamming distance between their binary descriptors. To remove ambiguous matches, we perform the distance-ratio test [19] and further select the inliers of a RANSAC homography computation. Fig. 5 shows feature matches between a stereo pair from our real-world dataset.

C. *Data Association*

Wrong correspondences affect the accuracy of the state estimate and landmark map. Thus, we need a reliable data association framework. Two operations—map update and landmark initialization—are explained below:

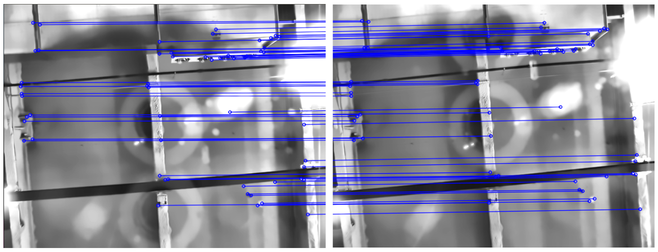


Fig. 5: Feature matching between a stereo pair of images from our real-world dataset (Section IV-D). Adaptive non-maximal suppression prevents clustering of feature points and gives good spatial distribution. In these frames, the vehicle is just below the water surface and the reflection of the stereo pair at the water interface is faintly visible.

1) ***Map update****:* The estimated positions of landmarks in the optimization are first corrected to their *apparent* positions for the current camera poses (Section II-B). They are then projected into the cameras of the stereo pair. A landmark is temporally matched with a stereo keypoint if its corresponding projection lies within an empirical gating threshold g_t in both cameras ($g_t = 5$ pixels). In the case of multiple matches, the closest projected landmark is considered.

2) ***Landmark initialization****:* If a stereo keypoint does not correspond to an existing landmark, it is considered for initialization as a new landmark. We only initialize landmarks when they are supported by observations from multiple viewpoints, similar to the distant stereo point triangulation in [23]. We triangulate a stereo keypoint upon first viewing it, but do not add it to the optimization yet. If a stereo keypoint lies within g_t of the projected landmark for the N consecutive frames, we add this landmark to the global map. We initialize it by triangulating over all the N views. It is only then that the landmark and its corresponding measurements are added to the optimization. The value of $N = 5$ is empirically selected, but this is often reduced in difficult visibility conditions.

D. Implementation

Our framework uses the GTSAM library [5] for factor-graph optimization. We use iSAM2 [17] for an efficient incremental solution using the Powell's dog-leg optimization algorithm. The experiments (Section IV-C and IV-D) are run offline on an Intel Core i7-7820HQ CPU @ 2.90GHz and 32GB RAM without GPU parallelization.

IV. Experimental Results

A. Vehicle Description

Our framework is intended for any AUV with vehicle odometry (Section III-A) and stereo sensing. We use the Hovering Autonomous Underwater Vehicle (HAUV) from Bluefin Robotics [33] (Fig. 1), with properties of the vehicle modeled in both our simulation and real-world experiments (see Table II). The vehicle's payload is comprised of a Doppler velocity log (DVL), attitude and heading reference system (AHRS) and depth sensor, with measurements characterized as follows:

(i) The depth sensor provides direct measurements of HAUV depth (Z).

(ii) The AHRS observes gravity to give drift-free pitch and roll estimates.

(iii) The X, Y and yaw quantities are obtained via dead reckoning. During long-term operation they drift unboundedly with time.

Using high-precision navigation sensors, the proprietary odometry of our vehicle exhibits very low drift over the relatively short time frames of operation. We treat this as the *ground truth*. We corrupt the relative odometry between poses with significant additive white Gaussian noise. This induces drift in the XY plane to mimic having a less accurate IMU + DVL payload, as usually seen in underwater applications. Noise is added in the XYH directions at every frame, with standard deviations $\sigma_x = \sigma_y = 0.01$ m and $\sigma_\phi = 0.01$ rad. In Section IV-C and IV-D, we compare this synthesized *dead reckoning* with our *SLAM solution*.

