Acoustic Neural 3D Reconstruction Under Pose Drift

Tianxiang Lin*[1], Mohamad Qadri*[1], Kevin Zhang[2], Adithya Pediredla[3], Christopher A. Metzler[2], Michael Kaess[1]

***Abstract*— We consider the problem of optimizing neural implicit surfaces for 3D reconstruction using acoustic images collected with drifting sensor poses. The accuracy of current state-of-the-art 3D acoustic modeling algorithms is highly dependent on accurate pose estimation; small errors in sensor pose can lead to severe reconstruction artifacts. In this paper, we propose an algorithm that jointly optimizes the neural scene representation and sonar poses. Our algorithm does so by parameterizing the 6DoF poses as learnable parameters and backpropagating gradients through the neural renderer and implicit representation. We validated our algorithm on both real and simulated datasets. It produces high-fidelity 3D reconstructions even under significant pose drift.**

I. Introduction

Autonomous Underwater Vehicles (AUVs) often carry imaging sonar, also known as forward-looking sonar (FLS). Unlike an optical camera, an imaging sonar is able to capture long-range information in turbid conditions. Thus, because of its robustness, FLS has been integrated into many underwater applications, such as underwater inspection, construction, ecology, archaeology, and surveillance [1]–[4].

Imaging sonar captures 2D measurements by emitting acoustic pulses and measuring the intensity and arrival time of reflections from 3D structures. Using beamforming and time-of-flight techniques, an imaging sonar can recover azimuth and range information. However, a key limitation is that it does not provide direct elevation measurements, making it inherently ill-suited for applications that require 3D information.

To reconstruct 3D structures, prior methods [5]–[7] tried to mitigate this limitation by integrating multiple imaging sonar measurements taken from known and precise poses. However, these approaches rely heavily on accurate poses and any errors in these poses will significantly degrade reconstruction quality.

To mitigate the impact of pose errors, we propose a technique that jointly optimizes the 3D structure and the sonar poses. By incorporating pose optimization directly into the reconstruction process, our method improves robustness to pose drift and sensor noise, enabling more reliable 3D reconstruction from imaging sonar. Our contributions are as follows:

- A framework for recovering 3D structure from acoustic sonar measurements while simultaneously optimizing sensor poses.
- A qualitative analysis of the convergence properties of the optimized pose solution set.
- Evaluation on both real and simulated datasets containing objects with varied geometries.

II. Related Work

A. 3D Reconstruction From Imaging Sonar

Prior works introduce a variety of techniques for reconstructing 3D structures from imaging sonar measurements, such as ICP-based alignment [8], space carving [6, 9], solving constraint equations [10], generative sensor modeling [11], graph-based processing [12, 13], convex optimization [14], and supervised learning [15]–[17].

More recently, differentiable rendering-based techniques have achieved state-of-the-art performance in the reconstruction of 3D structures from sonar imagery. Qadri et al. [5] propose combining a neural surface representation with a novel differentiable acoustic volume rendering process to recover 3D surfaces from imaging sonar data. In an extension [18] they combine optical camera information with sonar imagery to achieve improved 3D reconstruction performance in the small baseline setting. Reed et al. [19] recover 3D volumes from synthetic aperture sonar measurements using neural rendering. Xie et al. [20, 21] use neural rendering to perform bathymetry from imaging sonar and sidescan sonar, respectively. Qu et al. [22] derive a novel forward splatting process for Gaussian Splatting and use it to recover 3D structures from sonar imagery.

All of these methods rely critically upon reliable estimates of the poses at which the used sonar imagery is captured, but none as of yet concentrate on recovering 3D scenes from noisy poses.

B. Pose Optimization

Conventional underwater simultaneous localization and mapping (SLAM) methods have been widely explored and applied to solve underwater navigation problems. These methods rely on well-studied algorithms from state estimation [23]–[27]. Shin et al. [28] explore a pairwise bundle adjustment method by exploiting spatial constraints using KAZE features [29] between paired sonar images, which are refined by random sample consensus (RANSAC). Acoustic Structure-from-Motion (ASFM) algorithm to recover poses from multiple sonar images and drifting odometry [30]. Westman et al. [31] improve the performance of multiview pose optimization by adding two-view sonar constraints

* Indicates equal contribution

[1]T. Lin, M. Qadri, M. Kaess are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. {tianxian, mqadri, kaess}@cs.cmu.edu

[2]K. Zhang, C. Metzler are with the Department of Computer Science at the University of Maryland, College Park. {kzhang24, metzler}@umd.edu

[3]A. Pediredla is with the Computer Science Department at Dartmouth College. adithya.k.pediredla@dartmouth.edu

during loop closure detections. Loi et al. [32] introduce submap registration for point cloud alignment or feature matching. Xu et al. [33] propose a direct imaging sonar odometry system that minimizes the aggregated two-view reprojection errors of sonar pixels with high-intensity gradients. These prior works fail to recover poses from the drifting odometry when sonar images provide limited features. At the same time, sonar intensity values are determined by multiple factors such as the orientation of the sonar with respect to the object, objects' material, etc. Sonar images with significant speckle noise and multi-path effects can lead to the failure of classical underwater SLAM due to errors in feature matching.

Improving neural rendering reconstructions by simultaneously optimizing the reconstruction and poses is an active area of research. Wang et al. [34] propose using an axis-angle and translation parameterization of camera poses and optimizing them with a neural radiance field simultaneously. Lin et al. [35] use a coarse-to-fine optimization strategy along with a neural radiance field pipeline to recover image poses from a collection of images. Bian et al. [36] exploit monocular depth priors to improve joint estimation of poses and neural radiance fields from images. Chng et al. [37] use Gaussian activations to improve joint estimation of poses and neural radiance fields. Park et al. [38] precondition camera parameters leading to improved reconstruction during the joint estimation of poses and neural radiance fields. All these prior works concentrate on 3D reconstruction from optical images with noisy camera poses. In this work, we focus on 3D reconstruction from acoustic images with noisy sonar poses.

III. Background on Acoustic Neural Rendering

Given a training set of posed acoustic sonar images, the objective is to retrieve an accurate 3D reconstruction of an object of interest. We review *NeuSIS*, a neural rendering method presented by Qadri et al. [5] upon which our work is based. The technique allows for state-of-the-art performance for acoustic rendering by leveraging neural implicit surfaces and introduces a novel volumetric rendering equation.

A. Image Formation Model

Imaging sonars emit acoustic pulses and measure the intensity of the reflected signals to form a 2D acoustic image. While range and azimuth are resolved by the sensor, elevation remains ambiguous: The intensity of each pixel in the image is proportional to the cumulative sum of acoustic energy reflected by objects intersected by the acoustic arc at a specific range and azimuth. Hence, Qadri et al. [5] use the following image formation model where, for each pixel, the intensity I_p at pixel $p = (r_i, \theta_i)$ is modeled as the integral of the reflected acoustic energy along each ray over the acoustic arc:

$$I_p = \int_{\phi_{\text{min}}}^{\phi_{\text{max}}} \int_{r_i-\epsilon}^{r_i+\epsilon} \frac{E_e}{r} T(r, \theta_i, \phi)\sigma(r, \theta_i, \phi)\mathrm{d}r\mathrm{d}\phi. \quad (1)$$

where $\phi_{\text{min}}, \phi_{\text{max}}$ are the minimum and maximum elevation angles, E_e is the acoustic energy emitted by the sonar. $T = e^{-\int_0^{r_i} \sigma(r', \theta_i, \phi_i)\mathrm{d}r'}$ is the transmittance term, and σ is the particle density.

