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

# Towards Versatile Opti-Acoustic Sensor Fusion and Volumetric Mapping for Safe Underwater Navigation

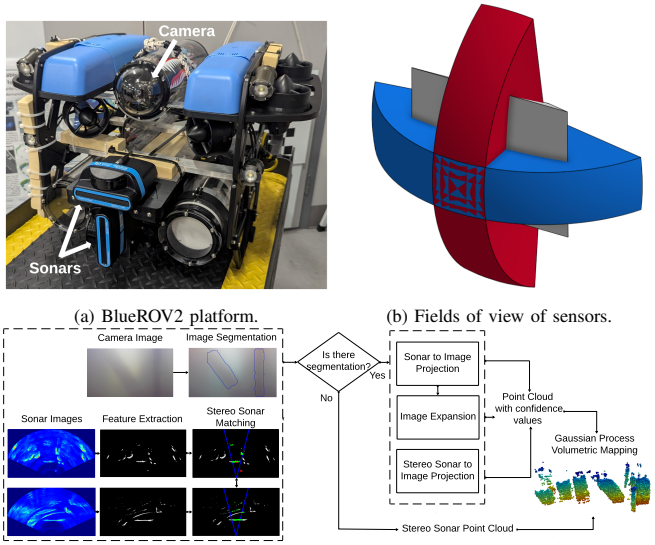
Ivana Collado-Gonzalez[1], John McConnell[2], and Brendan Englot[1]

***Abstract*— Accurate 3D volumetric mapping is critical for autonomous underwater vehicles operating in obstacle-rich environments. Vision-based perception provides high-resolution data but fails in turbid conditions, while sonar is robust to lighting and turbidity but suffers from low resolution and elevation ambiguity. This paper presents a volumetric mapping framework that fuses a stereo sonar pair with a monocular camera to enable safe navigation under varying visibility conditions. Overlapping sonar fields of view resolve elevation ambiguity, producing fully defined 3D point clouds at each time step. The framework identifies regions of interest in camera images, associates them with corresponding sonar returns, and combines sonar range with camera-derived elevation cues to generate additional 3D points. Each 3D point is assigned a confidence value reflecting its reliability. These confidence-weighted points are fused using a Gaussian Process Volumetric Mapping framework that prioritizes the most reliable measurements. Experimental comparisons with other opti-acoustic and sonar-based approaches, along with field tests in a marina environment, demonstrate the method's effectiveness in capturing complex geometries and preserving critical information for robot navigation in both clear and turbid conditions. Our code is open-source to support community adoption.**

## I. Introduction

Autonomous Underwater Vehicles (AUVs) are essential for tasks such as maintenance, mapping, and exploration, often in turbid waters. Information-gathering tasks often require large-scale environment mapping, while maintenance tasks demand accurate close-range reconstruction to guide manipulators and avoid unmodeled obstacles. Reliable sensing and mapping across diverse conditions and scales is therefore critical for AUV autonomy and safety.

Vision-based perception has shown partial success underwater [1], but its reliance on salient features makes it sensitive to illumination changes, blurring, and halo effects, resulting in unreliable performance in scattering media. Sonar sensors provide robustness in turbid water and under variable lighting, yet they have inherent limitations: profiling sonars capture narrow slices, 3D sonars are emerging [2] but remain range-limited, and imaging sonars produce 2D projections of 3D scenes, restricting vertical Field Of View (FOV) and introducing elevation ambiguity. Despite advances, high-resolution 3D sonar reconstruction remains challenging. Sonar's low resolution favors large-scale mapping but limits detection of fine features and close-range objects critical for collision avoidance and manipulation.




(a) BlueROV2 platform. (b) Fields of view of sensors.

(c) Proposed sensor fusion and mapping pipeline.

Fig. 1: **Overview.** (a) Our BlueROV2 platform; (b) an illustration of the fields of view of its camera and sonars, with the gray volume representing the camera and the blue/red volumes indicating the horizontal/vertical sonars respectively; (c) an illustration of our proposed volumetric mapping pipeline.

These limitations motivate the fusion of complementary sensing modalities. Optical imaging provides high-resolution detail, while acoustic imaging ensures robustness in turbid waters and extended range. Fusion can enhance 3D reconstruction, but differing sensor image-formation mechanisms across sensors complicate direct feature matching. Most approaches treat visual cues as primary input and use sonar for refinement [3], which fails in turbid environments. A recent opti-acoustic method [4] builds on sonar features, enhancing reconstructions with visual data to expand vertical FOV. Although it removes the dependency on reliable visual features, its application is limited to simple object geometries. These challenges highlight the need for a fusion strategy that prioritizes sonar while flexibly incorporating visual data for reliable 3D mapping in complex underwater scenes.

This work addresses underwater sensing and mapping under variable visibility conditions. This method leverages overlapping sensor FOVs to compensate for missing information in both camera and sonar, providing complementary distance and elevation cues. Confidence values derived from the fused 3D solution guide the mapping process by weight-

[1]I. Collado-Gonzalez, and B. Englot are with Stevens Inst. of Technology, Hoboken, NJ, USA, {icollado, benglot}@stevens.edu. [2]J. McConnell is with the U.S. Naval Academy, Annapolis, MD, USA, jmcconne@usna.edu. **This research was supported by grants NSF IIS-1652064, USDA-NIFA 2021-67022-35977, and ONR N00014-24-1-2522.** The views expressed in this paper are those of the author(s) and do not reflect the official policy or position of the U.S. Naval Academy, Department of the Navy, the Department of War, or the U.S. Government.

ing measurements according to reliability. The framework relies primarily on sonar while incorporating camera data only when available. A CNN-based segmentation module is trained efficiently using transfer learning and data augmentation on visual data, avoiding the need for large sonar datasets. This approach generates scene maps from a single pose, eliminating the need for specific motion patterns, and focuses on extracting information relevant for safe navigation. By addressing the limitations of existing approaches, this work improves AUV performance in complex underwater environments.

The key contributions of this paper are:

- To the best of our knowledge, this is the first underwater sensing strategy to fuse information from a stereo sonar pair and a monocular camera.
- A confidence-based volumetric mapping framework that prioritizes reliable measurements to preserve relevant information for safe robot navigation.
- Experimental comparisons in a tank environment demonstrate the accuracy of the proposed framework at mapping complex geometries.
- A real-world experiment showing our framework's adaptability to low-visibility conditions.
- Public release of our code and data to facilitate reproducibility: `https://github.com/ivanacollg/stereosonar_camera_mapping`

The remainder of the paper is structured as follows. Sec. II reviews related work, Sec. III presents the problem formulation, Sec. IV details the proposed method, Sec. V presents experiments and results, and Sec. VI concludes the paper.

