2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

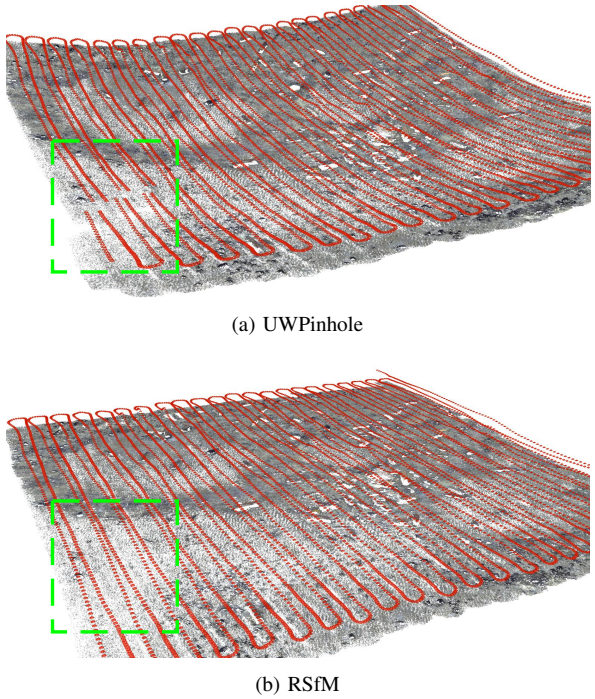
Refractive COLMAP: Refractive Structure-from-Motion Revisited

Mengkun She and Felix Seegräber and David Nakath and Kevin Köser

***Abstract*— In this paper, we present a complete refractive Structure-from-Motion (RSfM) framework for underwater 3D reconstruction using refractive camera setups (for both, flat- and dome-port underwater housings). Despite notable achievements in refractive multi-view geometry over the past decade, a robust, complete and publicly available solution for such tasks is not available at present, and often practical applications have to resort to approximating refraction effects by the intrinsic (distortion) parameters of a pinhole camera model. To fill this gap, we have integrated refraction considerations throughout the entire SfM process within the state-of-the-art, open-source SfM framework COLMAP. Numerical simulations and reconstruction results on synthetically generated but photo-realistic images with ground truth validate that enabling refraction does not compromise accuracy or robustness as compared to in-air reconstructions. Finally, we demonstrate the capability of our approach for large-scale refractive scenarios using a dataset consisting of nearly 6000 images. The implementation is released as open-source at: https://cau-git.rz.uni-kiel.de/inf-ag-koeser/colmap_underwater.**

I. INTRODUCTION

Simultaneous Localization and Mapping (SLAM) as well as Structure-from-Motion (SfM) are key technologies for inferring maps or 3D shapes from images. Their application in the underwater domain enables exploration of geological or archaeological sites on the seafloor, mapping or monitoring offshore installations, deposited munitions, or biological habitats, and visually aided autonomous underwater navigation in general. To protect cameras from water and high pressure in the ocean, they are enclosed in waterproof pressure housings and observe the environment through a transparent window, typically with a planar or spherical shape. Light rays from the underwater scene change direction when they travel through these interfaces in a non-orthogonal manner, leading to distortion in the acquired images. Although refraction is depth-dependant, in the past refraction effects have often been addressed by approximating the entire camera system, including the glass port, as a perspective camera [1]. This enables the use of standard 3D reconstruction software such as COLMAP [2] and Agisoft Metashape. Throughout this work, we refer to this approach as **UWPinhole**. However, this approximation is suitable only for certain refractive camera configurations and pre-defined working distances [3], and absorption of distance-dependent refraction into pinhole intrinsics can introduce bias and inconsistencies for large-scale reconstructions (see e.g. Fig. 1). As an alternative, underwater camera systems can be explicitly modeled with additional physical parameters describing the properties of the housing interface [4], [5], [6], [7]. While exact refraction modeling closely resembles physical effects, it invalidates classical pinhole-based multi-view geometry methods. Integrating these additional physical parameters into SfM therefore remains challenging.



(a) UWPinhole

(b) RSfM

Fig. 1. Results of the reconstruction on a rendered large-scale AUV-based seafloor mapping dataset containing 5740 refractive flat-port images. **Top**: Using the perspective camera model underwater creates a curved seafloor reconstruction. **Bottom**: Our proposed RSfM.

Over the past decade, several solutions have been proposed to address various aspects of the RSfM problem, such as refractive calibration [8], [9], [10], refractive motion estimation [11], [12], [13], [14], or even partial RSfM system demonstrators [15], [16], [17], however, limited to flat-ports, lacking open-source implementations, or demonstrated only on a small set of images. COLMAP [2] is widely recognized as a state-of-the-art open-source incremental SfM framework, upon which many downstream tasks like dense Multi-View Stereo [18] or NeRF [19] depend. Due to the lack of a suitable underwater alternative, the UWPinhole approximation often remains the commonly applied method in practice [3], and it is even sometimes considered as a reference ground truth in the literature [17], [20]. Hence, there remains a need for a complete, open-source, general

This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) Projektnummer 396311425, through the Emmy Noether Program

The authors are with the Department of Computer Science, Christian-Albrechts-University of Kiel, Neufeldtstraße 6, 24118 Kiel, Germany {mshe,fse,dna,kk}@informatik.uni-kiel.de

refractive Structure-from-Motion solution that is proven to be robust, accurate, and capable of handling a large number of images.

In this work, we make the following contributions:

- Integration of refraction into COLMAP, supporting generalized refractive camera setups with auto-optimizing the refractive parameters in the reconstruction.
- A robust relative pose estimation approach for geometric verification and SfM initialization.
- Extensive evaluations on the overall performance of the RSfM pipeline under various refractive camera setups.

