2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SonarSweep: Fusing Sonar and Vision for Robust 3D Reconstruction via Plane Sweeping

Lingpeng Chen[1], Jiakun Tang[1], Apple Pui-Yi Chui[2], Ziyang Hong[3*], and Junfeng Wu[1]

***Abstract*— Accurate 3D reconstruction in visually-degraded underwater environments remains a formidable challenge. Single-modality approaches are insufficient: vision-based methods fail due to poor visibility and geometric constraints, while sonar is crippled by inherent elevation ambiguity and low resolution. Consequently, prior fusion techniques rely on heuristics and flawed geometric assumptions, leading to significant artifacts and an inability to model complex scenes. In this paper, we introduce SonarSweep, a novel end-to-end deep learning framework that overcomes these limitations by adapting the principled plane sweep algorithm for cross-modal fusion between sonar and visual data. Extensive experiments in both high-fidelity simulation and real-world environments demonstrate that SonarSweep consistently generates dense and accurate depth maps, significantly outperforming state-of-the-art methods under challenging conditions, particularly in high turbidity. To foster further research, we publicly release our code and a novel dataset featuring synchronized stereo-camera and sonar data—the first of its kind—at `https://github.com/LIAS-CUHKSZ/SonarSweep`.**

I. Introduction

Accurate 3D scene reconstruction is a fundamental capability for Autonomous Underwater Vehicles (AUVs), enabling critical applications like infrastructure inspection and environmental mapping [1], [2]. These tasks demand operation in visually-degraded environments where turbid water and poor lighting pose formidable challenges to perception systems. Achieving dense, metrically accurate, and geometrically coherent reconstructions under such conditions remains a significant and unsolved problem in robotics. Vision-based methods, the standard for terrestrial 3D perception, are fundamentally unreliable underwater. First, the short baselines on compact AUVs render geometric triangulation ill-posed for objects beyond a few meters. Second, light scattering and absorption eradicate the high-frequency textures essential for robust stereo correspondence [3]. Active illumination techniques that project structured light can help [4], but they fail in the highly turbid waters typical of real-world operations. Sonar offers robustness where vision fails, but its utility for 3D reconstruction is crippled by the inherent elevation ambiguity of 2D scans. Attempts to resolve this ambiguity in a single-modality system introduce critical flaws. Approaches that rely on vehicle motion to create multi-view observations [5], [6] are critically dependent on accurate pose estimation—a fatal flaw in dynamic environments where odometry drift is unavoidable. Alternative solutions, such as using orthogonal sonars, are constrained by a severely limited overlapping field of view [7].

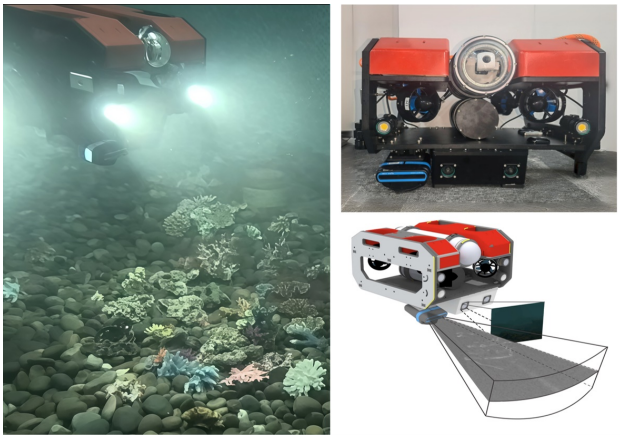

Fig. 1: **The SonarSweep System.** (Left) The experimental AUV in a challenging underwater environment. (Top Right) The integrated camera and sonar sensor suite. (Bottom Right) Conceptual diagram of the fusion approach.

Given the complementary limitations of single-modality systems, fusing optical (high-detail) and acoustic (range-accurate) sensing is a promising strategy. However, existing fusion techniques still fail to provide a complete solution. Vision-led SLAM systems that use sonar only for scale correction break down when visual features disappear in turbid water [8]. Neural rendering frameworks such as AONeuS [9] and Z-Splat [10] perform acoustic–optical fusion using NeRF or Gaussian Splatting (GS) representations. However, these methods assume known sensor poses, are computationally expensive, and focus on volumetric reconstruction rather than direct depth estimation. Real-time heuristic approaches avoid heavy computation but rely on simplified geometric assumptions that introduce significant artifacts on non-vertical surfaces [11].

To address this critical gap, we propose **SonarSweep**, a novel, end-to-end deep learning framework for dense and accurate underwater 3D reconstruction. Our method robustly fuses sonar and a monocular image by adapting the classic plane sweep algorithm to a learned, deep feature domain. We construct a multi-modal cost volume by differentiably warping sonar features into the camera's reference frame across a set of depth hypotheses, allowing our framework to regress a dense and geometrically coherent depth map.

[1]Lingpeng Chen, Jiakun Tang, and Junfeng Wu are with the Chinese University of Hong Kong, Shenzhen. {lingpengchen, jiakuntang, junfengwu}@link.cuhk.edu.cn
[2]Apple Pui-Yi Chui is with the Chinese University of Hong Kong, Hong Kong applepychui@cuhk.edu.hk
[3]Ziyang Hong is with the Department of Automation, Harbin Institute of Technology, Shenzhen, China hongzy@hit.edu.cn
*Corresponding author: Ziyang Hong.

Our contributions are threefold:

- The first adaptation of the deep plane sweep paradigm to the cross-modal fusion of sonar and visual data, overcoming the limitations of single-modality approaches.
- A comprehensive experimental validation showing that SonarSweep significantly outperforms state-of-the-art (SOTA) methods in challenging underwater conditions.
- The release of the first dataset of synchronized stereo-camera and imaging sonar data, along with our source code, to facilitate future research.

