

Bathymetric Surveying with Imaging Sonar Using Neural Volume Rendering

Yiping Xie¹, Giancarlo Troni², Nils Bore³ and John Folkesson¹

Abstract—This research addresses the challenge of estimating bathymetry from imaging sonars where the state-of-the-art works have primarily relied on either supervised learning with ground-truth labels or surface rendering based on the Lambertian assumption. In this letter, we propose a novel, self-supervised framework based on volume rendering for reconstructing bathymetry using forward-looking sonar (FLS) data collected during standard surveys. We represent the seafloor as a neural heightmap encapsulated with a parametric multi-resolution hash encoding scheme and model the sonar measurements with a differentiable renderer using sonar volumetric rendering employed with hierarchical sampling techniques. Additionally, we model the horizontal and vertical beam patterns and estimate them jointly with the bathymetry. We evaluate the proposed method quantitatively on simulation and field data collected by remotely operated vehicles (ROVs) during low-altitude surveys. Results show that the proposed method outperforms the current state-of-the-art approaches that use imaging sonars for seabed mapping. We also demonstrate that the proposed approach can potentially be used to increase the resolution of a low-resolution prior map with FLS data from low-altitude surveys.

I. INTRODUCTION

Acquiring high-resolution bathymetry is one of the most fundamental and crucial applications to underwater exploration. Traditionally, multibeam echo sounders (MBES) are the de facto sensors for collecting bathymetric data, however, recently there has been an increasing interest in the research topic of reconstructing bathymetry using imaging sonars, e.g., forward-looking sonars (FLS). FLS sensors have significant capabilities in underwater environments with many advantages, compared to other sensors, making them suitable for various applications. Compared to side-scan sonars (SSS) and MBES, FLS measurements have overlap between consecutive frames, and its high-resolution imagery makes it ideal for low-altitude surveys. While comparing to optical cameras, FLS has longer range measurements and can operate in murky underwater environments with low-light visibility. The main problem is that though FLS (as with SSS) resolves range and azimuth angle, there is an ambiguity (especially for wide-aperture sonars) in the elevation angle

*This work was supported in part by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation and in part by Stiftelsen för Strategisk Forskning (SSF) through the Swedish Maritime Robotics Centre (SMaRC) under Grant IRC15-0046. (Corresponding author: Yiping Xie.)

¹ Yiping Xie and John Folkesson are with Robotics Perception and Learning Division, KTH Royal Institute of Technology, Stockholm, Sweden. (e-mail: {yipingx, johnmf}@kth.se)

² Giancarlo Troni is with Monterey Bay Aquarium Research Institute, Moss Landing, CA, USA. (email: gtroni@mbari.org)

³ Nils bore is with Ocean Infinity. (email: nils.bore@oceaninfinity.com)

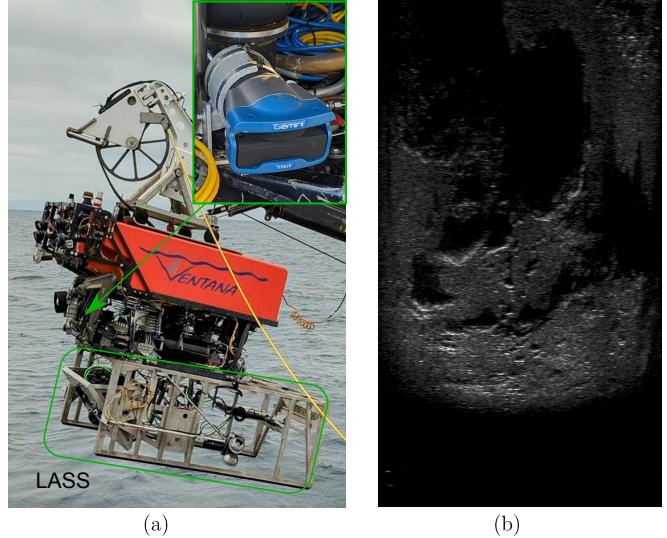


Fig. 1. (a) ROV Ventana with the Low Altitude Survey System (LASS) mounted underneath being deployed from R/V Rachel Carson. The top right corner shows the sonar to acquire the field images. (b) An example of a sonar image captured with Gemini 720is FLS.

in the sense that the return can come from anywhere along the elevation arc at given range and azimuth angle. Much research has been focusing on how to resolve the elevation ambiguity for object reconstruction, among which learning-based methods, more recently have achieved state-of-the-art results. However, existing algorithms are rarely suitable for this work’s application, which is, high-resolution, dense seabed mapping.

A lot of early works have mainly focused on feature-based, sparse 3D reconstruction [1]–[5], while other methods focusing on dense reconstruction suffer from various limitations, such as enforcing restrictions on elevation aperture [6], [7], the motion of the vehicle [8], [9], relying on information from object edges/shadows [10]–[12] or relying on volumetric grids [13]–[16] that are expensive when the scene is large with fine-grained geometry. Most of the learning-based methods require ground-truth labels [17]–[19] which are difficult to obtain in oceanic scenarios.

Very recently, self-supervised learning based on differentiable rendering and implicit neural representations has been introduced into 3D reconstruction from sonars, SSS [20], [21] and FLS [22]. This framework frees the need for labeled training datasets and the use of implicit neural representations avoids the memory overhead associated with volumetric methods. However, Qadri *et al.* [22] focus on

object reconstruction rather than seabed mapping which requires the vehicle hovering around the target to collect data. Although methods in [20], [21] are designed for seabed mapping with standard surveys, they use gradient descent algorithms to find the corresponding elevation angle of the intersection between the arc and seafloor and subsequently a Lambertian model for intensity modeling. Such a surface-rendering approach can neither model the shadows nor the layover phenomenon [23], [24], where different elevation angles are projected to the same pixel in the sonar image. Besides, all of [20]–[22] use frequency encodings in implicit neural representations, which are non-parametric and thus neither adaptive nor efficient [25], limiting their ability for high-resolution bathymetry reconstruction.