II. Related Work

Refractive Camera Modeling. Grossberg et al. [21] and Schöps et al. [22] utilize a generic ray-based camera model, which directly associates rays from the scene with the image coordinates. In principle, such models could also be used to encode refraction. In a more specific model for flat glass windows, Treibitz et al. [4] explicitly represent the flat-port interface with a plane and analyze the behavior of rays. Agrawal et al. [23] extend this model to a more general case involving tilted multi-layers interface and demonstrate that the system is an axial camera model. Jordt et al. [9] propose a more comprehensive calibration approach for such systems. Telem et al. [5] propose a varifocal model in which a feature-dependent focal length correction factor is applied to maintain the co-linearity of the ray. On the other hand, robots for deeper waters are often equipped with spherical glass windows (dome-ports), because they are mechanically much more stable for high water pressures and allow a large field of view. Additionally, refraction can be avoided if a pinhole camera is perfectly centered within the dome [24], [10]. However, in practice, de-centering the camera results in behavior akin to an axial camera model [7], similar in spirit to flat-port refraction, but at the sphere. Nevertheless, 3D reconstruction using non-central camera models requires additional effort. For special cases, a straightforward approach to avoid addressing this is to undistort refraction before reconstruction. This can be achieved by constructing a look-up table using the Pinax model [25] to map refracted image points back to un-refracted positions. However, this technique requires a small camera-to-interface distance in the order of millimeters, assumes a fixed scene distance. Moreover, the fixed look-up table does not allow refining the refractive calibration during bundle adjustment.

Refractive SfM. Several works exist on SfM using general camera models, such as those by Sturm et al. [26], [27], [28]. However, it has been discussed that this model is particularly sensitive to noise. Chari et al. [29] provide theoretical insights into multi-view geometry under planar refraction, although without numerical evaluations. Jordt-Sedlazeck et al. [15] is considered as the first approach that tackles the entire SfM problem for underwater imaging. Nevertheless, it is only demonstrated on a small-scale scene. Elnashef et al. [17] derive a differential motion model for an axial formulation of the continuous egomotion and propose a visual odometry pipeline.

Focusing on the motion estimation, Agrawal et al. [23] propose an 8-point algorithm to solve for the refractive interface and camera pose using the plane of refraction (POR) constraint. Kang et al. [30] present a two-view reconstruction approach for cameras under thin planar interfaces. Jordt-Sedlazeck et al. [15] introduce an alternating, iterative-based method for both absolute and relative pose estimation. Chadebecq et al. [16], [11] derive a refractive fundamental constraint for iterative refinement, mainly targeting thin flat-ports. However, in SfM, correspondences are often contaminated by outliers, necessitating robust estimation techniques such as RANSAC [31]. The aforementioned methods are not minimal solvers, but require a good initialization and are computationally slow when using RANSAC. Elnashef et al. [32] propose a linear approach to the absolute pose problem under flat refraction using the varifocal model. They later address relative pose estimation with a 3-point algorithm, highlighting the possibility to estimate the true scene scale [12]. However, determining the relative rotation requires performing a non-linear optimization to minimize the epipolar curve distances. While much attention has been given to flat-port cameras, little work has focused on decentered dome-ports. Hu et al. [14] employ a virtual perspective camera similar to the varifocal model, proposing pose refinement methods applicable to generalized refractive camera setups. They additionally propose a minimal solution to the relative pose estimation problem, requiring 17 point correspondences. However, we show in our evaluation that the algorithm can only be applied underwater in very low noise conditions. In more generalized cases, a minimal 3-point solution to absolute pose estimation is presented in [33], and Kneip et al. [34] propose an 8-point algorithm for solving the relative pose problem. Both approaches are designed for multi-camera systems in self-driving car scenarios, assuming relatively large baselines between the cameras and various directions they face. However, in the underwater-refraction induced axial camera model, the baselines between multiple virtual projection centers are small, typically in the millimeter range, which is a scenario where these approaches have not yet been tested.

Maybe surprisingly, we show that the former algorithm achieves comparable performance against the baseline, which uses standard Perspective-n-Point (PnP) algorithms [35], [36] on un-refracted data, whereas the latter approach is found to be inapplicable. Hence, we choose the 3-point algorithm for absolute pose estimation in our RSfM pipeline. Regarding the relative pose estimation problem, we have not found a satisfactory approach so far. Therefore, we propose a more practical approach that is more robust for geometric verification and SfM initialization., which will be elaborated in the next section.

III. Refractive Structure-from-Motion

Refractive Camera Models. To integrate refraction into the SfM process, we first make the following consideration: The refractive camera model should be generalizable to both thin/thick flat-port and dome-port, and extendable for

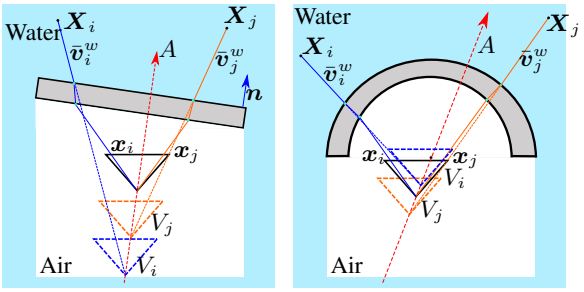


Fig. 2. A schematic illustration of the refractive camera models. The scene points $\boldsymbol{X}$ are observed by the camera at image points $\boldsymbol{x}$ through the interface. The virtual cameras V are depicted by differently colored dashed triangles situated along the refraction axis A. **Left**: Flat-port. **Right**: Dome-port.

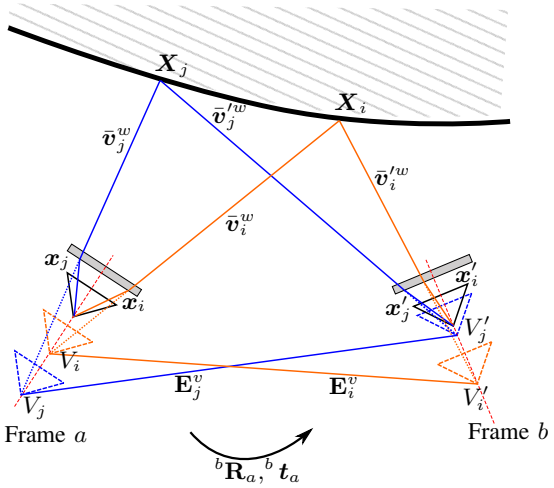


Fig. 3. A schematic illustration of the feature-dependent virtual epipolar geometry $\mathbf{E}^v$, which relates the relative pose ${}^b\mathbf{R}_a, {}^b\boldsymbol{t}_a$ of two frames.

potentially more scenarios; Additionally, the real physical camera, which is situated behind the refractive interface, should be interchangeable. A schematic illustration of the refractive imaging setup is depicted in Fig. 2.

