SonarCloud: A Simulated Forward Looking Sonar Dataset for Underwater Perception Tasks

Nael Jaber [*,‡], Qais Al-Ramahi [*,‡], Leif Christensen[‡], Bilal Wehbe [‡]

[‡]Robotics Innovation Center, German Research Center for AI (DFKI), Bremen, Germany

***Abstract*—With over 80% of the world's oceans remaining unmapped and unexplored, the advancement of robust underwater perception technologies is becoming more important than ever before. This vast frontier cannot be reliably observed with optical cameras mounted on Autonomous Underwater Vehicles (AUVs), which struggle in turbid, low-light conditions common to marine environments. Acoustic sensors like the Forward-Looking Sonar (FLS) are essential alternatives, yet progress in the field is significantly hampered by a profound scarcity of large-scale, publicly available sonar datasets. To address this critical gap, we introduce the SonarCloud Dataset, a comprehensive synthetic dataset generated to accelerate research in underwater perception. Our dataset consists of FLS and depth imagery of 19 distinct objects, totaling approximately 500,000 images, along with a 3D point cloud for each object generated from the corresponding depth maps in various orientations. As technical validation, we selected object detection and 3D reconstruction to evaluate the effectiveness of our dataset. We demonstrate that state-of-the-art models trained solely on simulated data from our dataset can successfully detect objects in real-world sonar images and reconstruct their 3D shapes. The SonarCloud Dataset is presented as a valuable tool for the research community, and it can be found publicly in: https://doi.org/10.5281/zenodo.16645568.**

I. Introduction

Advancements in underwater robotics increasingly rely on robust perception capabilities for different tasks ranging from object detection to detailed 3D reconstruction. However, the underwater environment presents significant challenges that complicate data collection and limit the effectiveness of common sensing techniques like optical imaging [1].

While optical cameras are widely used in terrestrial applications, they come with many impracticalities when utilized underwater. Certain factors frequently render visual inspection impracticable [2], such as: **1)** the rapid scattering of light, leading to color loss and distortion, **2)** the lack of natural lighting at depth, and **3)** the existence of turbidity and mobile sediments; particularly in coastal or disturbed waters, after storm events [3] and in harbor areas [4].

These limitations means it is necessary to use alternative sensors, with acoustic methods like 2D sonars being particularly well-suited for underwater imaging due to sound's superior propagation in water. Sonar systems can provide valuable information about underwater scenes regardless of optical clarity [5]. In our work, we look specifically at the Forward Looking Sonar (FLS) which is known for its high resolution in range-sensing and its compact size, allowing it to be mounted on underwater robots easily. However, the FLS has a major hardware limitation, which is the loss of the elevation angle. That is the reason why objects appear in 2D from the FLS. Moreover, progress in developing and evaluating sophisticated perception algorithms using sonar data is significantly brought back by a critical factor, which is the scarcity of comprehensive, publicly available datasets.

Real-world underwater data collection, whether optical or acoustic, remains a complex task, often constrained by high operational costs, logistical challenges associated with accessing marine environments, and potential legal or environmental regulations [6]. Furthermore, even when sonar data is acquired, existing datasets are often hidden from the public eye, lack the precise ground truth information (e.g., accurate sensor poses) required for rigorous algorithm validation, or cover only limited scenarios and object types [1].

To address all the previously mentioned limitations, we introduce a dataset, generated using the Stonefish simulator [7], focusing on different objects in realistic underwater scenarios to provide a reproducible platform for developing and validating underwater perception algorithms. Given that object representations in sonar images are highly dependent on the sensor's position, affecting both geometric appearance and shadow casting; this dataset provides simulated acoustic images captured from linear approaches to target objects, enabling the analysis of feature variations with viewpoint changes. We further provide simulated depth images of the scanned objects from multiple viewpoints, and a groundtruth reconstructed pointcloud. This dataset is created to provide the research community a valuable resource based on FLS data. The dataset can be utilized for training different machine learning models that involve sonars as a sensing modality, and in a wide range of perception applications such as: object detection, segmentation, pose estimation, and 3D reconstruction. As a technical validation of the dataset's real-world applicability, two of the most recent object detection models trained solely on the synthetic data were tested directly on real FLS images, which were able to detect the target objects. A 3D reconstruction framework was used to 3D reconstruct the objects from their 2D appearance in the sonar image. A python-based visualization tool was also produced as part of this work called 'SonarCloudViz'.

[*] share first authorship of this paper.

This work is supported by the "CleanSeas" project (funded by BMFTR, Grant ID: 01IW22003).

II. Related Work

There are only a select number of real-world, state-of-the-art Forward Looking Sonar datasets that have been made publicly available, and these are: *NKSID* [8], *UATD* [9], *UXO* [1], and *MDT* [10] datasets with each using real sonar systems to capture different underwater scenarios ranging from debris-laden, underwater infrastructure to search-and-rescue missions and spanning target objects of: propellers, boxes, bottles, mortar shells, mannequins, chains etc.

These datasets are extremely useful but also pinpoint a fundamental trade-off in underwater data collection. On one hand, datasets collected in uncontrolled sea environments like NKSID, accurately reflect real-world conditions but consequently suffer from severe class imbalance. For instance, common debris like "Tires" are found in abundance on the seafloor but the same thing cannot be said about "Fishing Nets". This makes it difficult to train robust models. On the other hand, when experiments are done in a controlled water tank like in *UXO* and *MDT*, this can ensure class balance and precise ground truth. However, they face different limitations; as noted by [11], objects are often captured from limited viewpoints, and the sterile tank environment fails to replicate the complex acoustic terrains and clutter found in natural bodies of water.

