Learning Pseudo Front Depth for 2D Forward-Looking Sonar-based Multi-view Stereo

Yusheng Wang[1], Yonghoon Ji[2], Hiroshi Tsuchiya[3], Hajime Asama[1], and Atsushi Yamashita[4]

Abstract— Retrieving the missing dimension information in acoustic images from 2D forward-looking sonar is a well-known problem in the field of underwater robotics. There are works attempting to retrieve 3D information from a single image which allows the robot to generate 3D maps with fly-through motion. However, owing to the unique image formulation principle, estimating 3D information from a single image faces severe ambiguity problems. Classical methods of multi-view stereo can avoid the ambiguity problems, but may require a large number of viewpoints to generate an accurate model. In this work, we propose a novel learning-based multi-view stereo method to estimate 3D information. To better utilize the information from multiple frames, an elevation plane sweeping method is proposed to generate the depth-azimuth-elevation cost volume. The volume after regularization can be considered as a probabilistic volumetric representation of the target. Instead of performing regression on the elevation angles, we use pseudo front depth from the cost volume to represent the 3D information which can avoid the 2D-3D problem in acoustic imaging. High-accuracy results can be generated with only two or three images. Synthetic datasets were generated to simulate various underwater targets. We also built the first real dataset with accurate ground truth in a large scale water tank. Experimental results demonstrate the superiority of our method, compared to other state-of-the-art methods.

I. INTRODUCTION

When water visibility is poor, sonar is usually the only viable modality for underwater sensing. High-frequency 2D forward-looking sonar, or acoustic cameras, can generate high-resolution 2D acoustic images in any body of water. They are compact in size and can be easily mounted on remotely operated vehicles (ROVs) or autonomous underwater vehicles (AUVs). In the last few years, the sensors have achieved remarkable results in tasks such as mosaicking, 3D mapping, and robot navigation [1]–[4]. Because sonar is a 2D imaging sensor, dimension missing problem also exists similar to optical cameras. Unlike the optical camera where the depth information is missing, the information in the elevation angle direction is missing during the image formation for acoustic cameras. Retrieving 3D information from 2D acoustic images is a nontrivial problem. The unique

imaging model, severe noise, and sonar artifacts make this problem difficult.

Early works on 2D forward-looking sonar utilized sparse features, such as corners or edges, for 3D reconstruction [5]–[7]. Points or lines were used to represent the sparse 3D model, which is not intuitive for human comprehension. More recently, dense 3D reconstruction methods have been proposed, based on multi-view stereo or photometric stereo. Multi-view stereo methods basically utilize an efficient space carving scheme, considering the known sensing range and limited field of view (FoV) of 2D forward-looking sonars [2], [8], [9]. However, the deterministic multi-view stereo methods require a large number of viewpoints; the 3D reconstruction result is extremely poor when only a few viewpoints (e.g., 2∼3) are available. Photometric stereo methods can generate a 3D model from a single acoustic image by modeling ultrasound propagation or using shadow information [10], [11]. However, such methods require ideal conditions, and usually have strong assumptions about the scene. Recently, learning-based methods have demonstrated successful results in computer vision problems. A convolutional neural network (CNN)-based method was proposed to estimate the elevation angle for each pixel from a single acoustic image [12]. However, for acoustic images, one pixel may correspond to multiple elevation angles, which is a *non-bijective 2D-3D correspondence problem*, denoted as 2D-3D problem in the following paper. To solve this problem, our research group proposed a CNN-based method (A2FNet) to learn the front view depth from a single acoustic image [3].

In fact, estimating 3D information from a single acoustic image is a highly ill-posed problem and cannot be solved for many cases, considering the ambiguity existing in acoustic imaging (see Section III.C). In addition, the accuracy of single view 3D reconstruction is limited. Recently, studies have focused on sonar stereo [13], [14]. Previous studies on sonar stereo still require handcrafted features that may not perform stably. In this work, we also emphasize the benefits of sonar stereo. Deterministic multi-view stereo methods can solve the ambiguity problem caused by single-view 3D reconstruction, but may require a large number of viewpoints. This paper aims to solve the 3D reconstruction problem by multi-view stereo but use only 2∼3 viewpoints.