B. Neural Representation

Similar to Yariv et al. [39], the object is represented as Signed Distance Function (SDF), $\mathbf{N}(\mathbf{x})$, which outputs the distance of each 3D point $\mathbf{x} = (X, Y, Z)$ to the nearest surface. A separate network, $\mathbf{M}$, computes $\mathbf{M}(\mathbf{x})$, the outgoing acoustic radiance at each spatial coordinate $\mathbf{x}$ which is then used to approximate the intensity of each pixel $\hat{I}_p$. The accuracy of this intensity estimate is correlated with the accuracy of the SDF representation: if the SDF is close to the ground-truth object then pixel p of the ith training image $\hat{I}_p^i \rightarrow I_p^i$.

Note that in this work, we assume noisy pose estimates. Hence, the approximated pixel intensity will depend on both the SDF representation as well as on our current estimate of the sonar poses.

C. Approximation of the Image Formation Integral

Eq. 1 is discretized and approximated by sampling 3D points along both acoustic rays and arcs.

$$\hat{I}_p = \sum_{\mathbf{x} \in \mathcal{A}_p} \frac{1}{r(\mathbf{x})} T[\mathbf{x}]\alpha[\mathbf{x}]\mathbf{M}(\mathbf{x}) \quad (2)$$

where $\mathcal{A}_p$ is the set of sampled points along the acoustic arc at pixel $p = (r_i, \theta_i)$ and $\mathbf{M}$ is the predicted radiance at $\mathbf{x}$. The computations of the discrete transmittance $T[\mathbf{x}]$ and opacity $\alpha[\mathbf{x}]$ terms require additionally sampling along acoustic rays. For any such spatial sample $\mathbf{x}_s$ on the acoustic ray, the discrete opacity at $\mathbf{x}_s$ can be approximated as

$$\alpha[\mathbf{x}_s] = \max\left(\frac{\mathbf{\Phi}_s(\mathbf{N}(\mathbf{x}_s)) - \mathbf{\Phi}_s(\mathbf{N}(\mathbf{x}_{s+1}))}{\mathbf{\Phi}_s(\mathbf{N}(\mathbf{x}_s))}, 0\right), \quad (3)$$

where $\mathbf{\Phi}_s(x) = (1 + e^{-sx})^{-1}$ is the sigmoid function and s is a trainable parameter while the discrete transmittance is modeled as

$$T[\mathbf{x_s}] = \prod_{\mathbf{x_r} \mid r<s} (1 - \alpha[\mathbf{x_r}]). \quad (4)$$

IV. Method

A. Loss Function

We define the following set of trainable parameters: Θ are the weights of the SDF network M, Φ are the weights of the neural renderer N, and $\mathcal{T} = \{T_i\}$ which parametrize the learnable sonar poses. Any intensity value computed via Eq. 2 is a function of Θ, Φ, and $\mathcal{T}$ - in other words, we can define $\hat{I}_p(\Theta, \Phi, \mathcal{T})$ as the predicted intensity of the pth pixel of a training image and express our rendering loss function in terms of these three sets of parameters.

Our loss function is composed of three terms: The intensity loss

$$\mathcal{L}_{\text{int}} \equiv \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} ||\hat{I}_p(\Theta, \Phi, \mathcal{T}) - I_p||_1, \quad (5)$$

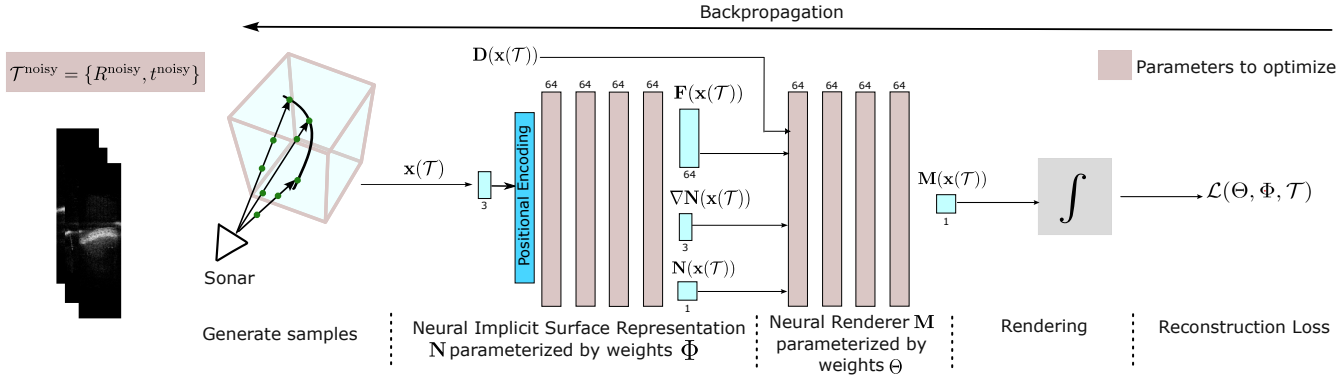


Fig. 1: The pipeline of our proposed method. Our approach jointly optimizes sonar neural implicit surface networks and pose parameters by minimizing the total reconstruction loss. It takes 3D samples and viewing directions—both dependent on pose estimates—as inputs and outputs the signed distance function (SDF) $\mathbf{N}$ and outgoing acoustic radiance $\mathbf{M}$ for sonar image rendering. This pipeline enables training with sonar images and odometry that may be subject to drift.

where $\mathcal{P}$ is the set of sampled pixels, which encourages the predicted intensity to match the intensity of the pth pixel in a training sonar image. The eikonal loss [40]

$$\mathcal{L}_{\text{eik}} \equiv \frac{1}{|\mathbf{X}|} \sum_{\mathbf{x} \in \mathbf{X}} (||\nabla \mathbf{N}[\mathbf{x}(\mathcal{T})]||_2 - 1)^2, \tag{6}$$

where $|\mathbf{X}|$ is the set of sampled 3D points, which is an implicit geometric regularization term used to regularize the SDF towards producing smooth reconstructions. Note that a 3D sample $\mathbf{x}(\mathcal{T})$ is dependent on our current estimate of the pose parameters. Finally, we add an ℓ_1 loss term

$$\mathcal{L}_{\text{reg}} \equiv \frac{1}{|\mathbf{X}|} \sum_{\mathbf{x} \in \mathbf{X}} ||\alpha[\mathbf{x}(\mathcal{T})]||_1, \tag{7}$$

to help produce favorable 3D reconstructions when we use sonar images from a limited set of view directions. Our final training loss term is therefore:

$$\mathcal{L}(\Theta, \Phi, \mathcal{T}) = \mathcal{L}_{\text{int}} + \lambda_{\text{eik}} \mathcal{L}_{\text{eik}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}. \tag{8}$$