The real camera which is described by its intrinsic parameters $\mathcal{P}_{\text{cam}}$, observes the scene points $\boldsymbol{X}$ from an image point $\boldsymbol{x}$ through a glass interface. The flat-port interface is defined by parameters including the unit normal vector of the interface $\boldsymbol{n}_{\text{int}} = (n_x, n_y, n_z)^\mathsf{T}$, the camera-to-interface distance d_{int}, and the thickness. These interface parameters are defined locally relative to the camera, with $\boldsymbol{n}_{\text{int}} = (0, 0, 1)^\mathsf{T}$ coinciding with the optical axis of the camera. The dome-port is characterized by its dome center (or decentering) $\boldsymbol{C}_d = (C_x, C_y, C_z)^\mathsf{T}$ in the local camera coordinate frame, along with the radius and thickness. The refraction axis A describes the camera ray passing through the interface perpendicularly. In the case of a flat-port, the refraction axis aligns with the interface normal $\boldsymbol{n}_{\text{int}}$, while in the case of a dome-port, it aligns with the normalized decentering direction.

According to Snell's law, the refracted normalized ray vector $\bar{\boldsymbol{v}}_{\text{refrac}}$ can be computed by:

$$\bar{\boldsymbol{v}}_{\text{refrac}} = r \cdot \bar{\boldsymbol{v}} - (rc - \sqrt{1 - r^2(1 - c^2)}) \cdot \boldsymbol{n} \qquad (1)$$

where $\bar{\boldsymbol{v}}$ is the normalized incident ray, $c = \boldsymbol{n} \cdot \bar{\boldsymbol{v}}$ and $r = n_1/n_2$ which represents the ratio of the two involved media's refraction indices.

In our convention, the normal vector $\boldsymbol{n}$ points from the surface towards the side where the ray is refracted. We then utilize the ray-tracing technique to obtain the refracted ray in water $\boldsymbol{v}^w$ starting from the outer interface. The implementation of the ray-plane/sphere intersection can be found in [37].

Virtual Camera Computation. Afterwards, we replace individual rays of the refractive camera with virtual pinhole cameras (depicted as dashed triangles in Fig. 2). These virtual cameras are feature-dependent and are positioned at the intersection of the refraction axis A with $\bar{\boldsymbol{v}}^w$, observing the same ray in water $\boldsymbol{v}^w$ perspectively. The pose of the virtual camera is described by a rigid transformation from the real to the virtual coordinate frame[1] ${}^v\mathbf{T}_r = ({}^v\mathbf{R}_r \mid {}^v\boldsymbol{t}_r)$. This technique is initially introduced in [5] for calibration and widely employed in [15], [32], [38].

However, unlike previous works where they align the virtual cameras with the refraction axis A and utilize the camera-to-interface distance d as the virtual focal length [15], we have discovered that this approach introduces instability in the forward projection of a 3D point onto the image plane refractively when the axis A is significantly off from the camera's optical axis (e.g. points are located behind the virtual camera). This situation can occur if the flat-port interface is severely tilted (which is unrealistic in the underwater imaging scenario), or if the decentering direction is oriented sideways in the case of a dome-port (which occurs more frequently). Furthermore, drastic changes in the virtual focal length can potentially lead to variations in the magnitude of the reprojection error, thereby introducing imbalance in the bundle adjustment process. We therefore suggest to keep the rotation of the virtual camera as identity ${}^v\mathbf{R}_r = \mathbf{I}$, and then take the mean focal length of the real camera as the virtual focal length $f_v = f_{\text{mean}}$. In addition, we determine the virtual principal points (c_{vx}, c_{vy}) in a way such that the original observed image point $\boldsymbol{x} = (x, y)^\mathsf{T}$ remains the same:

$$c_{vx} = x - f_v \cdot \bar{\mathbf{v}}^w_{\text{hnorm}}(x), \quad c_{vy} = y - f_v \cdot \bar{\mathbf{v}}^w_{\text{hnorm}}(y) \qquad (2)$$

Here, $\bar{\mathbf{v}}^w_{\text{hnorm}}(x)$ and $\bar{\mathbf{v}}^w_{\text{hnorm}}(y)$ represent the x-, and y-component of the homogeneous-normalized ray in water $\bar{\boldsymbol{v}}^w$. Finally, the virtual camera center ${}^r\boldsymbol{t}_v$ can be found by intersecting $\bar{\boldsymbol{v}}^w$ with A.

Absolute Pose Estimation. For absolute pose estimation, we utilize the generalized absolute pose estimator (referred to as GP3P), which is readily available in COLMAP [33]. We construct a set of virtual cameras $\mathcal{V}_{\text{cam}}$ from a set of image points and treat them as a rigidly mounted multi-camera

[1]Throughout this work, a rigid transformation ${}^b\mathbf{T}_a$ transforms a point in the a coordinate frame to the b coordinate frame.

rig. Estimating the absolute pose of the rig is equivalent to estimating the pose of the real camera. The algorithm requires minimally 3 point correspondences.

Relative Pose Estimation. The literature review has highlighted the inherent difficulty of relative pose estimation. In response, we propose a simplification strategy that involves a slight trade-off in accuracy. Rather than directly estimating the refractive relative pose, we opt to estimate the relative pose of the best-approximated perspective camera using the well-established 5-point algorithm [39]. To compute the best-approximated perspective camera model, we randomly sample 1000 image points and back-project them to 3D space at a distance of $5m$ using the original refractive camera model. The parameters are determined by minimizing the reprojection error of the 3D-2D points, but with the perspective camera model:

$$\mathcal{P}_{\text{prox}} = \underset{\mathcal{P}_{\text{prox}}}{\arg\min} \sum_i \|\pi(\mathcal{P}_{\text{prox}}, \boldsymbol{X}_i) - \boldsymbol{x}_i\|_2^2 \tag{3}$$

where $\mathcal{P}_{\text{prox}}$ denotes the parameters of the best-approximated camera, and $\pi(\cdot)$ represents the forward projection function. Such an approximation will never yield a perfect solution even under noise-free, outlier-free conditions, except in the case of a perfectly centered dome-port scenario. However, experimental results demonstrate that it generally performs adequately, with only a marginal loss of inlier correspondences (less than 2% in the worst case).