This is where simulation comes into play. It provides the ability to generate vast, diverse datasets and offers complete control over the object viewpoints and angles as well as the complexity of the environment and class distribution. This reliance on simulation is increasingly viable due to the emergence of advanced open-source tools [12]. Platforms, like the Stonefish simulator [7] offer robust customization for acoustic data. While simulators like HoloOcean [13] - [14] utilize powerful game engines like the Unreal Engine to generate highly realistic sonar and camera imagery bridging that 'Sim-to-real' gap more and more.

One such work that has utilised these sort of simulations is Oliveira et al. [15] who used the HoloOcean simulator to synthesize their own dataset consisting of primitive shapes, anchors and propellers in a water tank environment. Another piece of work that have synthesized their own data, but this time using the Gazebo simulator [16] is Ściegienka and Blachnik [17] who curated a large-scale dataset of 69,444 images but was limited to only Unexploded Ordnance models (UXO) objects. In the work by Wang et al. [18] and Jaber et al. [2] the respective authors tackle the inherent problem of missing 3D information in 2D sonar images. To train their distinct deep learning models, they developed a custom acoustic camera simulator. This simulator was used to generate a paired dataset where each standard 2D sonar image had a corresponding ground truth "pseudo front view," which is essentially equivalent to a depth map from a standard optical camera.

As established in the literature review, a notable disparity exists: despite the availability of several powerful simulators and a handful of real-world datasets, simulated datasets remain scarce. The main application for these datasets in the literature is to train robust deep learning models on several different tasks, specifically for perception in the field of Computer Vision. For example, these models can do tasks like: object detection, classification, image segmentation, simultaneous localization and mapping (SLAM), and 3D reconstruction [12].

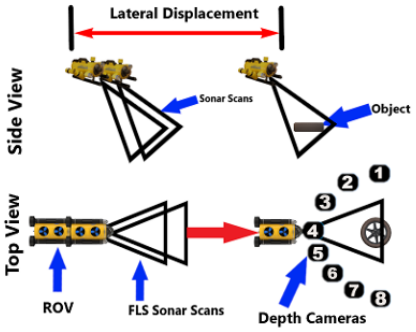

Fig. 1: A schematic illustrating the scanning procedure of target objects.

Focusing on object detection, there exists many works in the literature where several well-known 2D object detection models are used in the underwater domain to detect objects of different types from optical images. [19]–[23], all worked on the object detection of different sea creatures (starfish, sea cucumbers, sea urchins etc.) from the URPC dataset [24]. Many other papers are specialized in applying object detection underwater to man-made infrastructures like pipelines and cables which is useful for detecting leaks, defects or obstructions [25]–[27]. Other works in the literature are dedicated to the monitoring and analysis of the quality of marine ecosystems so in order to know the biodiversity of a species of fish for example, marine biologists would need to utilise deep learning techniques for fish detection and classification [28]–[30].

In the acoustic domain, 2D object detection is also well-established, with research heavily focused on imagery from two primary sensors: Side-Scan Sonar (SSS) and FLS. Applications involve several different objects, such as: pipes, chains, tires, mines, hulls, walls, etc. [31]–[36]. Zhang et al. focused his work on more primitive shaped objects in SSS scans [37]. Other work was done on unexploded ordnance and underwater mines [38], [39].

Another prominent perception task is the 3D reconstruction of objects underwater from both optical and sonar images. By combining the visual detail from optical cameras and the geometric accuracy of sonar, this method creates 3D models/pointclouds of submerged objects and terrain. This is invaluable to marine science as a whole and again in tasks like industrial inspection or underwater archaeology. This can be found in many papers including the following ones: [2], [18], [40]–[47].

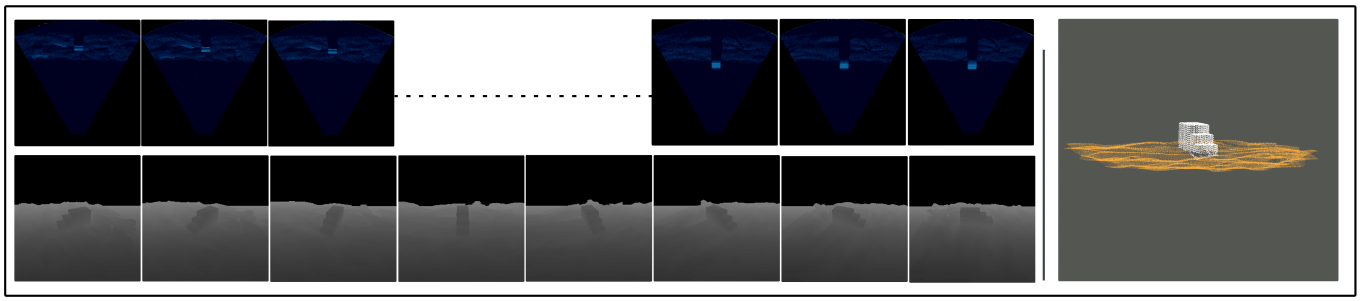

Fig. 2: Sample acoustics images, depthmaps, and their resulting reconstructed pointcloud from a scan of a stairs model. The first row displays samples of simulated acoustic images: the first three represent the initial scans, while the last three correspond to the final scans captured along the linear trajectory. The second row shows the eight corresponding depth maps captured over a span of 180 degrees. The figure on the right illustrates the 3D point cloud reconstructed from the eight acquired depth maps.

