3-D Reconstruction Using Monocular Camera and Lights: Multi-View Photometric Stereo for Non-Stationary Robots

Monika Roznere[1], Philippos Mordohai[2], Ioannis Rekleitis[3], and Alberto Quattrini Li[1]

***Abstract*— This paper proposes a novel underwater Multi-View Photometric Stereo (MVPS) framework for reconstructing scenes in 3-D with a non-stationary low-cost robot equipped with a monocular camera and fixed lights. The underwater realm is the primary focus of study here, due to the challenges in utilizing underwater camera imagery and lack of low-cost reliable localization systems. Previous underwater PS approaches provided accurate scene reconstruction results, but assumed that the robot was stationary at the bottom. This assumption is limiting, as many artifacts, reefs, and man-made structures are large and meters above the bottom. Our proposed MVPS framework relaxes the stationarity assumption by utilizing a monocular SLAM system to estimate small robot motions and extract an initial sparse feature map. To compensate for the scale inconsistency in monocular SLAM output, our MVPS optimization scheme collectively estimates a high-quality, dense 3-D reconstruction and corrects the camera pose estimates. We also present an attenuation and camera-light extrinsic parameter calibration method for non-stationary robots. Finally, validation experiments with a BlueROV2 demonstrated the low-cost capability of producing high-quality scene reconstructions. Overall, this work is the foundation of an active perception pipeline for robots (i.e., underwater, ground, and aerial) to explore and map complex structures in high accuracy and resolution with an inexpensive sensor-light configuration.**

I. Introduction

We present novel work for solving the Multi-View Photometric Stereo (MVPS) problem in the case of non-stationary underwater robots (see Fig. 1 for the main long-term vision). PS is a well-known computer vision technique for reconstructing high resolution scenes, considering stationary cameras and various lighting sources [1], [2]. While we primarily focus on the underwater domain due to its increased challenges, our MVPS framework can be easily generalized to above-water domains.

Scene reconstruction is an important aspect in many underwater robotic applications, particularly for inspecting man-made structures (e.g., oil rigs, ship hulls) [3], monitoring target biological locations [4], and exploring reefs [5] and archaeological sites [6]. Autonomous Underwater Vehicles (AUVs) are becoming more commonly dispatched to tackle these various tasks [7]. Not only can AUVs stay longer underwater than a diver, but they are also typically set up with a modular sensory suite – at the very least with an IMU, monocular camera, single-beam echosounder, and lights [8], and more extensively (and at higher cost) with a multibeam sonar, side scan sonar, and guidance-based equipment (i.e., fiber-optic gyroscope (FOG) IMU, acoustic Doppler Velocity Log (DVL)) [9], [10].

[1]Dartmouth College, Hanover, NH, USA, 03755 {monika.roznere.gr, alberto.quattrini.li}@dartmouth.edu

[2]Stevens Institute of Technology, Hoboken, NJ, USA, 07030, pmordoha@stevens.edu

[3] University of South Carolina, Columbia, SC, USA, 29208, yiannisr@cse.sc.edu

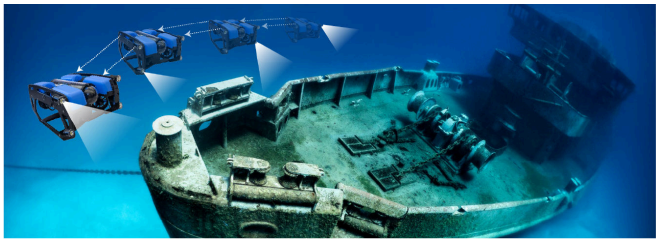

Fig. 1: How multi-view photometric stereo framework is applied to non-stationary robots (i.e., BlueROV2) for exploring shipwrecks and producing high-quality scene reconstruction models.

Multibeam and other sonars were shown to be extremely useful for accurate underwater scene reconstruction [3], [11]. However, sonars lack visual (e.g., color, texture) and resolution characteristics that cameras provide, which can be enriched by fusing sonar and camera data. Stereo vision setup is possible; however, not only does it require higher computation, but in scenes with significant uniformity (less varied or repeating textures – common in underwater environment), the left-right camera pixel correspondences can be erroneous or impossible [12]. On its own, monocular camera imagery input cannot provide accurate scene depth information [13]. IMU or DVL data can be integrated [14]–[16], as in Visual Inertial Odometry systems. However, these methods produce camera poses in a sparsely reconstructed scene, and we are interested in dense reconstruction.

PS relies only on camera imagery *and* light sources (artificial or natural). It is originally based [1] on the observation that an object's surface normals can be estimated by observing changes in the surface points' reflected light intensities among different images, where light source(s) change position, but the camera's position *always* stays in place. For a review of different PS methods, we refer readers to surveys [2], [17]. Lighting variations from moving light sources have been utilized to infer shape from shadows [18], while the interaction of video-lights with the walls of underwater caves produced 3D reconstructions of the cave passage [19].

MVPS for non-stationary underwater robots is to the best of our knowledge a novel, unsolved problem. Early approaches on MVPS in general environments [20]–[22] required silhouettes to initialize shape. (We assume that the object can be segmented for simplicity, but we do not use the silhouettes for shape estimation.) There is also work

benefiting from the complementary strengths of PS and MVPS [23]–[28].

Previous PS works [29]–[31] that tested with underwater robots required that the robots stay settled on the bottom to ensure that the camera does not move. This is undesirable as target objects (e.g., corals, parts of shipwrecks) may be located meters above the bottom, most underwater robots are setup to be neutrally or positively buoyant, thus requiring motor usage to stay at the bottom and that may cause sediment stirring and image hazing, and, lastly, it is common underwater practice to not touch the habitat in order to avoid accidental interference with sensitive organisms and artifacts. Furthermore, these prior PS works tested with small objects (i.e., seashells, plastic containers), whose sizes are small enough for short-range image capture, but different to the task of reconstructing shipwrecks and reefs.