Our objective function is therefore:

$$\Theta^*, \Phi^*, \mathcal{T}^* = \underset{\Theta, \Phi, \mathcal{T}}{\operatorname{argmin}} \, \mathcal{L}(\Theta, \Phi, \mathcal{T}) \tag{9}$$

Since our loss function is a fully differentiable function of all its parameters, we can perform parameter updates using iterative gradient-based methods such as ADAM. After convergence, we retrieve the surface by extracting the zero-level set of $\mathbf{N}$ using the Marching Cubes algorithm with a bounding box enclosing the object:

$$\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^3 : \mathbf{N}(\mathbf{x}) = 0\}. \tag{10}$$

B. Sensor Pose Parametrization

Each drifting sonar pose T_i is an element in $\text{SE}(3)$, the special Euclidean group in 3 dimensions. We parametrize it as a vector $(\omega_i, \mathbf{t}_i) \in \text{se}(3)$ where $\mathbf{t}_i \in \mathbb{R}^3$ represents the translation and $\omega_i \in \text{so}(3)$ represents the rotation in axis-angle form, and $\text{so}(3)$ is the Lie Algebra of rotations $\text{SO}(3)$. For each pose, we define a correction vector $(\delta\omega_i, \delta\mathbf{t}_i) \in \mathbb{R}^6$ as a learnable parameter. After each gradient update, the full corrective matrix is then recovered via the exponential map:

$$\delta T_i = \begin{bmatrix} \delta\hat{\omega} & \delta\mathbf{t} \\ 0 & 1 \end{bmatrix} \tag{11}$$

where $\delta\hat{\omega}$ is the skew-symmetric matrix given by:

$$\delta\hat{\omega} = \begin{bmatrix} 0 & -\delta\omega_3 & \delta\omega_2 \\ \delta\omega_3 & 0 & -\delta\omega_1 \\ -\delta\omega_2 & \delta\omega_1 & 0 \end{bmatrix} \tag{12}$$

The final corrected pose transformation corresponding to image i is then given by:

$$T_i \leftarrow T_i \cdot \delta T_i \tag{13}$$

V. Evaluation

We trained our models on an NVIDIA H100 80GB HBM3 GPU with Intel Xeon 8470 CPU. Each training runs for 100k iterations, which takes about 5 hours. Table I provides the total number of sonar/pose pairs for each simulated and real dataset.

For our comparison metric, we use the mean and root mean square (RMS) Hausdorff distances. The Hausdorff distance is defined as:

$$d_H(\mathcal{M}_1, \mathcal{M}_2) = \max\Big(\max_{\mathbf{p} \in \mathcal{M}_1} \min_{\mathbf{q} \in \mathcal{M}_2} ||p - q||_2, \max_{\mathbf{q} \in \mathcal{M}_2} \min_{\mathbf{p} \in \mathcal{M}_1} ||p - q||_2 \Big) \tag{14}$$

where $\mathcal{M}_1$ and $\mathcal{M}_2$ are the ground-truth (GT) and reconstructed meshes respectively.

A. Understanding the Drift

Similarly to ground robots, underwater vehicles often fuse measurements from multiple sensors to acquire accurate, drift-free odometry. Underwater vehicles are usually equipped with Doppler velocity logs (DVLs), inertial measurement units (IMUs), and depth sensors. A DVL provides low-noise measurements of vehicle velocity with respect to the sea floor in all three axes. An IMU typically consists of gyroscopes and accelerometers. By fusing DVL, IMU, and

Datasets		Real	Boat 1	Boat 2	Plane 1	Plane 2	Rock 1	Rock 2	Concrete column	Submarine
Elevation	14°	291	280	321	495	413	436	290	258	639
angle	28°	441	283	492	444	364	435	289	258	639

TABLE I: Total number of poses in each dataset.

depth sensor measurements, an underwater vehicle is capable of providing noisy but drift-free measurements of Z-axis translation and pitch (ϕ) and roll (θ) angles by measuring hydrostatic pressure and direction of gravity [41]. On the other hand, the X and Y translations and yaw angle (ψ) are estimated from integration over gyroscope and DVL measurements, which will accumulate drift over time due to the absence of an absolute reference from the measurements.

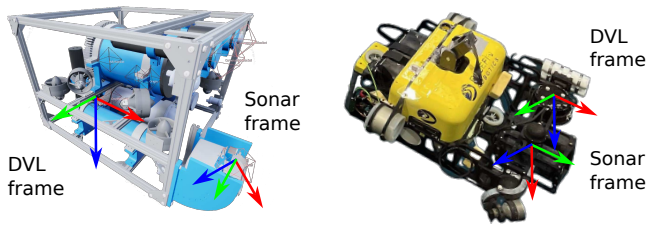


Fig. 2: Left: Simulated robot in HoloOcean [42] with the DVL and an example sonar frame visualized. Right: Real HAUV with the DVL and sonar frames overlaid. Note that the DVL frame is similarly oriented in both the simulated and real setups.

B. Modeling the Drift

Let $R_i = R_z(\psi_i)R_y(\phi_i)R_x(\theta_i)$ and $t_i = [x_i, y_i, z_i]^\top$ be the rotation matrix and translation vector of the 6DoF DVL pose T_i at time i. We model the unbounded drift along the x, y, ψ components of the AUV's odometric measurement as a stochastic process with a time-varying drift. Specifically, we assume that the difference between two consecutive DVL pose components, u_i and u_{i+1} for $u \in \{x, y, \psi\}$, follows:

$$u_{i+1} - u_i \sim \mathcal{N}(\Delta u_{i,i+1}, q^u) \tag{15}$$

where $\Delta u_{i,i+1}$ represents the true underlying relative motion between timesteps i and $i+1$ which depends on the AUV's control inputs. This can be rewritten as:

$$u_{i+1} - u_i = \Delta u_{i,i+1} + \varepsilon^u \tag{16}$$

where $\varepsilon^u \sim \mathcal{N}(0, q^u)$ is a zero-mean additive noise term with variance q^u which models odometric uncertainty. A similar formulation for drifting poses was proposed by Westman et al. [31]. The remaining components $u \in \{z, \theta, \phi\}$ are noisy but drift-free, and hence the noise in these components is modeled as zero-mean Gaussian noise.