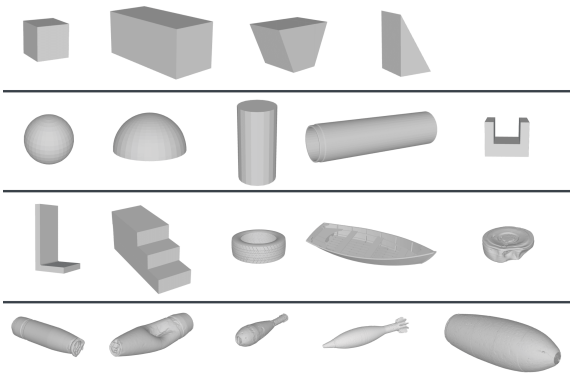

Fig. 3: Visual representations of the CAD models for all dataset objects

III. Dataset Generation

A. Stonefish Simulator

This work utilizes a synthetic dataset generated using the Stonefish underwater robotics simulator. Stonefish has the ability to model realistic underwater physics, including hydrodynamic effects, sensor behavior, and environmental conditions like turbidity and currents. It allows for precise definition of object properties and incorporates a GPU-based sonar simulation module for generating acoustic data. For this dataset, a FLS was simulated and configured with parameters matching the high-frequency mode of the **Oculus M1200d**: 0.1m minimum and 10m maximum range, 60° horizontal and 12° vertical aperture, and 512 beams.

B. Environment and setup

A forward-looking sonar performed a linear sweep toward each target in 24 equal steps, capturing one acoustic frame per step. The sequence begins when the object first appears in the sonar field of view and ends when it is no longer visible, ensuring that the target is fully scanned. Simultaneously, eight depth cameras (256×256 resolution, 60° FOV) were mounted at regular intervals on a circular arc of approximately 3 m radius around each object, providing eight distinct viewpoints for depth-map capture. This method of taking a batch of images in a linear manner is deliberately chosen to account for the changing shadow casts behind the objects based on the viewpoint. These changing shadow patterns, as well as the object geometry itself changing in the sonar images, provide valuable features to be learnt on the sonar image data by learning-based models. The 8 depth cameras are used to give a 180° perspective of the object so that it is not only limited to a single front-facing view. Given the fixed configuration of the depth cameras in simulation, a transformation matrix $(R_n, t_n)^{-1}$ was applied to the eight captured depth images creating a 3D pointcloud of the scanned object. Fig 1 gives an illustration of the data capturing process, while Fig 2 shows sample simulation data for a stairs model. The data collection configuration was adapted from our previous work on multi-view 3D reconstruction [48], where this setup was successfully employed and demonstrated its effectiveness.

To simulate realistic seabed conditions, target objects were situated on varied sand terrains, including rugged, moderately flat, and smooth topographies, alongside a baseline scenario of a completely flat and untextured terrain. This generated around 48,000 scenarios, with objects placed in various orientations and on varied terrains, corresponding in a total number of 142,080 acoustic images with 8 depth images per scenario case.

C. Target Objects

The collected dataset consists of acoustic scans captured for 19 different objects; ranging from basic shapes: sphere, semi-sphere, cylinder, pipe, cube, triangular prism, trapezoidal prism, rectangle; to more complex geometries: U-shape, L-shape, boat, stairs, and tire. The dataset was further enhanced with UXOs such as: Land-mine, mortar shells, deformed artillery-shell, 100lbs UXO, and 500lbs UXO. The aim of their inclusion is to improve their detection and identification, enabling more effective removal of unexploded ordnance from underwater environments. This application is critically important, not only for ensuring safer navigation and human activity in marine environments, but also for protecting marine

ecosystems from the long-term ecological damage caused by these hazardous remnants [49]. Table I shows the dimensions and sizes of the dataset's objects. Figure 3 shows the 3D CAD models of all dataset objects.

TABLE I: Overview of objects in the dataset with their labels, categories, dimensions, and corresponding image counts

Object Name	Category	Dimensions (m)	No. of Sonar Images
Sphere	Simple	d = 0.4, r = 0.2	3840
Semisphere	Simple	d = 0.6 , r = 0.3	3840
Cylinder	Simple	d = 0.8 , r = 0.2	3840
Pipe	Simple	d = 3.1 , r = 0.832	3840
Cube	Simple	l = 0.3 , w = 0.3 , h = 0.3	3840
Rectangle	Simple	l = 0.4 , w = 0.9 , h = 0.4	3840
Triangle	Simple	l = 0.47 , w = 0.5 , h = 0.8	3840
Trapezoid	Simple	l = 0.6 , w = 0.4 , h = 0.4	3840
U-Shape	Complex	l = 0.5 , w = 0.4 , h = 0.4	3840
L-Shape	Complex	l = 0.4 , w = 0.4 , h = 0.8	3840
Boat	Complex	l = 0.563 , w = 1.8 , h = 0.3	3840
Tire	Complex	d = 0.655 , r = 0.113	3840
Stairs	Complex	l = 0.35, w = 0.99 , h = 0.45	3840
Mine	UXO	d = 0.12, r = 0.15	15360
Artillery Shell	UXO	d = 0.58, r = 0.06	15360
Deformed Artillery Shell	UXO	d = 0.40 , r = 0.05	15360
Large Mortar Shell	UXO	d = 0.36, r = 0.06	15360
Small Mortar Shell	UXO	d = 0.31, r = 0.04	15360
500lbs	UXO	d = 0.56, r = 0.09	15360

D. SonarCloudViz

To facilitate the visualization of the dataset contents, a Python-based Graphical User Interface (GUI) named SonarCloudViz was developed. For each recorded scan within the dataset, the tool displays the 24 captured sonar images, their eight corresponding ground truth depthmaps, and a 3D plot of the corresponding pointcloud. The GUI allows users to interactively manipulate the 3D point cloud view to better inspect the object's geometry. This tool serves as a valuable utility for understanding the dataset's structure and visualizing its synchronized multi-modal components. Fig 4 showcases a sample of how SonarCloudViz shows the data of a scan.