The main contributions of this paper are: (1) a novel MVPS framework for computing 3-D scene models, considering camera non-stationarity, near-lighting model, and initial knowledge from a monocular SLAM system, (2) non-stationary calibration methods for underwater attenuation and camera-light extrinsic parameters, and (3) real-world experiments with a moving underwater robot integrated with four independently controlled lights. From our proposed framework, experiments, and qualitative/quantitative results, we show the capability of a non-stationary robot to reconstruct underwater 3-D objects, using only a monocular camera and lights.

This work represents the foundation for allowing inexpensive robots to explore unknown scenes and produce high-resolution models that are on par with what can be achieved with high-end sensors and robots.

II. Image Formation Model

Underwater scenarios are challenging for any image-based method. Light attenuates (reduces in intensity) more extensively in-water than in-air; as light travels in water, it is scattered and absorbed by colliding particles, see Fig. 2 (left). Many methods attempt to estimate the attenuation values [32]–[34], especially through a physics-based image formation model [35]. In the image formation model, an image I is captured by the camera's image sensor[1] and is described as a composition of direct signal D and backscatter B:

$$I = D + B \tag{1}$$

A. Direct Signal

Direct signal D corresponds to the amount of light that has traveled from the light source, reflected from the visible scene, and reached the camera's image sensor. During the light's travel, it is attenuated by β^D, based on the water medium's characteristics:

$$D = \frac{1}{z_i^2}\, a\, L_R\, e^{-\beta^D z_i} \tag{2}$$

[1]We assume the pinhole model. We will also refer to grayscale images for simplicity; color images require additional unknowns that will be considered in future work. Explanations will also be simplified to a single pixel x, corresponding to a unique surface point, such as $I = I_x$.

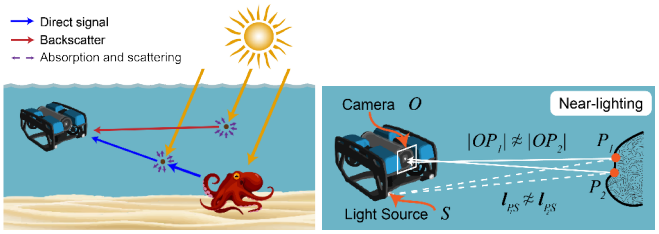

Fig. 2: *Left*: Visualization of how direct signal and backscatter are generated and attenuated. *Right*: Under the near-lighting model, all parameters correlated to surface points cannot be assumed to be the same.

where a is the surface point's albedo (or color), L_R is the light reflected from the surface point, and $z_i = |PS_i| + |OP|$, such that $|PS_i|$ is the distance from the light source S_i to the surface point P and $|OP|$ is the distance from the surface point to the camera O. Simply, D is the distorted image of $a\, L_R$, attenuated by $\beta^D\, z_i$. The inverse-square law ($\frac{1}{z_i^2}$) is applied, as light loses intensity over distance.

B. Backscatter

Backscatter B corresponds to the amount of light that *never* reached the visible scene surfaces, but was reflected by particles in the water and arrived at the camera's image sensor. Similar to D, it is attenuated by certain water medium's attenuation properties, denoted here as β^B:

$$B = \frac{1}{z_i^2}\, B^{\infty}(1 - e^{-\beta^B z_i}) \tag{3}$$

Note, β^D and β^B are not assumed to be the same, as studied in [35]. The veiling light B^{∞} is also a characteristic of the water medium, but it can be approximated as the color in the image's background (the far 'infinite' distance).

C. Complete Image Formation Model

The complete *underwater* image formation model for a light source i is as follows:

$$I_i = \frac{1}{k\, z_i^2}\,(\, a\, L_{R_i}\, e^{-\beta^D z_i} + B^{\infty}(1 - e^{-\beta^B z_i})) \tag{4}$$

where scalar k corresponds to image exposure.

D. Scene Reflection

The underwater scene is commonly assumed to be composed of Lambertian surfaces [5], [30], [31]. Thus, given a surface normal and light direction, the same amount of reflected light will be observed in any viewing direction:

$$L_{R_i} = L_{i_\phi}\, \mathbf{n}_i\, \hat{\mathbf{l}}_{\mathbf{PS_i}} = L_{i_\phi}\, \cos(\theta_i) \tag{5}$$

where θ_i is the angle between the surface normal $\mathbf{n}_i$ and any incident (incoming) light direction $\hat{\mathbf{l}}_{\mathbf{PS_i}}$, and L_{i_ϕ} is the light intensity.

E. Light Models

In daylight, ambient light is a significant source of illumination for the first 20-30 m deep in the water column[2]. If it is present, then one can capture an image of the scene

[2]With increasing depth, ambient light's intensity diminishes due to attenuation and the inverse-square law [35].

with only ambient light present and use that to subtract the following images that include artificial light sources.

Each artificial light source is represented as a point light with an original intensity L_{i_0} and modeled with a Gaussian diffuse filter. Following the model in [5], intensity is brightest along its center directional line $\hat{\mathbf{l}}_{\mathbf{S_i}}$, but decreases with angle ϕ from this line:

$$L_{i_\phi} = L_0 e^{-\frac{1}{2}\frac{\phi^2}{\sigma^2}}, \quad \sigma = \sqrt{\frac{\phi_{50\%}^2}{-2\log 0.5}} \tag{6}$$

where $\phi_{50\%}$ is the angle where the light's power is at 50%.

We applied the near-lighting model, see Fig. 2 (right), which is used in cases when the viewing/lighting distances are small [31], or in our case, when the target object is as large or larger than the viewing/lighting distances[3].