C. Simulation

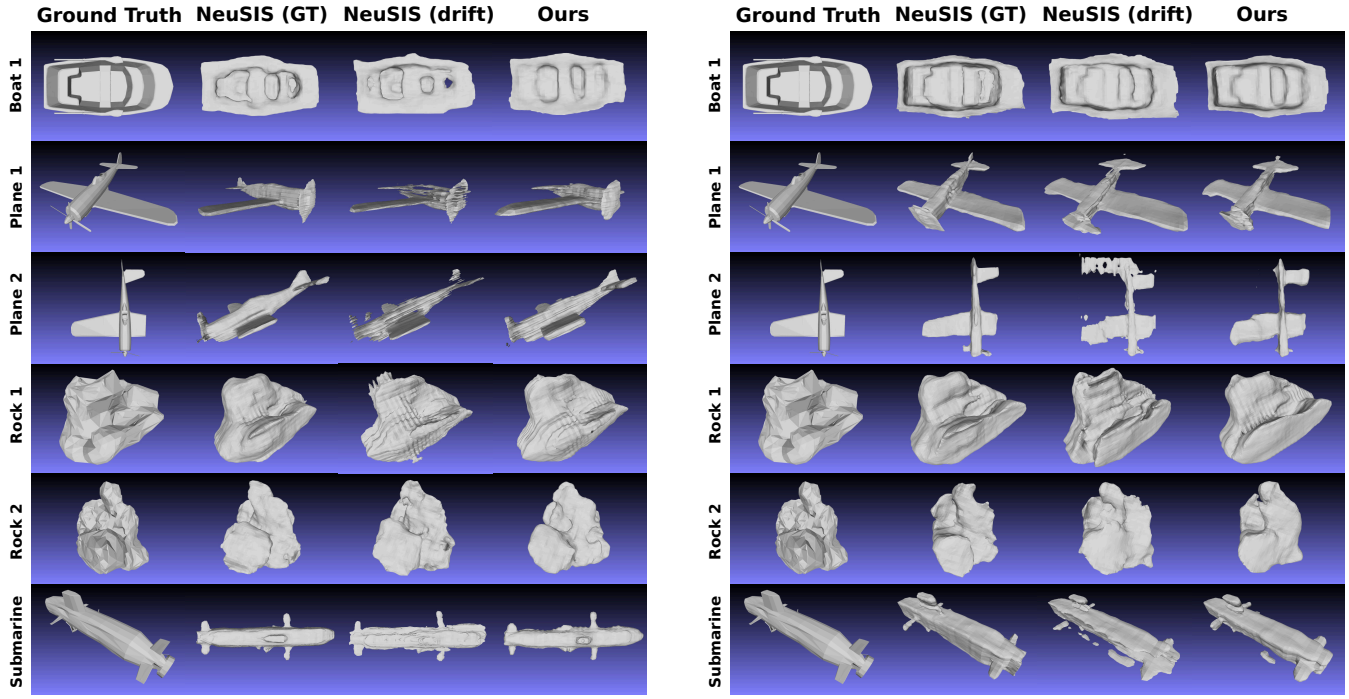
We used an imaging sonar dataset of objects of different shapes and sizes collected using HoloOcean [43], an underwater simulator. The dataset was collected with the simulation of multipath effects enabled and the inclusion of multiplicative noise $w^{\text{sm}} \sim \mathcal{N}(0, 0.15)$ and additive noise $w^{\text{sa}} \sim \mathcal{R}(0.2)$ where $\mathcal{R}$ is the Rayleigh distribution. The original dataset was collected with a frequency of 10Hz, which we downsample by a factor of two. Among other sensors, the simulated robot is equipped with a DVL whose frame is oriented similarly to the DVL on the real robot in Fig. 2. For each object, we collect two to three different trajectories while varying sonar orientation. For each sonar image, HoloOcean additionally returns the ground-truth sonar and DVL poses. Hence, to simulate realistic drifting pose patterns as described in subsection V-A, we adopt the following strategy: Let T_i^{dvl} and T_i^{sonar} be the ground-truth DVL and sonar poses at time i respectively.

1) Compute the relative DVL pose between timesteps i and $i+1$: $\Delta T_{i\to i+1}^{\text{dvl}} = (T_i^{\text{dvl}})^{-1} \cdot T_{i+1}^{\text{dvl}}$.
2) Add noise to the x, y, ψ axes of $\Delta T_{i,i+1}^{\text{dvl}}$ following Eq. 16 with $\varepsilon^x, \varepsilon^y \sim \mathcal{N}(0, 0.004\text{m})$ and $\varepsilon^\psi \sim \mathcal{N}(0, 0.004\text{rad})$. We obtain the noisy relative transform $\Delta \tilde{T}_{i\to i+1}^{\text{dvl}}$.
3) Compute the noisy DVL pose at timestep $i+1$ as $\tilde{T}_{i+1}^{\text{dvl}} = T_i^{\text{dvl}} \cdot \Delta \tilde{T}_{i\to i+1}^{\text{dvl}}$
4) To simulate the noisy but drift-free measurements over the z, ϕ, θ axes, we add Gaussian noise to $\tilde{T}_{i+1}^{\text{dvl}}$ to each of these axes with $\varepsilon^z \sim \mathcal{N}(0, 0.005\text{m})$ and $\varepsilon^\phi, \varepsilon^\theta \sim \mathcal{N}(0, 0.005\text{rad})$.
5) Obtain the corresponding noisy sonar pose by multiplying with the known DVL-to-sonar extrinsic matrix: $\tilde{T}_{i+1}^{\text{sonar}} = \tilde{T}_{i+1}^{\text{dvl}} \cdot T_{\text{dvl}\to\text{sonar}}$.

Figs. 3a and 3b present qualitative results for different simulated objects at elevation apertures of 14° and 28°. We compare (1) reconstructions using ground-truth (GT) odometry poses: NeuSIS (GT), (2) reconstructions with noisy pose measurements without optimization: NeuSIS (drift), and (3) our method, which jointly optimizes the SDF, renderer, and poses. For each object and each method, we select the

		NeuSIS (GT)		NeuSIS (drift)		Ours	
		RMS	Mean	RMS	Mean	RMS	Mean
Boat 1	14°	0.074	0.059	0.101	0.076	**0.089**	**0.065**
($3.8 \times 1.7 \times 0.84$)	28°	0.065	0.049	0.102	0.076	**0.067**	**0.049**
Boat 2	14°	0.093	0.064	0.146	0.100	**0.098**	**0.070**
($5.7 \times 2.3 \times 1.2$)	28°	0.108	0.079	0.179	0.127	**0.109**	**0.082**
Plane 1	14°	0.156	0.102	0.197	0.141	**0.159**	**0.122**
($13.5 \times 11.5 \times 3.6$)	28°	0.175	0.121	0.217	0.153	**0.183**	**0.126**
Plane 2	14°	0.117	0.091	0.400	0.162	**0.151**	**0.115**
($9.1 \times 12.6 \times 3.0$)	28°	0.150	0.115	1.551	0.737	**0.266**	**0.194**
Rock 1	14°	0.139	0.105	0.194	0.150	**0.154**	**0.113**
($5.7 \times 3.5 \times 2.8$)	28°	0.112	0.082	0.196	0.148	**0.129**	**0.098**
Rock 2	14°	0.110	0.083	0.148	0.112	**0.117**	**0.091**
($2.2 \times 2.2 \times 2.0$)	28°	0.135	0.102	0.169	0.127	**0.151**	**0.113**
Concrete column	14°	0.058	0.033	0.108	0.074	**0.080**	**0.059**
($1.9 \times 1.2 \times 4.3$)	28°	0.055	0.038	0.784	0.057	**0.052**	**0.039**
Submarine	14°	0.149	0.116	0.206	0.155	**0.150**	**0.117**
($5.1 \times 16.7 \times 4.7$)	28°	0.144	0.110	0.280	0.209	**0.186**	**0.141**

TABLE II: Size ($W \times L \times H$), root mean square (RMS), and mean Hausdorff distance (meters) for eight different objects from HoloOcean. The results show that our method produces more accurate 3D reconstructions compared to NeuSIS with drifting poses, demonstrating the effectiveness of our approach in handling pose drift.