II. Related Works

A. Opti-Acoustic Scene Reconstruction

Fusing optical and acoustic sensors is a key strategy for robust 3D perception in turbid underwater environments where a single modality is insufficient. Heuristic methods such as Opti-Acoustic [11] achieve real-time performance by avoiding direct cross-modal feature matching, instead associating segmented image regions with clustered sonar returns and back-projecting a single sonar depth to all pixels within the corresponding visual segment. However, this efficiency relies on a critical geometric assumption: all vertically aligned pixels share the same depth. This constrains the reconstruction to a series of "vertical curtains," introducing significant distortion on inclined or complex surfaces. Recent neural rendering approaches, including AONeuS [9] and Z-Splat [10], perform acoustic–optical fusion using NeRF or GS representations for 3D reconstruction. However, they assume known sensor poses and are not designed for direct pixel-wise depth estimation. Consequently, while useful for mapping vertical structures or improving volumetric reconstruction, these approaches remain unsuitable for accurate depth estimation in general underwater scenes.

B. Deep Plane Sweep Stereo

The dominant paradigm for dense multi-view 3D reconstruction is Deep Plane Sweep Stereo, which adapts a classical geometric algorithm into an end-to-end deep learning framework. Early works such as DPSNet [12] and MVSNet [13] established its core pipeline. The method first extracts features from reference and source images using a neural network, then discretizes scene depth into a set of hypothesized fronto-parallel planes. Source image features are subsequently projected ("warped") onto the reference view for each depth plane. By comparing the reference features to the warped source features, a cost volume is constructed that encodes matching similarity for every pixel at every potential depth. A deep network then regularizes this volume, learning spatial and contextual relationships to refine the costs. From this regularized volume, a dense depth map is computed. This end-to-end pipeline replaces handcrafted similarity metrics with a powerful, learned correspondence model, achieving SOTA performance in vision-only reconstruction.

III. Preliminaries and Notations

This section establishes the geometric models and notations for our system, which consists of a rigidly mounted and pre-calibrated pinhole camera and 2D forward-looking sonar (FLS). We define a 3D point in the sonar coordinate system (Frame S) as $\boldsymbol{P_s} = [X_s, Y_s, Z_s]^T$ and in the camera coordinate system (Frame C) as $\boldsymbol{P_c} = [X_c, Y_c, Z_c]^T$. The transformation between these frames is defined by a rotation $\boldsymbol{R}_s^c$ and translation $\boldsymbol{t}_s^c$, such that $\boldsymbol{P_c} = \boldsymbol{R}_s^c \boldsymbol{P_s} + \boldsymbol{t}_s^c$. The projection of $\boldsymbol{P_c}$ onto the camera's image plane is denoted by the pixel coordinates $\boldsymbol{p_c} = [u, v]^T$.

Given a camera image $\mathbf{I}_c$ and a corresponding 2D sonar scan $\mathbf{S}_s$, our goal is to estimate a dense depth map $\mathbf{D}_c$. This requires finding the depth value Z_c for every pixel $\boldsymbol{p_c}$ by resolving the sonar's inherent geometric ambiguity for all points in the scene.

A. Forward-Looking Sonar Model

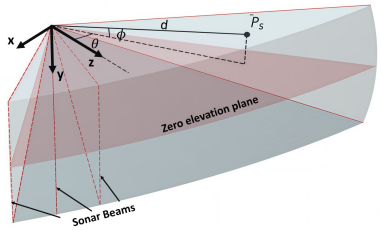

Fig. 2: The Forward-Looking Sonar (FLS) sensor model. A 3D point $\boldsymbol{P_s}$ is measured by its range d and bearing θ. The elevation angle ϕ is collapsed during the projection, leading to ambiguity along a circular arc.

The primary challenge of using 2D FLS for 3D reconstruction is its inherent elevation ambiguity. A 3D point $\boldsymbol{P_s}$ is described in spherical coordinates by its range d, bearing θ, and elevation ϕ. As illustrated in Fig. 2, the sonar measures the range and bearing but collapses all elevation information, mapping the 3D point to a single 2D polar coordinate $\boldsymbol{p_s} = (d, \theta)$. The relationship between spherical and Cartesian coordinates in the sonar frame is:

$$\boldsymbol{P_s} = \begin{bmatrix} X_s \\ Y_s \\ Z_s \end{bmatrix} = \begin{bmatrix} d\cos\phi\sin\theta \\ d\sin\phi \\ d\cos\phi\cos\theta \end{bmatrix} \tag{1}$$

Many FLS devices have a narrow vertical beamwidth, concentrating acoustic energy near the zero-elevation plane ($\phi \approx 0$). Following common practice [14], we therefore adopt an **orthographic projection approximation** by assuming $\cos\phi \approx 1$. This simplification effectively treats the 2D sonar image as a top-down view. Under this model, each sonar return $\boldsymbol{p_s}$ constrains the horizontal position of a 3D point, while its elevation Y_s remains unresolved. Resolving this ambiguity by fusing the sonar data with the camera image is the core task of our work.

IV. Proposed Algorithm

The SonarSweep pipeline, illustrated in Fig. 3, transforms the cross-modal reconstruction task into a structured, end-to-end learning problem. The framework consists of four main stages. First, two deep encoders extract multi-scale feature maps from the synchronized camera and sonar inputs respectively. For the camera, the image is cropped to the sonar's