When initializing SfM from the first image pair, we additionally refine the estimated relative pose by minimizing the refractive virtual epipolar cost, similar to the approach proposed in [14]. As depicted in Fig. 3, the refracted rays $\bar{\boldsymbol{v}}^w$ and $\bar{\boldsymbol{v}}'^w$, along with the vector connecting the two virtual camera centers, form an epipolar plane. Suppose the relative pose between the image pair is expressed as ${}^b\mathbf{T}_a = ({}^b\mathbf{R}_a \mid {}^b\boldsymbol{t}_a)$, and a feature point $\boldsymbol{x}_i$ in image a is matched to the feature point $\boldsymbol{x}'_i$ in image b. The transformations from the real camera to the virtual one at $\boldsymbol{x}_i$ and $\boldsymbol{x}'_i$ are ${}^{v_i}\mathbf{T}_r$ and ${}^{v'_i}\mathbf{T}_r$, respectively. Next, the transformation from the virtual camera to its corresponding virtual camera in frame b is concatenated as:

$${}^{v'_i}\mathbf{T}_{v_i} = ({}^{v'_i}\mathbf{R}_{v_i} \mid {}^{v'_i}\boldsymbol{t}_{v_i}) = {}^{v'_i}\mathbf{T}_r \cdot {}^b\mathbf{T}_a \cdot ({}^{v_i}\mathbf{T}_r)^{-1} \tag{4}$$

Then, for each feature point pair $\boldsymbol{x}_i$ and $\boldsymbol{x}'_i$, we have an epipolar constraint :

$$\hat{\mathbf{x}}_i'^{\mathsf{T}} \mathbf{E}_i^v \hat{\mathbf{x}}_i = 0 \;\text{ where }\; \mathbf{E}_i^v = [{}^{v'_i}\boldsymbol{t}_{v_i}]_\times \cdot {}^{v'_i}\mathbf{R}_{v_i} \tag{5}$$

Here, $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{x}}'_i$ are the normalized coordinates. Finally, the optimal form of ${}^b\mathbf{R}_a$ and ${}^b\boldsymbol{t}_a$ can be obtained by minimizing the virtual epipolar cost:

$${}^b\mathbf{R}_a, {}^b\boldsymbol{t}_a = \underset{{}^b\mathbf{R}_a, {}^b\boldsymbol{t}_a}{\arg\min} \sum_i \|\hat{\mathbf{x}}_i'^{\mathsf{T}} \mathbf{E}_i^v \hat{\mathbf{x}}_i\|^2 \tag{6}$$

An interesting aspect of refractive relative pose estimation is the potential to estimate the baseline length. However, the accuracy and reliability of scale estimation are not guaranteed across all refractive camera configurations, as reported by [15], [14], [12]. Therefore, we allow the optimizer to refine the full 6-DoFs relative pose estimation, but we do not attempt to recover the true scale. This decision is based on the observation that if the scale is observable (which occurs only in extreme refraction setups and very low noise conditions), it would indicate an accurate estimation of the baseline length and, consequently, a true scaled reconstruction. On the other hand, if the scale is not observable, it implies that there is no detriment to the final reconstruction, even if the scale is completely incorrect.

Triangulation. We keep the triangulation algorithm unchanged from its implementation in COLMAP, only modifying it to triangulate rays generated from their respective virtual perspective cameras. Therefore, such modification does not introduce any side effects on performance.

Bundle Adjustment. In classical bundle adjustment, 3D points are projected onto the image planes, and reprojection errors are minimized. However, in the refractive scenario, forward projection is computationally expensive. It involves either solving a 12^{th}-degree polynomial or iteratively back-projecting the currently estimated projection until the error in 3D is minimized [23], [24]. We therefore minimize the reprojection errors on the virtual image planes, similar to [15]. The refractive cost function is as follows:

$$E = \sum_j \rho_j(\|\pi_v(\mathcal{P}_{\text{cam}}, \mathcal{P}_{\text{refrac}}, {}^c\mathbf{T}_w, \boldsymbol{X}_k, \boldsymbol{x}_j) - \boldsymbol{x}_j\|_2^2) \tag{7}$$

where $\pi_v(\cdot)$ is a function that projects a 3D point $\boldsymbol{X}_j$ to the virtual image plane. This function first determines the virtual camera corresponding to the feature point $\boldsymbol{x}_j$, and then projects the 3D point $\boldsymbol{X}_j$ onto the virtual image plane perspectively. A loss function ρ_j is used to potentially down-weight outliers. Both the camera intrinsic parameters $\mathcal{P}_{\text{cam}}$ and the refractive parameters $\mathcal{P}_{\text{refrac}}$ can be jointly refined. However, for numerical stability reasons, we only refine the interface normal and camera-to-interface distance in the flat-port case and the decentering in the dome-port case.

IV. Evaluations

A. Numerical Evaluation

Before evaluating the proposed RSfM pipeline, we first analyze the performance of the absolute and relative pose estimation, as these two steps are very critical for SfM.

Absolute Pose Estimation. We construct a numerical setup where a camera observes a set of randomly generated 3D points from a random pose. We project these points onto the image plane both with and without refraction and introduce Gaussian-distributed noise. Additionally, a certain percentage of outliers are added to the data points. We then employ the GP3P method to estimate the camera pose within the RANSAC framework. As a baseline for comparison, we perform standard PnP pose estimation [35], [36] on the un-refracted data points. The simulated camera has an image size of 1920×1280 pixels and a field of view of $73°$. Each experiment consists of 200 points in total, with 30% of them being outliers. We conduct 1000 experiments, with the refractive parameters, 2D-3D points, and camera poses randomly generated within realistic bounds, for each

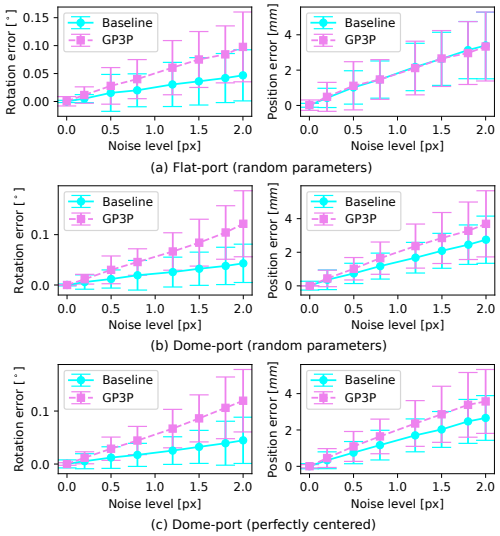


Fig. 4. Numerical evaluation results of the absolute pose estimation across various refractive camera configurations.