## II. Related Works

**Volumetric Mapping:** Volumetric occupancy mapping is an active research area [5]–[8], with voxelized maps widely used for real-time 3D planning [9], [10]. Above-water range sensors, such as LiDAR and RGB-D cameras, often provide dense, accurate, and evenly distributed measurements, making them well suited for signed distance surface reconstructions, such as Truncated Signed Distance Field (TSDF)-based mapping, which exploit dense data to produce continuous surfaces. In contrast, underwater sensing presents additional challenges: sonar returns are typically sparse and noisy, making probabilistic mapping approaches especially valuable. Consequently, occupancy-based methods, which explicitly model uncertainty, remain highly relevant for underwater applications.

Occupancy grid mapping represents the environment by maintaining an independent probability of occupancy for each cell, updated incrementally using Bayesian filtering. OctoMaps [11] improve scalability through octree structures, but they assume statistical independence between cells and update only those intersected by beams. As a result, sparse measurements can produce discontinuous maps, which may lead to unsafe path planning if gaps are misclassified as free space.

To address these limitations, implicit mapping approaches represent occupancy as a continuous distribution. Gaussian Process (GP) occupancy mapping models spatial correlations between cells, enabling probabilistic inference in unobserved regions [12]. Unlike grid-based methods, GPs can extrapolate across sparse data, filling sensor gaps with correlated occupancy estimates and improving classification accuracy, while GP kernel design allows balancing conservative versus aggressive predictions. The major drawback of GP regression is the high computational complexity. Wang presented an improved formulation of GP occupancy mapping using nested Bayesian Committee Machines (BCMs) and test data, which reduces computational complexity and enables real-time 3D mapping [13]. More generally, Gaussian Processes have also been applied to signed distance fields [14], demonstrating their suitability for volumetric mapping tasks.

**Sonar-based methods:** Sonars are widely used underwater because they are unaffected by lighting conditions and turbidity. Forward-looking multibeam imaging sonars provide wide Fields Of View (FOV) for navigation and collision avoidance, but they produce 2D projections of 3D scenes, without elevation information reconstruction is difficult. Several methods estimate missing elevation: Aykin [15] and Westman [16] use shadow edges but assume smooth, monotonic surfaces; space-carving approaches [17], [18] and multi-view reconstructions [19], [20] rely on observations from multiple angles, which are often infeasible in constrained or close-range scenarios.

Learning-based methods attempt to overcome these limits. Wang [21] and DeBortoli [22] infer elevation from sonar images, while Qadri [23] and Sethuraman [24] propose neural implicit and Gaussian splatting approaches. However, all require sonar datasets, which remain scarce. To reduce reliance on data or multiple images, McConnell [25] introduced stereo sonar, using orthogonally mounted sonars to jointly observe scenes without geometric assumptions or prior training, although limited by a small overlapping FOV. Later extensions [26], [27] expanded the FOV via Bayesian prediction but again required sonar training data.

**Opti-Acoustic approaches:** Combining sonar and camera data offers a way to resolve the depth and elevation ambiguities of each modality, but fusion is challenging due to their fundamentally different sensing mechanisms. Kim [28] proposed building separate optical and acoustic volumetric models that are iteratively merged, but the method is computationally expensive and not real-time. Roznere [29] and Cardaillac [30] matched sonar points to optical images to resolve scale ambiguity in visual Simultaneous Localization And Mapping (SLAM), but both approaches still depend on visual feature tracking, which is unreliable in turbid conditions.

Learning-based methods have also been explored: Qu [31] applied Gaussian splatting to acoustic–optical reconstruction, while Qadri [32] introduced AONeuS, a neural rendering framework for sonar–camera fusion. These techniques achieve promising results but require sonar datasets, which are scarce.

Other approaches focus on improving feature matching robustness in poor visibility. Babaee [33] used contours, and

Spears [34] restricted the feature search space to bounded optical regions, though both still rely on point features degraded by turbidity. Gutnik [35] avoided this limitation by detecting areas of interest in optical images and matching them directly to sonar depth values. Collado-Gonzalez [4] proposed an optical area-of-interest to sonar matching algorithm capable of reconstructing simple geometries in both clear and turbid conditions. However, this method assumes constant object depth along the vertical axis, and its segmentation process requires parameter tuning, limiting its generality.

## III. Problem Formulation

This work addresses the problem of **sensing and mapping** an underwater environment to support safe planning and decision-making in obstacle-rich conditions under varying visibility conditions.

### A. Gaussian Process Occupancy Mapping

We assume robot poses are known, so the occupancy probability of a map cell $m_j$ is

$$p(m_j \mid z_{1:t}, x_{1:t}), \tag{1}$$

where $z_{1:t}$ is the set of sensor observations and $x_{1:t}$ the set of robot poses. Gaussian Process (GP) occupancy mapping formulates mapping as a continuous probabilistic regression problem. The observation model is defined as

$$y = f(\mathbf{x}) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma_n^2), \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^3$ denotes an input location, $y \in \{0, 1\}$ is the observed latent occupancy, and $\epsilon$ is zero-mean Gaussian noise. A GP prior is placed over $f$:

$$f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')), \tag{3}$$

with covariance function $k(\mathbf{x}, \mathbf{x}')$. Sensor observations are used as training data $\{\mathbf{X}, \mathbf{y}\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, while map cells correspond to query (test) locations. The predictive distribution for a test location $\mathbf{x}_*$ is Gaussian

$$\begin{aligned} f(\mathbf{x}_*) \mid \mathbf{X}, \mathbf{y} &\sim \mathcal{N}(\mu_*, \sigma_*^2), \\ \mu_* &= k_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{y}, \\ \sigma_*^2 &= k(\mathbf{x}_*, \mathbf{x}_*) - k_*^\top (K + \sigma_n^2 I)^{-1} k_*, \end{aligned} \tag{4}$$

where $K_{ii'} = k(\mathbf{x}_i, \mathbf{x}_{i'})$, $k_* = \left[k(\mathbf{x}_1, \mathbf{x}_*), \ldots, k(\mathbf{x}_N, \mathbf{x}_*)\right]^\top$. This formulation enables smooth estimation of occupancy probabilities.

