IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

MV3D: Multi-View 3D Reconstruction of Objects Using Forward-Looking Sonar

Nael Jaber, Bilal Wehbe, Leif Christensen, and Frank Kirchner

***Abstract*—This work proposes a method for learning features from a batch of 2D sonar images to predict a multi-view point-cloud for achieving a dense 3D-reconstruction. In comparison to vision-based sensors, acoustics are considered a reliable sensing modality in underwater environments. The output of sonars is a 2D image which is unable to represent the scanned scene in all three dimensions. Estimation of this missing information, known as the elevation angle, is the key to performing 3d-reconstruction from acoustic images. One of the approaches is to predict a depth-map from the 2D sonar image, and transforming it into a point-cloud. In this letter, this idea is further improved into learning features from a batch of 2D acoustic images and predicting multiple depthmaps of the scanned object which covers it from different viewpoints. For training the deep learning model, and due to the lack of datasets from real environments, data was generated synthetically. For reducing the simulation-to-real gap, a Cycle-GAN was trained on real images for transferring the realistic style into the synthetically generated images. The conducted experiments in simulation showed that the proposed method is able to perform dense 3D reconstruction. The approach was then further tested in a real environment using an underwater vehicle, which accurately 3D-reconstructed the scanned objects achieving an average chamfer distance error of 0.06 meters when compared to a laser-scanned ground-truth.**

***Index Terms*—Marine robotics, deep learning methods, deep learning for visual perception.**

I. Introduction

UNDERWATER exploration has always been a challenging task for humans, whether it is a search and rescue mission, investigation of a new area, marine-life monitoring, and many others. AUVs being able to operate for long periods of time, help in covering large areas and in reaching inaccessible spots. They have different designs, and could have several sensors mounted onto them. This makes these vehicles suitable for various underwater applications such as seabed mapping and exploration, maintaining offshore infrastructure, marine life monitoring, and many others. Despite the aforementioned advantages, the underwater domain is still very challenging even for AUVs. Low visibility conditions and turbid currents limit the use of visual cameras to very short ranges, and even eliminates the use of them in some cases [1]. Furthermore, high attenuation of signals underwater obstruct the use of regular global positioning systems (GPS) and requires an acoustic-based positioning alternative sensors such as long-baseline (LBL) and ultra-short-baseline (USBL).

Sonars are often considered the most suitable sensing modality for underwater domains [2]. Sonars are the reflection of acoustic signals bouncing back from scanned objects. Relying on acoustics, sonars are able to operate in turbid waters, in deep and dark areas, and to provide accurate measurements (up to millimeter-level for close-ranged sonars). Although it is recognized among the best sources of information underwater, their output is a 2D-acoustic image which retrieves the azimuth angle and the measured range of a each emitted beam. The third dimension, known as the elevation angle, is lost which makes the use of 2D sonars for mapping and 3D reconstruction a very challenging task.

Performing 3D reconstruction from a 2D sonar is double-sided challenge. First, sonars retrieve only range and beam angle, while the elevation angle is lost. The second challenge is represented by the non-bijective 2D-3D correspondence, which means that a point in a sonar image could correspond to multiple points in the 3D world. Thus, research in this area was clearly focused on solving these tasks. Some of the approaches trying to estimate/recover this elevation angle are physics-based, others are geometry-based, and recently some learning-based methods are emerging [3], [4], [5].

In this work, a machine learning model is set to learn features from a batch of 2D-sonar images. The output of the model are several depthmaps generated from different viewpoints, which are then transformed into the world frame creating a complete 3D-reconstruction of the scanned object. The contributions of this letter are as follows:

- We propose the learning of features from a batch of 2D sonar images.
- A multi-view output is introduced to predict a more complete shape of the scanned object.
- A dataset which consists of various geometric objects was generated synthetically to train the proposed system.
- Results of the proposed model on real data showed accurate 3D reconstruction, being trained on synthetic data solely. The implementation of the network and sample from the generated dataset are publicly available: MV3D (https://zenodo.org/records/15797936).

II. Related Work

The loss of elevation angle by the acoustic sensors have pushed researchers to estimate this missing information using various approaches. A line of work performed sensor fusion of

Received 8 May 2025; accepted 23 June 2025. Date of publication 10 July 2025; date of current version 21 July 2025. This article was recommended for publication by Associate Editor P. Drews-Jr and Editor A. Valada upon evaluation of the reviewers' comments. This work was supported by European Union's Horizon 2020 Research and Innovation Programme, under Grant 956200. *(Corresponding author: Nael Jaber.)*

The authors are with DFKI - Robotics Innovation Center, 28359 Bremen, Germany (e-mail: nael.jaber@dfki.de; bilal.wehbe@dfki.de; leif.christensen@dfki.de; frank.kirchner@dfki.de).

Digital Object Identifier 10.1109/LRA.2025.3588062

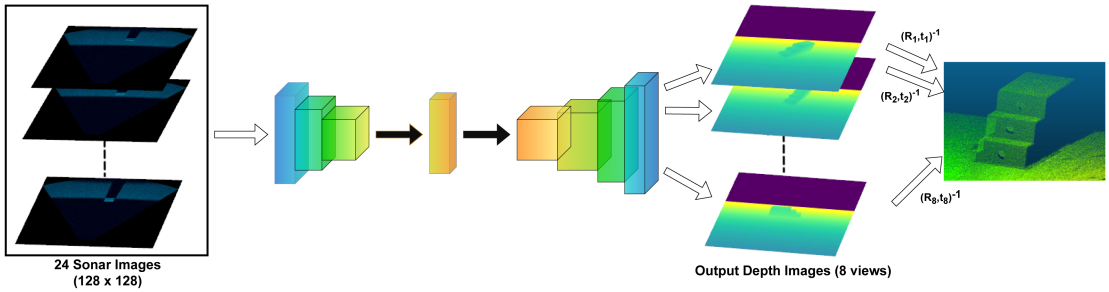

Fig. 1. Schematic showing the proposed deep learning model. The input of the encoder-decoder network are batches of 24 acoustic images, and the outputs are eight depthmaps. Transformation of the outputted depthmaps into the world frame creates a complete pointcloud.