III. Attenuation and Camera-Light Calibration

Attenuation coefficients and camera-light extrinsic parameters can be calibrated prior to main deployment, preferably in the same marine environment. Calibration techniques traditionally use ground truth targets, such as a white Lambertian board [31], a black-and-white checkerboard [32], or a chrome ball. Below, we explain how to perform both attenuation and camera-light extrinsic parameter calibration with a checkerboard, as a and $|OP|$ are known[4].

Attenuation coefficients in Equation (4) can be assumed to be constant throughout the general area and depth. At each depth, one can optimize the attenuation (β^D, β^B), veiling light B^∞, and camera exposure k by minimizing the difference between the observed pixels in I and the estimated pixels in I',

$$\underset{\beta^D,\beta^B,B^\infty,k}{\arg\min} \sum_{x=1}^{N_X} (I_x - I'_x(a_{W \text{or} B}, \beta^D, \beta^B, B^\infty, k))^2 \tag{7}$$

where N_X are all pixels in the image I used for calibration, corresponding to surface points with white W or black B albedo properties $a_{W \text{or} B}$. For white pixels use $a_W = 1$ and for black pixel use $a_B = \frac{1}{255}$ (to avoid using 0).

Calibrating camera-light extrinsic parameters prior to deployment helps minimize the overall number of unknowns in the PS objective function, later explained in Section IV. Let all light parameters – consisting of all light poses (S_i), light center directions ($\hat{\mathbf{l}}_{\mathbf{S_i}}$), original intensity L_0, and 50% power angle $\phi_{50\%}^2$ – be jointly denoted as $\mathbb{S}$. Exposure k is optional if it changes over time. As ambient light might be present in image I_i, let A_i be the corresponding pixel intensity of the image taken a few moments earlier with no artificial light. Therefore, after completing Equation (7), the parameters in $\mathbb{S}$ are optimized according to:

$$\underset{\mathbb{S},k}{\arg\min} \sum_{i=1}^{N_I} (I_i - (A_i + I'_i(a_{W \text{or} B}, \mathbb{S}, k)))^2 \tag{8}$$

[3]Contrarily, the distant-lighting model assumes that the distances between scene points and camera/lights are very large. While it sacrifices model accuracy, it decreases the complexity of the number of unknowns in the framework.

[4]Easier performed with no artificial lights on if ambient light is present. Calibration is jointly performed with N_I (3 or more) images, each with different camera-and-light pairings. The symmetry of the robot's lighting setup can be included as a constraint[5].

IV. Non-Stationary Photometric Stereo

The traditional PS problem consists of taking multiple N_I images with a stationary camera under different lighting conditions. The aim is to estimate the unknown parameters of each interested surface point in view, specifically its albedo a, surface normal $\mathbf{n}$, and depth Z. This is achieved with the image formation model by minimizing the difference between the predicted pixel and the observed pixel intensities:

$$o(a, \mathbf{n}, Z) = \sum_{i=1}^{N_I} (I_i - I'_i(a, \mathbf{n}, Z))^2 \tag{9}$$

Given frame i, known camera pose $O_i = (X_i^O, Y_i^O, Z_i^O)$ and light source $S_i = (X_i^S, Y_i^S, Z_i^S)$, other unknowns include: $|O_iP|$, $|PS_i|$, and $\hat{\mathbf{l}}_{\mathbf{PS_i}} = (S_i - P)/|PS_i|$. Here, $P = (\frac{uZ}{f}, \frac{vZ}{f}, Z)$, where f is the camera's focal length, and (u, v) is an image pixel coordinate. These unknowns can be estimated while solving for Z.

A. Camera Motion and Surface Point Correspondences

Unlike the conventional PS model, our proposed framework is applied to *non-stationary* robots. Specifically, while a robot is suspended underwater, it will not stay in place even in loiter mode; it will slightly move due to external water forces or motor usage. This breaks the main PS assumption that the camera is static at all times. Thus, O_1 can be assumed to be at the origin, and images $i > 1$ need to account for the relative pose changes in camera and light.

A monocular Simultaneous Localization and Mapping (SLAM) system, or a Structure-from-Motion (SfM) solver, can help detect small robot movements between image frames. Some methods include LSD-SLAM [36], DSO [37], monocular ORB-SLAM [13], and monocular SVO [38]. From an underwater domain study [39], it was concluded that DSO (direct method) and ORB-SLAM (indirect method) produced the most stable results for purely monocular setups. We chose to integrate monocular ORB-SLAM, as DSO was shown to have challenges in low gradient scenes and it requires more computation power.

However, ORB-SLAM provides a sparse depth/feature map, consisting of mostly edges and corners. To mitigate the sparsity, ORB-SLAM's map points and corresponding pixel coordinates are interpolated within the masked region of the target object in image I_1 to obtain an approximate, piecewise planar depth map. Then, the set of pixel/point correspondences across images I_2, I_3 and I_4 (with different lighting conditions) are matched by using the provided ORB-SLAM camera transformations.

Scale inconsistency is a known issue in monocular SLAM. Therefore, we assume that the pixel/point matches across the set of images are correct, but the camera poses (and, in

[5]E.g., with the BlueROV2, the top lights are symmetrical to one another in pose and direction, as is with the bottom lights.

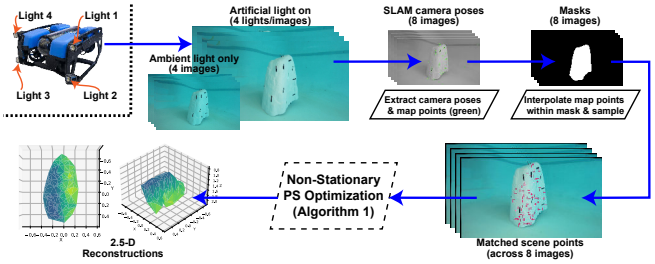

Fig. 3: Proposed pipeline for non-stationary MVPS.