The stereo pair consists of two Prosilica GC1380 cameras fixed adjacent to the DVL, oriented upwards (Fig. 1). It has a 0.078 m baseline and records 5 fps grayscale images (680×512). We calibrate the stereo camera underwater and manually measure the camera-robot transformation. Images are corrected for radial and tangential distortion.

B. Evaluation Metrics

We compare the dead reckoning and SLAM trajectories with the ground truth. We use standard trajectory evaluation metrics [30], namely the absolute trajectory error (ATE) and relative pose error (RPE). The ATE computes the difference between points on a query trajectory and the ground truth, while the RPE quantifies local drift [30]. In simulation, we also compute the mean and median absolute landmark error (ALE) of the final landmark map.

C. Simulated Experiments

For preliminary analysis, we run simulations with generated vehicle motions and assume known data association. We randomly initialize landmarks in space above the water surface, spread across the XY plane and between 4–5 m in the Z direction. We add Gaussian noise ($\sigma = 1$ pixel) to stereo landmark measurements. When projecting ground truth landmarks, we apply our refractive model to simulate looking through the water surface. Two scenarios are analyzed: a `square` and `corkscrew` trajectory. While `square` does not include motion in the Z direction or yaw rotation, `corkscrew` exercises all these degrees of freedom. To emulate the HAUV, we constantly vary the pitch and roll over the $\pm 5°$ range. Each dataset has 1200 poses, executing 7 loops of radius 2.5m in `corkscrew` and 10 loops of side length 3m in `square`.

In Table I, we quantitatively compare the dead reckoning and SLAM estimate trajectories against ground truth. It can be seen that we achieve substantial reduction in ATE and RPE with our framework for both trajectories. While the mean ALE is higher for `corkscrew`, the median verifies that it is due to outliers. Fig. 6 qualitatively compares both trajectories and estimated landmarks. The dead reckoning trajectory drifts significantly over time, while our solution

TABLE II: Covariance matrices (defined in Section III-A) used in simulation and real-world experiments. They are diagonal square matrices of the form $\text{diag}(M_0{}^2, M_1{}^2, \ldots)$. The units for translation, rotation and image measurements are meters, radians and pixels respectively.

Covariances	**Square roots of matrix diagonal elements** (M)
$\boldsymbol{\Sigma_0}$	10^{-4} m, 10^{-4} m, 10^{-4} m, 10^{-4} m, 10^{-4} m, 10^{-4} m
$\boldsymbol{\Psi_i}$	0.01 m, 0.01 m, 0.01 rad
$\boldsymbol{\Phi_i}$	0.01 m, 0.005 rad, 0.005 rad
$\boldsymbol{\Gamma_k}$	1 pix, 1 pix, 1 pix, 1 pix