two sensors in a stereo-acoustic configuration [6], [7], [8], [9], or an imaging and a profiling sensor [10], [11], [12], [13] as a way of retrieving the elevation angle. In this work, only one forward-looking sonar (FLS) is utilized. Westman et al. [14] aimed to reconstruct particular 3D surface points observed by an imaging sonar. However, this method restricts the vehicle's motion in a way that it needs a view ray perpendicular to the surface at each surface point, which means that this method needs numerous images to be collected from certain view points.

In their work, Aykin et al. [15], [16] rely on object edges and casted shadows in order to estimate the elevation angle of each pixel. Despite the assumptions, this method requires objects to be lying on the seabed. Westman et al. [17] later improved the idea and was able to eliminate the seafloor assumption. However, These methods ignore the non-bijective 2D-3D correspondence problem in sonars by estimate an elevation angle for each pixel.

Another line of work have proposed several volumetric methods [18], [19]. Wang et al. [20] proposed the idea of updating the occupancy in a voxel grid by introducing an inverse sonar sensor model. Later they used graph optimization as a way for aligning local sub-maps in order to minimize errors in pose estimates [21]. In comparison to space carving methods, these approaches can be more robust [15], [22], as they consider each voxel solely. The most recent volumetric method is the work done by [23], which addressed the elevation angle problem by refining an an occupancy grid by rendering echo probabilities. Although the proposed method showed promising results in simulation, real experiments were conducted in a controlled environment.

More recently, learning-based methods were proposed to resolve the elevation ambiguity. Arnold et al. [24] propose training a CNN to predict the signed distance and direction to the nearest surface for each cell in a 3D grid. However, the method requires ground truth Truncated Signed Distance Field (TSDF) information which can be difficult to obtain. DeBortoli et al. [3] proposed a self-supervised training procedure to fine-tune a Convolutional Neural Network (CNN) trained on simulated data with ground truth elevation information. Wang et al. [4] proposed a deep network (A2FNet) to transfer the acoustic view to a pseudo frontal view which was shown to help with estimating the elevation angle. However, these methods are limited to simple geometries or require collecting a larger dataset of real elevation data. In further work, Wang et al. [5] proposed the learning of pseudo front depth from 2–3 multi-acoustic-viewpoints following an elevation plane sweeping method. This method relies on the rotation of the FLS in the roll direction which requires the sensor to be stationary. Their work was tested in a real experiment after collecting and training their model on real data. In previous work [25], the use of conditional-GANs was proposed for transforming acoustic images into depth images for 3D reconstruction. This approach achieved accurate reconstruction results when tested in real experiments after being solely trained on simulated data. However, this method learns features from single acoustic images and predicts only one corresponding depthmap. One of the most recent methods is the work done by [26] which performs surface reconstruction from acoustic images, and their approach is to model the object geometry as a neural implicit function. This method requires capturing a lot of acoustic images of the target object by scanning it from several views to be able to reconstruct, unlike this work which learns features from only a small batch of images captured from a linear scan.

Overall, the work proposed in this letter builds on existing research adopting the methodology of predicting a depthmap from an acoustic image as a way of estimating the elevation angle, with an improvement in both feature learning and depthmap prediction. This approach is different from the existing line of work relying solely on shadow analysis for the estimation of the missing elevation angle. These methods require various predefined factors and field assumptions, and of course prominent shadow highlights to work; while this work utilizes shadow variation as an extra feature for reconstruction and requires considerably less field assumptions. Volumetric methods relying on filling volumetric grids with possible elevation angles using different approaches are also different than the proposed method of this letter, as most of them require scanning objects from all sides (considerably more viewpoints) with respect to straightforward linear scan in this letter, which makes the proposed method easier to implement in real-world scenarios.

III. Proposed Method

In this work, we propose a machine learning model which learns features from a batch of acoustic images and predicts multi-view depth images. Extracting features from sonar images has always been a challenge for the traditional computer vision methods, as objects appear very different than in RGB images. Basic features such as edges, corners, contours, etc. are in many cases very hard to identify and extract [27]. However, one of the main features to detect in acoustic images are the shadows cast behind the scanned objects generated from the blockage of the acoustic beams.

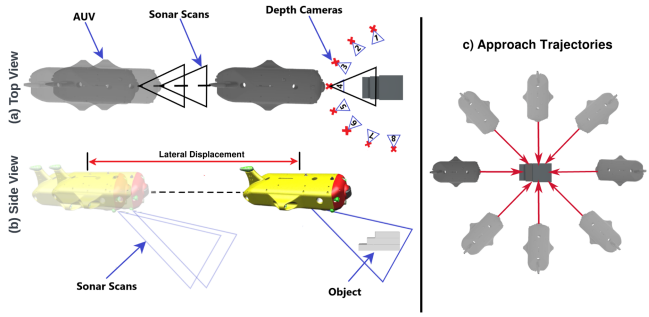

Fig. 2. Schematic illustrating the data capturing process. (a) corresponds to a top view showing the AUV approaching the object over a distance while capturing a batch of acoustic images. At the end of this displacement, eight depth cameras evenly distributed over the span of 180 degrees around the object are placed. (b) shows the capturing process from a side view for better visualization. (c) shows the chosen approach trajectories for capturing data for each object from all sides.