Algorithm 1 Non-Stationary Photometric Stereo Solver

Input: Images with lights on/off and corresponding mask $I_i, A_i, M_i \ \forall i \in N_I$; image formation model parameters and camera-light relative pose; and SLAM camera poses $\hat{O}_i \ \forall i \in N_I$ and set of map points $\mathbb{X}_0$ tracked from ORB-SLAM from I_1
Output: Denser set of scene points $P \in \mathbb{X}'$ with albedo a, depth Z, and surface normal $\mathbf{n}$, and camera pose translation corrections ΔO

/* samples of 3D Points visible across all images and within object mask */
1: $\mathbb{X} \leftarrow$ sampled from tracked SLAM points $\mathbb{X}_0$ corresponding to image pixels in I_0 within mask M_0 and linearly interpolated
2: $\mathbb{X}' \leftarrow$ points $\mathbb{X}$ projected in the subsequent images I_i and A_i that fall within the corresponding masks M_i
/* Initialization; assignment of guessed values to unknown parameters */
3: $\hat{S}_i \quad \forall i \in N_I \leftarrow$ camera poses with I_1 as origin, based on $\hat{O}_i$ and calibrated collocated setup
4: $a_P \leftarrow I_1 \quad \forall P \in \mathbb{X}'$ // albedo initialization with current intensity at corresponding pixel
5: $\mathbf{n}_P \leftarrow [0, 0, -1] \quad \forall P \in \mathbb{X}'$ // initialization of normals pointing to the camera
6: $Z_P \leftarrow Z_{0P} \quad \forall P \in \mathbb{X}'$ // initialization of depth values from the image corresponding to the first light or from single-beam echosounder
7: $\Delta O_i \leftarrow [0, 0, 0] \quad \forall i \in N_I$ // initialization of translation correction
/* Solve non-stationary MVPS objective function: Equation (10) */
8: $\arg\min_{a,\mathbf{n},Z,\Delta O} \sum_{i=1}^{N_I} (I_i - I'_i(a, \mathbf{n}, Z, \Delta O_i))^2$
/* with */
$P \leftarrow (\frac{uZ}{f}, \frac{vZ}{f}, Z)$ update $\forall P \in \mathbb{X}'$
$O_i \leftarrow \hat{O}_i + \Delta O_i \quad \forall i \in N_I$
$S_i \leftarrow \hat{S}_i + \Delta O_i \quad \forall i \in N_I$
Calculate $|O_i P|$, $|P S_i|$ and $\hat{\mathbf{I}}_{\mathbf{P}\mathbf{S}_\mathbf{i}}$, $\forall i \in N_I, \forall P \in \mathbb{X}'$
$I'_i \leftarrow$ via Equation (4) $\forall i \in N_I$
9: **return** a_P, $\mathbf{n}_P$, and $Z_P \ \forall P \in \mathbb{X}'$ and $\Delta O_i \ \forall i \in N_I$

parallel, the world coordinates of the associated scene points) are scaled incorrectly. We do assume that the rotation part of the camera transformations is correct. Hence, the additional translation correction ΔO of camera poses must also be optimized in the MVPS framework:

$$o(a, \mathbf{n}, Z, \Delta O_i) = \sum_{i=1}^{N_I} (I_i - I'_i(a, \mathbf{n}, Z, \Delta O_i))^2 \qquad (10)$$

B. Algorithm and Pipeline

Fig. 3 illustrates the proposed pipeline that uses an AUV with four independently-controlled lights and a monocular camera, and Algorithm (1) overviews the MVPS framework. Lines 1-2 interpolate the tracked SLAM points and ensures that they are located within the masks of all images. Here, we manually segment the object, but as future work we will apply automated semantic segmentation to obtain segmentation masks. Lines 3-7 initialize the unknown parameters with approximate guess values. Line 8 provides detail on how to solve the optimization function described in Equation (10).

V. Results

We performed experiments using data collected in a swimming pool and conducted our framework offline. Our software and plot visualizations are publicly available[6].

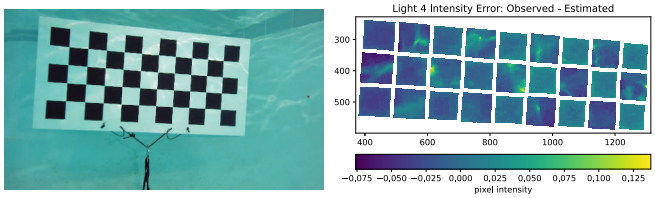

Fig. 4: Attenuation and light parameter calibration were performed with a black-and-white checkerboard. *Left*: Checkerboard illuminated by Light 4. *Right*: Plot on estimated intensity error for [inner] checkerboard with Light 4 on after calibration. Most of the error was due to ambient light and the caustic effect of the AUV's light reflecting from the above water surface.

We used the BlueROV2, equipped with the Sony IMX 322LQJ-C camera [40] with a 5 MP resolution, a horizontal field of view (FOV) of 80°, and a vertical FOV of 64°. Four lights [41] were installed, two on top and two on bottom. During the runs, one light would turn on for 5 s, then all lights would be off for 5 s, and the process would be repeated with a different light. For relative depth measurement, a forward-facing single-beam echosounder [42] (SBES) was installed. Also, an Intel RealSense Depth Camera D455 [43] was used separately to produce ground truth 3-D models, which can only be performed above-water.

A. Remarks on Photometric Stereo Tests

Our experiments were conducted following unorthodox PS standards, as they occurred in more challenging scenarios with uncontrollable factors and with larger target structures.