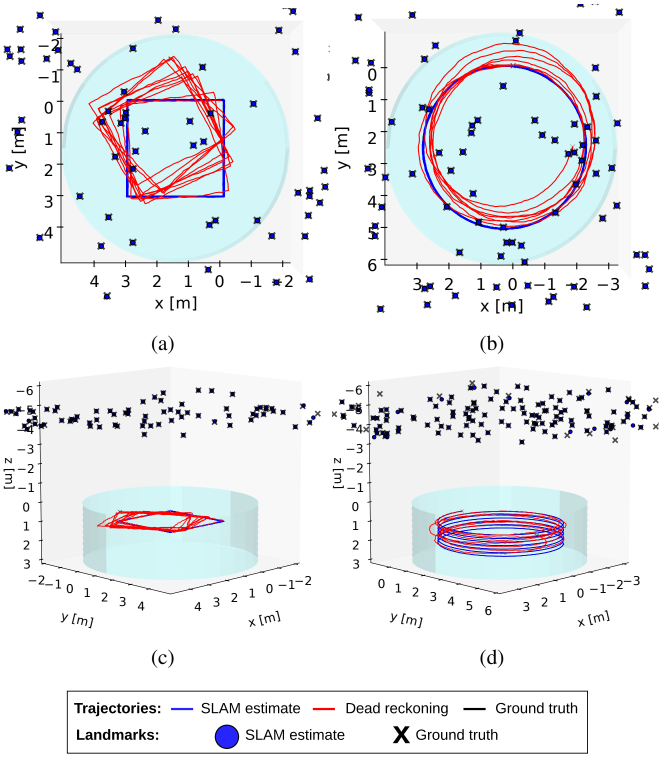


Fig. 6: Visualization of the SLAM trajectory and landmark estimates from simulation, overlaid with the tank environment. **(a)** and **(b)** show top-views while **(c)** and **(d)** are from the side. The SLAM solution coincides (and thus obscures) the ground truth, while the dead reckoning drifts. The estimated landmarks converge to near their ground truth positions.

roughly overlaps with the ground truth. In Table IV, we further compare these results with a modified implementation that does not account for refraction.

D. Real-world Experiments

Our SLAM framework is evaluated using the HAUV in an indoor test-tank. The tank has a depth of 3m and radius of 3.5m. Regions of the ceiling are not at the same height from the water surface due to piping, air ducts and girders. On measurement with survey equipment, they are found to be between 3.6–5.8m. Fig. 7 shows the ceiling and tank setup.

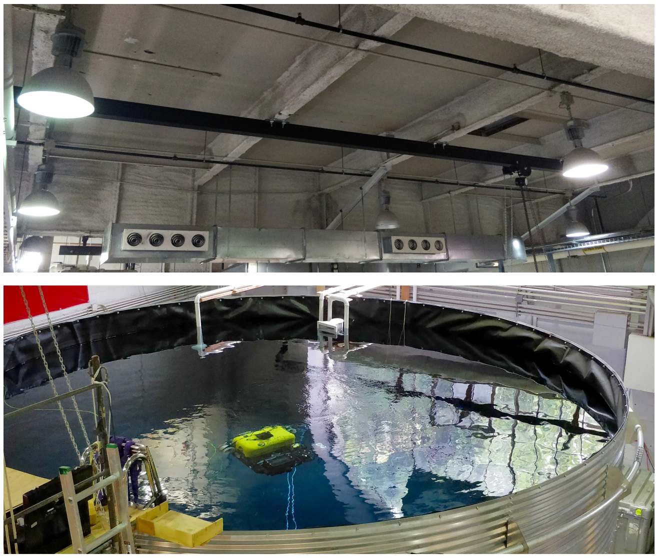


Fig. 7: **(top)** Ceiling present over the tank. Objects in the vehicle's field-of-view are between 3.6–5.8m in height from the water surface. **(bottom)** Tank setup with vehicle executing a trajectory at 1m depth.

We log 12 datasets for evaluation that encompass a wide range of scenarios the vehicle may encounter. They vary between 100–686 seconds in length and all but one execute pre-programmed loops in the tank. The vehicle translates in the X and Y directions at a fixed depth, along with rotation about the Z-axis (yaw rotation). The pitch and roll directions of the vehicle cannot be controlled, but fluctuate mildly underwater nevertheless. Upon receiving a valid pair of stereo frames, we use its timestamp to interpolate a state estimate. Challenges that can degrade the SLAM solution include water surface disturbances, motion blur, suspended particulates, light scattering and image washout (Fig. 1). The value of N (refer Section III-C.2) is reduced to 2 in datasets with larger disturbances. The datasets incorporate all these conditions *(brackets denotes number of such datasets)*:

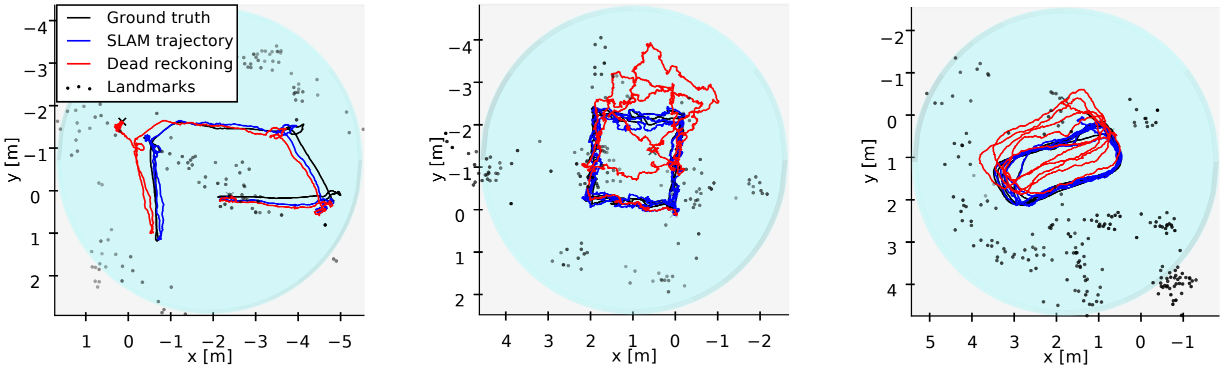
Depth: Just below surface *(4)*, 1m *(4)* and 2m depth *(4)*.
Visibility: With *(8)* and without *(4)* suspended particulates.
Lighting: With *(3)* and without *(9)* ceiling lights.

Table III lists the evaluation metrics for the dead reckoning and SLAM solution for all 12 datasets. We choose one representative dataset from each depth level—datasets **03**, **08** and **09**—and plot the trajectory estimates (Fig. 8). Our proposed method significantly reduces drift in all cases, as seen in the ATE and RPE metrics. This is most apparent in the longer datasets, **08** (Fig. 8 (b)) and 10.

We also compare the results from our real-world and simulation dataset with a modified implementation that does not account for refraction (refer Table IV). The results show reduced error when we account for refraction (RC), which reinforces our method. We also see a significant difference

TABLE I: Mean absolute trajectory error (ATE) and relative pose error (RPE) for the two simulation trajectories. Mean and median absolute landmark error (ALE) are also shown. We see a significant decrease in error in the SLAM solution as compared to the dead reckoning trajectory.

Dataset	**Dead reckoning**			**SLAM solution**				
	ATE (m)	$\text{RPE}_{\text{trans}}$ (m)	RPE_{rot} (°)	ATE (m)	$\text{RPE}_{\text{trans}}$ (m)	RPE_{rot} (°)	mean ALE (m)	median ALE (m)
`square`	0.458	0.661	16.444	0.012	0.018	0.130	0.015	0.008
`corkscrew`	0.415	0.593	10.931	0.011	0.017	0.112	0.107	0.005

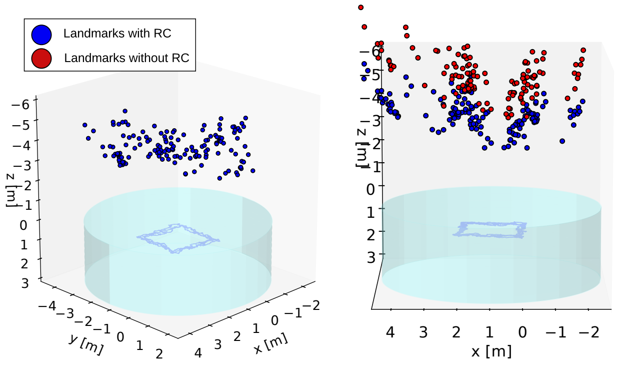


(a) Dataset **03** (0m depth) (b) Dataset **08** (1m depth) (c) Dataset **09** (2m depth)

Fig. 8: Qualitative comparison of trajectories from the representative datasets. We observe strong correspondence between our SLAM trajectory and the ground truth, while the dead reckoning trajectory drifts over time. The global coordinates (in the X and Y) vary between trajectories as the origin is defined by the vehicle start position prior to recording.

between the final landmark maps of both cases. Fig. 9 visualizes this result for dataset **08**.