(a) 3D reconstruction results for the 14° elevation aperture simulated sonar datasets.

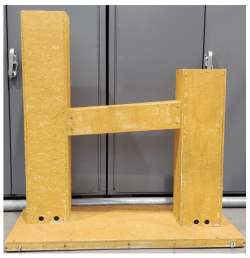
(b) 3D reconstruction results for the 28° elevation aperture simulated sonar datasets.

Fig. 3: 3D reconstruction results for (a) the 14° and (b) the 28° elevation aperture sonar datasets collected using the HoloOcean underwater simulator. From left to right, the images show ground-truth meshes of six different objects, followed by reconstructions from NeuSIS with ground-truth odometry, NeuSIS with drifting poses, and our proposed method. Our approach effectively restores dense 3D reconstructions despite drifting odometry, achieving results comparable to NeuSIS with ground-truth trajectories.

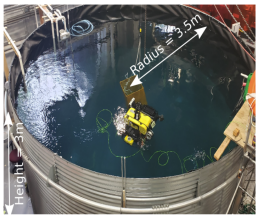
best level set, i.e. Marching Cubes threshold, $\epsilon \in [-0.2, 0.2]$.

As expected, the best results are achieved using the GT sonar poses from the HoloOcean simulator. These reconstructions represent an upper bound on reconstruction quality for each object. Our goal is to approach the accuracy of NeuSIS (GT), both qualitatively and quantitatively, even after injecting noise in the pose measurements. Qualitatively, our method consistently reduces reconstruction errors compared to using drifting poses across all objects. Notably, we observe less stratification in the *Plane 1* and *Plane 2* datasets at 14°, and a significant correction in the *Plane 2* dataset at 28°—particularly in the tail area. Similarly, the *submarine* reconstructions at 14° and 28° exhibit improvements along the entire shape. These qualitative results are supported by the quantitative results in Table II which demonstrate a significant improvement in reconstruction accuracy when using our method.

D. Water Tank Experiments



(a) Test structure.



(b) Water test tank.

Fig. 4: Real-world experimental setup.

We evaluate our proposed method on real-world datasets of a test structure submerged in a test tank (see Fig. 4) imaged with the two wide elevation apertures achievable by the sonar: 14° and 28°. Our experimental platform for dataset collections is a Bluefin Hovering Autonomous Underwater Vehicle (HAUV) [44] equipped with a 1.2MHz Teledyne/RDI Workhorse Navigator Doppler velocity log (DVL), a Honeywell HG1700 inertial measurement unit (IMU), a Paroscientific Digiquartz depth sensor, and a Sound Metric DIDSON imaging sonar [45] (please check [5, 10] for more details about the datasets and the collection setup). The frame rate of the real-world datasets is 2Hz.

We start by discarding sonar image/pose pairs that lack

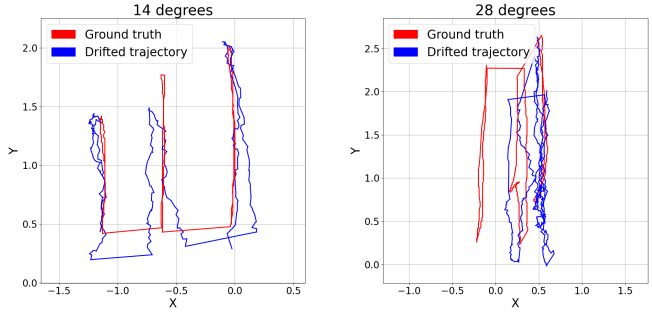
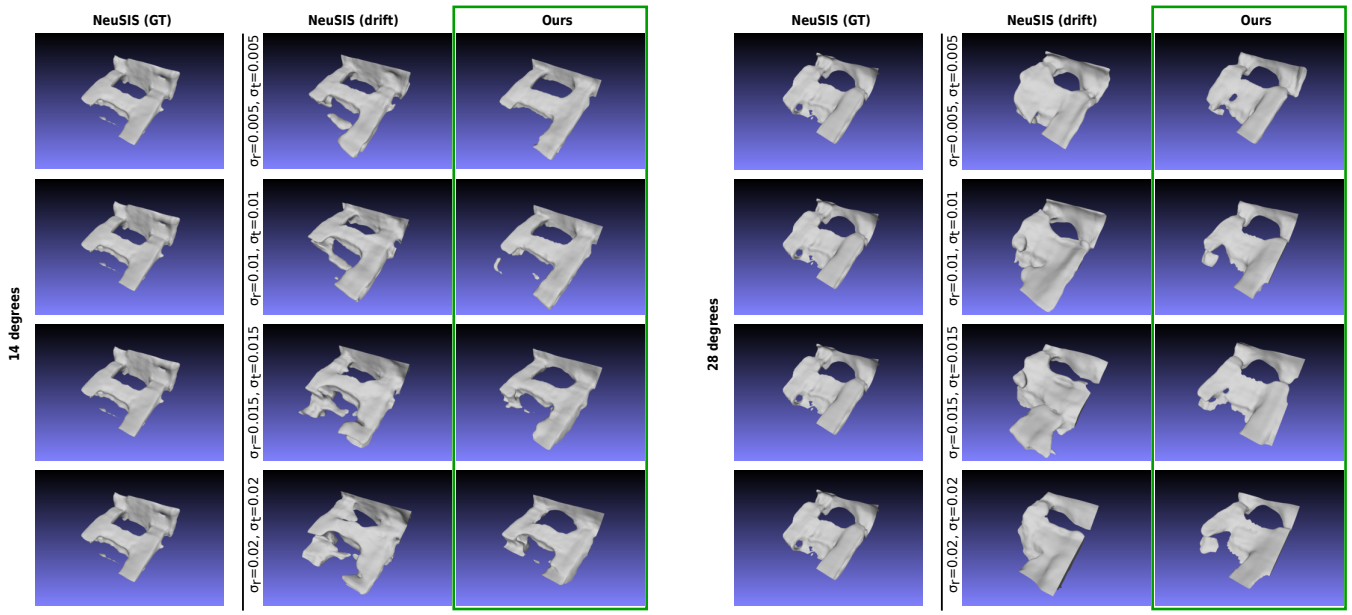


Fig. 5: Top-down view illustrating an example of a drifting DVL trajectory after noise injection (blue) alongside its corresponding DVL poses with no added noise (red) for the 14° and 28° elevation aperture real datasets, shown on the left and right, respectively. Noise is injected into the x, y, and ψ relative poses with $\varepsilon^x, \varepsilon^y \sim \mathcal{N}(0, 0.015 \text{ m})$ and $\varepsilon^\psi \sim \mathcal{N}(0, 0.015 \text{ rad})$.