After learning the features of the scanned objects from the batch of images, the model predicts depth images from different views. The aim of this multi-view output is to reconstruct a more dense and complete shape of the scanned object. Fig. 1 shows a schematic of the proposed system.

A. Dataset

In this work, Stonefish [28], which is an underwater simulator made for marine community, was used. It has the ability to simulate underwater sensor and actuators with real parameters, and render realistic environments by giving the user control in defining various environment and object characteristics. For simulating the multi-beam acoustic images, Stonefish embeds a GPU-based sonar simulator which is described in [28]. The use of Stonefish was maintained primarily due to the foundation built on prior work [25], which ensured compatibility and consistency in our simulation pipeline. However, other sonar simulation engines, such as HoloOcean [29], are actively emerging and offer promising capabilities in rendering realistic sonar images.

The simulation environment consists of a forward-looking sonar (FLS) tilted at an angle of 30 degrees, and eight depth cameras evenly distributed to cover a span of 180 degrees of the object being scanned. For each run, 24 sonar images are captured by linearly approaching the scanned object by a distance of one meter, which given the chosen sonar configuration it could cover the scanned object. This linear trajectory scan was chosen due to its ease of implementation and its practicality for later mitigation in real-world experiments. Fig. 2 illustrates the data capturing process. The FLS was defined in the simulation with the same specifications as a real Oculus Md1200 d when operating in its high frequency mode. Its minimum and maximum ranges are set to 0.1 and 10 meters respectively, its horizontal aperture is set to 60 degrees, vertical aperture to 12 degrees, and the number of beams is set to 512.

Data was captured for 13 different geometric objects: cube, rectangle, trapezoidal prism, triangular prism, sphere, semi-sphere, cylinder, pipe, U-shaped block, L-shaped block, stairs, tire, and a boat. As shown in Fig. 3, these objects were categorized into three datasets based on their geometric properties: edged objects, curved objects, and objects

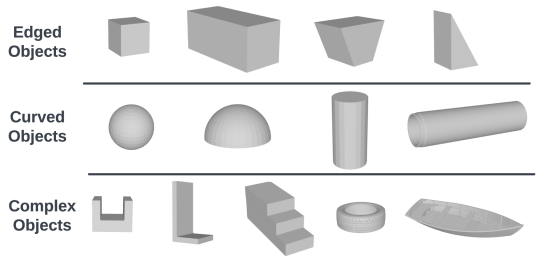

Fig. 3. 3D models of the dataset objects. The first row shows four different edged objects: cube, rectangle, trapezoidal prism, and a triangular prism. The second row shows the four curved objects: sphere, semi-sphere, cylinder, and a pipe. The third row shows objects of more compelx geometries: a U-shape, L-shape, stairs, tire, and a boat.

TABLE I
ARCHITECTURE DETAILS

Input size	Latent vector	Number of filters	
		image encoder	depth generator
128×128	512-D	conv: 96,128,192,256 linear: 2048, 1024, 512	linear: 1024, 2048, 4096 deconv: 192, 128, 96, 64, 48

with complex geometries. These three datasets are denoted by D_{edge}, D_{curve}, and $D_{complex}$ respectively.

This categorization is used to assess the model's ability to learn geometric features from sonar images, which can appear very different depending on the viewing angle. To address this challenge, we opted for a data split based on edge and curve characteristics, enabling the model to learn more robust and generalizable features. The captured datasets consist of batches of 24 sonar images of each specific object, along with its corresponding eight-view depth maps.

B. Deep Learning Model

The deep-learning model utilized is an encoder-decoder network [30]. As mentioned earlier, the input of the network is a batch of 24 images. The input goes through 2D convolutions to predict at the end eight depth images embedding the x,y,z characteristics. Table I shows the network's architecture in details. Trained on a fixed configuration, and by applying the transformation matrix $(R_n, t_n)^{-1}$ to each of the output depthmaps, eight point-clouds are generated which create a 3D shape of the scanned object.

The total loss function of this network relies on three optimizing losses. The first one is the L_{depth} which is an L1-loss represented by the following equation:

$$L_{\text{depth}} = \sum_{n=1}^{N} \|\widehat{Z}_n - Z_n\| \tag{1}$$

where $\widehat{Z}$ is the predicted depth-map and Z is the ground-truth. N is the total number of view-points which is equal 8.

For more optimization, a cross entropy loss is introduced into the equation in the aim of reducing errors in predictions. This error would be calculated by generating binary masks out of the predicted depthmaps and their ground-truths. The generated masks are generated after setting a maximum depth equal to the maximum range of the sonar, which is equal 10 meters. This helps in isolating the object and predicting more accurate results.

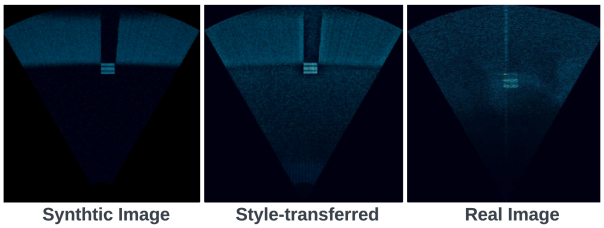

Fig. 4. Left image corresponds to a synthetically generated image for the stairs, the middle image is the style-transferred which is an output from the Cycle-GAN network, and the right image is the target real image.