In this work, we address these shortcomings by proposing a differentiable rendering-based framework (Fig. 2) that leverages the power of multi-resolution hash encodings [25], tailored to bathymetry reconstruction from FLS data during a standard survey with a "lawn-mower" pattern¹. The contributions of this work are as follows:

- We propose to use parametric multi-resolution hash encodings and a hierarchy sampling technique along the elevation arc that advance the efficiency of the volumetric renderer with low overhead, fast convergence and quick feedforward speed.
- We model the beam pattern in both horizontal and vertical directions that can be jointly learned with bathymetry.
- We evaluate the proposed method on both simulation and field data to illustrate the improvement over the state-of-the-art works. The implementation and simulation dataset are available in our Github²
- We demonstrate how the proposed framework can combine FLS and MBES data for super-resolution mapping, leveraging the advantages of both sensors.

II. RELATED WORKS

As aforementioned, different sparse 3D reconstructions [1]–[5] based on manually selected features [1], [2], AKAZE features [3], [5] and learned features [4] have been introduced to FLS images and many of them focus more on simultaneous localization and mapping (SLAM) [1], [3]–[5] rather than on 3D reconstructions.

As for dense 3D reconstructions, a number of works place restrictions or assumptions on the physical setup. Teixeira *et al.* [6], [7] use a FLS with 1° elevation aperture to reconstruct a 3D model of a ship hull, prohibiting the method from extending to wide-aperture sonars. Westman *et al.* [8] propose a non-light-of-sight (NLOS) method for 3D reconstruction on sonar based on Fermat paths, the use of which is considered to be an effective solution for the layover phenomenon. However, the method requires view rays perpendicular to the surface, which is impractical for bathymetry reconstruction.

¹A lawn-mower pattern that has the vehicle perform the survey as a series of long parallel lines.

²The source code link will be provided here upon acceptance.

Wang *et al.* [9] propose a method utilizing the motion field to estimate the missing elevation angle in a self-supervised learning framework, however, it places restrictions on the motion of the vehicle to avoid failures caused by degenerate motions.

A different group of methods is based on shape-from-shading (SFS) techniques with generative models (Lambertian) assuming diffuse reflection. For object-based reconstruction, [10]–[12] require estimates of object edges and/or shadows, making them unsuitable in real oceanic scenarios. For bathymetry reconstruction, neural FSF methods in [20], [21] propose to use Sinusoidal Representation Networks (SIRENs) [26] for the bathymetry representation and fit the sonar data from standard surveys through a gradient-based optimization. However, the use of the Lambertian model cannot explain the shadows in the sonar images thus introducing modeling errors especially during low-altitude surveys. As aforementioned, [20], [21] also assume there is only one elevation angle for each corresponding pixel in the sonar images, which ignores the layover phenomenon that happens commonly when the scene is complex such as surveying areas with boulders and rocks on the seafloor. Besides, the non-parametric frequency encoding used in SIRENs makes the representation inefficient, resulting in computational time and memory usage overhead.

Various supervised learning-based methods have also been proposed recently. DeBortoli *et al.* [17] propose to train a Convolutional Neural Network (CNN) on synthetic datasets and then fine-tune it on real datasets in a self-learning scheme, however the generalization ability and the sim-to-real gap limits such method's performance on field data. Xie *et al.* [27]–[29] use MBES data to create ground truth to directly train a CNN on field data using supervised learning and show that the trained CNN can generalize to some extent with the same sensor setup. But the methods require collecting MBES to create a large training set every time the sensor setup is different. Similarly, Wang *et al.* [18], [19] have found that transferring acoustic view to pseudo front view helps the performance by considering the layover phenomenon, but such methods still require collecting real datasets with ground truth.

Notably, inspired by NeuS [30], Qadri *et al.* [22] propose to replace the camera model with a sonar model and reconstruct a single object from multi-view FLS images, where a neural signed distance field (SDF) is used to represent the object. However, they rely on threshold-based filtering on the intensities to address the floater artifacts, similar to a background mask used in NeuS [30]. Besides, they suffer from the same limitations as [20] and [21], brought by frequency encodings in implicit neural representations. As for sampling along the arc, the lack of importance sampling makes it inefficient, especially when modeling wide-aperture sonars.

Implicit neural representations and differentiable rendering have gained much interest for 3D reconstruction, where NeRF [31] made a breakthrough by using volumetric rendering to learn a neural radiance and density field, achieving

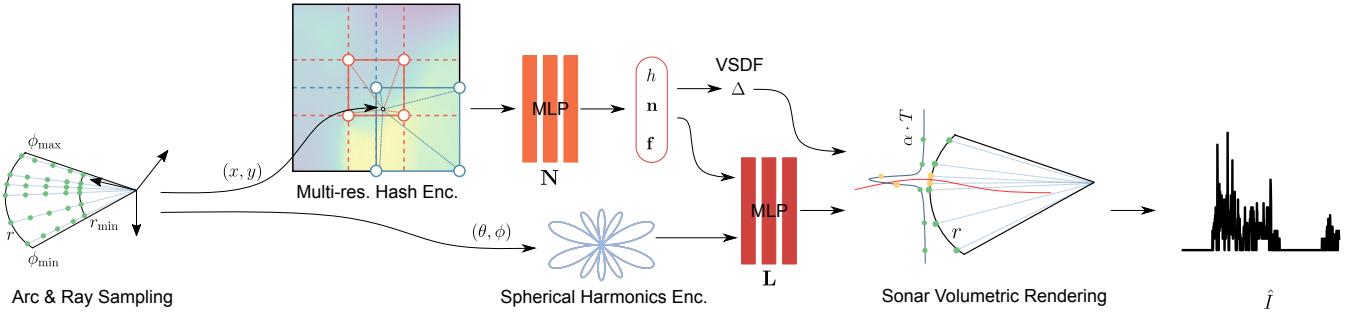


Fig. 2. System Overview: Given a range r and azimuth angle θ , we sample along the arc and for every sampled point on the arc, we sample along the acoustic ray. The spatial coordinates (x, y) are encoded and then fed into the neural heightmap \mathbf{N} outputting the height h , while at the same time the viewing angle (θ, ϕ) is encoded using a spherical harmonics basis. The rendering network \mathbf{L} takes spatial information, learned features \mathbf{f} , encoded viewing angles and surface normals \mathbf{n} to learn the radiance field, which is later used together with the vertical signed distance Δ to apply sonar volumetric rendering to predict the returned intensity.

state-of-the-art results. Wang *et al.* [30] propose to learn a neural SDF instead of the radiance and density field so that explicit surface constraints such as Eikonal loss can be added to achieve a high-quality surface reconstruction. Instead of using frequency encodings, which are non-parametric as in NeRF [31] and NeuS [30], Instant-NGP [25] proposes a multi-resolution hash encoding where a multi-resolution hash table of feature vectors is jointly trained with a smaller multi-layer perception (MLP), which increases the efficiency by reducing memory consumption and computational time.