To capture real-world sharp variations, we adopt the Matérn covariance function with smoothness parameter $\nu = 3/2$:

$$k(d) = \sigma_f^2 \left(1 + \frac{\sqrt{3}d}{l}\right) \exp\left(-\frac{\sqrt{3}d}{l}\right), \tag{5}$$

where $d = \|\mathbf{x} - \mathbf{x}'\|$, $\sigma_f^2$ the prior signal variance, and $l$ the length-scale.

Regression outputs are mapped to occupancy probabilities using the logistic function:

$$p(y = 1 | \mathbf{x}_j) = \frac{1}{1 + \exp(-\gamma \omega_j)}, \quad \omega_j = \frac{\sigma_{\min}^2 \mu_j}{\sigma_j^2}, \tag{6}$$

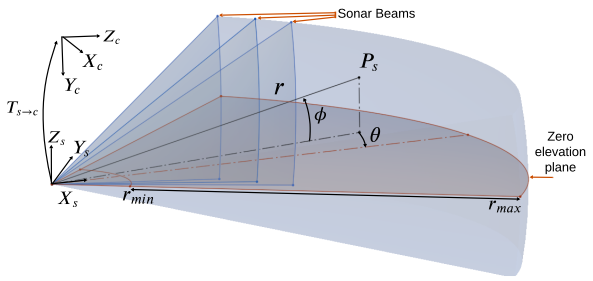



Fig. 2: **Forward looking imaging sonar model.** The point $\mathbf{P}_s$ can be represented by $[r, \theta, \phi]^T$ in a spherical coordinate frame. The range $r$ and the bearing angle $\theta$ of $\mathbf{P}_s$ are measured, while the elevation angle $\phi$ is not captured in the resulting 2D sonar image. The 3D transformation $\mathbf{T}_{s\to c}$ from sonar to camera frame is also depicted.

where $\sigma_{\min}^2$ is the minimum variance and $\gamma > 0$ is a scaling constant. Cells are classified as free, occupied, or unknown:

$$\text{state} = \begin{cases} \text{free}, & p < p_{\text{free}},\ {\sigma_j}^2 < \sigma_t^2 \\ \text{occupied}, & p > p_{\text{occupied}},\ {\sigma_j}^2 < \sigma_t^2 \\ \text{unknown}, & \text{otherwise} \end{cases} \tag{7}$$

with $\sigma_t^2$ serving as a variance threshold that expresses prediction confidence and balances the trade-off between predictive richness and classification accuracy.

### B. Sonar Model

Imaging sonars sense a 3D volume, as illustrated in Fig. 2. An imaging sonar measures points in spherical coordinates by emitting acoustic pulses and measuring the associated intensity $\beta \in \mathbb{R}_+$ from their returns. The transformation of a 3D point from spherical $[r, \theta, \phi]^T$ to Cartesian coordinates with respect to the sonar reference frame is given by

$$\mathbf{P}_s = \begin{pmatrix} x_s \\ y_s \\ z_s \end{pmatrix} = r \begin{pmatrix} \cos\phi\cos\theta \\ \cos\phi\sin\theta \\ \sin\phi \end{pmatrix}. \tag{8}$$

Here $r \in \mathbb{R}_+$ is the range, $\theta \in \Theta$ is the bearing, and $\phi \in \Phi$ is the elevation angle, with $\Theta, \Phi \subseteq [-\pi, \pi)$. Each acoustic ping return corresponds to a beam that spans elevation values $\phi \in [\phi_{\min}, \phi_{\max}]$, thus the sonar produces a 2D intensity image $I(r, \theta)$ that lacks explicit elevation information.

### C. Camera Model

Transforming a 3D point from the sonar reference frame $\mathbf{P}_s$ to the camera reference frame $\mathbf{P}_c$ is achieved using the extrinsic parameters $\mathbf{R}_{s\to c}$ the rotation matrix and $\mathbf{t}_{s\to c}$ the translation vector between the sonar and camera frames.

$$\mathbf{P}_c = \begin{pmatrix} x_c \\ y_c \\ z_c \end{pmatrix} = \mathbf{R}_{s\to c}\mathbf{P}_s + \mathbf{t}_{s\to c} \tag{9}$$

Using the pinhole camera model, a 3D point in the camera reference frame $\mathbf{P}_c$ is projected onto a 2D point $\mathbf{p}_c$ on the image plane using the intrinsic camera matrix $\mathbf{K}$.

$$\mathbf{p}_c = \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \mathbf{K}\frac{\mathbf{P}_c}{z_c}, \qquad \mathbf{K} = \begin{pmatrix} f_x & 0 & c_u \\ 0 & f_y & c_v \\ 0 & 0 & 1 \end{pmatrix}. \tag{10}$$

Here, $(f_x, f_y)$ are the focal lengths in pixels, and $(c_u, c_v)$ is the optical center of the camera.

### D. Data Association Problem

We assume a robot equipped with two forward-looking imaging sonars and one forward-looking camera, mounted with overlapping FOVs (Fig. 1). This configuration yields regions observed by all three sensors as well as areas covered only by the camera and a single sonar. Proper calibration allows measurements in overlapping regions to be associated to the same physical location.

Recovery of 3D sonar points requires knowledge of the elevation angle $\phi$. Following [25], features observed in orthogonal sonar images are fused in 3D as

$$\hat{\mathbf{P}}_s = \left( \frac{r^h + r^v}{2}, \theta^h, \theta^v \right), \tag{11}$$

where $r^h$, $r^v$ are range measurements from the horizontal and vertical sonars, and $\theta^h$, $\theta^v$ are the corresponding bearing angles. This association is only possible within the limited overlapping FOV, bounded by their vertical beam width. Outside this region, sonar measurements remain underconstrained and therefore undefined in 3D. Sonar provides range and bearing, while the camera contributes elevation. Fusing their complementary information enables full 3D reconstruction, but differing sensing mechanisms complicate integration and yield variable reliability.

This paper explores how overlapping sensor information can be leveraged to maximize the number of potential 3D measurements. Moreover, since measurements differ in reliability, we also address how to map the environment while incorporating information with varying levels of certainty.

## IV. Proposed Approach

In this section, we present our framework for underwater sensing and mapping. We first exploit the overlapping sensor FOVs and apply matching procedures to maximize the number of fully defined 3D measurements, each with an associated reliability. These measurements are then fused in a volumetric mapping process that prioritizes the most reliable data. An overview of the system is shown in Fig. 1c.