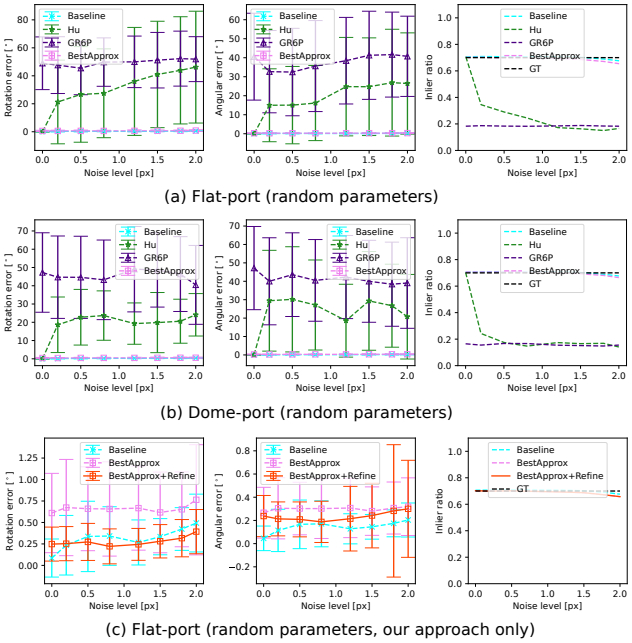
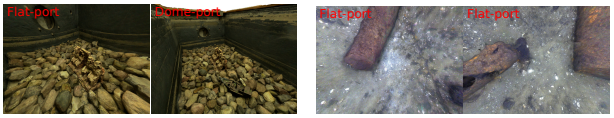


Fig. 5. Numerical evaluation results of the relative pose estimation under different noise and outlier conditions.

experiment, and measure the rotation error in degrees and position error in mm. Fig. 4 (a) and (b) show the flat- and dome-port setups. We also questioned ourselves whether the approach would become numerically unstable or degenerate in a scenario with minimal refractive effects, i.e. when the camera system is near central. To address this we conduct a third evaluation using a perfectly centered dome-port camera, shown in Fig. 4 (c). The GP3P pose estimator consistently demonstrates performance comparable to the baseline approach across various evaluation configurations. In nearly all cases, the GP3P exhibits only marginal differences from



(a) Tank Scene (b) AUV Scene

Fig. 6. Example images of the re-rendered refractive datasets.

the baseline, with maximum deviations of less than $0.2°$ in rotation error and $4mm$ in position error. Furthermore, the correct outlier ratio is reported by RANSAC in all cases. Note that there is no non-linear refinement involved in this evaluation. While the GP3P estimator was originally developed for multi-camera systems in self-driving car scenarios, we investigate its stability and performance when applied to our refractive camera setup. The results depicted in Fig. 4 (c) demonstrate that the approach remains robust and capable of handling such scenarios effectively. Based on these findings, we conclude that the GP3P estimator is sufficient for our application.

Relative Pose Estimation. We conduct the same experiments as the previous ones to evaluate relative pose estimation, except that random 2D-2D correspondences are generated. Since an accurate estimation of the baseline length cannot be guaranteed, we measure rotation error in degrees and the angular error between the estimated and ground truth relative translation direction, also in degrees, and the inlier ratio reported by RANSAC. We evaluate several minimal solvers: Hu's 17-point algorithm [14], Kneip's 8-point generalized relative pose solver (GR6P), and our proposed method (denoted as BestApprox). The baseline method for comparison is the 5-point algorithm [39] using un-refracted data points. Fig. 5 (a) and (b) present the evaluation results for the flat-port and dome-port setups, respectively, without involving any non-linear refinement. Fig. 5 (c) exclusively displays the evaluation outcomes of our approach alongside the non-linear refined results compared to the baseline in the flat-port setups.

As depicted in Fig. 5 (a) and (b), Hu's approach nearly recovers ground truth results under low noise conditions, but does not deliver satisfactory results when the noise level is increased, and the inlier ratio drops rapidly, similar to their report [14]. The GR6P algorithm performs poorly already under low noise conditions, meaning that the approach is inapplicable to the underwater-refraction induced axial camera model. Nevertheless, our approach performs stably and robustly under various conditions, with accuracy only marginally worse than the baseline method, as evident from Fig. 5 (c) where the other approaches are excluded. The maximum loss of inliers is only less than 2% in the worst case as compared to the baseline. Furthermore, refining the initial estimated relative pose by minimizing the virtual epipolar cost further improves the accuracy, ensuring an accurate and robust RSfM initialization.

TABLE I

EVALUATION RESULTS ON RE-RENDERED DATASETS. THE OPTIMAL RESULTS ARE HIGHLIGHTED IN **BOLD** TEXT.