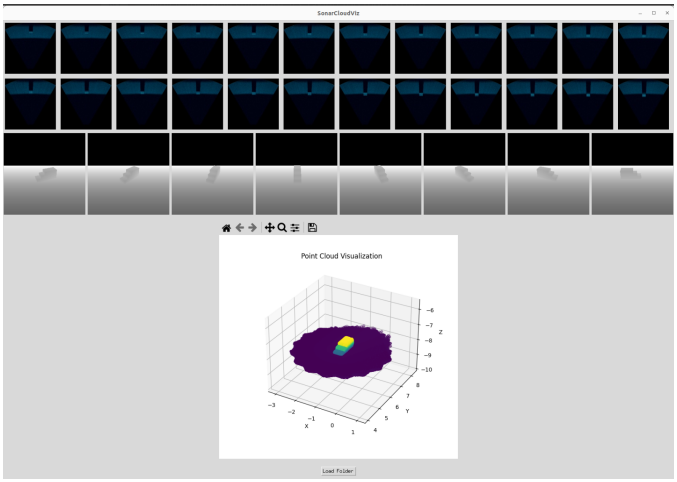

Fig. 4: An illustration of the GUI visualizing a scan of the stairs model.

IV. Technical Validation

In this work, we validate the effectiveness of our dataset through two key applications: 2D object detection and 3D reconstruction. This evaluation confirms the dataset's practical value in training models for multiple sonar-based perception tasks.

A. Object Detection

To validate the dataset's utility and benchmark its challenges for underwater perception, a technical evaluation was conducted using state-of-the-art (SOTA) 2D object detection methods. Modern deep learning architectures, particularly those based on Convolutional Neural Networks (CNNs) and Transformers, have become the standard for this task, largely supplanting earlier methods that relied on handcrafted features [50]. The You Only Look Once (YOLO) series [51] , for instance, works by splitting the image into $S \times S$ grid of cells with each grid cell responsible for detecting an object if the center of that object falls within it. The grid cells are then responsible for predicting the coordinates of the bounding box and its confidence score whilst simultaneously also predicting the probability of the detected object being part of each class. The results then undergo non-maximum suppression which eliminates the excess presence of bounding boxes. The final output is then a prediction vector containing the x, y values which are the coordinates of the centre of the bounding box with respect to the grid cells and w, h being the width and height of the box relative to the whole image, as well as probabilities of the detected object being part of each class. The evolution of 2D object detection models did not stop there however, since after the emergence of single-shot object detectors, research made another breakthrough when they introduced transformer-based 2D object detection models - taking these transformers from the Natural Language Processing world to the Computer Vision field. The DEtection TRansformer (DETR) [52] was one of the most important examples representing this class. DETR utilises a hybrid architecture of a Convolutional Neural Network which outputs a low-resolution feature map. This is then inputted into an encoder-decoder transformer. The encoder uses a self-attention mechanism to weigh the importance of the features found, while the decoder processes a fixed number of learnable "object queries" to probe for the presence of objects in images. This architecture is trained end-to-end using a bipartite matching loss, which enforces a one-to-one match between predicted and ground-truth objects. This means it does not need to carry out post-processing techniques like the non-maximum suppression.

1) Experiements and Results: In this paper, we specifically train the latest version of YOLO - the YoloV12 model [53] - which incorporates the self-attention mechanism into the classic YOLO architecture, as well as the RF-DETR model on our dataset to carry out our 2D object detection. The RF-DETR model combines a pretrained DINOv2 backbone with the structure of the LW-DETR [54]. This makes it highly effective at being fine-tuned and used for different tasks and real-world data. The models were trained specifically on the flat, untextured terrain portion of the dataset, which included all orientations of each object in simulation.

The dataset comprised 9,694 images, augmented through vertical flipping, 90° rotations (clockwise and counter-clockwise), and ±15° rotations. The data was split into training (80%), validation (10%), and testing (10%) sets. The test set included both simulated and real underwater images. The real images were collected in a water basin at the DFKI facility in Bremen, Germany, and featured three distinct objects: stairs, L-shape, and U-shape. Data acquisition was performed using an Oculus M1200 in high-frequency mode, scanning the objects placed on the basin floor. The YOLOv12 model achieved an average mAP of 94%, an F1-score of 88.8%, a precision of 93.6%, and a recall of 87.7% at a 50% confidence threshold. Its learning curve plateaued after approximately 50 epochs. In comparison, the RF-DETR model converged after just 23 epochs and achieved an average mAP of 82.8%, an F1-score of 86.2%, a precision of 88.5%, and a recall of 85.3% at the same threshold. Both models were trained using a batch size of 16 images. Fig 5 shows sample detection results from the evaluation of RF-DETR model on real sonar images of the U-shape, L-shape, and Stairs.