Following is the corresponding equation:

$$L_{\text{mask}} = \sum_{n=1}^{N} -M_n \log \widehat{M}_n - (1 - M_n) \log (1 - \widehat{M}_n) \quad (2)$$

where $\widehat{M}_n$ and M_n are the masks from prediction and ground-truth respectively. The overall loss function would then be defined as $L_{\text{total}} = L_{\text{depth}} + L_{\text{mask}}$

C. Cycle-GAN

As a way of minimizing the simulation-to-real gap, a Cycle-GAN network was adopted from the work of Isola P. et al. [31], and using the defined set of parameters in this implementation [32]. Cycle-GANs are style-transfer models which could learn a specific style from a set of images, and transfers this style into another set. This functionality fits perfectly for this application as we aim to transfer the realistic style into the generated synthetic images. The Cycle-GAN was trained in an unpaired manner on 1400 images, where 700 of them are synthetic and 700 real acoustic images are collected using the Oculus Md1200. Prior to the real-world experimentation, the learned style from this training was applied to initially captured dataset before training the network again on these style-transferred images. Fig. 4 shows a sample of the synthetically generated acoustic image of a stairs model, a real image captured by the real Oculus Md1200 FLS, and the output of the cycle-GAN. As shown in the figure, the Cycle-Gan learned the style of the real FLS while keeping the simulated geometric features the same, which is crucial in such applications.

IV. Simulation Experiments

The proposed network was evaluated after training and testing on the synthetically generated datasets. Results are interpreted at the end of this chapter.

A. Object Dataset

As mentioned in Section III-A, data was captured for thirteen different objects and categorized into three datasets. As shown in Fig. 2, data was generated by tilting the object and approaching it from eight different angles/sides. For this reason, the number of images captured for each object is 192 sonar images (8x24-sonar images batch). Thus, D_{edge} and D_{curve} consist of 768 sonar images, and $D_{complex}$ consists of 960 images. The size of the chosen objects vary between 0.3 and 3 meters in length, 0.3 and 1.76 meters in width, and 0.2 and 1 m in height. The biggest object in width is the pipe (3 meters), the biggest in terms of length is the boat (1.76 meters). The rest of the objects have average sizes of 0.5 meters in all dimensions.

B. Metrics

In order to evaluate the performance of the network in the most credible manner, four different metrics were used on two different stages of the workflow. Since the initial output of the network is depthmaps, evaluation of these results should take place before transforming them into point-clouds. For this stage, the mean average error (MAE) and the structural similarity index measure (SSIM) are applied. These two metrics are commonly used in literature for depthmap comparison. The MAE and SSIM formulas are as follow:

$$MAE = \sum_{i=1}^{D} |x_i - y_i| \quad (3)$$

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (4)$$

After transforming the depthmaps into point-clouds, statistical filtering is applied. This filtering process outputs cleaner results as it eliminates all outlier points from the final point-clouds. For evaluating the accuracy of these clouds, two of the most commonly used metrics: chamfer distance (CD) and hausdorff distance (HD), are used. For each point in each cloud, CD finds the nearest point in the other point set, and sums the square of distance up. The chamfer loss is represented by:

$$CD = \sum_{n=1}^{N} \left(\frac{\lambda}{S_1} \sum_{x \epsilon S_1} \min_{y \epsilon S_2} \|x - y\|_2^2 + \frac{\lambda}{S_2} \sum_{y \epsilon S_2} \min_{x \epsilon S_1} \|x - y\|_2^2 \right) \quad (5)$$

where λ was set to 1.

The HD is widely used for such task as it measures how far two subsets are from each other by calculating the greatest of all distances from a point in one cloud to the closest point in the other cloud. It is defined as follows:

$$HD(A, B) = \max \left(\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b) \right) \quad (6)$$

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.

C. Results

1) Training and Testing on Basic Objects: The proposed network is implemented using the PyTorch framework [38]. Training was first done on the datasets which consist of objects with basic geometries: D_{edge} and D_{curve} . Each dataset was split into 70% training, 10 % validation, and 20% testing. A Nvidia Titan-XP (12GB) GPU was utilized for training the network, starting with a learning rate of 0.001 which decays over epochs. The average inference time per sample is 309 milliseconds, with a standard deviation of 47.44 milliseconds.

The results of the these simulation experiments are presented in this section. Table II shows the average error achieved for each object, providing an in depth evaluation of the network's

TABLE II
MODEL'S PERFORMANCE ON D_{Edge} AND D_{Curve}

Dataset	Object	MAE	SSIM	CD	HD
D_{Edge}	Cube	0.05	0.96	0.026	0.057
	Rectangle	0.07	0.97	0.0145	0.041
	Trapezoidal prism	0.08	0.96	0.0212	0.092
	Triangular prism	0.09	0.94	0.0148	0.048
Avg. D_{Edge}		0.0725	0.9575	0.0188	0.0592
D_{Curve}	Cylinder	0.14	0.94	0.018	0.11
	Semi-sphere	0.11	0.96	0.013	0.13
	Sphere	0.10	0.96	0.0178	0.07
	Pipe	0.36	0.83	0.06	0.251
Avg. D_{Curve}		0.1775	0.9225	0.0272	0.1402

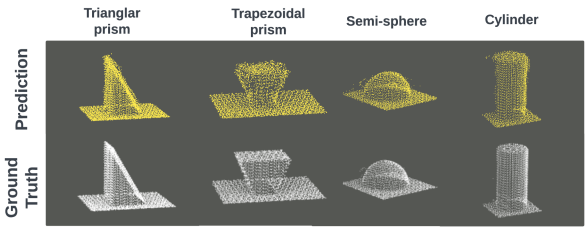

Fig. 5. Sample reconstruction results from the training and testing on D_{edge} and D_{curve}. The upper row shows the network's prediction and the lower one shows the ground-truth. Columns 1 through 4 correspond to the triangular prism, trapezoidal prism, semi-sphere, and cylinder respectively.

performance at both early and final stages of operation. Average MAE and SSIM values reflect the accuracy of the network in predicting the depthmaps, while the chamfer and hausdorff distances evaluate the final outputted pointclouds after outlier filtering. For the depthmaps' prediction, the model trained on D_{edge} achieved an average MAE of 0.07m and a similarity index of around 96%. The model trained on D_{curve} achieved an average MAE of 0.17 meters and an average SSIM of 92%. Transforming the predicted depthmaps into pointclouds, and after filtering outliers out, the chamfer and hausdorff distances were calculated. The model trained on edged objects achieved an average CD of 0.018 and a HD of 0.59, while the model trained on curved objects achieved CD of 0.02 and HD of 0.14 meters. Fig. 5 shows sample reconstruction results from testing on basic objects.