Retrieving 3D information using only a few viewpoints is still difficult, in this work, we propose a deep learning based method for the multi-view sonar stereo problem as shown in Fig. 1. We propose a novel Elevation Plane Sweeping Stereo Network (EPSSN). The major task is to retrieve the 3D information of the reference frame. The source frames are
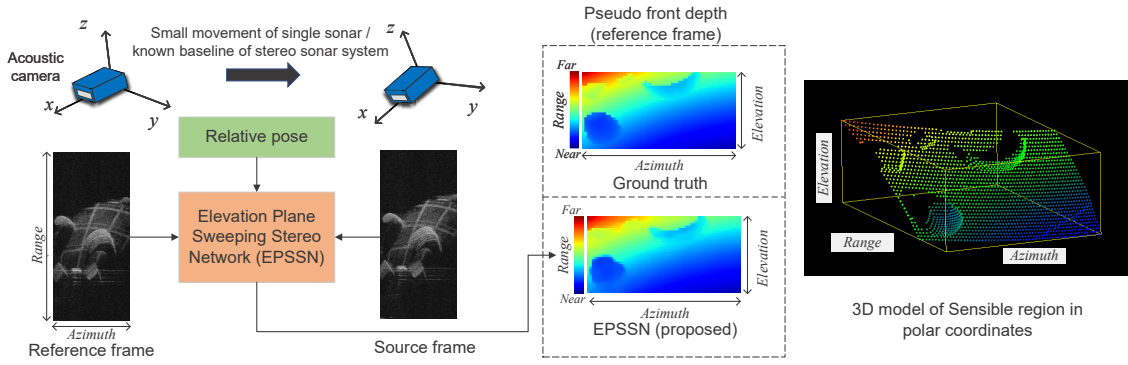
Fig. 1. Workflow of the proposed method. The reference frame and the source frames of sonar stereo are input to the elevation plane sweeping stereo network (EPSSN) with the corresponding relative poses. The EPSSN estimates the pseudo front depth of the reference frame. The acoustic image is an azimuth-range image in which the elevation information is missing. The 3D information can be retrieved by generating azimuth-elevation depth map.

the ones from viewpoints with small motion change to help the 3D reconstruction of the reference frame. The relative pose between the reference frame and the source frames are assumed given in this work, which can be acquired from control input of a motor, inertial measurement unit (IMU) or a known baseline of sonar stereo system. By providing reference and source frames with relative pose information, it is possible to generate a highly accurate pseudo front view depth for the reference frame. By warping the features of the source frame to virtual elevation planes in the reference frame, it is possible to generate a depth-azimuth-elevation cost volume. Finally, the front view depth can be directly generated from the cost volume according to the geometry relationship. It is noticed that not only the ambiguity can be solved by using 2∼3 viewpoints instead of a single viewpoint, the accuracy of 3D reconstruction results may increase drastically.

Our contributions are listed as follows.

- We propose a novel CNN-based method to generate a dense 3D reconstruction from few (e.g., two) acoustic images.
- We propose an elevation plane sweeping scheme to generate the depth-azimuth-elevation cost volume. The front view depth can be calculated from the cost volume. By using the front view depth, the 2D-3D problem can be avoided.
- We generate synthetic and real datasets for evaluation. The synthetic datasets include natural and artificial scenes, and have been released under an open-source license[1]. Our proposed method outperforms other state-of-the-art methods.

II. RELATED WORKS

A. 3D Reconstruction of Acoustic Images

Early methods used sparse points or lines for the 3D reconstruction of acoustic images. Mai et al. and Huang et al. used manually selected corner points combined with simultaneous localization and mapping (SLAM) systems for

3D information acquisition [5], [7]. For automatic feature detection, AKAZE features have been proven to be effective in acoustic images that have been used for sonar-based localization [15]. Wang et al. tracked AKAZE features based on optical flow and assumed the terrain as a Gaussian process random field on a tree structure [16]. The terrain is assumed to be smooth and the computation cost increases drastically when the feature points increase.

For better 3D representation, dense 3D reconstruction has been studied. For multi-view stereo methods, Aykin et al. applied a space carving method for small objects [8]. The acoustic images were binary-segmented into object and non-object regions. The space was then voxelized and projected onto the binary image. The voxels projected to the object region were kept and the others were carved. Guerneve et al. utilized a linearized sonar projection model, with min-filtering for space carving [9]. Our group proposed a probabilistic method using 3D occupancy mapping with an inverse sensor model [2]. It can be applied to more general scenes such as 3D mapping. However, the aforementioned methods require a large number of viewpoints and do not work well with only one or two viewpoints. Photometric stereo methods utilize shadows or other image cues to retrieve 3D information. Aykin et al. physically modeled ultrasound propagation and scenes, and proved that the physical model can be used for 3D reconstruction with object contours [10]. Westman et al. utilized a similar scheme and generated 3D model of continuous surface [11]. Non-line-of-sight (NLOS) methods have also been applied to acoustic images and proved to be feasible [17]. However, these methods usually have strong assumptions, e.g., that shadows or Fermat paths [17] can be accurately detected, which is difficult, owing to noise and artifacts.