	Datasets	N	UWPinhole RE	UWPinhole $\Delta\mathbf{R}$	UWPinhole $\Delta\boldsymbol{t}$	UWPinhole Δd	RSfM (Use GT Calib) RE	RSfM (Use GT Calib) $\Delta\mathbf{R}$	RSfM (Use GT Calib) $\Delta\boldsymbol{t}$	RSfM (Use GT Calib) Δd	RSfM (Refine) RE	RSfM (Refine) $\Delta\mathbf{R}$	RSfM (Refine) $\Delta\boldsymbol{t}$	RSfM (Refine) Δd
Tank	*Flat+Ortho+Close*	106	0.522	0.143	1.978	2.774	**0.347**	**0.016**	**0.230**	**2.098**	0.348	0.031	1.169	2.254
	Flat+Tilt+Close	106	1.147	6.442	19.432	19.060	**0.330**	**0.013**	**0.252**	**1.657**	0.332	0.027	1.077	1.844
	Flat+Ortho+Far	106	0.773	0.957	17.324	7.005	**0.335**	**0.021**	**0.236**	1.964	**0.335**	0.022	0.307	**1.939**
	Flat+Tilt+Far	106	1.180	5.774	23.276	16.052	**0.340**	**0.012**	**0.218**	**1.665**	0.344	0.031	1.212	1.918
	Dome+Backward+Close	106	0.313	0.017	0.516	1.915	**0.312**	**0.011**	0.577	1.852	**0.312**	0.027	**0.462**	**1.797**
	Dome+Backward+Far	106	0.390	0.181	6.193	3.854	**0.306**	**0.006**	**0.083**	2.225	**0.306**	0.013	0.440	**2.195**
	Dome+Sideward+Close	106	0.312	0.735	0.510	1.923	**0.306**	0.040	0.551	1.880	**0.306**	**0.026**	**0.227**	**1.863**
	Dome+Sideward+Far	106	1.124	5.247	6.246	7.595	**0.309**	**0.009**	0.092	**2.081**	**0.309**	0.012	**0.088**	2.088
AUV	*Flat+Ortho*	5740	0.277	10.272	893.730	816.131	**0.199**	**0.200**	**27.933**	**11.842**	**0.199**	0.217	29.441	13.778
	Flat+Tilt	5740	0.825	-	-	-	**0.196**	**0.238**	**29.182**	16.247	0.197	0.256	29.337	**16.020**

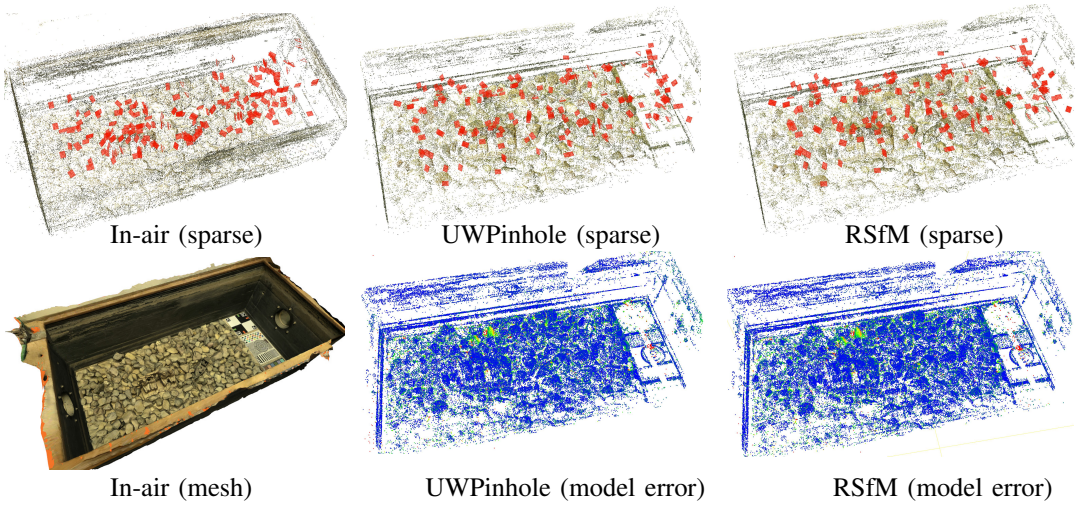


Fig. 7. Reconstruction results of the real water tank dataset.

TABLE II

RESULTS OF THE ESTIMATED REFRACTIVE PARAMETERS WHEN PERFORMING REFINEMENT IN RSFM.

	Datasets	GT $(n_x, n_y, n_z, d_{\text{int}})$				Est. $(n_x, n_y, n_z, d_{\text{int}})$			
Tank	*Flat+Ortho+Close*	0	0	1	0.01	8.80e-05	-3.51e-05	1.000	0.035
	Flat+Tilt+Close	0.166	0.148	0.975	0.01	0.166	0.148	0.975	0.033
	Flat+Ortho+Far	0	0	1	0.05	9.81e-05	-3.30e-05	1.000	0.053
	Flat+Tilt+Far	0.166	0.148	0.975	0.05	0.166	0.148	0.975	0.190
AUV	*Flat+Ortho*	0	0	1	0.02	-3.71e-05	8.11e-05	1.000	0.0201
	Flat+Tilt	0.166	0.148	0.975	0.02	0.166	0.148	0.975	0.0201
	Datasets	GT (C_x, C_y, C_z)				Est. (C_x, C_y, C_z)			
Tank	*Dome+Backward+Close*	0	0	0.003		5.14e-06	-8.14e-05	0.003	
	Dome+Backward+Far	0	0	0.03		6.66e-06	-3.13e-05	0.030	
	Dome+Sideward+Close	0.003	0	0		0.003	-6.05e-05	-8.47e-05	
	Dome+Sideward+Far	0.03	0	0		0.030	-4.07e-05	-4.96e-05	

B. Re-Render from Real-World

To benchmark our proposed RSfM approach, we render novel refractive images from 3D meshes reconstructed out of an existing refraction-free real-world dataset using a physically-based ray-tracer. We maintain identical camera poses during re-rendering to ensure a faithful emulation of the original photographic missions. The refractive effects are simulated during rendering by digitally placing a glass-material interface in front of the camera.

To evaluate the robustness of the system, we render the same scene with various refractive camera setups. These setups include orthogonal and tilted flat-port interfaces, as well as variations in the camera-to-interface distance and dome-port decentering. The parameters we use for re-rendering are shown in Tab. II. The first scene for re-rendering is a well-decorated test tank without water, reconstructed by COLMAP using a GoPro9 camera in-air under homogeneous illumination. The second scene contains a large-scale 3D reconstruction of seafloor (scene size $44m \times 35m$) obtained using the method described in [40]. The original images of this dataset were acquired by an AUV equipped with a calibrated dome-port camera system in a real-world mapping mission. This is to demonstrate the applicability of our approach on real-world AUV-based large-scale refractive seafloor reconstruction. Example images of the scenes are shown in Fig. 6.