Analysis of the MAE and SSIM metrics indicates that the model achieved a better performance in predicting depth maps from sonar images containing edge features. However, following transformation and filtering processes, the CD and HD metrics reveal a more comparable performance across both datasets, reflecting accurate reconstruction of object geometries. Visual inspection further confirmed that the higher MAE value observed in the dataset with curved objects is primarily attributed to inaccuracies in the background regions of the depth map predictions. Outlier points with distances larger than the average distance to their neighbors plus 2 times standard deviation are filtered out.

2) Evaluating on Unseen Objects: To further investigate the model's ability to extract meaningful features from sonar images, the model trained exclusively on the D_{edge} dataset was directly evaluated on objects from the $D_{complex}$ dataset. This experimental setup aimed to assess the model's capacity to reconstruct the overall geometric shape of objects it had not encountered during training. The two chosen objects were an inverted U-shape and the stairs. The evaluation results, summarized in Table III, are presented using the chosen metrics. Additionally, Fig. 6 visually illustrates the reconstruction results of the tested objects.

TABLE III
EVALUATION ON UNSEEN OBJECTS

Metric (m)	MAE	SSIM	CD	HD
Inverted U-shape	0.09	96 %	0.06	0.112
Stairs	0.127	93%	0.108	0.178

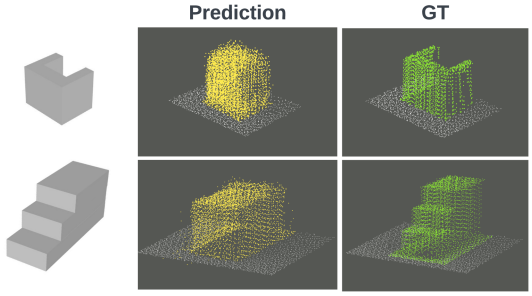

Fig. 6. Sampled reconstruction results from the evaluation on unseen objects from $D_{complex}$. Column 1 shows the 3D model of the evaluated object, column 2 shows the predicted pointcloud, and the final column shows the corresponding ground-truth pointcloud. The results of the inverted U-shape and stairs are shown in rows one and two respectively.

TABLE IV
SIMULATION RESULTS

Method	MV-3D				Sonar2Depth [25]		A2FNet [4]	
Metric(m)	MAE	SSIM(%)	CD	HD	CD	HD	CD	HD
Tire	0.001	99.8	**0.03**	0.13	0.1	0.13	0.1	0.16
U-shape	0.013	97.9	**0.07**	0.19	0.096	0.227	0.118	0.29
L-shape	0.13	93.9	**0.069**	0.2	0.084	0.37	0.29	0.41
Stairs	0.015	97.7	**0.05**	0.3	0.08	0.55	0.11	0.50
Boat	0.013	97.2	**0.09**	0.34	0.11	0.26	0.15	0.39

The network's ability to approximate the shape of unseen objects, represented by the low CD and HD values achieved, and as further illustrated in Fig. 6, highlights its capacity to capture both object scale and features from the sonar images. Based on the training data, the model successfully predicted the most similar shape in terms of scale and common features. For instance, the inverted U-shape was approximated by a box, and the stairs by a rectangle, both with proper scales. While the model cannot generalize to random unseen objects, the promising results encourage using the calculated weights for fine-tuning later over new complex objects.

3) Fine Tuning on Complex Objects: For reconstruction of objects having complex geometries, we utilized transfer learning to leverage the feature representations learned by our pre-trained models: $W_{D_{edge}}$ and $W_{D_{curve}}$. By initializing new networks with these pre-trained weights, training was done on each object of the $D_{complex}$ dataset, using the weights which better fit the trained object.

The results of these experiments are presented numerically through Table IV. For the depthmaps' prediction, the network achieved the lowest MAE value of 0.0012 m and the highest similarity index of 99.8% for the tire. This was probably the easiest to predict its depth-map due to its symmetrical geometry. However, the system further achieved very good results even