### A. Sonar Feature Extraction

Not all pixels in an acoustic image $I(r, \theta)$ contain meaningful information; noise and second returns are common. We therefore preprocess the data using the SOCA-CFAR filter [36], a variant of the Constant False Alarm Rate (CFAR) technique [37], to detect salient features. Let $F(r, \theta) = \text{SOCA-CFAR}(I(r, \theta))$ denote the filtered feature image, where only intensity values passing the CFAR threshold remain, an example of this can be seen in Fig. 1c. This approach effectively mitigates second returns, which frequently appear in sonar imagery [25]. Unlike some prior works [4], all detected features in the sonar image are retained in $F$, even if there are multiple range returns with intensities exceeding the CFAR threshold for a given bearing angle.

### B. Optical Image Segmentation

Instead of detecting individual point features in noisy, low-contrast underwater images, we propose detecting regions of interest (ROIs) corresponding to objects in the scene. Traditional segmentation methods often underperform underwater due to reliance on localized features and handcrafted extraction techniques, which are sensitive to poor contrast, color attenuation, and non-uniform lighting. Recent deep learning–based approaches have shown promising results in these conditions [38], [39].

We adopt YOLO11n-seg [40] for its real-time performance, robustness, and multi-scale object detection. The model is initialized with COCO-pretrained weights and fine-tuned on a small hand-annotated dataset of representative underwater imagery (distinct from validation sequences). Data augmentation, including random flips, rotations, saturation, brightness, and exposure, expands the training set and improves robustness to blur, low light, and color distortion. An example segmentation is shown in Fig. 1c. The resulting segmented regions of interest are denoted as $R \subset \mathbb{R}^2$.

### C. Stereo Sonar Matching

Given feature images $F_v$ and $F_h$ from the vertical and horizontal sonars, respectively, we need to associate features across the images to fully define them in 3D, as denoted in Eq. (11). Similar to [25], we perform range-based association, since range is the common denominator between the two sonars.

Unlike [25], we do not consider intensity information when performing feature matching. Sonar return intensity depends on several factors, including: reflectivity, incidence angle, and elevation ambiguity (i.e., the number of individual returns within the elevation angle of the beam.). According to the sonar image formation model formally presented in [23], the intensity of each pixel in the sonar image is proportional to the cumulative acoustic energy reflected by all objects intersected by the acoustic arc at that range and bearing. Because our sonars are oriented 90° apart, elevation effects vary significantly between them, making intensity unreliable for correspondence.

Consequently, matching is based solely on range and overlapping FOV. We only match features from the area inside the vertical FOV $\Phi \in [\phi_{min}, \phi_{max}]$ of each sonar's orthogonal companion, shown in red and blue in Fig. 1b. Features at the same range are paired as:

$$\mathcal{S} = \{(f_v, f_h) \mid r(f_v) = r(f_h), f_h, f_v \in \Phi\}. \tag{12}$$

These matches are then transformed into a 3D point cloud:

$$\mathcal{P}_{ss} = \{\pi(f_v, f_h) \mid (f_v, f_h) \in \mathcal{S}\}, \tag{13}$$

where $\pi(\cdot)$ denotes the stereo fusion function defined in Eq. (11).

The overlap region shrinks with proximity, so nearby objects not directly in the robot's line of sight are likely missed. In addition, sonar measurements are inherently sparse and noisy, so small objects may not be reliably detected by both sonars simultaneously. Yet, objects close to the robot are often the most critical for navigation, as they represent imminent collision threats. To compensate, we store close-range features not matched stereoscopically (features closer

than the minimum range among all valid stereo matches, $d_{\min}$):

$$\begin{aligned} \mathcal{C}_v &= \{f_v \in F_v \mid r(f_v) < d_{\min}\}, \\ \mathcal{C}_h &= \{f_h \in F_h \mid r(f_h) < d_{\min}\}. \end{aligned} \tag{14}$$

These sets of close-range, unmatched sonar features are later revisited, leveraging the higher-resolution information from the optical image (detailed in Secs. IV-E, IV-F).

### D. Stereo Sonar to Image Projection

$\mathcal{P}_{ss}$ is then projected onto the camera image using the intrinsic matrix $\mathbf{K}$ and the sonar-to-camera transform $\mathbf{T}_{s\to c}$:

$$\mathbf{p} = \begin{pmatrix} u \\ v \\ w \end{pmatrix} /w = \mathbf{K}\mathbf{T}_{s\to c} \begin{pmatrix} \mathbf{P}_s \\ 1 \end{pmatrix} = \tau(\mathbf{P}_s), \tag{15}$$

where $\tau : \mathbb{R}^3 \to \mathbb{R}^2$ is the projection from 3D points in the sonar reference frame to 2D image pixels $\mathbf{p} \in \mathbb{R}^2$.

Next, the segmented regions of interest $R$ from the RGB image are applied as a mask, discarding all projected points outside $R$. This step helps to remove erroneous matches. The resulting point cloud is: $\mathcal{Q}_{ss} = \{\mathbf{P}_s \in \mathcal{P}_{ss} \mid \tau(\mathbf{P}_s) \in R\}$.

If no region of interest is detected, we set $\mathcal{Q}_{ss} = \mathcal{P}_{ss}$ and skip the further inspection of close-range unmatched sonar features. This ensures mapping continues with sonar-only data, which may be necessary in low-visibility conditions.

### E. Sonar to Image Projection

Due to elevation ambiguity in sonar data, each unmatched feature $f_h \in \mathcal{C}_h$ and $f_v \in \mathcal{C}_v$ potentially corresponds to multiple elevations. Inspired by [4], we project each feature to all possible 3D points:

$$\mathcal{B}_j = \{\mathbf{P}_s\} = r \begin{pmatrix} \cos\phi_i \cos\theta \\ \cos\phi_i \sin\theta \\ \sin\phi_i \end{pmatrix}, \quad \phi_i \in [\phi_{min}, \phi_{max}]. \tag{16}$$

Each 3D point set $\mathcal{B}_j$ is then projected onto the camera image using Eq. 15. Because of our sensor configuration, $\mathcal{B}_j$ projects to vertical lines for horizontal sonar features $f_h$, and to horizontal lines for vertical features $f_v$. Let $\mathcal{O}$ represent the set of pixels $\mathbf{p}$ corresponding to the 3D point set $\mathcal{B}$ associated with the close-point feature sets $(\mathcal{C}_v \cup \mathcal{C}_h)$.