First, all runs were done in daylight with prevalent ambient light. Past PS works tested in very deep waters or at nighttime [31]; however outdoor nighttime tests are burdensome to arrange due to boat and diver availability. In clear waters, at depths of 30 m or below, ambient light might not be prevalent, but it is still present – thus one needs to account for ambient light in most cases.

The target objects that we used are substantially larger than what are typically used in underwater PS (e.g., hand-held barrel, seashell [30]). We have a black-and-white checkerboard (checkered area: L:1 m x W:0.4 m), two white-painted rocks (a rectangular column Rock A – H:0.94 m x L:0.51 m x W:0.49 m and a more irregular sloped Rock B – H:0.76 m x L:0.74 m x W:0.5 m) – which will mimic real-world reef structures – and a brown planter (Diam.:0.56 m x H:0.43 m). These larger objects will inherently cause an increase in the error of different camera-to-point and light-to-point measurements in the MVPS framework, but will replicate the larger coverage of object-in-view that we might encounter when exploring shipwrecks and other large scenes. While white objects are not expected in the real-world, here they provided helpful validation comparison ($a = 1$).

B. Model and Light Calibration

Image formation model and light parameters were calibrated with the checkerboard, see Fig. 4 left.

[6] https://github.com/dartmouthrobotics/psuw

Image formation model parameters (β^B, β^D, B^∞, and k) were calibrated first under ambient light. Following [33], β^B, β^D, and B^∞ were bounded by [0,5], [0,5], and [0,1], respectively. All pixel intensities are in the [0,1] range. After calibration (using Nelder-Mead method [44] for bound-constrained minimization based on the observation that it outperformed other methods like Powell), the pixel intensity error was 0.013 MAE (0.016 RMSE).

Light parameters (S_i, $\hat{\mathbf{I}}_{\mathbf{S_i}}$, L_{i_0}, and $\phi^2_{50\%}$) and camera exposure k were calibrated jointly with all 4 different camera-light pairings and their associated images. After calibration (using the Nelder-Mead method [44]), we calculated an intensity MAE of 0.012 with std. dev. of 0.004 (RMSE of 0.015 with std. dev. of 0.005).

With the calibrated parameters, we projected back the estimated checkerboard intensity values and noticed that scene ambient light and caustic effects from the AUV's lights reflected by the above water surface led to the most deviation. The right plot in Fig. 4 shows the estimated intensity error for Light 4 (0.018 MAE), which is notably due to the errors where the ambient light anomalies occurred in the captured image on the left. It is important to note that the following 3-D model reconstruction experiments were also affected by the ambient light anomalies that changed quickly over time and space.

C. Checkerboard Reconstruction

For quantitative validation, we conducted checkerboard reconstruction tests, see Table I, under cases where all parameters are known (even camera poses), albedo is unknown but assumed to be uniform across, albedo is unknown for each point, and lastly, albedo and the camera poses are unknown. As depth is unknown for all cases, we tested the framework with different initial depth guesses – $\times 0.5$, $\times 1$, and $\times 1.5$ scaling of ground truth depth values, corresponding to ORB-SLAM scale inconsistency. We utilized SciPy minimization [45], stopped after 10 calls, with no convergence guarantee [46], and took the best results. As expected, large initialization errors lead to larger reconstruction errors.

All cases produced models of similar characteristics, as seen in Fig. 5. Where the lights had more direct alignment with the surface normal, the depths were overestimated, though the error was low. This could indicate that the lighting model is too simple – as the AUV lights are circular, one might need to model them using cones of strongest intensity, not as lines.

TABLE I: Checkerboard reconstruction error in MAE (std. dev.) for cases where all parameters are known but depth, albedo is unknown but uniform across ($\sim a$), albedo is unknown for each point, and lastly, albedo and camera poses are unknown.

Initial Guess Depth Scale	a known O_i known	$\sim a$ unknown O_i known	a unknown O_i known	a unknown O_i unknown
$\times 0.5$	0.028 (0.04)	0.451 (0.03)	0.062 (0.04)	0.116 (0.05)
$\times 1$	0.029 (0.04)	0.029 (0.04)	0.051 (0.03)	0.042 (0.04)
$\times 1.5$	0.029 (0.04)	0.030 (0.04)	0.031 (0.04)	0.032 (0.05)

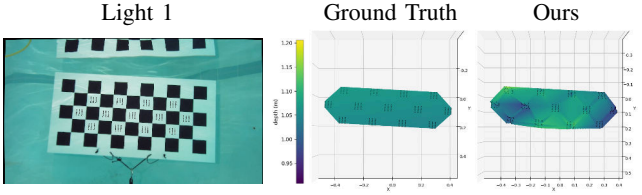

Fig. 5: Checkerboard reconstruction results, under Light 1 (left) viewpoint, from our MVPS framework with albedo a and camera pose O_i unknown. Black dots represent the sampling points. The plots on right are linearly interpolated.

D. Rock and Planter Reconstruction

To test our full proposed MVPS framework, we conducted two runs where the robot circled an object, Rocks **A** and **B**. Fig. 6 shows the results of a few views during these runs. Model reconstructions were performed with the provided map point values from monocular ORB-SLAM, our proposed MVPS framework with albedo known ($a = 1$), and our full MVPS framework with albedo unknown ($\neg a$). The numbers of samples used in the models are provided in Table II, where outliers whose depths were 2 std. dev. greater or less than the average were rejected. The SBES measurements are also provided in the table.

Overall, ORB-SLAM based reconstruction results were consistently flat, even in corner views (**A-2** and **B-2**), and provided small (closer to camera) depth values. Both of our MVPS frameworks were able to reconstruct the rocks well, capturing the shape, vertical slope, and corners. The image depths also show the relative correct depth gradients.