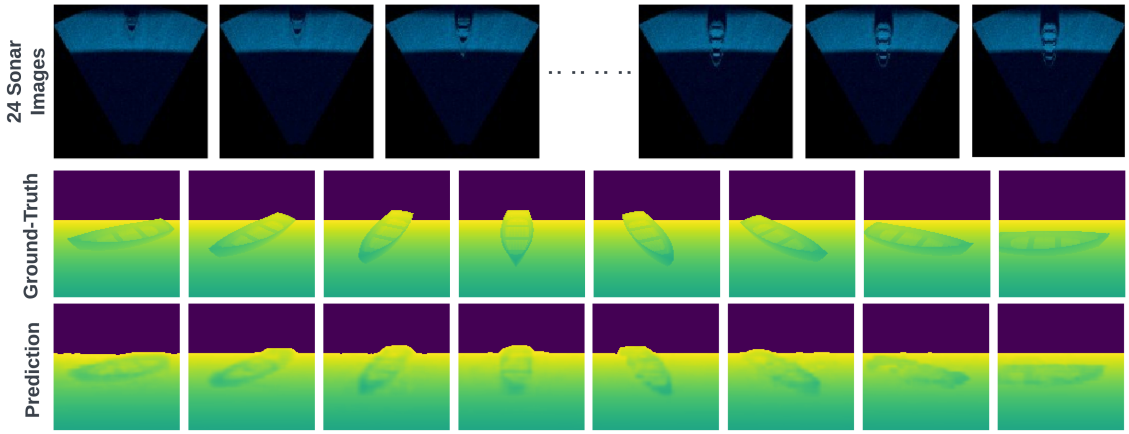

Fig. 7. Sample results of a boat. The first row shows samples from the batch of 24 acoustic images. The second row shows the corresponding eight ground-truth depthmaps. The third row shows the depthmaps predicted by the network.

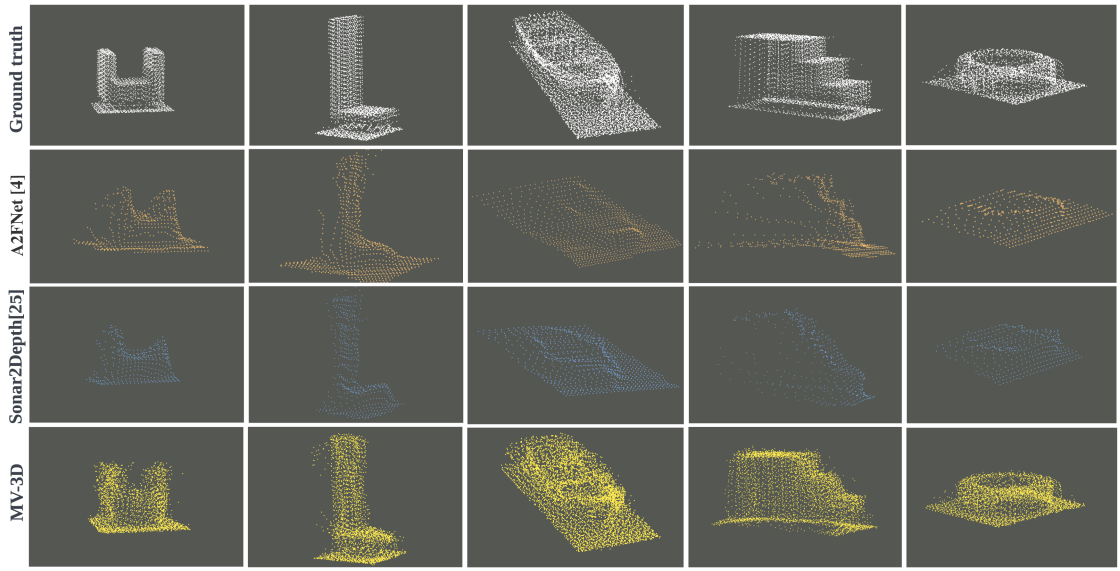

Fig. 8. 3D reconstruction results of the $D_{complex}$ objects. The first row shows the ground-truth pointclouds, the second shows the prediction results of A2FNet [4], the third shows the prediction results of Sonar2Depth [25],and the final row shows the predicted results of our proposed model. Columns 1 through 5 correspond to the U-shape, L-shape, boat, stairs, and tire respectively.

for the other complex objects such as the U-shape, stairs, and boat with MAE value ranging between 0.013 and 0.015 meters and very high SSIM values of around 98%. The hardest to predict was the L-shape geometry which the system achieved a MAE of 0.13 and an SSIM of 93.9%. The networks proposed in Sonar2Depth [25] and A2FNet [4] were trained and tested on $D_{complex}$. Since they predict a single depthmap compared to eight in this letter, evaluating a comparing performances using MAE and SSIM was not feasible. Consequently, only CD and HD were applied to assess their performance.

Fig. 7 shows sample of the simulation results for the boat. The first row shows some of the acoustic images as a sample from the complete input batch (24 images). The second and third rows show the ground-truth and the predicted depthmaps respectively. For better visualization, Fig. 8 shows sample 3D reconstruction results for all $D_{complex}$ objects. Looking at the reconstruction results, the predictions from MV-3D on all objects outperform state-of-the-art methods: Sonar2Depth [25] and A2FNet [4].

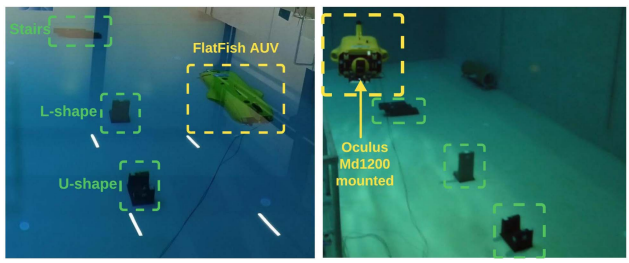

Fig. 9. Real experimental setup in the water tank showing the u-shape, l-shape, and stairs placed at the basin ground and the FlatFish AUV scanning with the Oculus Md1200 mounted to its bottom chassis.