Finally, the segmented RGB regions $R$ are applied as a mask, discarding all projections outside of $R$. The resulting point cloud is: $\mathcal{Q}_s = \{\mathbf{P}_s \in \mathcal{O} \mid \tau(\mathbf{P}_s) \in R\}$.

### F. Image Expansion

Imaging sonars have a limited and ambiguous vertical FOV. Similarly to [4], we expand this FOV using optical image information. Distance is derived from the original sonar measurement, while elevation is inferred from the optical image.

For the horizontal sonar, let $\mathcal{O}_h$ be the set of optical image pixels corresponding to the 3D points $\mathcal{B}_h$ from the close feature set $\mathcal{C}_h$, restricted to the region of interest $R$. For each column of pixels in the optical image, we compute the mean depth $\bar{z}_c$ and assign it to all pixels in that column that lie within $R$ but outside the sonar's vertical FOV (i.e., not in $\mathcal{O}_h$). This expansion assumes the surface of the object in view has a fixed standoff distance from the sonar, which is common for man-made structures (i.e. piers), but may not apply to objects of arbitrary geometry. These expanded 2D pixels $\mathbf{p}$ are then back-projected to 3D using:

$$\mathbf{P}_c = \bar{z}_c \mathbf{K}^{-1} \mathbf{p}. \tag{17}$$

The same process is applied to the vertical sonar, where $\mathcal{O}_v$ corresponds to $\mathcal{C}_v$ and expansion is performed along optical image rows instead of columns. The two resulting sets of 3D points are combined to form the expanded point cloud $Q_e$.

### G. Gaussian Process Occupancy Mapping

We base our GP mapping solution on the fast and accurate 3D mapping framework presented in [13], which has also demonstrated effectiveness in underwater sparse-data scenarios [41]. Standard GP mapping assumes a constant noise variance $\sigma_n^2$, treating all observations as equally reliable. Instead, we adopt a heteroscedastic GP, assigning each observation its own noise value $\epsilon$ that reflects the reliability of its source. The GP predictive distribution at $\mathbf{x}_*$ remains Gaussian with

$$\begin{aligned} \mu_* &= k_*^\top \big(K + \Sigma\big)^{-1} \mathbf{y}, \\ \sigma_*^2 &= k(\mathbf{x}_*, \mathbf{x}_*) - k_*^\top \big(K + \Sigma\big)^{-1} k_*, \end{aligned} \tag{18}$$

where $\Sigma = \mathrm{diag}\big(\sigma_{n,1}^2, \ldots, {\sigma_{n,N}}^2\big)$. This formulation weights each data point by its own uncertainty, with larger $\sigma_{n,i}^2$ reducing that observation's influence on the posterior.

In our framework, three point clouds are used, each with different reliability: $\mathcal{Q}_{ss}$ (stereo sonar) makes no geometric assumptions and is assigned high confidence $\alpha_{ss}$. $\mathcal{Q}_s$ (sonar-to-image projection) assumes sonar range applies within sonar vertical FOV, giving medium confidence $\alpha_s$. $\mathcal{Q}_e$ (image expansion) assumes sonar range applies across the optical elevation span. This makes the strongest assumption, receiving the lowest confidence $\alpha_e$. Finally, noise is set as $\sigma_{n,i}^2 = 1/\alpha_i$, linking GP uncertainty directly to measurement confidence.

## V. Experiments and Results

To validate the proposed sensing and mapping framework, we evaluate its performance on diverse structures in both clear-water tank and turbid-water marina experiments.

### A. Hardware Overview

The experimental platform is a heavy-configuration BlueROV2 (Fig. 1a) equipped with a VectorNav VN-100 IMU (200 Hz), KVH DSP-1760 FOG (250 Hz), Bar30 pressure sensor (5 Hz), Rowe SeaPilot DVL (5 Hz), a low-light Sony Exmor IMX322/323 camera (5 Hz), and two forward-looking multibeam imaging sonars: Oculus M750d and M1200d. Both sonars were operated in their low-frequency wide-aperture modes (750 kHz and 1200 kHz), providing a 20° vertical and 130° horizontal FOV at a 3 m range. The sensor mounting configuration, shown in Fig. 1, allows the use of measured sensor offsets as simple extrinsic calibration parameters, including a 10cm vertical offset between the sonar coordinate frames, as well as 15cm

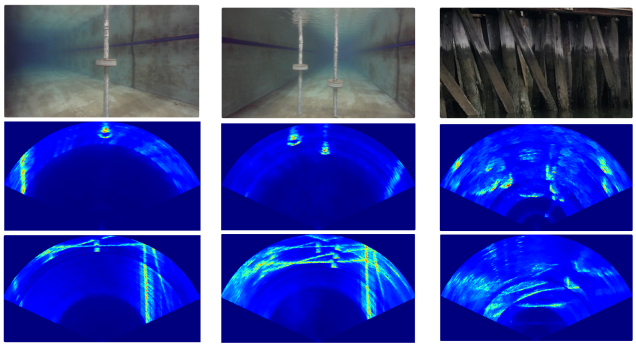


Fig. 3: **Testing environments and representative sonar images.** Environments include: Tank Single Disk (left column), Tank Double Disk (middle column), and Marina Pier (right column). The top row shows an example image of each environment, the middle row shows an example horizontal sonar image, and the bottom row shows an example vertical sonar image.

vertical and 15cm horizontal offsets between the sonar and camera frames.

Onboard computation is provided by a Pixhawk, Raspberry Pi, and Jetson Nano running ROS and used for data logging to a topside computer. Algorithms are executed via real-time playback on a workstation (Titan RTX GPU, Intel i9 3.6 GHz CPU), representing hardware that could be embedded on an AUV. The vehicle trajectory is estimated by integrating DVL velocities, FOG and IMU angular rates, and pressure-based depth measurements.

### *B. Benchmarks*

We compare our approach, Gaussian Process Confidence mapping with Stereo Sonar and RGB image (GPC SS RGB), against four baselines: (i) Standard GP occupancy mapping with Stereo Sonar and RGB (GP SS RGB), (ii) OctoMap with Stereo Sonar and RGB (Octo SS RGB), (iii) GP mapping with a single Sonar and RGB (GP S RGB) [4], and (iv) GP mapping with Stereo Sonar only (GP SS) [25]. The methods were chosen to showcase the impact of confidence values in the mapping process as well as the impact of using different sensor combinations. We exclude AONeuS [32] and Z-Splat [31], as they are not designed for real-time use and cannot generalize to scenes with unreliable visual information.