In this work, leveraging the power of multi-resolution hash encodings, we represent the bathymetry with a neural heightmap, which suits better and is more efficient for the application of bathymetry reconstruction than volume-based shape representations. A volumetric differentiable renderer is used to render sonar intensities at a given pose, range and azimuth angle, where a hierarchy sampling scheme is used rather than stratified sampling to improve efficiency. With a kernel-based beam pattern modeling for both vertical and horizontal directions, the difference between rendered and measured intensities is used to optimize the bathymetry and beam patterns jointly in a self-supervised manner.

III. NEURAL VOLUME RENDERING FOR RECONSTRUCTION

A. FLS Model

A multi-beam FLS's geometry can be described as in Fig. 3, where each transducer in the azimuth direction θ emits a 2D fan-shaped beam and records the time-of-flight (TOF) of the sound waves and the returned back-scattered intensities, I . Each beam is very wide in the elevation direction ϕ , where all the returns from the elevation arc ($\phi_{\min} \leq \phi \leq \phi_{\max}$) are projected onto the zero-elevation plane. This results in not only the ambiguity of the elevation angle for a given pixel corresponding with range r , azimuth angle θ with intensity I but also the layover phenomenon where one bin consists of the returns from multiple 3D points along the elevation arc, if the geometry is complex. The

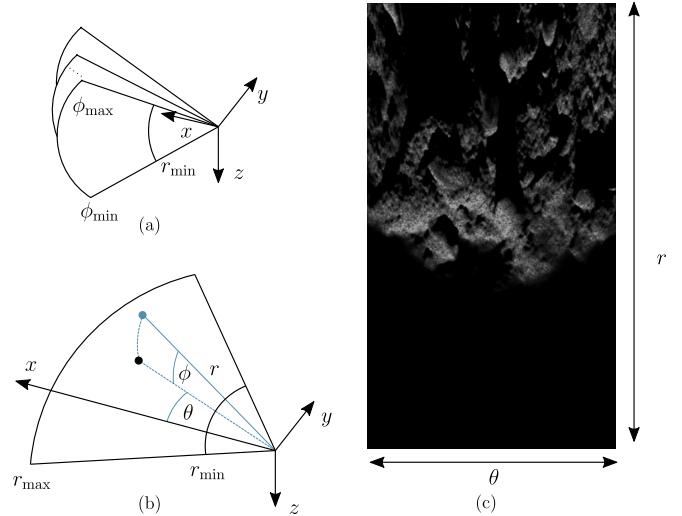


Fig. 3. (a) Each image column θ_i corresponds to a 2D fan-shape beam emitted from a transducer, recording returns between r_{\min} and r_{\max} . (b) Geometry model of the FLS, where the blue point (r, θ, ϕ) is a 3D point that is projected onto the image plane $z = 0$. (c) An example of a sonar image in simulation. Each pixel at (r, θ) contains the returned intensities of all points along the elevation arc.

returned intensity can be expressed as

$$I(r, \theta) = \int_{\phi_{\min}}^{\phi_{\max}} \beta(\theta, \phi) \sigma(r, \theta, \phi) T(r, \theta, \phi) L(r, \theta, \phi, \mathbf{v}) d\phi, \quad (1)$$

where T is the transmittance, σ is the particle volume density, similarly as in [30]. L is the radiance at a given 3D position (r, θ, ϕ) , from viewing direction \mathbf{v} , which depends on θ, ϕ and sensor position. $\beta(\theta, \phi)$ models the beam pattern at given θ and ϕ .

B. Efficient Neural Representations

Similar to [30] and [22], we represent the surface and radiance field using two MLPs, except that for the surface, we leverage a neural heightmap $\mathbf{N} : \mathbb{R}^2 \rightarrow \mathbb{R}$ that maps a 2D Cartesian position in world frame to its heights, $h = \mathbf{N}(x_w, y_w)$, instead of neural SDF to represent the bathymetry for the sake of efficiency. Instead of using

non-parametric, frequency encoding as in [22], we apply multi-resolution hash encoding [25] to the spatial positions. Specifically, L levels (two of which are shown as red and blue in Fig. 2) of grids store trainable encoding parameters at the vertices of the grid. For a spatial position, the features at the corners are looked up from the corresponding hash tables and then linearly interpolated. The feature vectors for different levels are concatenated before feeding into the MLP in \mathbf{N} . For the neural radiance field \mathbf{L} , we encode the radiance associated with a point $\mathbf{x} \in \mathbb{R}^3$ and a viewing direction $\mathbf{v} \in \mathbb{S}^2$ given its vertical signed distance, surface normal and encoding feature vector. Note that, given a 3D point (r, θ, ϕ) in polar coordinates in the local sonar frame, we need to transform it to Cartesian coordinates in the global reference frame $\mathbf{P}_p = [x_w, y_w, z_w]$ before calculating the vertical signed distance and normal. The transformation from polar coordinates to Cartesian coordinates is as follows:

$$\begin{aligned} x_s &= r \cos(\theta) \cos(\phi) \\ y_s &= r \sin(\theta) \cos(\phi) \\ z_s &= r \sin(\phi), \end{aligned} \quad (2)$$

and the transformation from local frame to world frame is

$$[x_w, y_w, z_w]^T = \mathbf{R}_W^s [x_s, y_s, z_s]^T + \mathbf{t}_W^s, \quad (3)$$

given rotation matrix \mathbf{R}_W^s and translation \mathbf{t}_W^s .