(a) 3D reconstruction results for the 14° elevation aperture real sonar datasets.

(b) 3D reconstruction results for the 28° elevation aperture real sonar datasets.

Fig. 6: 3D reconstructions from each method with two different elevation apertures and four sets of drifting noise added to the x, y, and yaw directions of the vehicle odometry. Our proposed method yields more accurate and cleaner reconstructions compared to NeuSIS with drifting noise added to the vehicle odometry. Notably, reconstructions from the proposed method with lower drifting noise are cleaner than those from NeuSIS with ground-truth odometry, demonstrating the capability of the proposed method to eliminate drifting noise in the odometry.

returns from the object of interest. We then inject relative pose noise as described in subsection 16, adding $\varepsilon^x, \varepsilon^y \sim \mathcal{N}(0, \sigma_t)$ to the relative DVL x and y measurements and $\varepsilon^\psi \sim \mathcal{N}(0, \sigma_r)$ to the relative yaw (ψ). Our method is evaluated across four increasing noise levels (listed in Table IV), with each experiment (i.e., each elevation angle and noise level) repeated using three different random seeds to simulate different noise patterns. Fig. 5 shows an example of a DVL trajectory before and after noise injection.

	NeuSIS (GT)		
	elevation	RMS	Mean
Real Datasets	14°	**0.052**	**0.036**
	28°	**0.071**	**0.049**

TABLE III: Root mean square (RMS) and mean Hausdorff distance in meters for NeuSIS with ground-truth odometry on real-world datasets with 14° and 28° elevation apertures.

		NeuSIS (drift)		Ours	
		RMS	Mean	RMS	Mean
$\sigma_r = 0.005$ m	14°	0.052 ± 0.001	0.039 ± 0.002	**0.046 ± 0.001**	**0.034 ± 0.001**
$\sigma_t = 0.005$ rad	28°	0.082 ± 0.003	0.058 ± 0.002	**0.073 ± 0.002**	**0.052 ± 0.001**
$\sigma_r = 0.01$ m	14°	0.061 ± 0.001	0.046 ± 0.004	**0.047 ± 0.005**	**0.035 ± 0.003**
$\sigma_t = 0.01$ rad	28°	0.085 ± 0.002	0.062 ± 0.000	**0.074 ± 0.002**	**0.054 ± 0.001**
$\sigma_r = 0.015$ m	14°	0.073 ± 0.001	0.056 ± 0.003	**0.054 ± 0.003**	**0.040 ± 0.003**
$\sigma_t = 0.015$ rad	28°	0.091 ± 0.002	0.069 ± 0.002	**0.079 ± 0.004**	**0.056 ± 0.002**
$\sigma_r = 0.02$ m	14°	0.081 ± 0.005	0.062 ± 0.006	**0.062 ± 0.004**	**0.047 ± 0.004**
$\sigma_t = 0.02$ rad	28°	0.093 ± 0.004	0.069 ± 0.004	**0.081 ± 0.003**	**0.060 ± 0.004**

TABLE IV: Root mean square (RMS) and mean Hausdorff distances for reconstructions of 14° and 28° real-world datasets using NeuSIS with drifting odometry and the proposed method. Translation is measured in meters and rotation in radians. We select three random seeds and compute the mean and standard deviation for each distance. The Hausdorff distance threshold is set to 0.2 m for the 14° elevation dataset and 0.25 m for the 28° dataset.

Fig. 6 shows qualitative results of reconstructions using the 4 different noise level and the 2 elevation apertures. We observe that for both elevation apertures, the reconstructions with NeuSIS (drift) degrade rapidly as the noise level increases. For example, at 14°, increasing artifacts can be observed near the shorter leg region, while at 28°, we observe the deterioration of the entire shape. In contrast, our method successfully recovers shapes in which the main components of the structure are preserved (a base, a smaller and larger leg, and a crossbar). We report in Table III, the result obtained using NeuSIS (GT) – i.e. reconstructions using the GT odometry. Table IV shows the metrics (average $\pm$ standard deviation) for both our method as well as NeuSIS (drift) after noise injection. As expected, the reconstruction quality of NeuSIS (drift) degrades with increasing noise. However, our method remains robust to significant pose drift, and only begins to struggle when the noise becomes particularly high ($\sigma_r = 0.02$rad and $\sigma_t = 0.02$m) between consecutive relative poses.

VI. Do we Recover the Odometry Poses?

We experimentally demonstrate that the set of possible poses, $\mathcal{T}$, that minimizes the reconstruction error is not unique. In other words, our algorithm can converge to a set of poses that is far from the odometry poses while still resulting in final 3D reconstructions that are perceptibly similar. We perform the following experiment on the 28° real data:

1) **Train using odometry poses:** Use the odometry poses, $\mathcal{T}_{\text{odom}}$, and train a 3D model $\mathcal{R}_{\text{odom}}$ until converge.
2) **Freeze network weights:** Freeze the weights of both the SDF network, **N**, and the neural renderer, **M**.
3) **Inject noise to relative pose:** Corrupt the odometry poses by injecting noise to the x, y, ψ relative poses with $\varepsilon^x, \varepsilon^y \sim \mathcal{N}(0, 0.01\text{m})$ and $\varepsilon^\psi \sim \mathcal{N}(0, 0.01\text{rad})$.

We obtain a noisy drifting set of poses: $\mathcal{T}_{\text{noisy}}$.

4) **Optimize the noisy poses:** Optimize $\mathcal{T}_{\text{noisy}}$ using the loss function in Eq. 8 while keeping the weights of networks **N** and **M** from step 2 frozen. In other words, solve the following optimization problem:

$$\mathcal{T}^*_{\text{noisy}} = \underset{\mathcal{T}}{\operatorname{argmin}}\, \mathcal{L}(\Theta, \Phi, \mathcal{T})$$

5) **Train a new 3D model:** Finally, freeze $\mathcal{T}^*_{\text{noisy}}$ and train a new 3D model $\mathcal{R}^*_{\text{noisy}}$.