The solve time for each dataset (Table III) depends on how densely connected the underlying factor graph is. Most of the execution time is devoted to the optimization; the refraction module requires only a smaller proportion of the compute time. For example, the entire optimization for dataset **08** (724 seconds dataset duration, 144 landmarks) takes 884.3 seconds to solve with the refraction module, and 789.4 seconds without. We can achieve real-time performance through keyframing or fixed-lag smoothing.



(a) (b)

Fig. 9: (a) Final landmark map of dataset **08**. (b) The landmark map with refraction correction is compared with that without refraction correction.

V. Conclusion and Future Work

We have presented a novel localization framework for underwater vehicles in nuclear pools and other analogous environments. There exists no prior work that takes cues from above the water surface for underwater visual SLAM. By utilizing an onboard upward-facing stereo camera, our method is less prone to failure in cluttered environments as compared to traditional line-of-sight methods. We detail the challenges that refraction presents and develop a correction module. Previously, refraction correction had only been

TABLE III: Mean ATE and RPE for the 12 underwater datasets. Details about each dataset—operation depth, runtime duration and solve time—are shown. 0m indicates a depth *just* below the water surface. Datasets in **bold** are the representative datasets, which further appear in Fig. 8 and Table IV.

Dataset				Dead reckoning			SLAM solution		
#	depth (m)	duration (s)	solve time (s)	ATE (m)	RPE$_{\text{trans}}$ (m)	RPE$_{\text{rot}}$ (°)	ATE (m)	RPE$_{\text{trans}}$ (m)	RPE$_{\text{rot}}$ (°)
01	1	133.8	476.3	0.069	0.112	4.572	0.053	0.072	2.198
02	0	99.6	188.7	0.122	0.149	3.301	0.067	0.079	1.480
03	0	202.2	505.5	0.280	0.370	5.976	0.115	0.090	2.539
04	2	121.8	63.0	0.103	0.145	4.121	0.058	0.125	2.822
05	1	192.6	23.4	0.076	0.112	2.600	0.046	0.062	1.273
06	2	203	13.5	0.095	0.137	2.328	0.051	0.068	1.520
07	1	238.8	329.9	0.181	0.248	5.839	0.074	0.096	2.886
08	1	724.0	884.3	0.568	0.818	21.451	0.073	0.098	2.267
09	2	260.0	449.2	0.265	0.343	5.696	0.086	0.105	2.216
10	2	686.2	1409.0	0.327	0.402	20.583	0.068	0.082	1.365
11	0	446.8	2088.0	0.259	0.329	9.050	0.037	0.051	1.175
12	2	200.0	91.4	0.096	0.160	2.972	0.050	0.065	1.164

TABLE IV: ATE of real-world **(left)** and simulation datasets **(right)** with/without refraction correction (RC). It reduces when RC is present in the framework.

Dataset	01	02	**03**	04	05	06	07	**08**	**09**	10	11	12	*average*	`square`	`corkscrew`
ATE with RC (m)	0.053	0.067	0.115	0.058	0.046	0.051	0.074	0.073	0.086	0.068	0.037	0.050	*0.065*	0.012	0.011
ATE without RC (m)	0.057	0.067	0.155	0.060	0.047	0.053	0.075	0.074	0.102	0.071	0.041	0.058	*0.072*	0.015	0.014

addressed for aerial photogrammetry and lens housing compensation. We formulate the landmark-based stereo SLAM problem and address the challenges faced by the frontend. We evaluate the method through simulation and a dozen real-world underwater experiments. These validate our method's ability to achieve a drift-free state estimate in the presence of significant noise.