We also calculated the error in albedo estimation, as shown in the right column of Fig. 6 and in Table II. Generally, areas that are flat and parallel to the image frame or are close to the camera's immediate direction corresponded to lower albedo error. Despite the slight error, the shape of the reconstructions are very similar to cases when albedo was known.

Another reason for error could be due to self-shadow. In the case of **B-2**, Light 1 and 2 (left of the AUV) never reached the right side of the rock, causing self-shadow in those images. If the AUV continued its trajectory to the right and collected further images, such reconstruction errors could be mitigated – an idea for future work.

In addition, we conducted two runs with a large non-white planter. Fig. 7 shows the results. While we do not know the albedo of the object, we can compare the results where the albedo is assumed to be uniform across. Here as well, our MVPS framework reconstructed the structure well compared to ORB-SLAM. We also tested the case where the AUV was stationary. The results indicated that there is still room for improving the non-stationarity issue when correctly matching points across all images under different lighting conditions.

TABLE II: *Top*: Number of sample points used in the reconstruction and number of rejected outliers (points whose depth is 2 std. dev. greater or less than the average). *Middle*: SBES depth measurement at Light 1 viewpoint. *Bottom*: Albedo ([0,1]) error for each view of the plastic rocks.

	A-1	**A-2**	**B-1**	**B-2**
Total Points (Outliers)	206 (1)	215 (5)	196 (2)	214 (7)
SBES Measurements	1.20 m	1.14 m	1.07 m	0.83 m
Mean a Error (std. dev.)	0.291 (0.06)	0.367 (0.08)	0.192 (0.12)	0.369 (0.09)

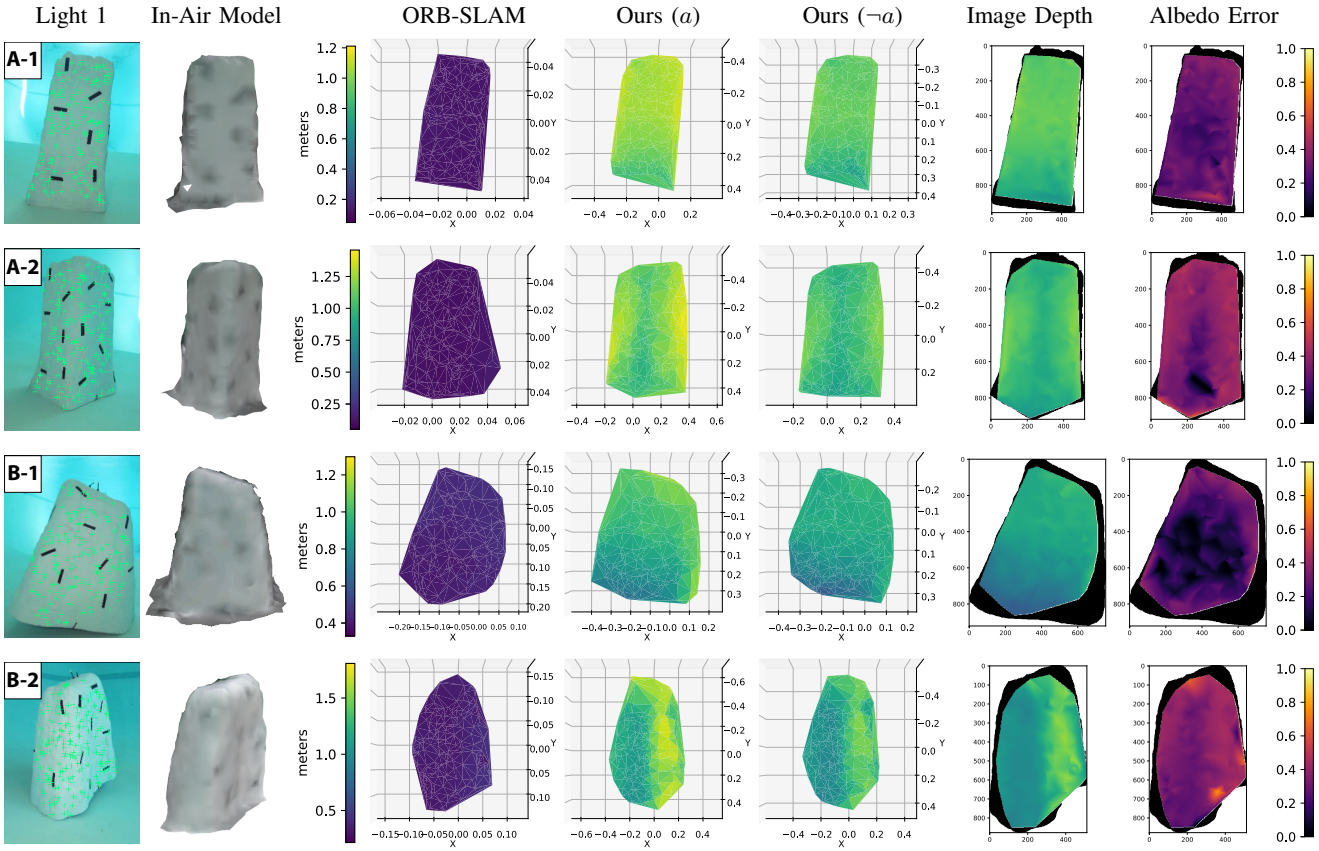

Fig. 6: Reef rocks **A** and **B** reconstruction results provided from the RealSense camera (In-Air Model), monocular ORB-SLAM [13], our MVPS framework with albedo known ($a = 1$), and our MVPS framework with albedo unknown ($\neg a$). The green crosses in the images are the sampled points used in optimization. The colorbar on left applies to all graphs in that row. Projected image depth and albedo error are provided for $\neg a$ case. Note, black tape was placed on the rocks to help ORB-SLAM detect features, considering that the rocks are completely white. Further visualizations are available in our code.