The vertical signed distance is simply $\Delta := z_w - \mathbf{N}(x_w, y_w)$ and the normal is calculated from the two gradient components ∇_x, ∇_y of Δ :

$$\mathbf{n} = [-\nabla_x \mathbf{N}(x_w, y_w), -\nabla_y \mathbf{N}(x_w, y_w), 1]^T. \quad (4)$$

C. Hierarchical Sampling

As pointed out in [22], to render sonar images, other than sampling points along the acoustic ray for corresponding pixels to compute T and σ in Eq. (1), we also need to sample points along the elevation arc since we do not know where are the intersections between the arc and the surface. In this work, we use a hierarchical sampling strategy (illustrated in Fig. 4) when sampling along the arc by first stratified sampling $N_{A,s}$ points along the elevation arc A_p and then conducting importance sampling to obtain another $N_{A,i}$ elevation angles based on the coarse probability estimation, which is computed using the S-density $\phi_s(\Delta) = se^{-s\Delta}/(1+e^{-s\Delta})^2$ with standard deviations s . To compute the S-density, we need to feedforward these $N_{A,s}$ samples to the neural heightmap \mathbf{N} , which increases computation overhead slightly. Note that, the S-density function is the derivative of the sigmoid function $\Phi_s(\Delta) = (1+e^{-s\Delta})^{-1}$. For the sampling along the ray, we follow a similar strategy as [22], where the samples along the ray R_p contain exactly one point on the arc (at the end of the ray) and $N_R - 1$ points uniformly sampled along the ray.

D. Beam Pattern

To model the beam pattern, a characteristic property of the sensor, we follow the idea in [20], [21] except that we estimate for both vertical and horizontal direction, $\beta(\theta, \phi) =$

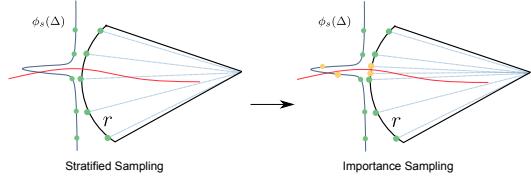


Fig. 4. Illustration of hierarchical sampling along the arc. Assuming the red curve is the seafloor, first $N_{A,s} = 5$ stratified samples (in green) are drawn on the arc and subsequently $N_{A,i} = 2$ importance samples (in yellow) are obtained using the S-density of the 5 green samples. The whole 7 samples are used to construct acoustic rays where we further sample along the rays to compute opacity and transmittance of the 7 samples along the arc.

$\beta_\Theta(\theta)\beta_\Phi(\phi)$. We use kernel densities whose kernel weights Θ_K, Φ_K are estimated jointly during the optimization. We set kernels at fixed positions θ_{K_1}, ϕ_{K_1} evenly spread across the range:

$$\beta_\Theta(\theta) = \sum_{K_1} \Theta_{K_1} \exp\left(-\frac{(\theta_{K_1} - \theta)^2}{2\sigma_\Theta^2}\right), \quad (5)$$

$$\beta_\Phi(\phi) = \sum_{K_2} \Phi_{K_2} \exp\left(-\frac{(\phi_{K_2} - \phi)^2}{2\sigma_\Phi^2}\right), \quad (6)$$

where $\sigma_\Theta, \sigma_\Phi$ are the spreads of the kernels, set to be the range divided by the number of kernels K_1, K_2 .

E. Sonar Volumetric Rendering

The discrete form of intensity modeling in Eq. (1) is:

$$\hat{I}(r, \theta) = \sum_{\mathbf{P}_p \in \mathcal{A}_p} \beta(\mathbf{P}_p) T(\mathbf{P}_p) \alpha(\mathbf{P}_p) \mathbf{L}(\mathbf{P}_p), \quad (7)$$

where \mathcal{A}_p is the elevation arc located at (r, θ) , $\mathbf{L}(\mathbf{P}_p)$ is the predicted intensity at \mathbf{P}_p predicted by the neural renderer \mathbf{L} , $\alpha(\mathbf{P}_p)$ is the discrete opacity at \mathbf{P}_p which can be computed following [30]:

$$\alpha(\mathbf{p}_i) = \max\left(\frac{\Phi_s(\Delta(\mathbf{p}_i)) - \Phi_s(\Delta(\mathbf{p}_{i+1}))}{\Phi_s(\Delta(\mathbf{p}_i))}, 0\right), \quad (8)$$

where $\mathbf{p}_i, \mathbf{p}_{i+1}$ are consecutive samples along the acoustic ray. Finally we have the discrete transmittance at the endpoint of the ray \mathbf{P}_p shown to equal [22]:

$$T(\mathbf{P}_p) = \prod_{\mathbf{p} \in \mathcal{R}_{\mathbf{P}_p} \setminus \mathbf{P}_p} (1 - \alpha(\mathbf{p})), \quad (9)$$

where $\mathcal{R}_{\mathbf{P}_p}$ is the acoustic ray ends at \mathbf{P}_p and the transmittance is computed using the set of points on that ray excluding the endpoint.

F. Loss Functions

The loss function consists of two terms, the intensity loss

$$\mathcal{L}_{\text{int}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \left\| \hat{I}(p) - I(p) \right\|_1 \quad (10)$$

and the regularization term

$$\mathcal{L}_{\text{reg}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} (\|\mathbf{n}(p)\|_2 - 1)^2 \quad (11)$$

TABLE I
DATASET DETAILS

Dataset	Aerial Rocks	Sponge Ridge 1	Sponge Ridge 2
Vehicle	Simulator	Ventana+LASS	Ventana
Avg altitude (m)	4.84	3.67	3.78
Image res.	919×512	1887×512	920×512
Survey area	~70m×70m	~80m×80m	~80m×40m
No. of images	42354	12851	9302
FLS range (m)	30	31	30
Total traj. (m)	408.75	387.63	365.05
Duration (min)	71 (10Hz)	38 (5Hz)	10 (15Hz)

to encourage surface smoothness. \mathcal{P} is the set of sampled pixels in a mini-batch. Optionally an altimeter loss could be added to accelerate convergence:

$$\mathcal{L}_{\text{alt}} = \frac{1}{|\mathcal{P}_{\text{alt}}|} \sum_{p \in \mathcal{P}_{\text{alt}}} \|\Delta(p)\|_1, \quad (12)$$

where \mathcal{P}_{alt} is the set of single-beam altimeter readings along the ROV transit. The total loss is a weighted sum of the loss terms above.