We evaluate our system in three runs, always initializing the intrinsics with the ground truth in-air calibration. Specifically, we compare 1) UWPinhole which refines the intrinsics and distortion parameters; 2) RSfM using ground truth refractive parameters and keeping them constant; 3) RSfM using incorrect refractive parameters as initialization, and only refining them during bundle adjustment. For the AUV scene, we set the baseline of the initial registered two images as the baseline measured by the navigation data to constrain the true scene scale. Nevertheless, the navigation data is not used for RSfM reconstruction. The results are presented in Tab. I, where RE stands for the reprojection error in pixels. $\Delta\mathbf{R}$, $\Delta\mathbf{t}$ represent the rotation error in degrees, position errors in mm respectively. All error measures are averaged across all images. The 3D model error Δd is

measured as the average closest distance of the reconstructed sparse point cloud to the 3D mesh from which images are rendered in mm. In addition, Tab. II shows the estimated refractive parameters against the ground truth values when refining the parameters in bundle adjustment, where the optimizable parameters are highlighted in **Bold** text.

It is evident from Tab. I that our proposed RSfM approach consistently yields the best results in terms of the accuracy of both camera poses and the 3D model across various refractive camera configurations. However, it is also interesting to note that absorbing refraction in distortion parameters is not necessarily a bad practice in some scenarios. For instance, when the flat-port interface is orthogonal, the refraction effects are mostly symmetric, and radial distortion can effectively absorb these effects for reasonable distance ranges. Similarly, in the case of a slightly decentered dome-port system where the refraction effects are not pronounced, ignoring refraction can still yield decent results. However, when dealing with large datasets, such as in the AUV mapping scenario, the limitations of the UWPinhole approach become apparent. As shown in Fig. 1 (top), despite achieving a low reprojection error of only 0.277 pixels, the UWPinhole approach leads to a severely distorted reconstruction. This occurs even when the interface normal is orthogonal, and the approach fails to produce meaningful results when the normal is tilted. In contrast, the proposed RSfM approach can handle various dataset types and camera configurations effectively.

Scale Awareness. Tab. II demonstrates that refining refractive parameters in RSfM effectively recovers incorrectly initialized values except for the camera-to-interface distance d_{int}, which is only estimated up to scale. This observation also aligns with the findings of [41], because scaling the entire scene, including d_{int}, does not alter the angle of incident rays at the interface. Therefore, constraining the scene scale can lead to true d_{int} calibration as evident from the AUV scene results where external information such as navigation data is utilized to initialize the scene scale, resulting in the calibration of d_{int} to its true value.

C. Real-World Experiments

To obtain ground truth for evaluating the RSfM approach, we employ a GoPro9 camera to perform an in-air scan of the decorated tank without water using standard COLMAP. An AruCo checkerboard is positioned on the floor as a reference target for alignment based on similarity transforms, and the resulting in-air scanned model is considered the ground truth. The in-air reconstruction of the tank is shown in Fig. 7 (left). Subsequently, the tank is filled with water, and an underwater dataset is captured by the same GoPro camera with a flat-port case. The flat-port parameters are obtained through underwater calibration [9]. The acquired images have dimensions of 5184×3888 pixels. Similar to previous experiments, we reconstruct the model once using the UWPinhole approach and once with our RSfM approach. A view of the reconstructed results are presented in Fig. 7 (center and right). The reprojection errors of the reconstructions are 1.109 pixels for the in-air scan, 1.050 pixels using the UWPinhole approach, and 1.037 pixels with the RSfM approach. In addition, all reconstructions are aligned, and the model error is measured as the cloud-to-mesh distance using CloudCompare. The model error for the UWPinhole reconstruction is $2.061mm$, and for the RSfM approach, it is $2.103mm$. This experiment demonstrates that the RSfM approach can achieve ground truth-level reconstruction even without accurate measures of the refractive interface. However, in this specific setup where the flat-port case for the GoPro camera is only around $2mm$ thick and the camera-to-interface distance is even less than $2mm$, there is no clear advantage over simply ignoring refraction and reconstructing using standard COLMAP. Additionally, the relatively small scene size of about $2m \times 1m$ and small altitude variations may not fully exploit the advantages of RSfM. Nonetheless, Fig. 1 demonstrates the necessity of the RSfM approach when mapping a large area of the seafloor using a refractive camera. Therefore, we present this approach to the community for situations where considering refraction is necessary.

V. CONCLUSIONS

We have introduced a comprehensive refractive Structure-from-Motion (RSfM) pipeline for underwater 3D reconstruction, which has been integrated into the widely used open-source SfM framework COLMAP. Our proposed components enable robust and accurate geometric verification and SfM initialization. Through comprehensive evaluations, we have demonstrated the accuracy and robustness of each individual component as well as the overall system performance. Our implementation is publicly available as an underwater extension of COLMAP.

REFERENCES

[1] M. Shortis, "Calibration techniques for accurate measurements by underwater camera systems," *Sensors*, vol. 15, no. 12, pp. 30 810–30 826, 2015. [Online]. Available: http://www.mdpi.com/1424-8220/15/12/29831

[2] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4104–4113.

[3] R. Rofallski, O. Kahmen, and T. Luhmann, "Investigating distance-dependent distortion in multimedia photogrammetry for flat refractive interfaces," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 48, pp. 127–134, 2022.

[4] T. Treibitz, Y. Y. Schechner, and H. Singh, "Flat refractive geometry," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition CVPR 2008*, 2008, pp. 1–8.

[5] G. Telem and S. Filin, "Photogrammetric modeling of underwater environments," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 5, pp. 433–444, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/B6VF4-50F9H66-1/2/d8dba566f79b0a207e13a6aa2bf3f69d

[6] A. Jordt, K. Köser, and R. Koch, "Refractive 3d reconstruction on underwater images," *Methods in Oceanography*, vol. 15-16, pp. 90 – 113, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2211122015300086

[7] M. She, D. Nakath, Y. Song, and K. Köser, "Refractive geometry for underwater domes," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 183, pp. 525–540, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092427162100304X

[8] A. Agrawal, Y. Taguchi, and S. Ramalingam, "Analytical forward projection for axial non-central dioptric and catadioptric cameras," in *European Conference on Computer Vision*. Springer, 2010, pp. 129–143.