TABLE II: Average Precision (mAP@50) by object class for the two evaluated models

Object Class	YOLOv12 mAP (%)	RF-DETR mAP (%)
All Classes (Overall)	**94.0**	82.8
Sphere	**85.8**	36.8
Semisphere	**76.8**	39.8
Cylinder	**97.8**	96.8
Pipe	**99.8**	96.8
Cube	**99.8**	93.0
Rectangle	**96.8**	94.8
Triangle	**100**	36.8
Trapezoid	**100**	92.8
U-shape	78.8	**80.8**
L-shape	**89.8**	66.8
Boat	**100**	**100**
Tire	91.8	**93.8**
Stairs	**100**	89.8
Mine	**100**	89.8
Artillery Shell	**96.8**	92.8
Deformed Artillery Shell	**98.8**	**98.8**
Large Mortar Shell	**99.8**	**99.8**
Small Mortar Shell	**93.8**	85.8
500lbs	**97.8**	95.8

2) Discussion: Looking over Table II, a clear trend emerges when analyzing performance based on object geometry and size. For classes with simple, well-defined geometric shapes; such as Cube, Trapezoid, and Triangle; both models achieved perfect or near-perfect scores. These shapes are characterized by sharp edges and clear boundaries, which likely make them easier to detect consistently.

Both models also performed exceptionally well on the Big UXO class and achieved perfect scores on the Boat class. These objects are large and prominent in the imagery, which minimizes ambiguity and allows for highly reliable detection. Similarly, the consistently high scores for the Deformed Artillery Shell class indicate that, despite their irregular structure, these objects contain distinctive features that make them easily distinguishable from the background.

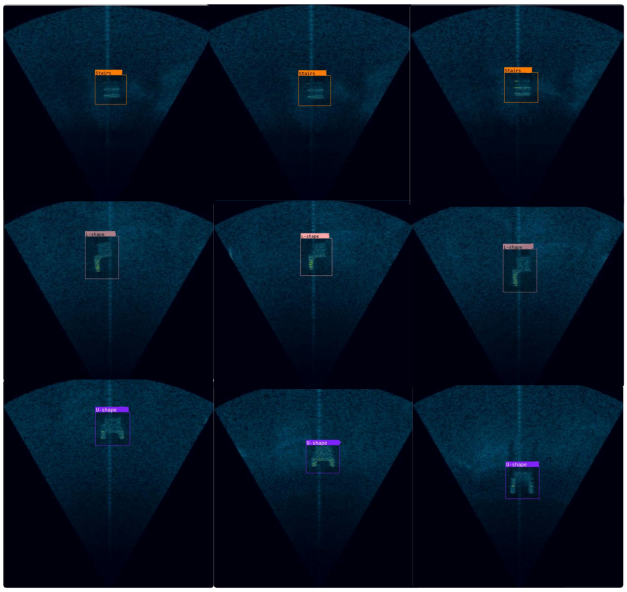

Fig. 5: The figure showcases sample 2D detections from testing the RF-DETR model on real sonar images of each of the 3 objects: Stairs, L-shape and U-shape corresponding to rows 1,2, and 3 respectively

The Stairs class is another noteworthy example of high performance. Although stairs are structurally more complex than primitive shapes, their strong, repetitive, and periodic pattern appears to provide a highly reliable signal for detection.

In contrast, smaller or more geometrically ambiguous objects, such as Semisphere and Sphere, showed a noticeable drop in performance. These objects have smoother surfaces and fewer distinctive edges, making them harder to separate from the background and more susceptible to minor variations in viewpoint or noise.

The Tire and U-shape classes also highlight the effect of geometry on detection performance. Their hollow or open structures introduce additional complexity, which in some cases appears to reduce detection accuracy. Similarly, the confusion between L-shape and U-shape can be attributed to their geometric similarity; both share two prominent right-angled segments, with the absence of the third side in the L-shape providing only a subtle geometric difference that is sometimes misclassified.

The evaluation results for the U-shape, L-shape, and Stairs classes are especially noteworthy, as these are the only classes for which the test set included a substantial number of real images alongside synthetic data. Despite the presence of real-world variations not seen during training, both models performed well, demonstrating strong generalization capabilities.

YOLOv12 achieved high mAP scores on all three classes, notably reaching 100 on Stairs and above 78 on both U-shape and L-shape. RF-DETR also showed competitive results, particularly excelling on U-shape with an 80.8 mAP, although its performance on L-shape was somewhat lower.