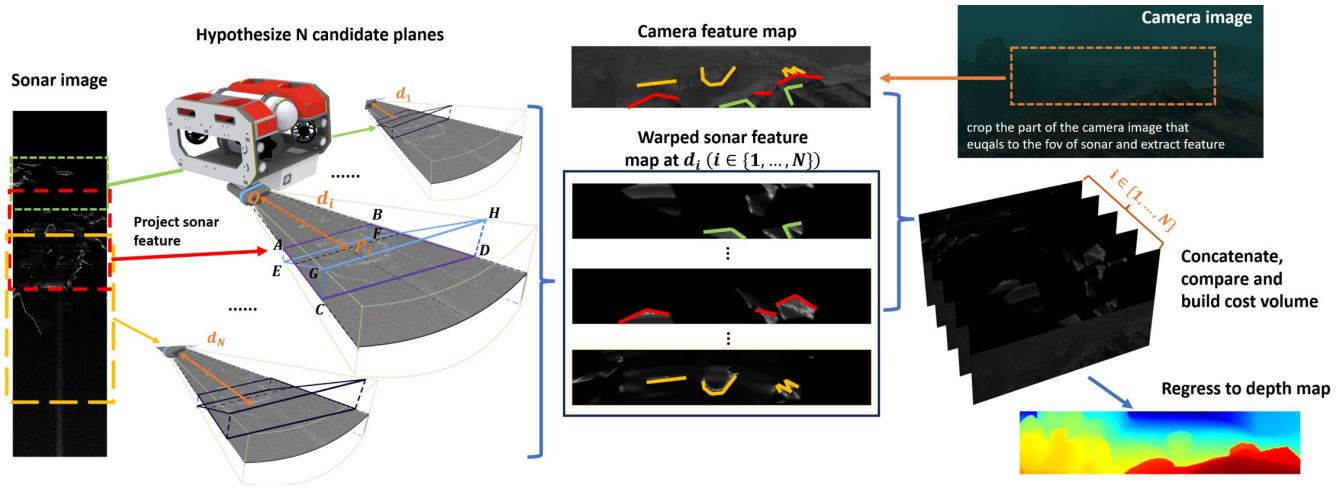

Fig. 3: SonarSweep Pipeline. From a sonar and camera image pair, we extract feature maps using two encoders. The core of our method involves hypothesizing N candidate planes, onto which 2D sonar features are back-projected and differentiably warped into the camera's view. These warped feature maps are concatenated with the camera feature map to construct a multi-modal cost volume, which is regularized and regressed to a dense depth map. The colored lines in the central feature maps highlight the fundamental matching principle, where a structure finds its strongest correspondence only at the correct depth plane (d_i); the feature maps are illustrative visualizations of the high-dimensional vectors learned by the encoders.

field of view and converted to grayscale; this encourages the network to learn robust, cross-modal **geometric similarities** rather than relying on spurious color-based correlations. Second, the 3D space is discretized into a set of N candidate planes, onto which 2D sonar features are back-projected and then differentiably warped into the camera's view. Third, a multi-modal cost volume is constructed by concatenating the camera feature map with the N warped sonar feature maps, encoding feature similarity across all hypothesized depths. Finally, a regularization network processes the cost volume, and a dense, metrically accurate depth map is regressed.

A. Sonar-Aligned Plane Hypothesization

Instead of using the camera-centric, fronto-parallel planes common in stereo vision, our method discretizes the 3D space in alignment with the sonar's imaging geometry. From Fig. 3, we hypothesize a set of N candidate planes within the sonar's field of view. As detailed in Fig. 4(a), each plane is parameterized by a fixed inclination angle α and a unique distance $\|OP_i\| = d_i$ from the sonar origin, for $i \in \{1, ..., N\}$.

This parameterization allows us to define a back-projection that maps any 2D sonar measurement onto a 3D point on each of these N planes. Leveraging the orthographic projection assumption (Section III-A), a sonar measurement $\boldsymbol{p}_s = [d, \theta]^T$ is back-projected to its corresponding 3D point $\boldsymbol{P}_s^i$ on the i-th plane as follows:

$$X_s^i = d\sin(\theta) \tag{2a}$$

$$Y_s^i = (d_i - Y_s^i)\tan(\alpha) = (d_i - d\cos(\theta))\tan(\alpha) \tag{2b}$$

$$Z_s^i = d\cos(\theta) \tag{2c}$$

Here, X_s^i and Y_s^i are computed directly from the polar measurement, while the elevation Z_s^i is determined by the plane's geometry. This geometric transformation is the foundation of our learning approach; it is applied to rich feature maps extracted from parallel, multi-scale encoders for both the sonar and camera. This feature pyramid structure allows the network to learn robust correspondences by capturing both coarse and fine details. The process, therefore, effectively lifts the 2D sonar features into 3D, endowing them with the spatial information necessary for the subsequent warping stage.

B. Projective-Consistent Plane Sampling

The strategic selection of the N plane distances is critical for effective feature matching. Standard methods that sample planes uniformly in depth or inverse-depth space [13] do not guarantee uniform pixel displacements for non-frontal scenes, which can impair the learned matching process. To address this, we introduce a **projective-consistent sampling** strategy designed to create uniform search steps in the camera's pixel space.

Our core objective is to maintain a constant geometric transformation between consecutive planes from the camera's perspective[1]. As depicted in Fig. 4(b), this is achieved by enforcing projective consistency. Consider the 3D points I^m, J^m and K^m, which lie on the same camera ray and thus project to the same image point m. Likewise, I^n and J^n project to image point n. The consistency condition requires that the pixel displacement from m to n observed for the transition from plane $i-1$ to i (via points I^m and I^n) is the same as for the transition from plane i to $i+1$ (via points J^m and J^n). This is met only if the distances from the origin to these points form a geometric progression, i.e., $\frac{\|OI^m\|}{\|OJ^m\|} = \frac{\|OI^n\|}{\|OJ^n\|} = \frac{\|OJ^m\|}{\|OK^m\|}$. Therefore, the ratio of the distances to

[1]This derivation assumes a zero-baseline configuration between the camera and sonar for illustrative clarity. The resulting geometric progression principle holds as a strong approximation for the practical, non-zero baseline case.