### *C. Indoor Tank Experiments*

The testing structures were designed to be geometrically more complex than typical pier pilings (Fig. 3, left and middle columns), with surface discontinuities (sudden changes in dimensions) and few distinctive features, making feature matching and 3D reconstruction challenging. The first setup consists of a single pole-mounted disk, while the second includes two pole-mounted disks at different heights. A 200-image, hand-annotated dataset was used to train a pretrained YOLO11n-seg model for detecting pilings and disks.

Key frames were sampled every 5 cm or 10° for all mapping results. Voxel resolution was set to 2.5 cm. GP regression and classification hyperparameters were manually tuned: $\sigma_n^2 = 0.01$, $l = 0.025$, $\sigma_f^2 = 0.1$, $\sigma_{min}^2 = 0.001$, $\sigma_t^2 = 50$, and $\gamma = 100$. GPC confidence values for tank tests were set to $\alpha_{ss} = 20$, $\alpha_s = 1$, and $\alpha_e = 0.5$.

Error metrics were computed by comparing generated voxel maps against a ground-truth CAD model. Absolute error is defined as the shortest distance from each voxel center to the CAD model, with a bounding box used to exclude irrelevant objects (e.g., tank walls, tank floor, etc.). Coverage was quantified via voxel count. Robot trajectories and full experiment playback are shown in the video attachment.

| Method | MAE (cm) ↓ | RMSE (cm) ↓ | SD (cm) ↓ | Precision (%) ↑ | Inlier Voxels ↑ |
|---|---|---|---|---|---|
| Octo SS RGB | 5.9844 | 0.8106 | 0.5468 | 34.5989 | **1212** |
| GP SS RGB | 6.6030 | 0.8816 | 0.5841 | 35.7711 | 450 |
| GPC SS RGB | **1.1712** | **0.1579** | **0.1059** | **88.1844** | 306 |
| GP S RGB | 6.0007 | 0.8305 | 0.5741 | 38.8336 | 273 |
| GP SS | 1.3590 | 0.1819 | 0.1209 | 81.5789 | 31 |

TABLE I: Performance of volumetric mapping methods over the single disk structure in tank setting - proposed method is highlighted in gray. Best results per metric are shown in bold.

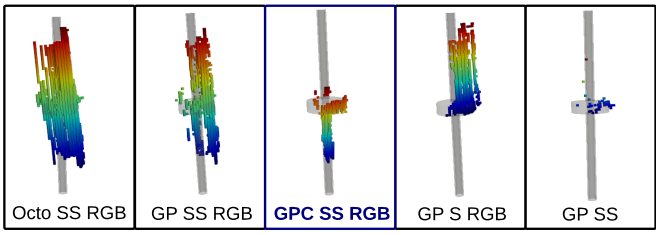



Fig. 4: **Tank single disk mapping results.** Each column in the image shows the voxel map result for a different method. From left to right the order is: Octo SS RGB, GP SS RGB, **GPC SS RGB** (proposed approach highlighted in blue), GP S RGB, and GP SS. Voxel colors depict height.

*1) **Tank Single Disk Results**:* Numerical results are reported in Table I, with visual results in Fig. 4. GP S RGB assumes a single distance from the robot applies to the entire structure. As a result, it reconstructs the scene as a cylinder with disk width, which is erroneous in this case where distances vary, leading to low accuracy. Octo SS RGB and GP SS RGB also show low accuracy, as they treat all data equally and cannot leverage confidence values to distinguish reliable measurements. GP SS achieves reasonable accuracy but has very limited coverage. In contrast, the proposed method, GPC SS RGB, attains the highest accuracy while maintaining substantial coverage.

| Method | MAE (cm) ↓ | SD (cm) ↓ | RMSE (cm) ↓ | Precision (%) ↑ | Inlier Voxels ↑ |
|---|---|---|---|---|---|
| Octo SS RGB | 2.6064 | 0.3141 | 0.4081 | 64.1379 | **1302** |
| GP SS RGB | 2.5962 | 0.3338 | 0.4229 | 68.8243 | 521 |
| GPC SS RGB | 1.6219 | 0.1651 | 0.2315 | **81.5825** | 598 |
| GP S RGB | **1.3885** | **0.1306** | **0.1906** | 79.5385 | 517 |
| GP SS | 2.3853 | 0.3511 | 0.4244 | 68.4685 | 76 |

TABLE II: Performance of volumetric mapping methods over the double disk structures in tank setting - proposed method is highlighted in gray. Best results per metric are shown in bold.

*2) **Tank Double Disk results**:* Numerical results are reported in Table II, with visual results in Fig. 5. GP S RGB achieves the lowest errors, but its single-distance assumption causes it to capture the poles while completely missing the disks. GP SS still exhibits very limited coverage. Octo SS RGB and GP SS RGB capture some of the varying distance information but also introduce errors. Once again,

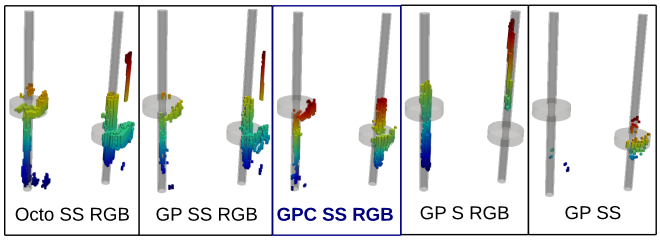



Fig. 5: **Tank double disk mapping results.** Each column in the image shows the voxel map result for a different method. From left to right the order is: Octo SS RGB, GP SS RGB, **GPC SS RGB** (proposed approach highlighted in blue), GP S RGB, and GP SS. Voxel colors depict height.

our proposed method, GPC SS RGB, achieves the highest accuracy while maintaining substantial coverage. Overall, the proposed framework demonstrates superior capability to map complex geometries, prioritizing critical information for robotic decision-making and path planning.

### *D. Outdoor Field Experiments*

To evaluate our approach in the field, we deployed our robot at the U.S. Merchant Marine Academy (USMMA) marina in King's Point, NY, a highly turbid environment, as can be seen in the example image in Fig. 1c. The observed wood piling pier is shown in the right column of Fig. 3. A 120-image hand-annotated dataset was used to train a pretrained YOLO11n-seg model to detect pilings. GPC confidence values were set to $\alpha_{ss} = 20$, $\alpha_s = 2$, and $\alpha_e = 1$. Due to the lack of ground-truth data, only qualitative results are presented in Fig. 6.