This can be attributed to the series of input sonar images and the multi-view output in MV-3D predicting eight depth images, in comparison to the single output from Sonar2Depth and A2FNet, which can output a prediction in 3.5 ms and 2.9 ms

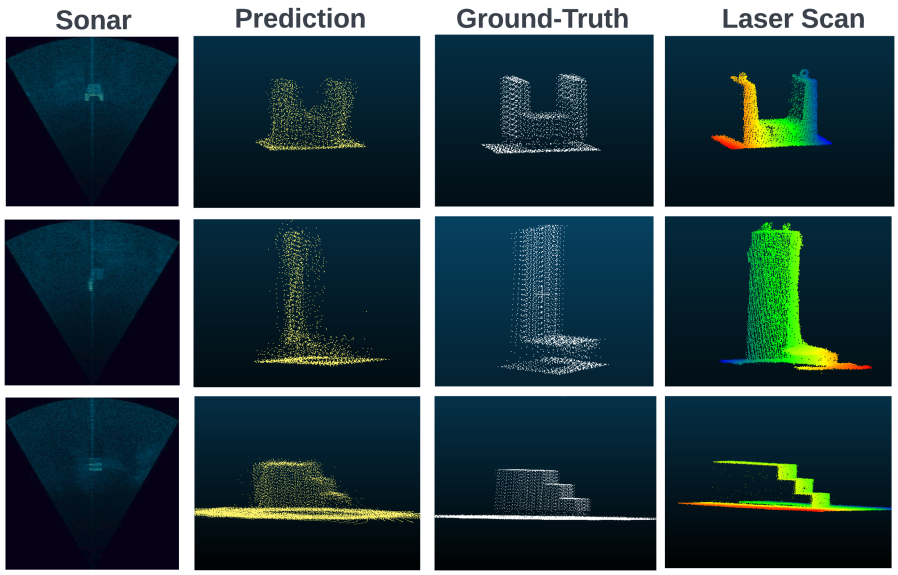

Fig. 10. 3D reconstruction results of the scanned objects from the real experiment. Rows 1 to 3 correspond to the results of the u-shape, l-shape, and stairs respectively. Column 1 shows a sample from the batch of real acoustic images. Columns 2 through 4 show the network's prediction, simulation ground-truth, and the SeaVision laser scan respectively.

respectively. This resulted in denser representations of objects, and in predicting finer geometrical details, giving a more detailed and complete 3D shape.

V. Real-World Experiment

A. Setup

The real-world experiment was conducted in a big water tank which is 8 meters deep. The FLS used for in this experiment is the Oculus Md1200. The FLS was mounted to the FlatFish AUV [33] which was operated manually for scanning the desired objects. Fig. 9 shows the experimental setup.

The sonar was operating in its high frequency mode of 2.1 MHz, with a minimum and maximum ranges of 0.1 and 10 meters respectively and an angular resolution of 0.4 degrees. In order to show that the proposed system works in real experiment, three concrete objects identical in shape to the ones trained and tested on in simulation were placed on the ground of the water tank for scanning. The chosen objects are the U-shape block, L-shape block, and the stairs. As a ground-truth, the SeaVision Laser imaging-scanner system was installed at the area of operation as it delivers dense 3D point-clouds in sub-millimeter resolution.

B. Results

Evaluation of the system's performance was done after training on the style-transferred synthetic images only, and testing on real acoustic images. The batches of 24 real images were chosen from approaching each object linearly, following same setup as for the simulation. The two metrics used for evaluation are the CD and HD, as described in Section IV-B. since the laser scanner position was fixed, only a frontal view of the objects was captured. Hence, the predicted 8-view point-cloud by the network could not be directly compared to the laser's data but was instead compared to the synthetically generated ground truth. In lieu, the front-view predicted depth-map (the fourth of the eight predicted views) was compared to the laser's output point-cloud. Results are shown in Table V. The low CD and HD achieved for all tested objects reflected accurate 3D reconstruction of the scanned objects' geometrical shape and scale, as shown in Fig. 10. The low average error values of CD and HD reflect accurate depthmap prediction for each of the eight predicted views. This is further supported by the low CD and HD values observed when comparing the predicted front-view depthmap to the ground-truth single-view laser scan. Despite the real experiment using the same objects as in the simulation, the model demonstrated the capability of reconstructing objects without requiring real data.

TABLE V
Real Experiments Results

	Simulation GT		Laser GT	
Metric (m)	CD	HD	CD	HD
U-shape	0.04	0.11	0.07	0.146
L-shape	0.048	0.2	0.068	0.275
Stairs	0.0538	0.158	0.06	0.2

VI. Conclusion

The idea of predicting multi-view depthmaps from an acoustically scanned object is proposed in this work. Taking the idea of depth-map prediction from a single image into inputting a batch enhanced the feature extraction process, learning not only objects' geometrical features, but also scene variables such as casted shadows. Our method improves upon existing approaches by predicting multiple depthmaps, reconstructing more complete and detailed shapes of scanned objects. Data for training and testing this network was generated synthetically for different objects and a Cycle-GAN network was trained in the purpose of transferring the realistic style into the generated images. Trained on the style-transferred synthetic images only, a real experiment

was conducted for evaluation of the network's performance. Results were compared to a laser generated ground-truth, and an average chamfer distance of 0.06 was achieved indicating a high accuracy in reconstruction relative to objects' sizes. Future work includes training on more objects and varying terrains, improving the multi-view depth prediction network to accommodate more generalizable features, and testing on a broader range of real-world data.

ACKNOWLEDGMENT

The authors would like to thank Kraken Robotics for providing and operating the SeaVision laser scanner.

REFERENCES

[1] H. Johannsson, M. Kaess, B. Englot, F. Hover, and J. Leonard, "Imaging sonar-aided navigation for autonomous underwater harbor surveillance," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2010, pp. 4396–4403.

[2] E. O. Belcher, W. L. Fox, and W. H. Hanot, "Dual-frequency acoustic camera: A candidate for an obstacle avoidance, gap-filler, and identification sensor for untethered underwater vehicles," in *Proc. OCEANS'02 MTS/IEEE*, 2002, pp. 2124–2128.