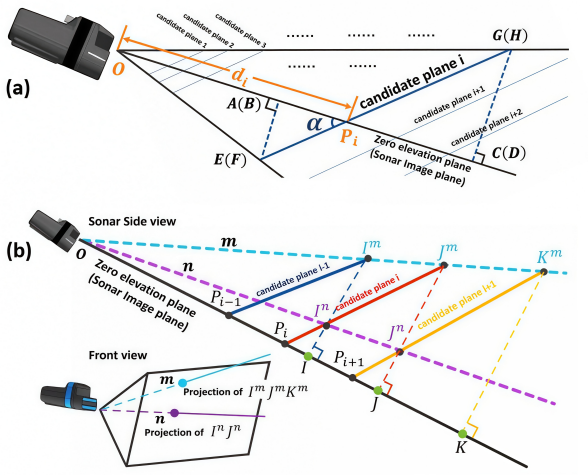

Fig. 4: (a) Geometric parameterization of a candidate plane, defined by an inclination angle α and a distance d_i.
(b) Illustration of our projective-consistent sampling. I^m and I^n are 3D points back-projected from the 2D sonar measurement I. This principle is applied consistently for all measurements (e.g., J, K). To create uniform steps in the camera's pixel space, the hypothesized planes must be sampled in a geometric progression ($d_{i+1} = k \cdot d_i$).

consecutive planes must be constant:

$$\frac{\|OP_i\|}{\|OP_{i-1}\|} = \frac{\|OP_{i+1}\|}{\|OP_i\|} = k \tag{3}$$

where k is a constant scaling factor. This directly leads to our sampling formula, a geometric progression, where the distance to the i-th plane is given by:

$$d_i = d_0 \cdot k^{i-1}, \quad \text{for } i = 1, \ldots, N \tag{4}$$

where $d_0 = d_{\text{min}}$. By creating uniform steps in this projective space, our strategy provides a more robust and stable basis for the learned feature matching. In our implementation, we set $N = 48$, $d_0 = 0.5\,\text{m}$, and $k = 1.05$, spanning the sonar's effective sensing range from 0.5 m to approximately 5 m.

C. Differentiable Warping via Ray-Plane Intersection

To construct the cost volume, we must warp the 3D sonar features, which reside on the hypothesized planes, into the camera's reference frame. This requires a differentiable mapping that, for each camera pixel $\boldsymbol{p_c}$, finds its corresponding 3D location $\boldsymbol{P_s^i}$ on the i-th candidate plane. Back-projecting a 2D pixel to a 3D point is inherently ill-posed, as the pixel's viewing ray contains infinite points. However, our plane hypotheses resolve this ambiguity by enforcing that the 3D point must lie on one of the known planes.

We find this unique intersection point by formulating and solving a system of linear equations derived from two geometric constraints. This approach ensures the transformation is a closed-form, differentiable solution suitable for end-to-end learning.

1) The Planar Constraint: First, the point $\boldsymbol{P_s^i}$ must lie on the i-th candidate plane. From Section IV-A, this plane has a normal vector $\boldsymbol{n_s} = [0, \cos(\alpha), \sin(\alpha)]^T$ and a known distance d_i. This geometric fact yields our first linear equation for $\boldsymbol{P_s^i}$:

$$\cos(\alpha) Y_s^i + \sin(\alpha) Z_s^i = d_i \sin(\alpha) \tag{5}$$

2) The Camera Projection Constraint: Second, the point $\boldsymbol{P_s^i}$ must project to the specified pixel coordinates $\boldsymbol{p_c} = [u, v]^T$. This is described by the camera projection model: $s[u, v, 1]^T = \mathbf{K}_c(\mathbf{R}_s^c \boldsymbol{P_s^i} + \mathbf{t}_s^c)$, where s is the unknown depth in the camera frame. By defining $\mathbf{M} = \mathbf{K}_c\mathbf{R}_s^c$ and $\mathbf{C} = \mathbf{K}_c\mathbf{t}_s^c$, and letting $\mathbf{m_j}$ be the j-th row of $\mathbf{M}$, we can eliminate the unknown scalar s to obtain two additional linear equations for $\boldsymbol{P_s^i}$:

$$(u\mathbf{m_3} - \mathbf{m_1})^T \boldsymbol{P_s^i} = C_1 - uC_3 \tag{6}$$

$$(v\mathbf{m_3} - \mathbf{m_2})^T \boldsymbol{P_s^i} = C_2 - vC_3 \tag{7}$$

3) Solving the Linear System: Combining Equations (5), (6), and (7) for each plane i, we form a standard linear system $\mathbf{A}_i \boldsymbol{P_s^i} = \mathbf{b}_i$:

$$\underbrace{\begin{bmatrix} 0 & \cos(\alpha) & \sin(\alpha) \\ (u\mathbf{m_3} - \mathbf{m_1})^T \\ (v\mathbf{m_3} - \mathbf{m_2})^T \end{bmatrix}}_{\mathbf{A}_i} \underbrace{\begin{bmatrix} X_s^i \\ Y_s^i \\ Z_s^i \end{bmatrix}}_{\boldsymbol{P_s^i}} = \underbrace{\begin{bmatrix} d_i \sin(\alpha) \\ C_1 - uC_3 \\ C_2 - vC_3 \end{bmatrix}}_{\mathbf{b}_i} \tag{8}$$

This system has a unique solution, $\boldsymbol{P_s^i} = \mathbf{A}_i^{-1}\mathbf{b}_i$, which can be computed efficiently. Performing this calculation for every pixel $\boldsymbol{p_c}$ and every candidate plane i yields the complete set of 3D sampling coordinates. This grid is the crucial component that allows for the differentiable warping of sonar features into the camera's perspective.

D. Cost Volume Regularization and Depth Estimation

With the differentiable warping grid established, the final stage of the pipeline transforms the geometric problem into a probabilistic estimation task, culminating in a dense depth map. This process involves constructing a multi-modal cost volume, regressing a probable depth for each pixel, and transforming this estimate into the final metric depth.

1) Cost Volume Construction and Regularization: The sampling grid $\boldsymbol{P_s^i}$ computed in Section IV-C provides the precise coordinates to warp the sonar features into the camera's view. For each of the N candidate planes, we use this grid with a differentiable bilinear sampling mechanism to sample from the sonar feature map $\mathcal{F}_s$. This generates N warped sonar feature maps, $\mathcal{F}_s^i$, each aligned with the camera's perspective. A 4D cost volume $\mathcal{C}$ of size $H \times W \times N \times F$ is then constructed by concatenating the camera feature map $\mathcal{F}_c$ with each of the N warped sonar feature maps. For a given pixel (u, v), the feature vector at the i-th depth hypothesis is:

$$\mathcal{C}(u, v, i) = \text{Concat}(\mathcal{F}_c(u, v), \mathcal{F}_s^i(u, v)) \tag{9}$$

This volume, which encodes cross-modal similarity, is then processed by a 3D CNN. This network regularizes the costs by learning to enforce spatial and geometric consistency,

producing a refined, single-channel cost volume $\mathcal{C}'$ of size $H \times W \times N$.