IV. EXPERIMENTS

In this section, we evaluate the performance of the proposed method on both simulation and field datasets from low-altitude surveys. We demonstrate two applications, namely bathymetry reconstruction from FLS and super-resolution bathymetric mapping from FLS with the aid of a low-resolution prior map. For the former application, we use the simulation dataset and one of the field datasets, Sponge Ridge 1. We compare the proposed model against three baselines: 1) *Lambert.+GD*: a Lambertian model is used for the intensity modeling and gradient descent (GD) to compute the one intersection between the elevation arc and the seabed as in [20], [21]. A threshold set empirically is used to remove shadows in the loss computation. 2) *Lambert.+Sampl.*: the same Lambertian model for the sonar scattering modeling but only using stratify sampling strategy as in [22] is used. 3) *Freq.+Sampl.*: the full model in [22] with the exception that we use a neural heightmap instead of a full 3D SDF. For the super-resolution mapping application, we use the other field dataset, Spondge Ridge 2 where we have a prior bathymetric map from an AUV survey at a much higher altitude. We use the gridded bathymetric data from the low-resolution map (1 m) in the altimeter loss to directly supervise the neural heightmap and demonstrate that with the aid of FLS from low-altitude surveys, we can create a super-resolution (5 cm) bathymetry with high fidelity.

A. Data

1) *Simulation Data: Aerial Rocks*: The simulation dataset was generated using the Stonefish [32] where the seabed terrain is based on public heightmaps to simulate a rocky area. The ROV was programmed to survey in a variation of the common lawn-mower trajectory [33] (seen Fig. 5) with specifications shown in Table I.

2) *Field Data*: Two field datasets were collected using the ROV Ventana aboard the R/V Rachel Carson from the Monterey Bay Aquarium Research Institute (MBARI), located in Moss Landing, California. Sponge Ridge 1 was collected with the Low Altitude Survey System (LASS) mounted on ROV Ventana during an expedition in 2023 in Monterey Bay. The navigation was controlled by a scripted mission plan using a 10 m line spacing. A Kefratt SeaDevil inertial navigation system (INS) aided by a RDI Workhorse Doppler Velocity Log (DVL) is used for navigation estimation. Data from the Reson SeaBat 7125 MBES on LASS is used to provide 10 cm resolution map as ground truth. Sponge Ridge 2 was collected with only ROV Ventana during an expedition in 2022 in Monterey Bay, where a pilot was flying the ROV. An Octans Fiber Optic Gyrode and a RDI Workhouse DVL combo is used to estimate the 6 degree-of-freedom (DOF) pose of the ROV, which is treated as the ground truth for navigation. A prior map with 1 m resolution collected with MBARI mapping AUV with MBES is available as a reference.

For both datasets, a GPS receiver is used to initialize the navigation prior to deployment and ultra-short baseline (USBL) tracking is used to stabilize the navigation fix during ROV descent. Both datasets were collected using a Gemini 720is FLS with a slightly different setup, as shown in Table I. Both simulation and field sonar have a horizontal field of view (HFOV) of 120° and vertical field of view (VFOV) of 20°. For both simulation and field data, the ROV were running low-altitude surveys, as shown in Table I.

B. Metrics

To evaluate the performance, we use the mean absolute errors (MAE) and the standard deviation (STD) of the signed errors between the reconstructed bathymetry and ground truth, both in meters. Additionally, we also computed the structural similarity index measure (SSIM) between the reconstructed bathymetry and ground truth by treating the heightmaps as 16-bit gray-scale images. SSIM, [0, 1], is a metric for evaluating the quality of the reconstructed images, with 1 indicating the same as the reference image.

C. Implementation Details

For simulation and field datasets, we use an MLP with 2 hidden layers of size 64 with multi-resolution hash encoding to model N. The hash table size is 2^{15} ($L = 15$) and the maximum resolution $N_{\max} = 1024$. For L, we again use a 2-hidden-layer MLP of 64 neurons wide but with sphere harmonics encoding (up to 3 degrees) for the view directions. For the baselines that use frequency encodings, we use 6 frequencies and keep the same sphere harmonics encoding to view directions with up to 3 degrees. For simulation experiments, we did not estimate the beam pattern but for experiments on field datasets, we use $\mathcal{K}_1 = 30$ kernels over 120° for the horizontal beam pattern modeling and $\mathcal{K}_2 = 10$ kernels over 40° for vertical beam pattern, given the 3dB beam width of Gemini 720is is 20°. As for the hierarchical sampling, for simulation dataset, we assume 20°

TABLE II
SIMULATION QUANTITATIVE RESULTS

Model	Area	70×70 [m]			50×50 [m]		
		MAE ↓	STD ↓	SSIM ↑	MAE ↓	STD ↓	SSIM ↑
Lambert.+GD		0.305	0.209	0.875	0.254	0.209	0.863
Lambert.+Sampl.		0.385	0.355	0.799	0.241	0.212	0.812
Freq.+Sampl.		0.282	0.235	0.813	0.210	0.200	0.818
Ours		0.240	0.187	0.905	0.133	0.115	0.906

vertical sensor opening and we first sample $N_{A,s} = 15$ points along the elevation arc and then another $N_{A,i} = 15$ importance samples. For both field datasets, we assume 40° sensor opening with $N_{A,s} = 30$ and $N_{A,i} = 30$. We set $N_R = 60$ for samples along each ray in all experiments. All models are trained for 60 k steps with Adam optimizer with a learning rate 5×10^{-2} that linearly decays by a factor of 0.97 every 600 steps. For each mini-batch we randomly select one beam out of 512. To enforce and accelerate convergence, we conduct Progressive Training strategy as in [34] and we also use the altimeter loss. The altimeter readings are simulated for the simulation dataset and as for the Sponge Ridge 1 dataset, we use the actual readings from the onboard DVL. For the super-resolution mapping experiment, we convert the gridded low-resolution prior map to point clouds, simulating dense altimeter readings.