Our underlying approximation of water surface planarity can be improved by modeling for waves and ripples [9]. The generic point feature frontend can be improved by taking ideas from the state-of-the-art in visual SLAM. It can be replaced by a dense or semi-dense method for mapping applications, or combined with lines for robust detection [12]. In larger environments, we can also integrate loop closure detection. For computational efficiency, an over-compensation factor can be used in the refraction module, or it can be completely replaced by a lookup-table [21]. A large baseline stereo pair will guarantee better results for distant stereo points. Further, we can also rectify stereo images to exploit epipolar constraints for faster matching. Point correspondences are restricted to epipolar *curves* due to the refractive interface, as detailed by [11]. The SLAM framework may also be extended to support the use of monocular cameras.

REFERENCES

[1] A. Agrawal, S. Ramalingam, Y. Taguchi, and V. Chari, "A theory of multi-layer flat refractive geometry," *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, pp. 3346–3353, 2012.

[2] M. Agrawal, "A Lie algebraic approach for consistent pose registration for general Euclidean motion." *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pp. 1891–1897, 2006.

[3] O. Bailo, F. Rameau, K. Joo, J. Park, O. Bogdan, and I. S. Kweon, "Efficient adaptive non-maximal suppression algorithms for homogeneous spatial keypoint distribution," *Pattern Recognition Letters*, vol. 106, pp. 53–60, 2018.

[4] B.-H. Cho, S.-H. Byun, C.-H. Shin, J.-B. Yang, S.-I. Song, and J.-M. Oh, "KeproVt: Underwater robotic system for visual inspection of nuclear reactor internals," *Nuclear engineering and design*, vol. 231, no. 3, pp. 327–335, 2004.

[5] F. Dellaert, "Factor graphs and GTSAM: A hands-on introduction," Georgia Institute of Technology, Tech. Rep., 2012.

[6] F. Dellaert and M. Kaess, "Factor graphs for robot perception," *Foundations and Trends in Robotics*, vol. 6, no. 1-2, pp. 1–139, 2017.

[7] D. Fox, W. Burgard, F. Dellaert, and S. Thrun, "Monte Carlo localization: Efficient position estimation for mobile robots," *AAAI Conf. on Artificial Intelligence*, vol. 1999, no. 343-349, pp. 2–2, 1999.

[8] J. G. Fryer, "Photogrammetry through shallow water," *Australian J. of Geodesy, Photogrammetry and Surveying*, vol. 38, pp. 25–38, 1983.

[9] J. G. Fryer and H. T. Kniest, "Errors in depth determination caused by waves in through-water photogrammetry," *The Photogrammetric Record*, vol. 11, no. 66, pp. 745–753, 1985.

[10] R. Garcia and N. Gracias, "Detection of interest points in turbid underwater images," *OCEANS 2011*, pp. 1–9, 2011.

[11] J. Gedge, M. Gong, and Y.-H. Yang, "Refractive epipolar geometry for underwater stereo matching," *Computer and Robot Vision (CRV), 2011 Canadian Conference on*, pp. 146–152, 2011.

[12] R. Gomez-Ojeda, D. Zuñiga-Noël, F.-A. Moreno, D. Scaramuzza, and J. Gonzalez-Jimenez, "PL-SLAM: a stereo SLAM system through the combination of points and line segments," *arXiv:1705.09479*, 2017.

[13] W. Jeong and K. M. Lee, "CV-SLAM: A new ceiling vision-based SLAM technique," *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pp. 3195–3200, 2005.

[14] J. Jung, Y. Lee, D. Kim, D. Lee, H. Myung, and H.-T. Choi, "AUV SLAM using forward/downward looking cameras and artificial landmarks," *Underwater Technology, 2017 IEEE*, pp. 1–3, 2017.

[15] J. Jung, J.-H. Li, H.-T. Choi, and H. Myung, "Localization of AUVs using visual information of underwater structures and artificial landmarks," *Intelligent Service Robotics*, vol. 10, no. 1, pp. 67–76, 2017.