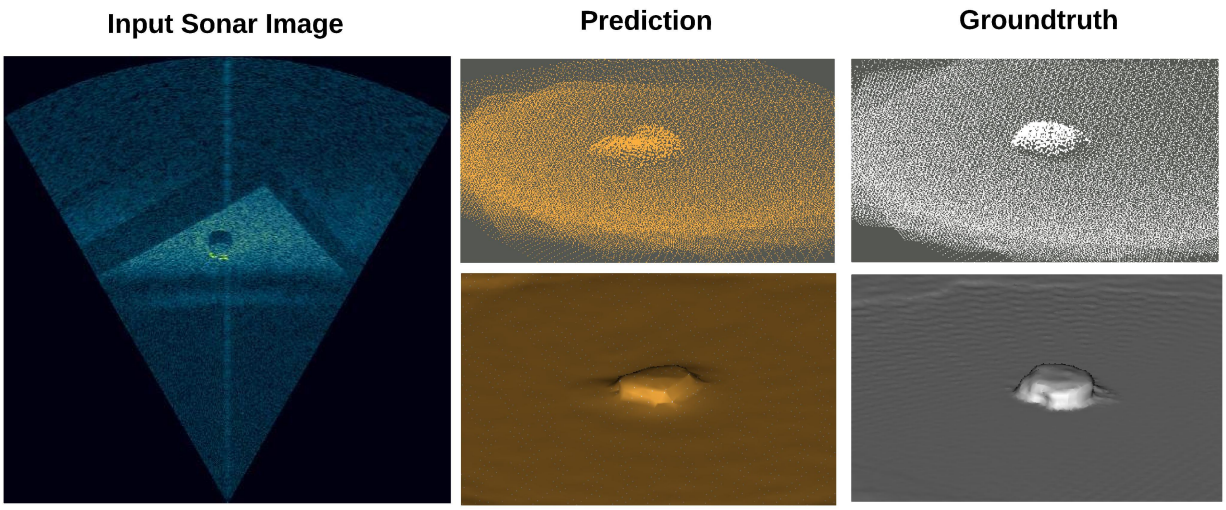

Fig. 6: 3D reconstruction result from evaluation on real sonar images of the mine. The left panel shows a representative real sonar image. The middle panel displays the predicted 3D reconstruction, with the point cloud on top and the corresponding mesh below. The right panel shows the ground-truth point cloud (top) and its generated mesh (bottom).

The ability of both models to maintain solid detection accuracy on these mixed real-synthetic datasets underscores the quality and diversity of the dataset and suggests that the models can effectively transfer knowledge learned from synthetic data to real-world sonar imagery. These results reflect the strength and representativeness of the dataset used for training, demonstrating that the models were able to learn robust features that generalize well to real sonar images. The consistent performance on these classes during testing indicates that the dataset provides effective coverage of both synthetic and real-world variations, supporting reliable detection across diverse object geometries and conditions.

B. 3D Reconstruction

To validate the usefulness of the dataset, we reference its prior application in a deep learning–based 3D reconstruction framework, MV3D [48]. In that work, an encoder–decoder network was trained and evaluated on a subset of the dataset, using 24-view sonar image sequences as input and the corresponding 8-view depth maps as ground truth. This data format proved valuable for 3D reconstruction from 2D FLS images, enabling the model to extract meaningful spatial features from batched multi-view inputs and predict complete object shapes through its multi-depth map supervision. Accurate and dense reconstructions of selected underwater objects demonstrated the effectiveness of the dataset structure. Notably, the model was trained exclusively on the synthetic dataset and successfully generalized to real FLS images, highlighting both the realism of the simulated data and the robustness of the dataset design for learning-based 3D reconstruction.

In this dataset paper, we extend this validation by applying the same reconstruction framework to a new object: a landmine UXO, which was not included in the original publication. This complementary evaluation further demonstrates the dataset's consistency, diversity, and applicability across a broader range of object shapes and categories. For this experiment, a subset of the dataset containing the mine data was used for training. To assess the model's performance on real sonar images, additional data of a mine placed on the basin floor was acquired using the same water basin and sensor setup as for the U-shape, L-shape, and stairs.

The metrics commonly used for evaluating 3D reconstruction results are Chamfer Distance (CD) and Hausdorff Distance (HD), defined as follows:

$$CD = \sum_{n=1}^{N} \left(\frac{\lambda}{S_1} \sum_{x \epsilon S_1} \min_{y \epsilon S_2} \|x - y\|_2^2 + \frac{\lambda}{S_2} \sum_{y \epsilon S_2} \min_{x \epsilon S_1} \|x - y\|_2^2\right) \quad (1)$$

where λ was set to 1.

$$HD(A, B) = \max\left(\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b)\right) \quad (2)$$

where $D_h(A, B)$ represents the Hausdorff distance between sets A and B. $d(a, b)$ is the distance function used to measure the distance between elements a and b in the underlying metric space. sup denotes the supremum (least upper bound) of a set, and inf denotes the infimum (greatest lower bound) of a set.

After being trained exclusively on synthetic data and evaluated directly on real sonar images of the mine, the model achieved a CD of 0.028 m and an HD of 0.09 m, reflecting a dense and accurate reconstruction of the 3D shape. Figure 6 illustrates representative real sonar test images, the predicted point cloud, and the corresponding ground-truth point cloud, with both point clouds also converted to meshes for improved visualization. These reconstruction results further confirm the dataset's suitability for training and benchmarking 3D perception models under realistic sonar-based sensing conditions.

V. Conclusion

In this paper, we introduced a novel, publicly available synthetic dataset designed to address the critical scarcity of comprehensive data for underwater perception. By pairing simulated FLS imagery captured along a linear trajectory with ground truth depth maps from a 180-degree field of view, our dataset provides rich, multi-modal information that captures viewpoint-dependent geometric features and shadow dynamics. This resource offers a robust platform for developing and validating a wide range of algorithms, including object detection, pose estimation, and 3D reconstruction.

To demonstrate the dataset's practical value, we validated it on two key applications: 2D object detection and 3D reconstruction. Models trained solely on the dataset's synthetic data were evaluated on real sonar images and achieved strong performance in both tasks. These results highlight the dataset's robustness and generalizability, underscoring its potential to accelerate research in critical underwater applications such as the safe removal of Unexploded Ordnance (UXO), marine habitat monitoring, restoration, and beyond.