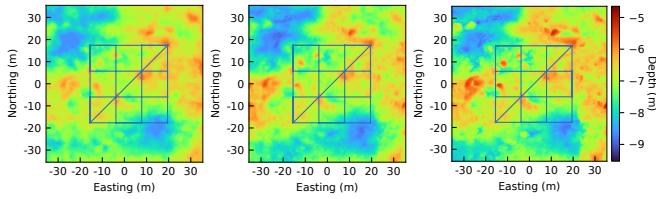


Fig. 5. Reconstructed bathymetry for *Freq.+Sampl.* (left) and for the proposed method (middle) and ground truth (right) for simulation dataset, at 10 cm resolution (ROV trajectory in blue).

V. RESULTS

A. Bathymetry Reconstruction

1) *Simulation dataset results:* The MAE, STD and SSIM for the simulation datasets computed for all aforementioned models are shown in Table II. In addition to the metrics in the survey area (70×70 m) we also compute the metrics in the inner area (50×50 m). Of all the models, the proposed method has the best performance in all categories, demonstrating the importance of relaxation on the Lambertian assumption as well as the advantages of parametric encodings that could automatically focus on relevant details with similar amount of trainable parameters (MLP weights + encoding parameters). Note that across all models, the errors in the inner area are lower than the whole survey area, which are congruent with our expectations since at the outer area, the number of observations from the sonar is lower, resulting in weaker constraints. Furthermore, around the perimeter, the increased distance between the seabed and the sonar, results in higher noise levels and lower resolution, inevitably leading to lower quality reconstruction. This creates a trade-off between efficiency (survey coverage) and reconstruction

TABLE III
FIELD RESULTS - ABSOLUTE ERROR

Model	Area	80×80 [m]		60×60 [m]	
		MAE ↓	STD ↓	MAE ↓	STD ↓
Lambert.+GD		0.922	0.838	0.808	0.785
Lambert.+Sampl.		0.810	0.593	0.738	0.570
Freq.+Sampl.		0.847	0.635	0.714	0.582
Ours ⁻		0.727	0.604	0.643	0.515
Ours		0.531	0.495	0.450	0.464

quality. The predicted bathymetry at 10 cm resolution from baseline *Freq.+Sampl.* (left) and our method (middle) together with the ground truth (right) are shown in Fig. 5. The plot shows that the baseline method manages to reconstruct the topographic details of the surveyed rocky area fairly well but our approach is able to capture finer geometric details with higher fidelity, providing a sharper heightmap. This is consistent with the SSIM metric where ours (0.905) has a 11.3% improvement compared to the baseline (0.813). Furthermore, we can observe from Fig. 5 that the outer area, for example at the upper left corner and lower right corner, the baseline *Freq.+Sampl.* loses much more details than the proposed method, mainly because of the limited non-parametric frequency encodings.

2) *Sponge Ridge 1:* Table III shows the MAE and STD for the Sponge Ridge 1 dataset. All models except *Ours⁻* use the altimeter data from the onboard DVL for bathymetric constraints and we estimate the beam pattern the same way across all models. For the setup in *Ours⁻*, we only used intensity and regularization loss. SSIM is not computed since we do not have full-coverage ground truth bathymetry, as shown in Fig. 6 (right). Again the best results are obtained from the proposed method, with 0.531 m MAE for the whole survey area and 0.450 m at the inner area, demonstrating the proposed method's advantages over the baselines. Furthermore, compared to Lambertian-based methods, our approach relaxes the pure diffusion reflectance assumption and is able to better model the sonar ensonification process. Note that with the proposed method without altimeter loss, i.e., *Ours⁻*, the MAE is still lower than all baselines. Fig. 6 shows the reconstructed bathymetry from the proposed approach gridded at 10 cm resolution, showing that the shape of the reconstructed ridges. As one can notice that the reconstructed quality of the topology deteriorates as it gets further away from sonar. This is more noticeable than the case in simulation due to the higher level noise in field data. Note that we masked out the places that have been ensonified less than twice, which are generally at the perimeter. The plot and the table demonstrate the quality of the bathymetry we could obtain from field FLS using the proposed approach.

Fig. 7 shows the beam pattern estimated together with the bathymetry during the optimization. The estimated vertical beam pattern [Fig. 7(a)] is modelled from $\phi_{\max} = -5^\circ$ to $\phi_{\min} = -45^\circ$, where from the plot we can see the 3dB beam width is roughly 20°, which agrees with the datasheet of Gemini 720is. Fig. 7(b) shows the estimated horizontal beam pattern and the “calibration”, which is obtained by

averaging all intensities over azimuth angles for the entire dataset, excluding the water column parts. This is a crude approximation of the calibration, which can be used to qualitatively evaluate the estimated beam pattern.

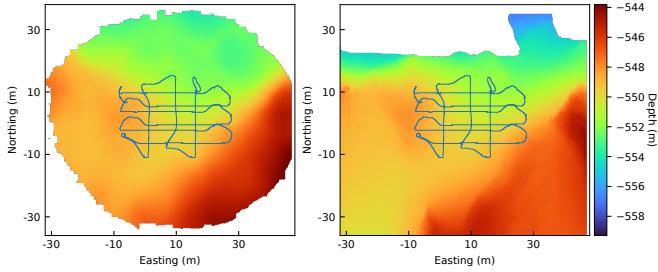


Fig. 6. Reconstructed bathymetry for the proposed method (left) and ground truth (right) for Sponge Ridge 1 dataset collected with LASS, at 10 cm resolution (ROV trajectory in blue).

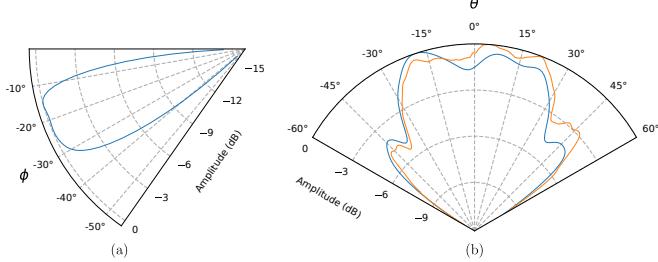


Fig. 7. Normalized beam pattern. (a) Estimated beam pattern in vertical direction. (b) Beam pattern in horizontal direction, where the blue line is estimated during the optimization and the orange line is the average intensities over azimuth angles across all FLS images in the dataset, as a crude reference.