[16] M. Kaess, A. Ranganathan, and F. Dellaert, "iSAM: Incremental smoothing and mapping," *IEEE Trans. Robotics*, vol. 24, no. 6, pp. 1365–1378, 2008.

[17] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert, "iSAM2: Incremental smoothing and mapping using the Bayes tree," *Intl. J. of Robotics Research*, vol. 31, no. 2, pp. 216–235, 2012.

[18] T. E. Lee and N. Michael, "State estimation and localization for ROV-based reactor pressure vessel inspection," *Field and Service Robotics*, pp. 699–715, 2018.

[19] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[20] T. Łuczyński, M. Pfingsthorn, and A. Birk, "The pinax-model for accurate and efficient refraction correction of underwater cameras in flat-pane housings," *Ocean Engineering*, vol. 133, pp. 9–22, 2017.

[21] H.-G. Maas, "On the accuracy potential in underwater/multimedia photogrammetry," *Sensors*, vol. 15, no. 8, pp. 18 140–18 152, 2015.

[22] S. E. Masry, "Measurement of water depth by the analytical plotter," *The International Hydrographic Review*, vol. 52, no. 1, 2015.

[23] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[24] T. Murase, M. Tanaka, T. Tani, Y. Miyashita, N. Ohkawa, S. Ishiguro, Y. Suzuki, H. Kayanne, and H. Yamano, "A photogrammetric correction procedure for light refraction effects at a two-medium boundary," *Photogr. Eng. & Remote Sensing*, vol. 74, no. 9, pp. 1129–1136, 2008.

[25] S. Nawaz, M. Hussain, S. Watson, N. Trigoni, and P. N. Green, "An underwater robotic network for monitoring nuclear waste storage pools," *International Conference on Sensor Systems and Software*, pp. 236–255, 2009.

[26] L. Paull, S. Saeedi, M. Seto, and H. Li, "AUV navigation and localization: A review," *IEEE Journal of Oceanic Engineering*, vol. 39, no. 1, pp. 131–149, 2014.

[27] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," *Intl. Conf. on Computer Vision (ICCV)*, pp. 2564–2571, 2011.

[28] I. Rust and H. Asada, "A dual-use visible light approach to integrated communication and localization of underwater robots with application to non-destructive nuclear reactor inspection," *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pp. 2445–2450, 2012.

[29] M. R. Shortis and E. S. Harvey, "Design and calibration of an underwater stereo-video system for the monitoring of marine fauna populations," *International Archives of Photogrammetry and Remote Sensing*, vol. 32, pp. 792–799, 1998.

[30] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pp. 573–580, 2012.

[31] G. C. Tewinkel, "Water depths from aerial photographs," *Photogrammetric Engineering*, vol. 29, no. 6, pp. 1037–1042, 1963.

[32] T. Treibitz, Y. Schechner, C. Kunz, and H. Singh, "Flat refractive geometry," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 34, no. 1, pp. 51–65, 2012.

[33] J. Vaganay, M. Elkins, D. Esposito, W. O'Halloran, F. Hover, and M. Kokko, "Ship hull inspection with the HAUV: US Navy and NATO demonstrations results," *OCEANS 2006*, pp. 1–6, 2006.

[34] N. Weidner, S. Rahman, A. Q. Li, and I. Rekleitis, "Underwater cave mapping using stereo vision," *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pp. 5709–5715, 2017.

[35] R. M. Westaway, S. N. Lane, and D. M. Hicks, "The development of an automated correction procedure for digital photogrammetry for the study of wide, shallow, gravel-bed rivers," *Earth Surface Processes and Landforms*, vol. 25, no. 2, pp. 209–226, 2000.

[36] E. Westman and M. Kaess, "Underwater AprilTag SLAM and calibration for high precision robot localization," Carnegie Mellon University, Tech. Rep. CMU-RI-TR-18-43, October 2018.