2) Differentiable Depth Regression: To obtain a continuous, sub-pixel accurate depth estimate from the discrete cost volume $\mathcal{C}'$, we employ a soft-argmin operation. For each pixel, the matching costs across all N candidate depths are first converted into a probability distribution using the softmax function:

$$P(d_i|u,v) = \frac{\exp(-\mathcal{C}'(u,v,i))}{\sum_{j=1}^{N} \exp(-\mathcal{C}'(u,v,j))} \tag{10}$$

The final estimated plane distance $\hat{d}(u,v)$ for the pixel is then computed as the expected value over this probability distribution:

$$\hat{d}(u,v) = \sum_{i=1}^{N} d_i \cdot P(d_i|u,v) \tag{11}$$

3) Final Metric Depth Transformation: The regression stage yields a per-pixel estimate of the most probable plane distance, $\hat{d}(u,v)$, which exists in the sonar's geometric space. The final step is to transform this value into a dense, metric depth map in the camera's reference frame. For each pixel $\boldsymbol{p_c}$, its corresponding 3D point $\hat{\boldsymbol{P}}_c$ must simultaneously lie on its camera viewing ray and on the sonar plane defined by $\hat{d}(u,v)$.

By substituting the camera ray constraint ($\hat{\boldsymbol{P}}_c = Z_c \cdot \mathbf{K}_c^{-1}[u,v,1]^T$) into the sonar plane constraint, we can solve directly for the unknown camera-frame depth, Z_c. This yields a closed-form solution:

$$Z_c(u,v) = \frac{\hat{d}(u,v)\sin(\alpha) + (\mathbf{R}_s^c \boldsymbol{n_s})^T \mathbf{t}_s^c}{(\mathbf{R}_s^c \boldsymbol{n_s})^T (\mathbf{K}_c^{-1}[u,v,1]^T)} \tag{12}$$

Since Z_c represents the depth along the camera's principal axis, we convert it to a true Euclidean distance to generate the final metric depth map $\mathcal{D}$:

$$\mathcal{D}(u,v) = |Z_c(u,v)| \cdot \left\| \mathbf{K}_c^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \right\|_2 \tag{13}$$

This entire pipeline, from feature extraction to the final depth computation, is fully differentiable, enabling end-to-end training of the SonarSweep network.

V. Experiments and Results

This section validates our proposed camera-sonar fusion framework through a series of experiments. We analyze the model's performance in varied simulated and real-world conditions, compare it against SOTA baselines, and evaluate its robustness to environmental degradations such as turbidity.

A. Experimental Setup

1) Platform and Sensor Suite: Our experimental platform is a custom underwater vehicle equipped with a belly-mounted sensor pod, as shown in Fig. 1. The pod, which houses the primary perception sensors, is angled 15 degrees downwards to prioritize detailed seafloor mapping while maintaining forward-looking obstacle awareness. The sensor payload includes:

- **Stereo Camera:** A time-synchronized stereo camera with a 15 cm baseline supports traditional stereo-vision algorithms, enabling fair comparison with ours.
- **Imaging Sonar:** An Oculus M1200d forward-looking sonar operating in its high-frequency mode (2.1 MHz) to maximize geometric detail. It provides a 60° horizontal and 12° vertical FOV within a 5m sensing range.

The extrinsic calibration parameters between the camera and sonar are directly derived from the robot's precise CAD design model and remain fixed in our algorithm. Model training was conducted offline on a desktop with an NVIDIA RTX 4090 GPU, while all inference and validation were performed exclusively on an NVIDIA Jetson Orin AGX. This setup mirrors the computational constraints of a field-deployable autonomous system.

2) Baselines for Comparison: To provide a comprehensive benchmark, we evaluate SonarSweep against three SOTA baselines, each representing a distinct sensing modality for dense depth estimation:

- **FoundationStereo [15]:** A SOTA vision-only baseline for stereo depth estimation. Renowned for its strong zero-shot generalization capabilities.
- **Multi-view Sonar Stereo [5]:** A learning-based, sonar-only method that adapts multi-view stereo principles to acoustic imagery.
- **Opti-Acoustic Fusion [11]:** A representative geometry-based fusion method that offers a robust, real-time solution by matching visual segments with sonar returns.

These baselines were selected to ensure a fair evaluation against leading specialized methods in the vision-only, sonar-only, and heuristic fusion domains. We excluded methods like AONeuS [9] and Z-Splat [10] from our comparison, as they are designed for neural rendering and are not suited for real-time and generalizable depth estimation.

B. Data Collection and Training Strategy

We adopted a dual-pronged data collection strategy, using a high-fidelity simulator for large-scale data generation and a physical testbed for real-world validation and fine-tuning.

1) Real-World Water Tank Data: Our physical experiments were conducted in a 4.5 × 10m water tank containing various objects like rocks, corals, and boulders to create a complex environment. Ground truth camera poses and depth maps were generated using the commercial photogrammetry software Agisoft Metashape [16]. This methodology is validated by its successful application in generating ground truth in the FLSea dataset [17]. The reconstruction quality was validated by an average RMS reprojection error of 1.5 pixels, confirming its suitability as reliable ground truth. From this physical setup, we collected a total of 2,543 data points.