B. Super Resolution Mapping

Fig. 8 shows the estimated bathymetry gridded at 5 cm resolution (b) and the prior map with 1 m resolution (a). It is difficult to see the effect of super resolution from heightmaps directly, other than the prior map is more pixelated compared to our estimation. To demonstrate the details of the super resolution bathymetry, we also compute the gradients of the heightmaps [Fig. 8 (c)-(e)]. Note that since the estimated bathymetry is represented by an MLP, we can compute the exact gradients. As for the gradient map for the prior map, we use finite difference to approximate the gradients after interpolating the map to 5 cm resolution. The gradient maps reveal that the reconstructed bathymetry not only maintains the shape of the general structure but also shows many details such as small rocks and ripples on the seafloor. The plot reflects that the high resolution FLS images due to its high frequency and low altitudes provide rich information on the micro-topology of the seabed which is not available from MBES data from AUVs flying at higher altitudes. Note that Fig. 8 (d) illustrates that if we directly fit the prior map to the MLP, we could leverage the continuous property of implicit neural representations and obtain a heightmap with sharper details [(d)] than interpolating the gridded data [(c)]. However, rocks that are smaller than MBES's footprint can only be resolved with the aid of FLS data, as shown in (e).

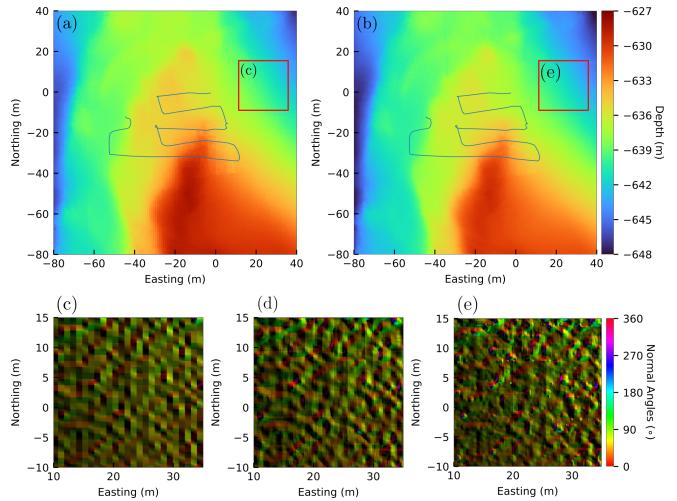


Fig. 8. Bathymetry (top) for the low-resolution prior map (a) and the reconstructed super-resolution map using proposed method gridded at 5 cm resolution (b) for Sponge Ridge 2 dataset collected with Ventana (ROV trajectory in blue). At the bottom we zoom in and show the gradients of the map (the red square) for interpolation of prior map in (c), training NNs only with point clouds from the prior map in (d), and training NNs with FLS and point clouds from the prior map in (e).

VI. CONCLUSIONS AND FUTURE WORK

We have presented a learning-based approach based on neural rendering for reconstructing bathymetry from FLS data in a self-supervising manner. The proposed framework has been tested on three surveys, two of them collected with field robots. When compared to the current state-of-the-art solutions for bathymetry reconstruction from imaging sonars, our method is capable of estimating high-resolution bathymetry with lower errors and higher fidelity, thanks to the novel contributions proposed in our approach. The results have shown that the proposed method could be applied to FLS field data from standard surveys for bathymetric mapping. Additionally the super-resolution experiment shows the application of using FLS from ROVs or small AUVs in low-altitude surveys to increase the resolution of bathymetry from a higher altitude survey, from for example, large AUVs or surface vessels.

The current major limitation of the proposed method is that we assume to have very accurate navigation estimates, which rarely holds in long missions if only relying on dead reckoning (DR). The unbounded drift over time will limit the quality of the reconstructing map. More research focusing on FLS SLAM could help to reduce the drift in the DR estimates, which could be used prior to applying the proposed method for seabed mapping. Another possible future work is to incorporate the neural rendering-based mapping into the FLS SLAM framework, for example, commonly a tracking thread to estimate the $SE(3)$ pose of the sonar and a parallel mapping thread to build the bathymetric maps only using key frames. Another possible future work is that, if a prior map is available, how to use FLS for localization and substantially creating a super-resolution bathymetry.

Another limitation is that our method is mostly suited to

offline optimization. One of the key reasons is that we need to sample along the elevation arc, resulting in more samples being fed into the MLPs and subsequently their gradient computation. This increases the training time compared to works using camera images for 3D reconstruction in real time [25]. However, recent advances based on rasterization rather than ray casting such as 3D Gaussian Splatting [35] are significantly faster than methods based on [25], and can potentially be used for FLS mapping or even SLAM [36] in real time.

ACKNOWLEDGMENT

The authors gratefully thank Bastián Muñoz for creating the simulation data. The authors also acknowledge the Monterey Bay Aquarium Research Institute (MBARI) engineering and technical team, especially Dave Caress, Kent Headley, Eric Martin, Kevin Barnard and Sebastián Rodríguez, the R/V Rachel Carson officers and crew for their support in collecting the field data. This work was a contribution of the CoMPAS laboratory at MBARI and was supported by the David and Lucile Packard Foundation. The network training was partly enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