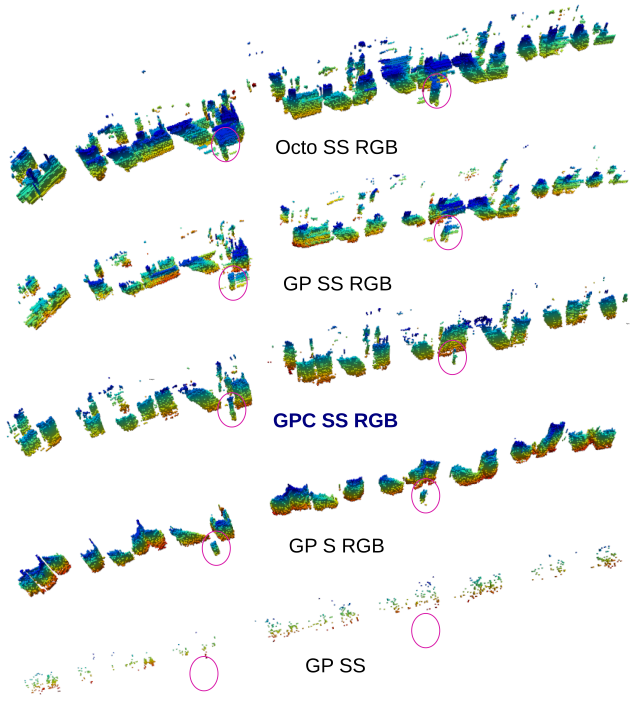



Fig. 6: **Marina Pier mapping results.** Each row in the image shows voxel map results for a different method. From top to bottom the order is: Octo SS RGB, GP SS RGB, **GPC SS RGB** (proposed approach highlighted in blue), GP S RGB, and GP SS. Voxel colors depict height. Pink circles emphasize the presence of two small metal pipes in front of the wooden pilings.

The resulting pier voxel maps are shown in Fig. 6. The pier includes two small pipes in front of the wooden pilings, circled in pink in the maps. GP SS is very sparse and fails to capture the pipes. GP S RGB produces a clean map with minimal outliers and captures the pipes perfectly, but it provides no information beyond the first sonar return at each range, and diagonal pier pilings appear triangular. GP SS RGB and Octo SS RGB clearly overestimate occupancy. Our method, GPC SS RGB, successfully captures the two small pipes, preserves the diagonally-oriented beams, and maps an inner row of pier pilings, even when partially occluded. As intended, the method prioritizes relevant information over raw coverage, capturing close objects essential for safety while maintaining background information important for path planning.

Mapping time results for GP, GPC, and Octomap across all experiments are reported in Table III. For fairness, only SS RGB methods are compared, since they map the same information. GPC and GP mapping show nearly identical computation times, indicating that adding confidence values has minimal impact. In contrast, Octomap has the highest computation times, in some cases more than double those of the GP-based approaches.

| Method | Tank Single Disk | Tank Double Disk | Marina Pier |
|---|---|---|---|
| Octo SS RGB | 0.1495 ± 0.0838 | 0.1168 ± 0.0414 | 0.2797 ± 0.3474 |
| GP SS RGB | **0.0551** ± 0.0634 | **0.0699** ± 0.0363 | 0.1040 ± 0.0365 |
| GPC SS RGB | 0.0555 ± **0.0197** | 0.0712 ± **0.0311** | **0.0970** ± **0.0324** |

TABLE III: Mapping time (in seconds) with standard deviation. The proposed approach is highlighted in gray. Best-performing results are shown in bold.

## VI. Conclusion

This paper presents a sensor fusion and volumetric mapping framework for underwater robots designed to operate reliably under varying visibility to support safe navigation in cluttered environments. The approach leverages overlapping sonar and camera fields of view, assigns confidence values to each fully defined 3D measurement, and fuses them in a volumetric mapping framework that prioritizes the most reliable data. The proposed framework is optimized for challenging underwater environments by relying primarily on sonar, incorporating camera data only when available, while falling back to stereo sonar-only mode if visual information is unavailable. CNN-based segmentation enables efficient training via transfer learning and data augmentation on visual datasets, avoiding manual parameter tuning and dependence on sonar datasets. The sensor fusion prioritizes mapping close-range, critical features for safe navigation.

We evaluate the method against state-of-the-art techniques across scenes with varying geometric complexity. Our approach outperforms existing methods in accuracy while maintaining adequate scene coverage. Field experiments further demonstrate the system's ability to map both small and large structures in low-visibility conditions. While the framework demonstrates robustness and high accuracy, future work includes extending the method to dynamic and large-scale environments, including improvements in sonar–camera calibration, and exploring adaptive sensing under different visibility conditions.

## REFERENCES


[1] Y. Cong, C. Gu, T. Zhang, and Y. Gao, "Underwater robot sensing technology: A survey," *Fundamental Research*, vol. 1, no. 3, pp. 337–345, 2021.

[2] Water Linked, "Discover the advantages of real-time 3D sonar," 2025. [Online]. Available: https://waterlinked.com/3dsonar

[3] K. Hu, T. Wang, C. Shen, C. Weng, F. Zhou, M. Xia, and L. Weng, "Overview of underwater 3D reconstruction technology based on optical images," *J. Marine Science and Engineering*, vol. 11, no. 5, 2023.

[4] I. Collado-Gonzalez, J. McConnell, P. Szenher, and B. Englot, "Opti-acoustic scene reconstruction in highly turbid underwater environments," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025, pp. 1282–1289.

[5] M. Gomes, M. Oliveira, and V. Santos, "Volumetric occupancy detection: A comparative analysis of mapping algorithms," *arXiv preprint arXiv:2307.03089*, 2023.

[6] M. Grimaldi, N. Palomeras, I. Carlucho, Y. R. Petillot, and P. Ridao Rodriguez, "FRAGG-Map: Frustum accelerated GPU-based grid map," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 1138–1144.

[7] J. Jung, S. Boche, S. B. Laina, and S. Leutenegger, "Uncertainty-aware visual-inertial SLAM with volumetric occupancy mapping," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 14 550–14 556.