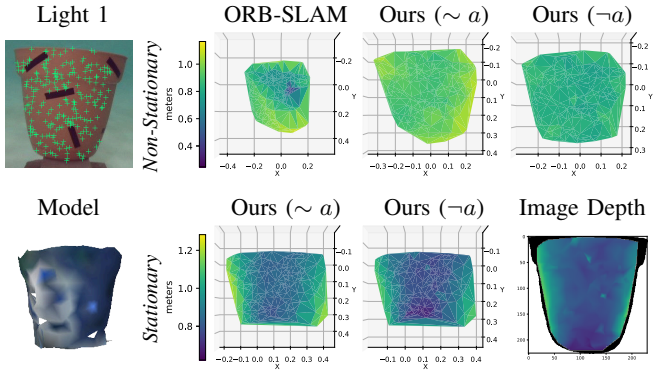

Fig. 7: Planter reconstruction results during *Non-Stationary* (*top*) and *Stationary* (*bottom*) cases, including models from the RealSense camera (In-Air Model), monocular ORB-SLAM [13], our MVPS framework with one uniform albedo unknown ($\sim a$), our MVPS framework with all albedo unknown ($\neg a$), and the image depth for *Stationary* ($\neg a$) case. SBES depth measurements: *Non-Stationary* 1.03 m and *Stationary* 0.91 m.

VI. Conclusion and Future Work

We presented a Multi-View Photometric Stereo (MVPS) framework for non-stationary underwater robots – a common case when the robot is neutrally/positively buoyant, avoiding environmental impact, or exploring large structures – which to our knowledge is a novel and unsolved problem. By expanding the traditional PS framework to include monocular SLAM for extracting camera poses and feature/map points, our MVPS framework is able to calculate a reliable 3-D model of the target object while also correcting the scale inaccuracy afflicting monocular SLAM. Moreover, we presented an easy attenuation and camera-light extrinsic parameter calibration method for non-stationary robots.

Our next step is to design an online framework that builds a 3-D model of a target object as the AUV journeys around it. As the AUV is non-stationary even during the few seconds with one light on, an estimated model could be initially built with this one-light arrangement. Then, the model can be improved and further refined with the other light arrangements. In addition, our MVPS framework produces scene model and camera pose corrections which can be fed back into the SLAM system for AUV trajectory improvement, mitigating the drift and depth scale inconsistency.

Ultimately, an MVPS approach for non-stationary underwater robots has significant impacts – it allows for inexpensive AUVs to accomplish scene reconstruction tasks with results on the same level as using high-end sonars.

Acknowledgment

We thank Devin Balkcom for experiment and intellectual help. This work is supported by the Dartmouth Burke Research Initiation Award, NSF CNS-1919647, 2024541, 2024653, 2144624, 2024741, 1943205, OIA-1923004.

REFERENCES

[1] R. J. Woodham, "Photometric method for determining surface orientation from multiple images," *Optical Engineering*, vol. 19, no. 1, pp. 139 – 144, 1980. 1

[2] J. Ackermann and M. Goesele, "A survey of photometric stereo techniques," *Foundations and Trends® in Computer Graphics and Vision*, vol. 9, no. 3–4, p. 149–254, 2015. 1

[3] F. S. Hover, R. M. Eustice, A. Kim, B. Englot, H. Johannsson, M. Kaess, and J. J. Leonard, "Advanced perception, navigation and planning for autonomous in-water ship hull inspection," *Int. J. Robot. Res.*, vol. 31, no. 12, pp. 1445–1464, 2012. 1

[4] O. Hoegh-Guldberg and J. F. Bruno, "The impact of climate change on the world's marine ecosystems," *Science*, vol. 328, no. 5985, 2010. 1

[5] M. Bryson, M. Johnson-Roberson, O. Pizarro, and S. B. Williams, "True color correction of autonomous underwater vehicle imagery," *J. Field Robot.*, vol. 33, no. 6, pp. 853–874, 2016. 1, 2, 3

[6] "The world's underwater cultural heritage," http://www.unesco.org/new/en/culture/themes/underwater-cultural-heritage/underwater-cultural-heritage/, Accessed 02/20/2020 2020. 1

[7] Y. R. Petillot, G. Antonelli, G. Casalino, and F. Ferreira, "Underwater robots: From remotely operated vehicles to intervention-autonomous underwater vehicles," *IEEE Robotics & Automation Magazine*, vol. 26, no. 2, pp. 94–101, 2019. 1

[8] M. Roznere and A. Quattrini Li, "Underwater monocular image depth estimation using single-beam echosounder," in *IROS*, 2020. 1

[9] K. Richmond, C. Flesher, L. Lindzey, N. Tanner, and W. C. Stone, "SUNFISH®: A human-portable exploration AUV for complex 3D environments," in *MTS/IEEE OCEANS Charleston*, 2018, pp. 1–9. 1

[10] S. Rahman, A. Quattrini Li, and I. Rekleitis, "SVIn2: An Underwater SLAM System using Sonar, Visual, Inertial, and Depth Sensor," in *IROS*, 2019, pp. 1861–1868. 1

[11] H. Cho, B. Kim, and S.-C. Yu, "Auv-based underwater 3-d point cloud generation using acoustic lens-based multibeam sonar," *IEEE Journal of Oceanic Engineering*, vol. 43, no. 4, pp. 856–872, 2018. 1

[12] A. Quattrini Li, A. Coskun, S. M. Doherty, S. Ghasemlou, A. S. Jagtap, M. Modasshir, S. Rahman, A. Singh, M. Xanthidis, J. M. O'Kane, and I. Rekleitis, "Experimental comparison of open source vision based state estimation algorithms," in *ISER*, 2016. 1