REFERENCES

- [1] T. A. Huang and M. Kaess, “Incremental data association for acoustic structure from motion,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2016, pp. 1334–1341.
- [2] N. T. Mai, H. Woo, Y. Ji, Y. Tamura, A. Yamashita, and H. Asama, “3-D reconstruction of underwater object based on extended Kalman filter by using acoustic camera images,” *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 1043–1049, 2017.
- [3] E. Westman, A. Hinduja, and M. Kaess, “Feature-based SLAM for imaging sonar with under-constrained landmarks,” in *Proc. IEEE Int. Conf. Robot. Autom.* IEEE, 2018, pp. 3629–3636.
- [4] J. Li, M. Kaess, R. M. Eustice, and M. Johnson-Roberson, “Pose-graph SLAM using forward-looking sonar,” *IEEE Robot. Automat. Lett.*, vol. 3, no. 3, pp. 2330–2337, 2018.
- [5] E. Westman and M. Kaess, “Degeneracy-aware imaging sonar simultaneous localization and mapping,” *IEEE J. Ocean. Eng.*, vol. 45, no. 4, pp. 1280–1294, 2019.
- [6] P. V. Teixeira, M. Kaess, F. S. Hover, and J. J. Leonard, “Underwater inspection using sonar-based volumetric submaps,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2016, pp. 4288–4295.
- [7] P. V. Teixeira, D. Fourie, M. Kaess, and J. J. Leonard, “Dense, sonar-based reconstruction of underwater scenes,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2019, pp. 8060–8066.
- [8] E. Westman, I. Gkioulekas, and M. Kaess, “A theory of fermat paths for 3D imaging sonar reconstruction,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2020, pp. 5082–5088.
- [9] Y. Wang, Y. Ji, C. Wu, H. Tsuchiya, H. Asama, and A. Yamashita, “Motion degeneracy in self-supervised learning of elevation angle estimation for 2D forward-looking sonar,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2023, pp. 6133–6140.
- [10] M. D. Aykin and S. Negahdaripour, “On 3-D target reconstruction from multiple 2-D forward-scan sonar views,” in *Proc. IEEE OCEANS Conf.* IEEE, 2015, pp. 1–10.
- [11] M. D. Aykin and S. S. Negahdaripour, “Modeling 2-D lens-based forward-scan sonar imagery for targets with diffuse reflectance,” *IEEE J. Ocean. Eng.*, vol. 41, no. 3, pp. 569–582, 2016.
- [12] E. Westman and M. Kaess, “Wide aperture imaging sonar reconstruction using generative models,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2019, pp. 8067–8074.
- [13] T. Guerneve and Y. Petillot, “Underwater 3D reconstruction using BlueView imaging sonar,” in *Proc. IEEE OCEANS Conf.* IEEE, 2015, pp. 1–7.
- [14] T. Guerneve, K. Subr, and Y. Petillot, “Three-dimensional reconstruction of underwater objects using wide-aperture imaging sonar,” *J. Field Robot.*, vol. 35, no. 6, pp. 890–905, 2018.
- [15] 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, 2018.
- [16] E. Westman, I. Gkioulekas, and M. Kaess, “A volumetric albedo framework for 3D imaging sonar reconstruction,” in *Proc. IEEE Int. Conf. Robot. Autom.* IEEE, 2020, pp. 9645–9651.
- [17] 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 Int. Conf. Intell. Robots Syst.* IEEE, 2019, pp. 8040–8047.
- [18] 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 Robot. Automat. Lett.*, vol. 6, no. 2, pp. 1535–1542, 2021.
- [19] Y. Wang, Y. Ji, H. Tsuchiya, H. Asama, and A. Yamashita, “Learning pseudo front depth for 2D forward-looking sonar-based multi-view stereo,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2022, pp. 8730–8737.
- [20] N. Bore and J. Folkesson, “Neural shape-from-shading for survey-scale self-consistent bathymetry from sidescan,” *IEEE J. Ocean. Eng.*, vol. 48, no. 2, pp. 416–430, 2023.
- [21] Y. Xie, N. Bore, and J. Folkesson, “Sidescan only neural bathymetry from large-scale survey,” *Sensors*, vol. 22, no. 14, p. 5092, 2022.
- [22] M. Qadri, M. Kaess, and I. Gkioulekas, “Neural implicit surface reconstruction using imaging sonar,” in *Proc. IEEE Int. Conf. Robot. Autom.* IEEE, 2023, pp. 1040–1047.
- [23] P. Wook and C. Frey, “Deep-sea AUV navigation using side-scan sonar images and SLAM,” in *Proc. IEEE OCEANS Conf.* IEEE, 2010, pp. 1–8.
- [24] Y. Xie, N. Bore, and J. Folkesson, “Towards differentiable rendering for sidescan sonar imagery,” in *Proc. IEEE/OES Auton. Underwater Veh. Symp.* IEEE, 2022, pp. 1–6.
- [25] T. Müller, A. Evans, C. Schied, and A. Keller, “Instant neural graphics primitives with a multiresolution hash encoding,” *ACM Trans. Graph.*, vol. 41, no. 4, pp. 1–15, 2022.
- [26] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, “Implicit neural representations with periodic activation functions,” *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020.
- [27] Y. Xie, N. Bore, and J. Folkesson, “Inferring depth contours from sidescan sonar using convolutional neural nets,” *IET Radar, Sonar & Navigation*, vol. 14, no. 2, pp. 328–334, 2020.
- [28] ———, “Bathymetric reconstruction from sidescan sonar with deep neural networks,” *IEEE J. Ocean. Eng.*, vol. 48, no. 2, pp. 372–383, 2022.
- [29] ———, “Neural network normal estimation and bathymetry reconstruction from sidescan sonar,” *IEEE J. Ocean. Eng.*, vol. 48, no. 1, pp. 218–232, 2022.
- [30] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, “Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction,” *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 27171–27183, 2021.
- [31] 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,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 405–421.
- [32] P. Cieślak, “Stonfish: An advanced open-source simulation tool designed for marine robotics, with a ros interface,” in *Proc. IEEE OCEANS Conf.* IEEE, 2019, pp. 1–6.
- [33] B. Muñoz and G. Troni, “Learning the ego-motion of an underwater imaging sonar: A comparative experimental evaluation of novel cnn and rcnn approaches,” *IEEE Robot. Automat. Lett.*, 2024.
- [34] Y. Wang, Q. Han, M. Habermann, K. Daniilidis, C. Theobalt, and L. Liu, “NeuS2: Fast learning of neural implicit surfaces for multi-view reconstruction,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3295–3306.
- [35] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, “3D Gaussian splatting for real-time radiance field rendering,” *ACM Trans. Graph.*, vol. 42, no. 4, 2023.
- [36] H. Matsuki, R. Murai, P. H. J. Kelly, and A. J. Davison, “Gaussian splatting SLAM,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024.