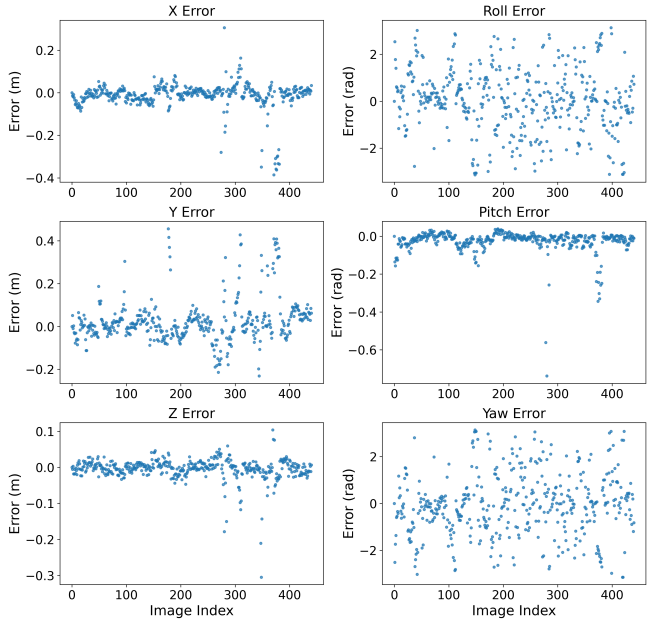


Fig. 7: Error in each pose component between the odometry poses, $\mathcal{T}_{\text{odom}}$ and the optimized poses after noise injection, $\mathcal{T}^*_{\text{noisy}}$. We note that the two sets of poses differ significantly.

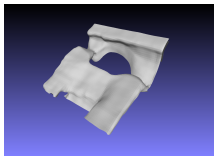


Fig. 8: The reconstruction $\mathcal{R}^*_{\text{noisy}}$ obtained from step 5: Freeze $\mathcal{T}^*_{\text{noisy}}$ and train **M** and **N**.

Fig. 7 shows the errors between $\mathcal{T}_{\text{odom}}$ and $\mathcal{T}^*_{\text{noisy}}$ in the sonar frame, revealing significant differences between the two sets of poses. This indicates that if we freeze the converged mesh $\mathcal{R}_{\text{odom}}$ and optimize only the poses, the optimization can still converge to a solution set far from the odometric poses. Fig. 8 shows the resulting mesh reconstruction when we freeze $\mathcal{T}^*_{\text{noisy}}$ and train **M** and **N** to obtain $\mathcal{R}^*_{\text{noisy}}$. The reconstruction $\mathcal{R}^*_{\text{noisy}}$ captures the major parts of the structure (two legs and the middle tile), supporting the observation that even significantly different pose sets can produce plausible 3D reconstructions. This result is supported by the quantitative metrics for $\mathcal{R}^*_{\text{noisy}}$ (RMS = 0.077m, mean = 0.058m), which are close to the metrics reported when using $\mathcal{T}_{\text{odom}}$ in Table III.

VII. Conclusion and Future Work

We proposed an approach for reconstructing objects using imaging sonar, even in the presence of significant pose drift. Our method jointly optimizes both the sonar poses and the 3D model parameters using only the reconstruction loss, without relying on external sensors. That is, it learns directly using only the sonar images and their corresponding noisy poses. Through extensive experiments across different objects and elevation apertures, we demonstrated that our method is robust to high noise levels and adapts well to diverse target geometries.

For future work, we see multiple promising directions. First, incorporating additional sensor modalities available on the vehicle, such as Doppler Velocity Logs (DVL), IMUs, or optical cameras, could provide stronger constraints for pose optimization, further improving reconstruction accuracy. Second, our current approach is designed for offline 3D reconstruction. To enable real-time applications, we plan to explore acceleration techniques such as Instant-NGP [46], which uses neural graphics primitives for highly efficient scene optimization. These enhancements would make our method more practical for real-world autonomous underwater operations.

VIII. Acknowledgment

T.L., M.Q., and M.K. were partially supported by the Office of Naval Research (ONR) grant N00014-24-1-2272. K.Z. and C.A.M. were partially supported by AFOSR award no. FA9550-22-1-0208, NSF award no. 2339616, and ONR award no. N000142312752.

References

[1] J. Wang, T. Shan, and B. Englot, "Underwater terrain reconstruction from forward-looking sonar imagery," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, Montreal, Canada, May 2019, pp. 3471–3477.

[2] S. Negahdaripour, "Application of forward-scan sonar stereo for 3-D scene reconstruction," *IEEE J. of Oceanic Engineering (JOE)*, vol. 45, no. 2, pp. 547–562, Oct. 2018.

[3] J. Albiez, S. Joyeux, C. Gaudig, J. Hilljegerdes, S. Kroffke, C. Schoo, S. Arnold, G. Mimoso, P. Alcantara, R. Saback *et al.*, "Flatfish-a compact subsea-resident inspection AUV," in *Proc. IEEE/MTS OCEANS Conf. and Exhibition*, DC, USA, Oct. 2015, pp. 1–8.

[4] T. Lin, A. Hinduja, M. Qadri, and M. Kaess, "Conditional GANs for sonar image filtering with applications to underwater occupancy mapping," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, London, UK, May 2023, pp. 1048–1054.

[5] M. Qadri, M. Kaess, and I. Gkioulekas, "Neural implicit surface reconstruction using imaging sonar," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, London, UK, May 2023, pp. 1040–1047.

[6] M. D. Aykin and S. Negahdaripour, "Three-dimensional target reconstruction from multiple 2-D forward-scan sonar views by space carving," *IEEE J. of Oceanic Engineering (JOE)*, vol. 42, no. 3, pp. 574–589, Sep. 2016.

[7] Y. Feng, W. Lu, H. Gao, B. Nie, K. Lin, and L. Hu, "Differentiable space carving for 3D reconstruction using imaging sonar," *IEEE Robotics and Automation Letters (RAL)*, vol. 9, no. 11, pp. 10 065–10 072, Sep. 2024.

[8] P. V. Teixeira, M. Kaess, F. S. Hover, and J. J. Leonard, "Underwater inspection using sonar-based volumetric submaps," in *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Daejeon, South Korea, Oct. 2016, pp. 4288–4295.

[9] M. D. Aykin and S. Negahdaripour, "On 3-D target reconstruction from multiple 2-D forward-scan sonar views," in *Proc. IEEE/MTS OCEANS Conf. and Exhibition*, Genova, Italy, May 2015, pp. 1–10.

[10] E. Westman, I. Gkioulekas, and M. Kaess, "A theory of Fermat paths for 3D imaging sonar reconstruction," in *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Las Vegas, NV, USA, Oct. 2020, pp. 5082–5088.

[11] E. Westman and M. Kaess, "Wide aperture imaging sonar reconstruction using generative models," in *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Macau, China, Nov. 2019, pp. 8067–8074.

[12] Y. Wang, Y. Ji, H. Woo, Y. Tamura, A. Yamashita, and A. Hajime, "3D occupancy mapping framework based on acoustic camera in underwater environment," *IFAC-PapersOnLine*, vol. 51, no. 22, pp. 324–330, Dec. 2018.

[13] Y. Wang, Y. Ji, H. Woo, Y. Tamura, A. Yamashita, and H. Asama, "Three-dimensional underwater environment reconstruction with graph optimization using acoustic camera," in *In Proc. IEEE/SICE Intl. Symp. on System Integration (SII)*, Paris, France, Jan. 2019, pp. 28–33.

[14] E. Westman, I. Gkioulekas, and M. Kaess, "A volumetric albedo framework for 3D imaging sonar reconstruction," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, Paris, France, May 2020, pp. 9645–9651.