Recent deep learning-based methods have also been utilized in this field. Debortoli et al. used a UNet-like structure for pixel-wise elevation angle estimation of a single acoustic image [12]. For the synthetic dataset, the network was trained in a supervised manner. For the real dataset, the network was first trained using synthetic images and fine-tuned using a self-supervised scheme with small motions. The study

---

[1]https://github.com/sollynoay/EPSSN

assumes that each pixel only corresponds to one elevation angle which is not always true. To overcome this limitation, our group proposed a network to learn the front view depth, which is a better representation with no 2D-3D ambiguity problem [3]. The network implicitly learns to transfer an azimuth-range image to an azimuth-elevation depth map. Retrieving 3D information from a single acoustic image is highly ill-posed, and cannot be solved, owing to the ambiguity in many cases that are difficult to apply in real underwater scenarios.

Negahdaripour first proposed sonar stereo, established sonar epipolar geometry theory and proved the optimal configuration of two forward-looking sonars [13]. McConnell et al. mounted two orthogonally configured sonars to an ROV to generate dense point clouds [14]. Such methods still require handcrafted features in the acoustic images. In this work, we are the first to propose a learning-based method for multi-view sonar stereo. It can be used for structured sonar stereo and multiple images from a single sonar. It can solve the problems caused by sonar imaging and generate accurate dense 3D reconstruction results even for complex underwater scenarios.

B. Learning Multi-view Stereo

For optical cameras, retrieving 3D information from stereo cameras is a long-standing problem. Recently, learning-based end-to-end depth from stereo matching methods has shown outstanding abilities [18]–[20]. By building 3D cost volumes to fuse the information from stereo images, and using soft argmax to estimate the depth of each pixel, the neural networks show high performance for depth estimation problems [20]. However, they usually assume that the two optical cameras are well calibrated and the image pairs are rectified.

Recently, studies have been conducted on unstructured multi-view stereo [21], [22]. To build the cost volume, plane sweeping scheme [23] is applied, by projecting images or features to virtual depth planes in the reference camera coordinates. The projection process was calculated using homography. After acquiring the cost volume, multi-view stereo usually follows the same structure as stereo matching methods. Acoustic cameras have a different projection theory compared to optical cameras, so theories such as homography cannot be applied. In addition, the depth estimation of optical images directly estimates the depth of each pixel in the optical image. However, instead of estimating the elevation angles for each pixel, we estimate the pseudo front depth from the acoustic images.

III. PRELIMINARIES

A. Sonar Projection Theory

A 3D point in the sonar coordinate system is usually represented as (r, θ, ϕ) in the polar coordinate system, which can be transformed to the Euclidean coordinate as follows.

$$\begin{bmatrix} X_c & Y_c & Z_c \end{bmatrix}^\top = \begin{bmatrix} r\cos\phi\cos\theta & r\cos\phi\sin\theta & r\sin\phi \end{bmatrix}^\top.$$
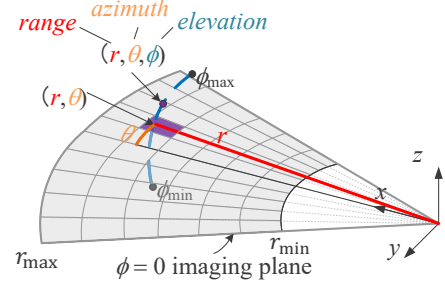(1)

Fig. 2. Geometry model of 2D forward looking sonar. It can be equivalently seen as a projection of a 3D point to the $\phi = 0$ plane, where the elevation information is missing.
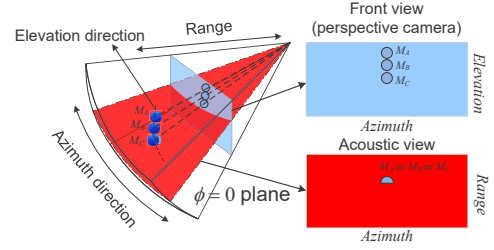

Fig. 3. An example of the ambiguity problem. The spheres from M_A to M_C may generate the same acoustic image. This may cause ambiguity for inverse problem.

During the imaging process, the ϕ angle information is missing, and the corresponding 2D point in the image coordinates can be written as (r, θ) as shown in Fig. 2. Different from optical imaging, each pixel in the acoustic images may correspond to multiple 3D points. The backscattered intensity of the 3D points with the same (r, θ) accumulates. For each (r, θ) in the acoustic image I_a, discretizing ϕ as $\{\phi_1, \phi_2, \ldots, \phi_n\}$ and the i-th corresponding backscattered intensities as $I(r, \theta, \phi_i)$ gives $I_a(r, \theta) = \sum_{i=1}^{n} I(r, \theta, \phi_i)$.