2) Simulated Data Generation: We used the OceanSim [18] simulator, creating a digital twin of our robot that replicates its physical dimensions and sensor configurations. A large-scale underwater map was designed with diverse and

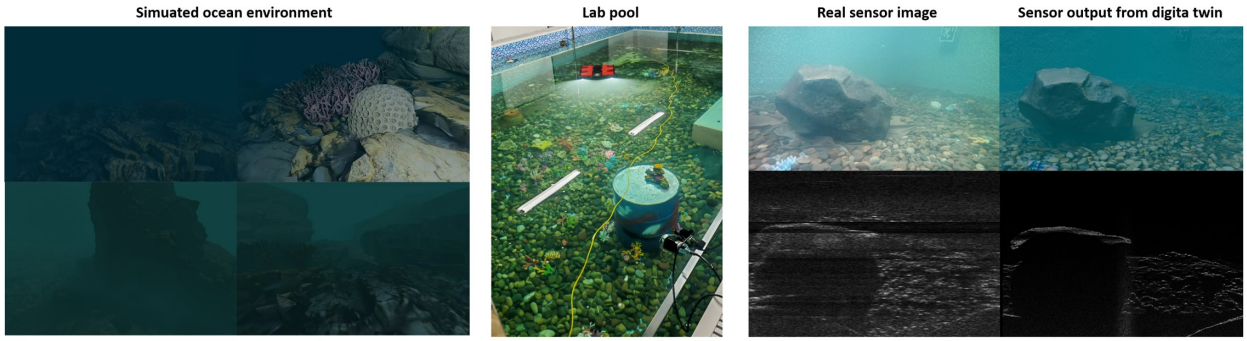

Fig. 5: From left to right: the simulated underwater world in OceanSim with varied water conditions; the physical lab pool setup; a real-world sensor suite; and the corresponding high-fidelity output from our digital twin.

complex terrain (e.g., rocks, corals, undulating seafloor). To increase data diversity, we rendered this environment under two distinct water conditions, simulating both clear inland and lower-visibility oceanic water, as shown in Fig. 5. We collected 7686 synchronized data points from simulation, each including ground truth pose, a sonar image, a stereo image pair, and a dense depth map.

3) Data Preprocessing: All data underwent a standardized preprocessing pipeline. Camera images were cropped to the sonar's FOV, converted to grayscale, and enhanced using histogram equalization to encourage the learning of geometric features. Sonar data was processed using a pipeline inspired by [19] to mitigate noise and artifacts, as shown in Fig. 6. First, a background model is created by averaging numerous frames captured without object in the scene. Next, both the individual target frames and this background model are independently denoised using a pre-trained Swin-Conv-Unet [20]. Finally, the denoised background is subtracted from each denoised target frame, significantly improving the signal-to-noise ratio by isolating the acoustic returns.

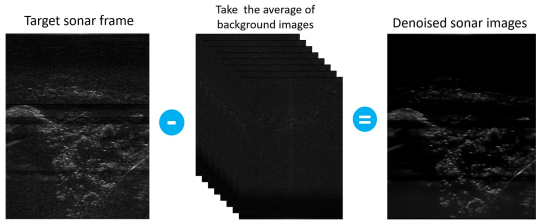

Fig. 6: The sonar image preprocessing pipeline.

4) Sim-to-Real Training Strategy: To bridge the substantial domain gap between simulated and real-world data (see Fig. 5, right panel), we employ a two-stage training strategy. The model is first pre-trained on the large-scale simulated dataset to learn generalizable cross-modal features. Subsequently, it is fine-tuned on the smaller, real-world dataset to adapt to the specific noise characteristics and sensor artifacts of the physical system.

C. Performance Analysis

We conducted a rigorous comparative analysis of SonarSweep against the three baselines across both simulated and real-world underwater environments. The evaluation demonstrates our method's superior performance in generating dense, accurate, and reliable depth maps.

1) Qualitative Comparison: Fig. 7 provides a qualitative comparison of the depth maps and error maps produced by each method. The results highlight the distinct failure modes of single-modality and heuristic fusion approaches. The **Opti-Acoustic** method produces sparse and often geometrically incorrect results, as its performance is limited by a fragile segmentation algorithm and a flawed "vertical curtain" assumption. It struggles to reconstruct continuous surfaces like the seafloor and fails to detect obstacles in the real-world cases.

The sonar-only **Stereo Sonar** baseline is limited by the low resolution of acoustic imagery, resulting in depth maps with blurry edges and a lack of fine detail. Conversely, the vision-only **Foundation Stereo** excels at reconstructing nearby objects with sharp edges, but its accuracy rapidly degrades with distance due to the constrained stereo baseline. This leads to large-scale distortions on more distant surfaces.

In contrast, **SonarSweep** consistently produces complete and geometrically accurate depth maps in all scenarios. By synergistically fusing the two modalities, our method leverages the camera's high-resolution texture for precise edge definition while relying on the sonar's robust range measurements to ensure accuracy over distance. This fusion overcomes the fundamental limitations of the individual sensors, resulting in a more reliable depth estimation.

TABLE I: Quantitative Depth Estimation Performance

	Abs Rel ↓	Abs Diff ↓	RMSE ↓	a1 Acc. ↑
Simulation				
Foundation Stereo	0.0509	0.1316	0.3807	0.9960
Stereo Sonar	0.0533	0.1387	0.2194	0.9622
Ours	**0.0231**	**0.0577**	**0.0928**	**0.9951**
Real-World				
Foundation Stereo	0.0757	0.2437	0.3279	0.9506
Stereo Sonar	0.0691	0.1970	0.3131	0.9425
Ours	**0.0382**	**0.1064**	**0.1479**	**0.9922**

2) Quantitative Evaluation: The quantitative results, evaluated on 2,000 simulated and 1,000 real-world test sets, confirm the qualitative observations. We evaluate depth quality using four standard metrics: Absolute Relative error (Abs Rel) and Absolute Difference (Abs Diff) for mean deviations, Root Mean Square Error (RMSE) for large-error penalties, and threshold accuracy ($a_1 < 1.25$).