[15] R. DeBortoli, F. Li, and G. A. Hollinger, "ElevateNet: A convolutional neural network for estimating the missing dimension in 2D underwater sonar images," in *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Macau, China, Nov. 2019, pp. 8040–8047.

[16] Y. Wang, Y. Ji, D. Liu, H. Tsuchiya, A. Yamashita, and H. Asama, "Elevation angle estimation in 2D acoustic images using pseudo front view," *IEEE Robotics and Automation Letters (RAL)*, vol. 6, no. 2, pp. 1535–1542, Feb. 2021.

[17] S. Arnold and B. Wehbe, "Spatial acoustic projection for 3D imaging sonar reconstruction," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, Philadelphia, PA, USA, May 2022, pp. 3054–3060.

[18] M. Qadri, K. Zhang, A. Hinduja, M. Kaess, A. Pediredla, and C. A. Metzler, "AONeuS: A neural rendering framework for acoustic-optical sensor fusion," in *Proc. SIGGRAPH*, Denver, CO, USA, Jul. 2024, pp. 1–12.

[19] A. Reed, J. Kim, T. Blanford, A. Pediredla, D. Brown, and S. Jayasuriya, "Neural volumetric reconstruction for coherent synthetic aperture sonar," *ACM Trans. on Graphics (TOG)*, vol. 42, no. 4, pp. 1–20, Jul. 2023.

[20] Y. Xie, G. Troni, N. Bore, and J. Folkesson, "Bathymetric surveying with imaging sonar using neural volume rendering," *IEEE Robotics and Automation Letters (RAL)*, vol. 9, no. 9, pp. 8146–8153, Sep. 2024.

[21] Y. Xie, J. Zhang, N. Bore, and J. Folkesson, "NeuRSS: Enhancing AUV localization and bathymetric mapping with neural rendering for sidescan SLAM," *IEEE J. of Oceanic Engineering (JOE)*, pp. 1–10, Jan. 2025.

[22] Z. Qu, O. Vengurlekar, M. Qadri, K. Zhang, M. Kaess, C. Metzler, S. Jayasuriya, and A. Pediredla, "Z-Splat: Z-axis gaussian splatting for camera-sonar fusion," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1–12, Sep. 2024.

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

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

[25] M. Qadri, P. Sodhi, J. G. Mangelson, F. Dellaert, and M. Kaess, "InCOpt: Incremental constrained optimization using the bayes tree," in *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Kyoto, Japan, Oct. 2022, pp. 6381–6388.

[26] M. Qadri, Z. Manchester, and M. Kaess, "Learning covariances for estimation with constrained bilevel optimization," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, Yokohama, Japan, May 2024, pp. 15 951–15 957.

[27] M. Qadri and M. Kaess, "Learning observation models with incremental non-differentiable graph optimizers in the loop for robotics state estimation," in *Intl. Conf. on Machine Learning (ICML) 2023 Workshop on Differentiable Almost Everything: Differentiable Relaxations, Algorithms, Operators, and Simulators*, Jul. 2023.

[28] Y.-S. Shin, Y. Lee, H.-T. Choi, and A. Kim, "Bundle adjustment from sonar images and SLAM application for seafloor mapping," in *Proc. IEEE/MTS OCEANS Conf. and Exhibition*, DC, USA, Oct. 2015, pp. 1–6.

[29] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "KAZE features," in *Proc. Eur. Conf. on Computer Vision (ECCV)*, Florence, Italy, Oct. 2012, pp. 214–227.

[30] T. A. Huang and M. Kaess, "Towards acoustic structure from motion for imaging sonar," in *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Hamburg, Germany, Sep. 2015, pp. 758–765.

[31] E. Westman and M. Kaess, "Degeneracy-aware imaging sonar simultaneous localization and mapping," *IEEE J. of Oceanic Engineering (JOE)*, vol. 45, no. 4, pp. 1280–1294, Oct. 2020.

[32] N. Loi, Y. Z. Tan, E. W. Goh, and M. H. Ang, "Sonar SLAM in structured underwater environments," in *Proc. IEEE/MTS OCEANS Conf. and Exhibition*, Singapore, Apr. 2024, pp. 1–10.

[33] S. Xu, K. Zhang, Z. Hong, Y. Liu, and S. Wang, "DISO: Direct imaging sonar odometry," in *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, Yokohama, Japan, May 2024, pp. 8573–8579.

[34] Z. Wang, S. Wu, W. Xie, M. Chen, and V. A. Prisacariu, "NeRF--: Neural radiance fields without known camera parameters," 2022. [Online]. Available: https://arxiv.org/abs/2102.07064

[35] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, "BARF: Bundle-adjusting neural radiance fields," in *Proc. Intl. Conf. on Computer Vision (ICCV)*, Montreal, Canada, Oct. 2021, pp. 5721–5731.

[36] W. Bian, Z. Wang, K. Li, J. Bian, and V. A. Prisacariu, "NoPe-NeRF: Optimising neural radiance field with no pose prior," in *Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, Jun. 2023, pp. 4160–4169.

[37] S.-F. Chng, S. Ramasinghe, J. Sherrah, and S. Lucey, "Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation," in *Proc. Eur. Conf. on Computer Vision (ECCV)*, Tel Aviv, Israel, Oct. 2022, pp. 264–280.

[38] K. Park, P. Henzler, B. Mildenhall, J. T. Barron, and R. Martin-Brualla, "CamP: Camera preconditioning for neural radiance fields," *ACM Trans. on Graphics (TOG)*, vol. 42, no. 6, pp. 1–11, Dec. 2023.

[39] L. Yariv, Y. Kasten, D. Moran, M. Galun, M. Atzmon, B. Ronen, and Y. Lipman, "Multiview neural surface reconstruction by disentangling geometry and appearance," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, Vancouver, Canada, Dec. 2020, pp. 2492–2502.

[40] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, "Implicit geometric regularization for learning shapes," 2020. [Online]. Available: https://arxiv.org/abs/2002.10099

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

[42] Field Robotic Systems Lab (FRoStLab), Brigham Young University, "HoveringAUV - HoloOcean 1.0.0 documentation," https://byu-holoocean.github.io/holoocean-docs/UE5.3_Prerelease/agents/hovering-auv-agent.html.

[43] E. Potokar, K. Lay, K. Norman, D. Benham, S. Ashford, R. Peirce, T. Neilsen, M. Kaess, and J. Mangelson, "HoloOcean: A full-featured marine robotics simulator for perception and autonomy," *IEEE J. of Oceanic Engineering (JOE)*, vol. 49, no. 4, pp. 1322–1336, Oct. 2024.

[44] General Dynamics Mission Systems, "Bluefin HAUV," https://gdmissionsystems.com/products/underwater-vehicles/bluefin-hauv.

[45] Sound Metrics, "DIDSON 300m: OBSERVE AND CONQUER," https://blueprintsubsea.com/oculus/oculus-m-series.

[46] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. on Graphics (TOG)*, vol. 41, no. 4, pp. 1–15, Jul. 2022.