B. Pseudo Front Depth

Considering the ultrasound as rays, each ray strikes the object surface and reflected to sensor. Before integration in the elevation angle direction, it can be seen as a common time-of-flight (ToF) sensor, from which a depth map can be acquired. Theoretically, a front depth map contains the entire 3D information of the sensible region as shown in Fig. 1. Instead of estimating the elevation angles directly for each pixel, the front depth map representation can avoid the 2D-3D problem [3]. However, it is less constrained than the direct elevation angle estimation problem.

C. Ambiguity Problem

Sonar imaging faces severe ambiguity problems. For example, when sensing a single spherical object, theoretically, all spheres from M_A to M_C will generate the same acoustic image as shown in Fig. 3. Estimating the sphere position from a single image is highly ill-posed. In addition, objects or scenes symmetric to the imaging plane will generate the same image. For CNN-based regression methods, if one input

corresponds to multiple ground truth labels, the network cannot learn the problem well. Therefore, it is necessary to study sonar stereo to overcome the aforementioned problems.

IV. LEARNING SONAR STEREO

A. Overview

Figure 4 shows the network structure overview of EPSSN. The inputs are the reference frame and the source frames with corresponding poses. The output is the pseudo front depth map of the reference frame. We propose two major novel blocks: the elevation plane warping and front depth generation blocks. A 2D CNN is first used to extract deep features $\mathbf{F}$ from the acoustic images. The features in the source frames are warped to the reference frame based on elevation plane sweeping to generate cost volumes as explained in Section IV.B. In this study, we use a depth-azimuth-elevation volume. After fusing the cost volumes from the reference frame and the source frames, a following 3D CNN is used to refine and upsample the initial cost volume. Finally, the front depth generation block is used to calculate the front depth map from the cost volume, as explained in Section IV.C.

B. Elevation Plane Warping

Plane sweeping stereo has been commonly used in multi-view stereo for optical cameras [23]. Multi-view images are projected onto virtual depth planes in the reference image coordinates to generate a cost volume. The depth maps are then estimated using this cost volume. This type of theory has never been used for acoustic images because of the difference in projection theory. In this work, we propose an elevation plane sweeping scheme for acoustic images to generate a depth-azimuth-elevation volume and use the volume to further generate front depth maps. We first generate N virtual elevation planes along the elevation angle direction as shown in Fig. 5. The elevation angle ϕ of the j-th plane can be described as follows.

$$\phi_j = \phi_{\min} + \frac{j(\phi_{\max} - \phi_{\min})}{(N-1)}, \tag{2}$$

where j is an integer from 0 to $N-1$. For each virtual elevation plane, denoting a 3D point i in plane j as $\mathbf{p}_{i,j}^{src} = (r_{i,j}, \theta_{i,j}, \phi_j)$, the point is first transformed into the Euclidean coordinate as $\mathbf{e}_{i,j}^{src}$. The point in the reference frame is then transformed into the source coordinate as $\mathbf{e}_{i,j}^{ref} = \mathbf{T}_{ref}^{src}\mathbf{e}_{i,j}^{src}$. Here, $\mathbf{T}_{ref}^{src}$ denotes the relative motion between the reference and source frames. Then, we transform the point to the polar coordinate as $\mathbf{p}_{i,j}^{ref}$ and project the point onto the imaging plane in the source frame. We use bilinear sampling to acquire the corresponding feature $\mathbf{F}$ in source frame. Then, we can warp feature $\mathbf{F}$ to elevation plane j and the result is denoted as $\mathbf{F}_j^{wp}$. The warped source frame cost volume $\mathbf{V}_{src}$ can be generated as follows.

$$\mathbf{V}_{src} = CAT(\mathbf{F}_0^{wp}, \dots, \mathbf{F}_{N-1}^{wp}), \tag{3}$$

where $CAT(.)$ refers to the concatenation operation along a new dimension corresponding to the elevation direction.

The initial cost volume is a 4D tensor, for which the shape is $C \times N \times H/4 \times W/4$, where C refers to the length of feature channels. The cost volume for the reference frame itself is denoted as $\mathbf{V}_{ref}$. The volumes are aggregated and the initial cost volume $\mathbf{C}$ is calculated as follows.

$$\mathbf{C} = \frac{(\mathbf{V}_{ref} - \bar{\mathbf{V}})^2 + (\mathbf{V}_{src} - \bar{\mathbf{V}})^2}{2}, \tag{4}$$

where $\bar{\mathbf{V}}$ is the element-wise arithmetic mean of $\mathbf{V}_{src}$ and $\mathbf{V}_{ref}$. This can be easily extended if there are multiple source frames.