Future work will focus on further enhancing the dataset's realism by simulating more complex underwater environments and sensor conditions. Additionally, we plan to expand the object variety to cover a broader range of shapes and materials, improving the dataset's diversity and applicability.

References

[1] N. Dahn, M. B. Firvida, P. Sharma, L. Christensen, O. Geisle, J. Mohrmann, T. Frey, P. K. Sanghamreddy, and F. Kirchner, "An acoustic and optical dataset for the perception of underwater unexploded ordnance (uxo)," in *OCEANS 2024-Halifax*. IEEE, 2024, pp. 1–6.

[2] N. Jaber, B. Wehbe, and F. Kirchner, "Sonar2depth: Acoustic-based 3d reconstruction using cgans," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 5828–5835.

[3] M. J. DeWeerts, "Detection of underwater military munitions by a synoptic airborne multi-sensor system," *Honolulu (HI)*, 2010.

[4] D. G. Gallagher, "Diver-based rapid response capability for maritime-port security operations," in *OCEANS'11 MTS/IEEE KONA*, 2011, pp. 1–10.

[5] E. Belcher, W. Hanot, and J. Burch, "Dual-frequency identification sonar (didson)," in *Proceedings of the 2002 interntional symposium on underwater technology (Cat. No. 02EX556)*. IEEE, 2002, pp. 187–192.

[6] R. L. P. de Lima, F. C. Boogaard, and R. E. de Graaf-van Dinther, "Innovative water quality and ecology monitoring using underwater unmanned vehicles: Field applications, challenges and feedback from water managers," *Water*, vol. 12, no. 4, p. 1196, 2020.

[7] P. Cieślak, "Stonefish: An advanced open-source simulation tool designed for marine robotics, with a ros interface," in *OCEANS 2019 - Marseille*, 2019, pp. 1–6.

[8] W. Jiao, J. Zhang, and C. Zhang, "Open-set recognition with long-tail sonar images," *Expert Systems with Applications*, vol. 249, p. 123495, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417424003609

[9] K. Xie, J. Yang, and K. Qiu, "A dataset with multibeam forward-looking sonar for underwater object detection," *Scientific Data*, vol. 9, no. 1, p. 739, 2022.

[10] M. Valdenegro-Toro, D. C. Padmanabhan, D. Singh, B. Wehbe, and Y. Petillot, "The marine debris forward-looking sonar datasets," 2025. [Online]. Available: https://arxiv.org/abs/2503.22880

[11] D. Singh and M. Valdenegro-Toro, "The marine debris dataset for forward-looking sonar semantic segmentation," 2021. [Online]. Available: https://arxiv.org/abs/2108.06800

[12] M. Aubard, A. Madureira, L. Teixeira, and J. Pinto, "Sonar-based deep learning in underwater robotics: Overview, robustness, and challenges," *IEEE Journal of Oceanic Engineering*, pp. 1–19, 2025.

[13] E. Potokar, S. Ashford, M. Kaess, and J. G. Mangelson, "Holoocean: An underwater robotics simulator," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 3040–3046.

[14] E. Potokar, K. Lay, K. Norman, D. Benham, T. B. Neilsen, M. Kaess, and J. G. Mangelson, "Holoocean: Realistic sonar simulation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 8450–8456.

[15] G. de Oliveira, M. M. d. Santos, and P. L. Drews-Jr, "Synthetic enclosed echoes: A new dataset to mitigate the gap between simulated and real-world sonar data," *arXiv preprint arXiv:2505.15465*, 2025.

[16] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *2004 IEEE/RSJ international conference on intelligent robots and systems (IROS)(IEEE Cat. No. 04CH37566)*, vol. 3. Ieee, 2004, pp. 2149–2154.

[17] P. Ściegienka and M. Blachnik, "On the development of an acoustic image dataset for unexploded ordnance classification using front-looking sonar and transfer learning methods," *Sensors*, vol. 24, no. 18, p. 5946, 2024.

[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 Robotics and Automation Letters*, vol. 6, no. 2, pp. 1535–1542, 2021.

[19] L. Chen, Z. Liu, L. Tong, Z. Jiang, S. Wang, J. Dong, and H. Zhou, "Underwater object detection using invert multi-class adaboost with deep learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[20] S. Cai, G. Li, and Y. Shan, "Underwater object detection using collaborative weakly supervision," *Computers and Electrical Engineering*, vol. 102, p. 108159, 2022.

[21] M. Zhang, S. Xu, W. Song, Q. He, and Q. Wei, "Lightweight underwater object detection based on yolo v4 and multi-scale attentional feature fusion," *Remote Sensing*, vol. 13, no. 22, p. 4706, 2021.

[22] C. Tan, C. DanDan, H. Huang, Q. Yang, and X. Huang, "A lightweight underwater object detection model: Fl-yolov3-tiny," in *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2021, pp. 0127–0133.

[23] H. Yang, P. Liu, Y. Hu, and J. Fu, "Research on underwater object recognition based on yolov3," *Microsystem Technologies*, vol. 27, no. 4, pp. 1837–1844, 2021.

[24] C. Liu, H. Li, S. Wang, M. Zhu, D. Wang, X. Fan, and Z. Wang, "A dataset and benchmark of underwater object detection for robot picking," in *2021 IEEE international conference on multimedia & expo workshops (ICMEW)*. IEEE, 2021, pp. 1–6.