[8] Y. Cai, F. Kong, Y. Ren, F. Zhu, J. Lin, and F. Zhang, "Occupancy grid mapping without ray-casting for high-resolution LiDAR sensors," *IEEE Transactions on Robotics*, vol. 40, pp. 172–192, 2024.

[9] K. Ebadi, L. Bernreiter, H. Biggie, G. Catt, Y. Chang, A. Chatterjee, C. E. Denniston, S.-P. Deschênes, K. Harlow, S. Khattak, L. Nogueira, M. Palieri, P. Petráček, M. Petrlík, A. Reinke, V. Krátký, S. Zhao, A.-a. Agha-mohammadi, K. Alexis, C. Heckman, K. Khosoussi, N. Kottege, B. Morrell, M. Hutter, F. Pauling, F. Pomerleau, M. Saska, S. Scherer, R. Siegwart, J. L. Williams, and L. Carlone, "Present and future of SLAM in extreme environments: The DARPA SubT challenge," *IEEE Transactions on Robotics*, vol. 40, pp. 936–959, 2024.

[10] B. Lindqvist, A. Patel, K. Löfgren, and G. Nikolakopoulos, "A tree-based next-best-trajectory method for 3-D UAV exploration," *IEEE Transactions on Robotics*, vol. 40, pp. 3496–3513, 2024.

[11] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: an efficient probabilistic 3D mapping framework based on octrees," *Autonomous Robots*, vol. 34, no. 3, pp. 189–206, 2013.

[12] S. T. O'Callaghan and F. T. Ramos, "Gaussian process occupancy maps*," *The International Journal of Robotics Research*, vol. 31, no. 1, pp. 42–62, 2012.

[13] J. Wang and B. Englot, "Fast, accurate gaussian process occupancy maps via test-data octrees and nested bayesian fusion," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 1003–1010.

[14] L. Wu, C. Le Gentil, and T. Vidal-Calleja, "VDB-GPDF: Online gaussian process distance field with VDB structure," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 374–381, 2025.

[15] M. D. Aykin and S. Negahdaripour, "Forward-look 2-D sonar image formation and 3-D reconstruction," in *2013 OCEANS - San Diego*, 2013, pp. 1–10.

[16] E. Westman and M. Kaess, "Wide aperture imaging sonar reconstruction using generative models," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 8067–8074.

[17] M. D. Aykin and S. Negahdaripour, "Three-dimensional target reconstruction from multiple 2-D forward-scan sonar views by space carving," *IEEE Journal of Oceanic Engineering*, vol. 42, no. 3, pp. 574–589, 2017.

[18] T. Guerneve, K. Subr, and Y. Petillot, "Three-dimensional reconstruction of underwater objects using wide-aperture imaging sonar," *Journal of Field Robotics*, vol. 35, no. 6, pp. 890–905, 2018.

[19] J. Park, H. Baek, B. Jun, and P. 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.

[20] 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 *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 8730–8737.

[21] 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*, vol. 6, no. 2, pp. 1535–1542, 2021.

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

[23] M. Qadri, M. Kaess, and I. Gkioulekas, "Neural implicit surface reconstruction using imaging sonar," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 1040–1047.

[24] A. V. Sethuraman, M. Rucker, O. Bagoren, P.-C. Kung, N. N. Amutha, and K. A. Skinner, "SonarSplat: Novel view synthesis of imaging sonar via gaussian splatting," *IEEE Robotics and Automation Letters*, vol. 10, no. 12, pp. 13 312–13 319, 2025.

[25] J. McConnell, J. D. Martin, and B. Englot, "Fusing concurrent orthogonal wide-aperture sonar images for dense underwater 3D reconstruction," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2020, pp. 1653–1660.

[26] J. McConnell and B. Englot, "Predictive 3D sonar mapping of underwater environments via object-specific bayesian inference," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 6761–6767.

[27] J. McConnell, I. Collado-Gonzalez, P. Szenher, and B. Englot, "Large-scale dense 3-D mapping using submaps derived from orthogonal imaging sonars," *IEEE Journal of Oceanic Engineering*, vol. 50, no. 1, pp. 354–369, 2025.

[28] J. Kim, M. Lee, S. Song, B. Kim, and S.-C. Yu, "3-D reconstruction of underwater objects using image sequences from optical camera and imaging sonar," in *OCEANS 2019 MTS/IEEE SEATTLE*, 2019, pp. 1–6.

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

[30] A. Cardaillac and M. Ludvigsen, "Camera-sonar combination for improved underwater localization and mapping," *IEEE Access*, vol. 11, pp. 123 070–123 079, 2023.

[31] 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 Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–12, 2024.

[32] 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. ACM SIGGRAPH*, 2024.

[33] M. Babaee and S. Negahdaripour, "3-D object modeling from occluding contours in opti-acoustic stereo images," in *OCEANS - San Diego*, 2013.

[34] A. Spears, A. M. Howard, M. West, and T. Collins, "Acoustic sonar and video sensor fusion for landmark detection in an under-ice environment," in *OCEANS - St. John's*, 2014.

[35] Y. Gutnik, I. Fabian, N. Zagdanski, O. Gal, T. Treibitz, and M. Groper, "Enhancing AUV 3D obstacle avoidance: A novel approach with self-supervised network for fusion of forward-looking camera and sonar data," in *OCEANS - Halifax*, 2024.

[36] K. El-Darymli, P. McGuire, D. Power, and C. R. Moloney, "Target detection in synthetic aperture radar imagery: a state-of-the-art survey," *Journal of Applied Remote Sensing*, vol. 7, no. 1, p. 071598, 2013.

[37] M. Richards, "Fundamentals of radar signal processing," *McGraw-Hill Education*, 2005.

[38] W. Akram, A. Baidar Bakht, M. Ud Din, L. Seneviratne, and I. Hussain, "Enhancing aquaculture net pen inspection: A benchmark study on detection and semantic segmentation," *IEEE Access*, vol. 13, pp. 3453–3474, 2025.

[39] M. Mohammadi, A. Abdullah, A. Juneja, I. Rekleitis, M. J. Islam, and R. Zand, "Edge-centric real-time segmentation for autonomous underwater cave exploration," in *2024 International Conference on Machine Learning and Applications (ICMLA)*, 2024, pp. 1404–1411.

[40] G. Jocher and J. Qiu, "Ultralytics YOLO11," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[41] J. Wang, T. Shan, and B. Englot, "Underwater terrain reconstruction from forward-looking sonar imagery," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 3471–3477.