C. Front Depth Generation

The initial cost volume is noisy and not regularized. A 3D UNet-like structure [24] is used to refine the initial cost volume, followed by an extra 3D convolution transpose layer to further upsample the cost volume. The last convolution layer makes the size of feature channels to one.

The refined cost volume can be considered as a probabilistic volumetric representation of the target as shown in Fig. 6. Each voxel has a probability value of occupancy and corresponds to the 3D space in the reference coordinate. In particular, in our cases, the boundaries of the cost volume refer to the scope of the sonar. In this work, we assume that the front depth representation is ideal so that the probability distribution for the range of a (θ, ϕ) pair is always unimodal. On the other hand, the probability distribution for the elevation angle of a (r, θ) pair is more likely to be multimodal due to the non-bijective projection as explained in Section III.A. Theoretically, the front depth map can be retrieved by winner-take-all (i.e., argmax) [20]. However, it is difficult to use the same scheme for the multi-modal distribution of the elevation map case. Because argmax cannot offer subpixel results and is not differentiable for network training, it is usually replaced by soft argmax in the neural network. The depth $\hat{\mathbf{D}}$ for (θ, ϕ) is calculated as follows using softmax and the depth value for each cell.

$$\hat{\mathbf{D}}(\theta, \phi) = \frac{1}{\sum_{j=r_{\min}}^{r_{\max}} e^{\mathbf{C}(j,\theta,\phi)}} \sum_{d=r_{\min}}^{r_{\max}} d \times e^{\mathbf{C}(d,\theta,\phi)}. \tag{5}$$

D. Loss Functions

In this study, we estimate the depth information instead of reverse depth. This is mainly because the sensing range $|r_{\max}\text{-}r_{\min}|$ is not a large value (e.g. 2 m) in this work. This may lose precision, owing to float-point numbers when using reverse depth. The shortcoming of using direct depth is that some of the values in the ground truth depth value may exceed $r_{\max}$, especially for submerging objects, and some depth values may be nearly infinite. During the training time, we provide a binary mask $\mathcal{M}$ from depth labels $\mathbf{D}$ as follows.

$$\mathcal{M}(\theta, \phi) = \begin{cases} 0 & \mathbf{D}(\theta, \phi) \leq r_{\min} \vee \mathbf{D}(\theta, \phi) \geq r_{\max} \\ 1 & r_{\min} < \mathbf{D}(\theta, \phi) < r_{\max} \end{cases}. \tag{6}$$

The loss function is written as follows.

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \lambda \mathcal{M}_i |\hat{\mathbf{D}}_{\mathbf{i}} - \mathbf{D}_{\mathbf{i}}| + \bar{\mathcal{M}}_i |\hat{\mathbf{D}}_{\mathbf{i}} - \mathbf{D}_{\mathbf{i}}|, \tag{7}$$
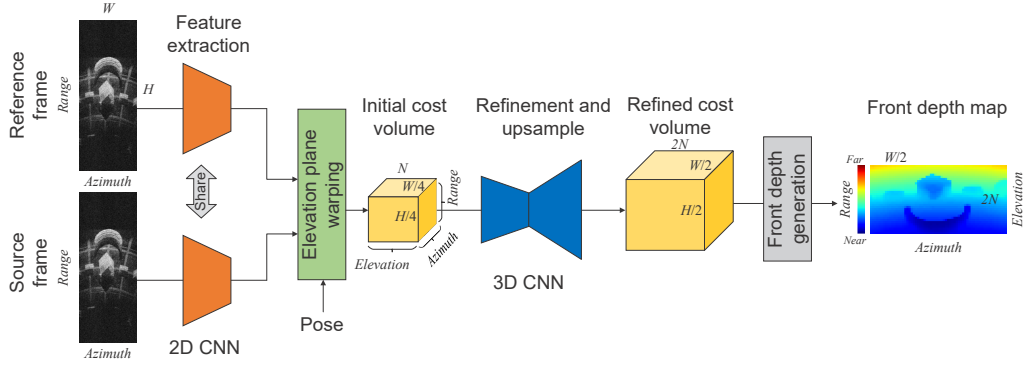
Fig. 4. Overview of EPSSN. The reference and source frames are passed through a 2D CNN for feature extraction. We propose a novel feature-level elevation plane warping method to generate the depth-azimuth-elevation cost volume. A 3D CNN is used to refine and upsample the cost volume. Then, a front depth generation block is proposed to generate the front depth map.
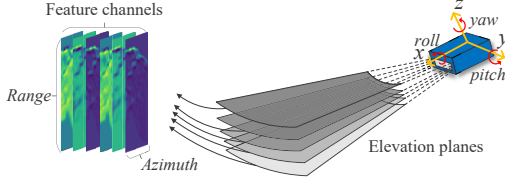