[3] R. DeBortoli, F. Li, and G. A. Hollinger, "ElevateNet: A convolutional neural network for estimating the missing dimension in 2D underwater sonar images," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 8040–8047.

[4] Y. Wang, Y. Ji, D. Liu, H. Tsuchiya, A. Yamashita, and H. Asama, "Elevation angle estimation in 2D acoustic images using pseudo front view," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 1535–1542, Apr. 2021.

[5] Y. Wang, Y. Ji, H. Tsuchiya, H. Asama, and A. Yamashita, "Learning pseudo front depth for 2D forward-looking sonar-based multi-view stereo," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 8730–8737.

[6] S. Negahdaripour, "On 3-D reconstruction from stereo FS sonar imaging," in *Proc. OCEANS 2010 MTS/IEEE SEATTLE*, 2010, pp. 1–6.

[7] V. Murino and A. Trucco, "Three-dimensional image generation and processing in underwater acoustic vision," *Proc. IEEE*, vol. 88, no. 12, pp. 1903–1948, Dec. 2000.

[8] N. Brahim, D. Guériot, S. Daniel, and B. Solaiman, "3D reconstruction of underwater scenes using DIDSON acoustic sonar image sequences through evolutionary algorithms," in *Proc. OCEANS, 2011 IEEE-Spain*, 2011, pp. 1–6.

[9] A. Trucco and S. Curletto, "Extraction of 3D information from sonar image sequences," *IEEE Trans. Syst., Man, Cybern., Part B*, vol. 33, no. 4, pp. 687–699, Aug. 2003.

[10] H. Joe, J. Kim, and S.-C. Yu, "Probabilistic 3D reconstruction using two sonar devices," *Sensors*, vol. 22, no. 6, 2022, Art. no. 2094.

[11] H. Joe, H. Cho, M. Sung, J. Kim, and S.-C. Yu, "Sensor fusion of two sonar devices for underwater 3D mapping with an AUV," *Auton. Robots*, vol. 45, no. 4, pp. 543–560, 2021.

[12] J. McConnell, J. D. Martin, and B. Englot, "Fusing concurrent orthogonal wide-aperture sonar images for dense underwater 3D reconstruction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 1653–1660.

[13] J. McConnell and B. Englot, "Predictive 3D sonar mapping of underwater environments via object-specific Bayesian inference," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 6761–6767.

[14] E. Westman, I. Gkioulekas, and M. Kaess, "A theory of fermat paths for 3D imaging sonar reconstruction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 5082–5088.

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

[16] M. D. Aykin and S. S. Negahdaripour, "Modeling 2-D lens-based forward-scan sonar imagery for targets with diffuse reflectance," *IEEE J. Ocean. Eng.*, vol. 41, no. 3, pp. 569–582, Jul. 2016.

[17] E. Westman and M. Kaess, "Wide aperture imaging sonar reconstruction using generative models," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 8067–8074.

[18] P. V. Teixeira, M. Kaess, F. Hover, and J. Leonard, "Underwater inspection using sonar-based volumetric submaps," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 4288–4295.

[19] M. Franchi, A. Bucci, L. Zacchini, E. Topini, A. Ridolfi, and B. Allotta, "A probabilistic 3D map representation for forward-looking sonar reconstructions," in *Proc. IEEE/OES Auton. Underwater Veh. Symp.*, 2020, pp. 1–6.

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

[21] Y. Wang, Y. Ji, H. Woo, Y. Tamura, A. Yamashita, and H. Asama, "Three-dimensional underwater environment reconstruction with graph optimization using acoustic camera," in *Proc. IEEE/SICE Int. Symp. Syst. Integration*, 2019, pp. 28–33.

[22] M. D. Aykin and S. Negahdaripour, "Three-dimensional target reconstruction from multiple 2-D forward-scan sonar views by space carving," *IEEE J. Ocean. Eng.*, vol. 42, no. 3, pp. 574–589, Jul. 2017.

[23] Y. Feng, W. Lu, H. Gao, B. Nie, K. Lin, and L. Hu, "Differentiable space carving for 3D reconstruction using imaging sonar," *IEEE Robot. Automat. Lett.*, vol. 9, no. 11, pp. 10065–10072, Nov. 2024.

[24] S. Arnold and B. Wehbe, "Spatial acoustic projection for 3D imaging sonar reconstruction," in *Proc. Int. Conf. Robot. Automat.*, 2022, pp. 3054–3060.

[25] N. Jaber, B. Wehbe, and F. Kirchner, "Sonar2depth: Acoustic-based 3D reconstruction using cGANs," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 5828–5835.

[26] M. Qadri, M. Kaess, and I. Gkioulekas, "Neural implicit surface reconstruction using imaging sonar," in *Proc. 2023 IEEE Int. Conf. Robot. Autom.*, 2023, pp. 1040–1047.

[27] H. Wang, N. Gao, Y. Xiao, and Y. Tang, "Image feature extraction based on improved FCN for UUV side-scan sonar," *Mar. Geophysical Res.*, vol. 41, pp. 1–17, 2020.

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

[29] E. Potokar, S. Ashford, M. Kaess, and J. G. Mangelson, "HoloOcean: An underwater robotics simulator," in *Proc. Int. Conf. Robot. Automat.*, 2022, pp. 3040–3046.

[30] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proc. 8th Workshop Syntax, Semantics Struct. Stat. Transl.*, 2014, pp. 103–111.

[31] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.

[32] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[33] J. Albiez et al., "Flatfish-A compact subsea-resident inspection AUV," in *Proc. OCEANS 2015-MTS/IEEE Washington*, 2015, pp. 1–8.