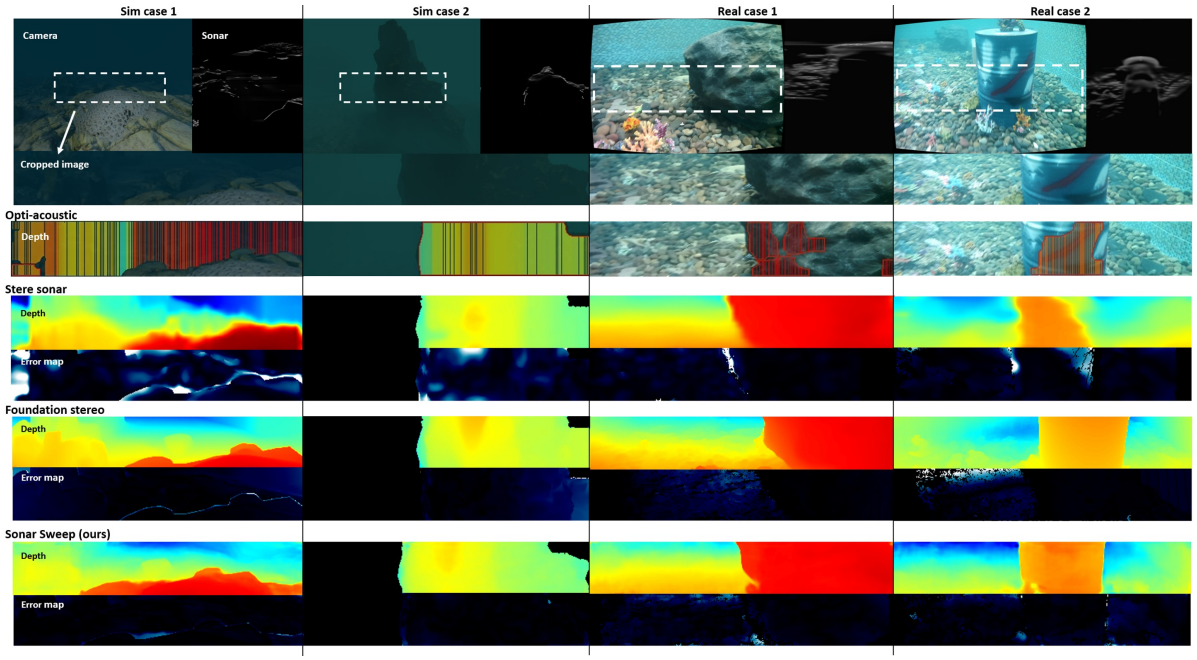

Fig. 7: Qualitative comparison of each methods in simulated (Sim case 1 & 2) and real-world (Real case 1 & 2) scenarios. In the depth maps, warmer colors (red) are closer, while cooler colors (blue) are farther. In the error maps, brighter regions signify larger errors. Note that the output perspectives differ, as each algorithm uses a different native reference frame.

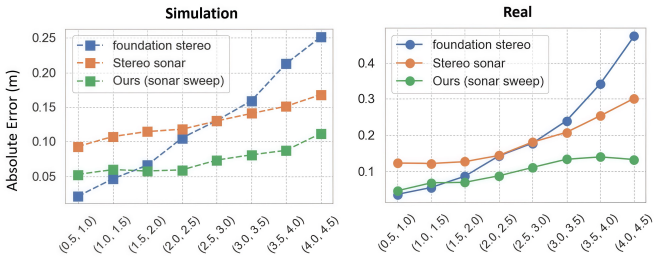

Fig. 8: Quantitative comparison of absolute error versus distance in simulated (left) and real-world (right) datasets.

As shown in Tab. I, our method significantly outperforms all baselines across all metrics on both datasets. The performance gap is particularly pronounced on the more challenging real-world data, highlighting our model's robustness to sensor noise and environmental variability.

For a more granular analysis, Fig. 8 plots the estimation error as a function of depth. This reveals the complementary nature of our fusion approach. At close ranges ($< 2m$), Foundation Stereo has highest accuracy, but its error grows sharply with distance. In contrast, the sonar-based methods exhibit more stable performance across the full range. SonarSweep effectively achieves the best of both worlds: it leverages high-resolution visual features to attain an accuracy comparable to Foundation Stereo at close ranges, while capitalizing on sonar's inherent precision to mitigate error growth over distance, maintaining a stability similar to the pure sonar method.

D. Robustness to Turbidity

A key advantage of opti-acoustic fusion is its potential for robust performance in degraded visual conditions. To validate this, we evaluated our model's resilience to turbidity by synthesizing poor visibility conditions on our real-world test data.

1) Experimental Design: We synthesized turbidity on clear test tank images using the underwater Image Formation Model (IFM) [21]. Three conditions were simulated based on Jerlov coastal water types [22]: Type-1C (mild), Type-3C (moderate), and Type-5C (high turbidity). The turbid image $I^c(x)$ was generated from the clear image $J^c(x)$ using:

$$I^c(x) = J^c(x)(T_1^c)^d + (1 - (T_1^c)^d)B^c, \tag{14}$$

where c is the color channel, B^c is the ambient background light, d is object distance, and T_1^c is the spectral transmission rate. We used a representative distance of $d = 2.5\,\text{m}$, with the specific transmission rates for each water type detailed in Tab. II.

TABLE II: Spectral Transmission Rates (T_1^c) for Jerlov Types

Water Type	Red Channel	Green Channel	Blue Channel
Type-1C	0.75	0.87	0.88
Type-3C	0.71	0.80	0.82
Type-5C	0.67	0.67	0.73

2) Results: As illustrated in Fig. 9, the performance of the Foundation Stereo degrades significantly as turbidity increases, as its feature matching relies entirely on visual clarity. In contrast, the acoustics-only Stereo Sonar is immune to this optical degradation, exhibiting a constant error rate across all conditions. Our SonarSweep method consistently outperforms both baselines. While its error increases slightly in the most turbid scenario, it remains significantly more accurate than the individual modalities. This result confirms that our fusion strategy effectively leverages sonar data to

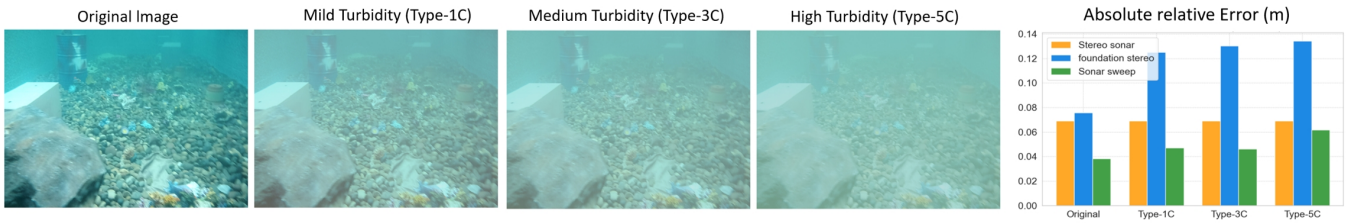

Fig. 9: Synthesized images for mild, moderate, and high turbidity, followed by performance analysis

compensate for the loss of visual information, ensuring reliable performance in challenging underwater environments.