[13] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, 2015. 1, 3, 6

[14] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, 2015. 1

[15] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, 2018. 1

[16] S. Hong, D. Chung, J. Kim, Y. Kim, A. Kim, and H. K. Yoon, "In-water visual ship hull inspection using a hover-capable underwater vehicle with stereo vision," *J. Field Robot.*, vol. 36, no. 3, pp. 531–546, 2019. 1

[17] O. Drbohlav and M. Chaniler, "Can two specular pixels calibrate photometric stereo?" in *ICCV*, vol. 2, 2005, pp. 1850–1857. 1

[18] M. Daum and G. Dudek, "On 3-d surface reconstruction using shape from shadows," in *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231)*, 1998, pp. 461–468. 1

[19] N. Weidner, S. Rahman, A. Q. Li, and I. Rekleitis, "Underwater cave mapping using stereo vision," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 5709–5715. 1

[20] C. Hernandez, G. Vogiatzis, and R. Cipolla, "Multiview photometric stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 548–554, 2008. 1

[21] D. Vlasic, P. Peers, I. Baran, P. Debevec, J. Popović, S. Rusinkiewicz, and W. Matusik, "Dynamic shape capture using multi-view photometric stereo," in *ACM SIGGRAPH Asia*, 2009, pp. 1–11. 1

[22] G. Oxholm and K. Nishino, "Multiview shape and reflectance from natural illumination," in *CVPR*, 2014, pp. 2155–2162. 1

[23] F. Langguth, K. Sunkavalli, S. Hadap, and M. Goesele, "Shading-aware multi-view stereo," in *ECCV*, 2016, pp. 469–485. 2

[24] J. Park, S. N. Sinha, Y. Matsushita, Y.-W. Tai, and I. S. Kweon, "Robust multiview photometric stereo using planar mesh parameterization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1591–1604, 2016. 2

[25] F. Logothetis, R. Mecca, and R. Cipolla, "A differential volumetric approach to multi-view photometric stereo," in *ICCV*, 2019, pp. 1052–1061. 2

[26] B. Kaya, S. Kumar, C. Oliveira, V. Ferrari, and L. Van Gool, "Uncertainty-aware deep multi-view photometric stereo," in *CVPR*, 2022, pp. 12 601–12 611. 2

[27] G. Nam, J. H. Lee, D. Gutierrez, and M. H. Kim, "Practical SVBRDF acquisition of 3d objects with unstructured flash photography," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–12, 2018. 2

[28] K. Zhang, F. Luan, Z. Li, and N. Snavely, "IRON: Inverse rendering by optimizing neural sdfs and materials from photometric images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5565–5574. 2

[29] Z. Murez, T. Treibitz, R. Ramamoorthi, and D. Kriegman, "Photometric stereo in a scattering medium," in *ICCV*, 2015. 2

[30] C. Tsiotsios, T. Kim, A. Davison, and S. Narasimhan, "Model effectiveness prediction and system adaptation for photometric stereo in murky water," *Computer Vision and Image Understanding*, vol. 150, pp. 126–138, 2016. 2, 4

[31] C. Tsiotsios, A. J. Davison, and T.-K. Kim, "Near-lighting photometric stereo for unknown scene distance and medium attenuation," *Image and Vision Computing*, vol. 57, pp. 44–57, 2017. 2, 3, 4

[32] M. Roznere and A. Quattrini Li, "Real-time model-based image color correction for underwater robots," in *IROS*, 2019. 2, 3

[33] D. Akkaynak and T. Treibitz, "Sea-thru: A method for removing water from underwater images," in *Proc. CVPR*, 2019. 2, 5

[34] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE Transactions on Image Processing*, vol. 29, pp. 4376–4389, 2020. 2

[35] D. Akkaynak and T. Treibitz, "A revised underwater image formation model," in *Proc. CVPR*, 2018, pp. 6723–6732. 2

[36] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *ECCV*, 2014. 3

[37] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2018. 3

[38] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO : Fast semi-direct monocular visual odometry," in *ICRA*. IEEE, 2014. 3

[39] B. Joshi, S. Rahman, M. Kalaitzakis, B. Cain, J. Johnson, M. Xanthidis, N. Karapetyan, A. Hernandez, A. Quattrini Li, N. Vitzilaios, and I. Rekleitis, "Experimental comparison of open source visual-inertial-based state estimation algorithms in the underwater domain," in *IROS*. IEEE, 2019. 3

[40] "Bluerobotics low-light hd usb camera," https://www.bluerobotics.com/store/sensors-sonars-cameras/cameras/cam-usb-low-light-r1/, Accessed 09/14/2022 2022. 4

[41] "Bluerobotics lumen subsea light," https://bluerobotics.com/store/thrusters/lights/lumen-sets-r2-rp/, Accessed 09/14/2022 2022. 4

[42] "Bluerobotics ping sonar altimeter and echosounder," https://bluerobotics.com/store/sensors-sonars-cameras/sonar/ping-sonar-r2-rp/, Accessed 09/14/2022 2022. 4

[43] "Intel realsense depth camera d455," https://www.intelrealsense.com/depth-camera-d455/, Accessed 09/14/2022 2022. 4

[44] J. A. Nelder and R. Mead, "A simplex method for function minimization." *The Computer Journal*, vol. 7, pp. 308–313, 1965. 5

[45] "Scipy optimize minimize documentation," https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html, Accessed 09/14/2022 2022. 5

[46] Y. Quéau, B. Durix, T. Wu, D. Cremers, F. Lauze, and J.-D. Durou, "Led-based photometric stereo: Modeling, calibration and numerical solution," *Journal of Mathematical Imaging and Vision*, vol. 60, pp. 313–340, 2018. 5