Fig. 5. Elevation plane warping. We generate virtual elevation planes along elevation angle direction. Each plane warps the deep feature and all the warped features are concatenated in a new dimension which refers to the elevation direction.
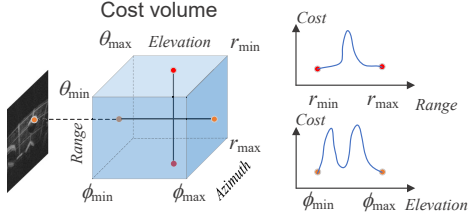

Fig. 6. Refined cost volume. The refined cost volume can be considered as a probabilistic volumetric representation of the target. In our problem, the probability distribution along the elevation direction may be multi-modal. On the other hand, we assume that the distribution along the range direction is always uni-modal.

(a)   (b)

Fig. 7. Water tank experiment. (a) Moving device for collecting data from different positions. (b) Target in the large scale water tank.

where $\bar{\mathcal{M}}$ refers to $1 - \mathcal{M}$ and λ is a hyperparameter that balances the two terms in the equation. Noting that for ground truth depth maps, during training, we set the range larger than $r_{\max}$ to $r_{\max}$ and the range smaller than $r_{\min}$ to $r_{\min}$. We assume that the object or scene targets are away from the boundaries. It is then easy to filter the non-informative estimations during the test times.

V. EXPERIMENT

To evaluate the proposed methods, we generated both simulation and real datasets. Usually, roll rotation is considered to be efficient for retrieving 3D information. Existing sonar stereo systems use concurrent orthogonal sonars for sensing which can be approximately considered as a roll rotation of
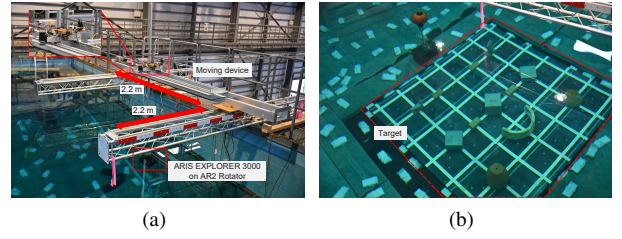
$90°$ [13], [14]. However, to ensure a sufficient overlap of the images, we used known small roll rotations ($5° \sim 10°$) to generate image pairs in this work. The influence of motion will be investigated in future studies.

A. Simulation Datasets

The simulation datasets were generated in Blender environment which can be found on our GitHub[2]. We prepared three datasets with different scenes, including artificial environment, terrain, and sphere. The artificial environment was built using a 3D CAD model which is also the ground truth of our real dataset. Image pairs were generated at random positions throughout the scene. We generated 4,000 image pairs for training and 2,000 pairs for test. For the terrain dataset, we generated terrains with the A.N.T. Landscape add-on in Blender by using hetero noise, which simulated undulating seabed. We trained the network on one terrain and tested it on another terrain. We generated 1,000 image pairs for training and 1,000 pairs for test. The sphere dataset was used to verify the ambiguity problem explained in Section III.C. To generate the dataset, the sphere was located at a random position within the sonar scope.

B. Real Datasets

We collected a real dataset in a large scale water tank with Sound Metrics ARIS EXPLORER 3000 imaging sonar. The configuration of the objects was designed and then set up by a diver in the water tank. The ground truth of the

[2]https://github.com/sollynoay/Sonar-simulator-blender

objects was acquired using Agisoft Metashape in land. As shown in Fig. 7, a moving device was built to move the sonar to different positions. The sonar was mounted on a Sound Metrics AR2 rotator for roll rotations. To acquire the ground truth pose of the sonar in water tank coordinates, we calibrated the initial pose of the sonar by matching the edges of the real and synthetic images [25]. The rest of the poses were calculated from the control input of the moving device and the roll rotations. The ground truth depth labels were generated by inputting the sonar pose into the simulator.

The sonar with the rotator was moved by the device to 51 positions. For each position, the sonar performed a roll rotation from -35° to 35° and an acoustic video was recorded. Denoting the roll angle for the reference frame as $\alpha°$, the source frame is taken at the roll angle of approximately $(\alpha + 7)°$. In total, we generated 1,493 pairs of reference and source images. The dataset was then separated randomly at an 8:2 ratio for training and test. Thus, there were 1,194 pairs for training and 299 pairs for test. The resolution in the range direction for the real images was downsampled by twice. All the heights and widths for the synthetic images and real images were 512×128.