VI. Conclusion

In this paper, we introduced **SonarSweep**, a novel, end-to-end deep learning framework that overcomes key challenges in underwater 3D reconstruction by fusing imaging sonar and camera data. Our core contribution is the successful adaptation of the deep plane sweep paradigm to this cross-modal problem. Extensive experiments demonstrated that SonarSweep significantly outperforms SOTA baselines, showing exceptional robustness at extended ranges and in turbid conditions. This work represents a significant step towards more reliable autonomous perception; to accelerate progress, we will release our source code and dataset, with future work aimed at integration into a full SLAM system for globally consistent mapping.

Acknowledgment

This work was partly supported by the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2024A1515240009, and the Shenzhen Science and Technology Program under Grant No. JCYJ20240813113609013.

References

[1] M. VanMiddlesworth, M. Kaess, F. Hover, and J. J. Leonard, "Mapping 3d underwater environments with smoothed submaps," in *Field and Service Robotics: Results of the 9th International Conference*, Springer, 2015, pp. 17–30.

[2] F. Nauert and P. Kampmann, "Inspection and maintenance of industrial infrastructure with autonomous underwater robots," *Frontiers in Robotics and AI*, vol. 10, p. 1 240 276, 2023.

[3] M. E. Angelopoulou, C. Tsiotsios, and M. Petrou, "Limitations of vision guided underwater navigation," *IFAC Proceedings Volumes*, vol. 45, no. 5, pp. 312–317, 2012.

[4] R. Detry et al., "Turbid-water subsea infrastructure 3d reconstruction with assisted stereo," in *2018 OCEANS-MTS/IEEE Kobe Techno-Oceans (OTO)*, IEEE, 2018, pp. 1–6.

[5] Y. Wang, Y. Ji, H. Tsuchiya, H. Asama, and A. Yamashita, *Learning pseudo front depth for 2d forward-looking sonar-based multi-view stereo*, 2022. arXiv: 2208.00233 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2208.00233.

[6] J.-Y. Park, H. Baek, B.-H. Jun, and P.-M. Lee, "3d reconstruction using multiple acoustic images under roll motion based on backprojection techniques," in *OCEANS 2023 - MTS/IEEE U.S. Gulf Coast*, 2023, pp. 1–4. DOI: 10.23919/OCEANS52994.2023.10337360.

[7] J. McConnell, J. D. Martin, and B. Englot, "Fusing concurrent orthogonal wide-aperture sonar images for dense underwater 3d reconstruction," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 1653–1660. DOI: 10.1109/IROS45743.2020.9340995.

[8] M. Roznere and A. Q. Li, "Underwater monocular image depth estimation using single-beam echosounder," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2020, pp. 1785–1790.

[9] M. Qadri, K. Zhang, A. Hinduja, M. Kaess, A. Pediredla, and C. A. Metzler, *Aoneus: A neural rendering framework for acoustic-optical sensor fusion*, 2024. arXiv: 2402.03309 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2402.03309.

[10] Z. Qu et al., "Z-splat: Z-axis gaussian splatting for camera-sonar fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 47, no. 9, pp. 7255–7267, 2024.

[11] I. Collado-Gonzalez, J. McConnell, P. Szenher, and B. Englot, *Opti-acoustic scene reconstruction in highly turbid underwater environments*, 2025. arXiv: 2508.03408 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2508.03408.

[12] S. Im, H.-G. Jeon, S. Lin, and I. S. Kweon, *Dpsnet: End-to-end deep plane sweep stereo*, 2019. arXiv: 1905.00538 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1905.00538.

[13] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, *Mvsnet: Depth inference for unstructured multi-view stereo*, 2018. arXiv: 1804.02505 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1804.02505.

[14] N. Hurtos, D. Ribas, X. Cufi, Y. Petillot, and J. Salvi, "Fourier-based registration for robust forward-looking sonar mosaicing in low-visibility underwater environments," *Journal of Field Robotics*, vol. 32, no. 1, pp. 123–151, 2015.

[15] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, *Foundationstereo: Zero-shot stereo matching*, 2025. arXiv: 2501.09898 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2501.09898.

[16] Agisoft LLC, *Agisoft metashape professional*, https://www.agisoft.com/, Version 2.0.2, 2023.

[17] Y. Randall, "Flsea: Underwater visual-inertial and stereo-vision forward-looking datasets," M.S. thesis, University of Haifa (Israel), 2023.

[18] J. Song, H. Ma, O. Bagoren, A. V. Sethuraman, Y. Zhang, and K. A. Skinner, "Oceansim: A gpu-accelerated underwater robot perception simulation framework," *arXiv preprint arXiv:2503.01074*, 2025.

[19] 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*, vol. 9, no. 11, pp. 10 065–10 072, 2024. DOI: 10.1109/LRA.2024.3469778.

[20] K. Zhang et al., "Practical blind image denoising via swin-conv-unet and data synthesis," *Machine Intelligence Research*, vol. 20, no. 6, pp. 822–836, Sep. 2023, ISSN: 2731-5398. DOI: 10.1007/s11633-023-1466-0. [Online]. Available: http://dx.doi.org/10.1007/s11633-023-1466-0.

[21] J. R. Ahamed, P. E. Abas, and L. C. De Silva, "An image synthesis method generating underwater images," *Advances in Technology Innovation*, vol. 7, no. 3, p. 195, 2022.

[22] N. G. Jerlov and F. F. Koczy, *Photographic measurements of daylight in deep water*. Elanders boktr., 1951.