[9] A. Jordt-Sedlazeck and R. Koch, "Refractive calibration of underwater cameras," in *Computer Vision - ECCV 2012*, ser. Lecture Notes in Computer Science, A. Fitzgibbon, S. Lazebnik, P. Pietro, Y. Sato, and C. Schmid, Eds. Springer Berlin Heidelberg, 2012, vol. 7576, pp. 846–859.

[10] M. She, Y. Song, J. Mohrmann, and K. Köser, "Adjustment and calibration of dome port camera systems for underwater vision," in *German Conference on Pattern Recognition*. Springer, 2019, pp. 79–92.

[11] F. Chadebecq, F. Vasconcelos, R. Lacher, E. Maneas, A. Desjardins, S. Ourselin, T. Vercauteren, and D. Stoyanov, "Refractive two-view reconstruction for underwater 3d vision," *International Journal of Computer Vision*, pp. 1–17, 2019.

[12] B. Elnashef and S. Filin, "A three-point solution with scale estimation ability for two-view flat-refractive underwater photogrammetry," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 198, pp. 223–237, 2023.

[13] H. Li, R. Hartley, and J.-H. Kim, "A linear approach to motion estimation using generalized camera models," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 7 2008, pp. 1 –8.

[14] X. Hu, F. Lauze, and K. S. Pedersen, "Refractive pose refinement: Generalising the geometric relation between camera and refractive interface," *International Journal of Computer Vision*, vol. 131, no. 6, pp. 1448–1476, 2023.

[15] A. Jordt-Sedlazeck and R. Koch, "Refractive structure-from-motion on underwater images," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2013, pp. 57–64.

[16] F. Chadebecq, F. Vasconcelos, G. Dwyer, R. Lacher, S. Ourselin, T. Vercauteren, and D. Stoyanov, "Refractive structure-from-motion through a flat refractive interface," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5315–5323.

[17] B. Elnashef and S. Filin, "Drift reduction in underwater egomotion computation by axial camera modeling," *IEEE Robotics and Automation Letters*, 2023.

[18] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision*. Springer, 2016, pp. 501–518.

[19] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[20] G. Billings, R. Camilli, and M. Johnson-Roberson, "Hybrid visual slam for underwater vehicle manipulator systems," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6798–6805, 2022.

[21] M. D. Grossberg and S. K. Nayar, "The raxel imaging model and ray-based calibration," *International Journal of Computer Vision*, vol. 61, no. 2, pp. 119–137, 2005.

[22] T. Schops, V. Larsson, M. Pollefeys, and T. Sattler, "Why having 10,000 parameters in your camera model is better than twelve," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2535–2544.

[23] A. Agrawal, S. Ramalingam, Y. Taguchi, and V. Chari, "A theory of multi-layer flat refractive geometry," in *CVPR*, 2012.

[24] C. Kunz and H. Singh, "Hemispherical refraction and camera calibration in underwater vision," in *OCEANS 2008*. IEEE, 2008, pp. 1–7.

[25] T. Luczynski, 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. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0029801817300434

[26] P. Sturm, "Multi-view geometry for general camera models," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 206–212.

[27] P. Sturm, S. Ramalingam, and S. Lodha, "On calibration, structure from motion and multi-view geometry for generic camera models," in *Imaging Beyond the Pinhole Camera*, ser. Computational Imaging and Vision, K. Daniilidis and R. Klette, Eds. Springer, aug 2006, vol. 33.

[28] S. Ramalingam, S. K. Lodha, and P. Sturm, "A generic structure-from-motion framework," *Computer Vision and Image Understanding*, vol. 103, no. 3, pp. 218–228, Sept. 2006.

[29] V. Chari and P. Sturm, "Multiple-View Geometry of the Refractive Plane," in *BMVC 2009 - 20th British Machine Vision Conference*, A. Cavallaro, S. Prince, and D. C. Alexander, Eds. London, United Kingdom: The British Machine Vision Association (BMVA), Sept. 2009, pp. 1–11. [Online]. Available: https://hal.inria.fr/inria-00434342

[30] L. Kang, L. Wu, and Y.-H. Yang, "Two-view underwater structure and motion for cameras under flat refractive interfaces," in *Computer Vision - ECCV 2012*, ser. Lecture Notes in Computer Science, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Springer Berlin / Heidelberg, 2012, vol. 7575, pp. 303–316.

[31] M. Fischler and R. Bolles, "RANdom SAmpling Consensus: a paradigm for model fitting with application to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 6 1981.

[32] B. Elnashef and S. Filin, "Direct linear and refraction-invariant pose estimation and calibration model for underwater imaging," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 154, pp. 259–271, 2019.

[33] G. Hee Lee, B. Li, M. Pollefeys, and F. Fraundorfer, "Minimal solutions for pose estimation of a multi-camera system," in *Robotics Research: The 16th International Symposium ISRR*. Springer, 2016, pp. 521–538.

[34] L. Kneip and H. Li, "Efficient computation of relative pose for multi-camera systems," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 446–453.

[35] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng, "Complete solution classification for the perspective-three-point problem," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 8, pp. 930–943, 2003.

[36] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o (n) solution to the pnp problem," *International journal of computer vision*, vol. 81, no. 2, p. 155, 2009.

[37] M. Pharr, W. Jakob, and G. Humphreys, *Physically based rendering: From theory to implementation*. MIT Press, 2023.

[38] F. Seegräber, P. Schöntag, F. Woelk, and K. Köser, "Underwater multiview stereo using axial camera models."

[39] D. Nistér, "An efficient solution to the five-point relative pose problem," *TPAMI*, vol. 26, pp. 756–777, 2004.

[40] M. She, Y. Song, D. Nakath, and K. Köser, "Semihierarchical reconstruction and weak-area revisiting for robotic visual seafloor mapping," *Journal of Field Robotics*, 2023.

[41] B. Elnashef and S. Filin, "Target-free calibration of flat refractive imaging systems using two-view geometry," *Optics and Lasers in Engineering*, vol. 150, p. 106856, 2022.