C. Training

The proposed method was implemented in PyTorch. The networks were trained and tested on NVIDIA GeForce RTX 3090 graphics card. For each dataset, we trained 50 epochs. The learning rate began at 0.001 and was reduced by half after 25 epochs. The batch size was set to 16 for training and test. Adam optimizer was used for optimization. λ was set to 3.

D. Metrics

To evaluate the results, two metrics were used: the mean absolute error (MAE) for the front view depth map, and chamfer distance (CD) for the point cloud.

$$\text{MAE} = \frac{\beta}{HW} \times \sum_{i=1}^{H} \sum_{j=1}^{W} |\hat{D}(i,j) - D(i,j)|, \quad (8)$$

$$\text{CD} = \frac{\gamma}{S_1} \sum_{x \in S_1} \min_{y \in S_2} ||x-y||_2^2 + \frac{\gamma}{S_2} \sum_{y \in S_2} \min_{x \in S_1} ||x-y||_2^2, \quad (9)$$

where β and γ are scale parameters, which refer to 1,000 and 500, respectively. S_1 and S_2 refer to the two point sets.

E. Results

To evaluate our method, we also compared our results with state-of-the-art methods, ElevateNet [12] and A2FNet [3]. To implement ElevateNet, we chose UNet [24] as the backbone for training, with a sigmoid layer before the final output. Elevation maps were also generated using our simulator. The network was trained in a supervised manner using L1 loss. We also prepare a binary mask to segment the signal/non-signal region and the loss for the non-signal region was set to zero. During test time, the mask was multiplied to the network output. For A2FNet, the inverse pixel shuffle block was replaced by a four-layer CNN for feature extraction. We

TABLE I
SIMULATION RESULTS

	A2FNet	ElevateNet	Ours
Artificial (simulation)			
CD [m]	0.2707±0.1053	0.3263±0.1301	**0.1368±0.0473**
MAE [m]	10.41 ± 3.09	–	**4.90±1.89**
Terrain			
CD [m]	0.9962 ± 0.6155	0.5102 ± 0.2394	**0.2200±0.0837**
MAE [m]	42.19±16.59	–	**14.07±4.67**
Sphere			
CD [m]	24.7811 ± 27.1205	9.6571 ± 9.4248	**1.4603±2.4521**
MAE [m]	71.56±20.46	–	**9.60±5.87**

TABLE II
REAL EXPERIMENT RESULTS

	A2FNet	ElevateNet	Ours
Artificial (real water tank)			
CD [m]	0.8677±0.2583	0.7605±0.2234	**0.2822±0.0879**
MAE [m]	30.54 ± 7.07	–	**13.62±2.94**

used the same hyper-parameters to train the baseline methods as our EPSSN.

The results of the simulation datasets are listed in Table I. We show the average error and the standard deviation. All the metrics are the smaller the better. For artificial (simulation) and terrain datasets, our proposed method can generate more accurate and precise results compared to the baselines. For sphere datasets, the baseline methods using a single image can hardly deal with the problem; however, our proposed method can successfully estimate the 3D information. Figure 8 shows an example of the visualization of the results. For ElevateNet, although elevation maps were estimated, they were transferred to front view depth maps. The color refers to distance, where blue and red refer to the minimum and maximum distances, respectively. It can be seen from the results that our proposed method is closest to the ground truth. ElevateNet suffers from the incompleteness caused by non-bijective correspondence. For sphere datasets, A2FNet tends to estimate all the possible positions of the sphere; on the other hand, ElevateNet tends to estimate one of the possible results. By using two or more images, such ambiguity can be resolved and our proposed method can estimate the true results. Later, we will discuss how viewpoint numbers will influence our proposed method and the baseline methods.

The results for the real dataset are listed in Table II. Our proposed method is still far more better than the baseline methods. Examples of visualizing the results are shown in Fig. 9. We show the results of the mesh model for comparison. The mesh models were generated in MeshLab by first estimating normals and using ball pivoting. Clearly, the proposed method generated the 3D model with the best quality.