[25] X. Zhao, X. Wang, and Z. Du, "Research on detection method for the leakage of underwater pipeline by yolov3," in *2020 IEEE international conference on mechatronics and automation (ICMA)*. IEEE, 2020, pp. 637–642.

[26] B. Gašparović, J. Lerga, G. Mauša, and M. Ivašić-Kos, "Deep learning approach for objects detection in underwater pipeline images," *Applied artificial intelligence*, vol. 36, no. 1, p. 2146853, 2022.

[27] S. K. Kartal and R. F. Cantekin, "Autonomous underwater pipe damage detection positioning and pipe line tracking experiment with unmanned underwater vehicle," *Journal of Marine Science and Engineering*, vol. 12, no. 11, p. 2002, 2024.

[28] M. Pedersen, J. Bruslund Haurum, R. Gade, and T. B. Moeslund, "Detection of marine animals in a new underwater dataset with varying visibility," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 18–26.

[29] Z. Zhao, Y. Liu, X. Sun, J. Liu, X. Yang, and C. Zhou, "Composited fishnet: Fish detection and species recognition from low-quality underwater videos," *IEEE Transactions on Image Processing*, vol. 30, pp. 4719–4734, 2021.

[30] H. Huang, H. Zhou, X. Yang, L. Zhang, L. Qi, and A.-Y. Zang, "Faster r-cnn for marine organisms detection and recognition using data augmentation," *Neurocomputing*, vol. 337, pp. 372–384, 2019.

[31] P. Jonsson, I. Sillitoe, B. Dushaw, J. Nystuen, and J. Heltne, "Observing using sound and light–a short review of underwater acoustic and video-based methods," *Ocean Science Discussions*, vol. 6, no. 1, pp. 819–870, 2009.

[32] M. Aubard, L. Antal, A. Madureira, and E. Ábrahám, "Knowledge distillation in yolox-vit for side-scan sonar object detection," *arXiv preprint arXiv:2403.09313*, 2024.

[33] M. Aubard, A. Madureira, L. Madureira, and J. Pinto, "Real-time automatic wall detection and localization based on side scan sonar images,"

in *2022 IEEE/OES Autonomous Underwater Vehicles Symposium (AUV)*. IEEE, 2022, pp. 1–6.

[34] L. R. Fuchs, A. Gällström, and J. Folkesson, "Object recognition in forward looking sonar images using transfer learning," in *2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV)*. IEEE, 2018, pp. 1–6.

[35] X. Wang, S. Liu, and Z. Liu, "Underwater sonar image detection: A combination of non-local spatial information and quantum-inspired shuffled frog leaping algorithm," *PLoS One*, vol. 12, no. 5, p. e0177666, 2017.

[36] S. Lee, B. Park, and A. Kim, "Deep learning from shallow dives: Sonar image generation and training for underwater object detection," *arXiv preprint arXiv:1810.07990*, 2018.

[37] F. Zhang, W. Zhang, C. Cheng, X. Hou, and C. Cao, "Detection of small objects in side-scan sonar images using an enhanced yolov7-based approach," *Journal of Marine Science and Engineering*, vol. 11, no. 11, p. 2155, 2023.

[38] N. P. Santos, R. Moura, G. S. Torgal, V. Lobo, and M. de Castro Neto, "Side-scan sonar imaging data of underwater vehicles for mine detection," *Data in Brief*, vol. 53, p. 110132, 2024.

[39] K. Denos, M. Ravaut, A. Fagette, and H.-S. Lim, "Deep learning applied to underwater mine warfare," in *OCEANS 2017-Aberdeen*. IEEE, 2017, pp. 1–7.

[40] F. Bruno, G. Bianco, M. Muzzupappa, S. Barone, and A. V. Razionale, "Experimentation of structured light and stereo vision for underwater 3d reconstruction," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 4, pp. 508–518, 2011.

[41] 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.

[42] M. Massot-Campos and G. Oliver-Codina, "Optical sensors and methods for underwater 3d reconstruction," *Sensors*, vol. 15, no. 12, pp. 31 525–31 557, 2015.

[43] 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)*. IEEE, 2019, pp. 8040–8047.

[44] S. Arnold and B. Wehbe, "Spatial acoustic projection for 3d imaging sonar reconstruction," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 3054–3060.

[45] M. D. Aykin and S. Negahdaripour, "On 3-d target reconstruction from multiple 2-d forward-scan sonar views," in *OCEANS 2015-Genova*. IEEE, 2015, pp. 1–10.

[46] M. D. Aykin and S. S. Negahdaripour, "Modeling 2-d lens-based forward-scan sonar imagery for targets with diffuse reflectance," *IEEE journal of oceanic engineering*, vol. 41, no. 3, pp. 569–582, 2016.

[47] 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, 2016.

[48] N. Jaber, B. Wehbe, L. Christensen, and F. Kirchner, "Mv3d: Multi-view 3d reconstruction of objects using forward-looking sonar," *IEEE Robotics and Automation Letters*, 2025.

[49] E. Maser and J. S. Strehse, ""don't blast": blast-in-place (bip) operations of dumped world war munitions in the oceans significantly increase hazards to the environment and the human seafood consumer," *Archives of toxicology*, vol. 94, pp. 1941–1953, 2020.

[50] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[51] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[52] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[53] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524*, 2025.

[54] P. Robicheaux, J. Gallagher, J. Nelson, and I. Robinson. (2025, mar) RF-DETR: A SOTA Real-Time Object Detection Model. [Online]. Available: https://blog.roboflow.com/rf-detr/