F. Discussions

1) Viewpoint Numbers: The viewpoint numbers may influence the results. For a fair comparison, we also added viewpoints for ElevateNet and A2FNet by concatenating

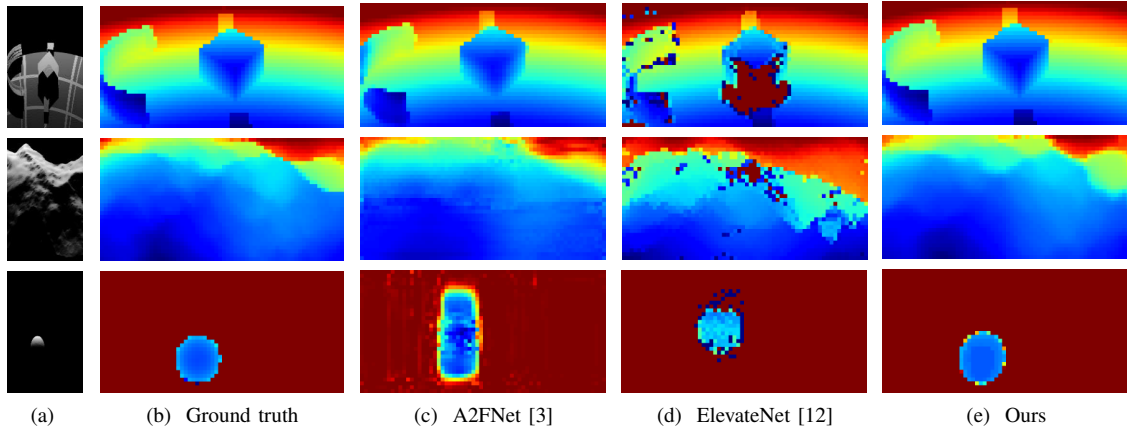| | (a) | (b) Ground truth | (c) A2FNet [3] | (d) ElevateNet [12] | (e) Ours |

Fig. 8. Visualization examples of simulation datasets. (a) Reference frames. Each row refers to a simulation dataset, which includes artificial (simulation), terrain, and sphere. The results are shown in depth maps. The color refers to depth values. Red and blue refer to the minimum and maximum, respectively.

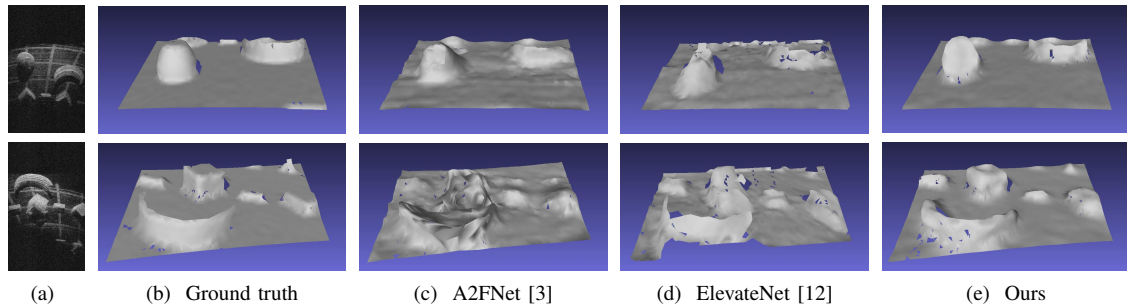| | (a) | (b) Ground truth | (c) A2FNet [3] | (d) ElevateNet [12] | (e) Ours |

Fig. 9. Visualization examples of real datasets. (a) Reference frames. We first transfer the estimation results to point clouds and then generate mesh models for comparison.

TABLE III

VIEWPOINT NUMBERS

Artificial (simulation)			
	A2FNet (two)	ElevateNet (two)	Ours
CD [m]	0.2707±0.1053	0.3745±0.1597	**0.1368±0.0473**
Terrain			
	A2FNet (two)	ElevateNet (two)	Ours
CD [m]	0.9097±0.3439	0.3541±0.1040	**0.2200±0.0837**

Fig. 10. Ablation test results using the CD. Doing regression on the depth outperforms the regression on elevation. Increasing viewpoints may improve the performance of the proposed method.

the reference and source images as inputs and modifying the channels of the networks. We chose the artificial and terrain datasets for training and test. The performance of ElevateNet seems unstable which significantly improved on terrain dataset but became worse on artificial dataset. This may because the 2D-3D problem is more obvious on artificial dataset. Our proposed method still outperforms the baseline. We also tied to increase the viewpoint number for EPSSN to three on terrain dataset, the CD was **0.1333±0.0573** m, which shows significant improvements.

2) Regression on Elevation Map: Front depth generation from the cost volume and regression on the front depth map is one of the main contributions of this work. We tried to replace this part by generating elevation map and performing regression on it. Figure 10 shows the performance of elevation map regression and depth map regression on CD. For the same number of viewpoints, depth map regression significantly outperforms elevation map regression.
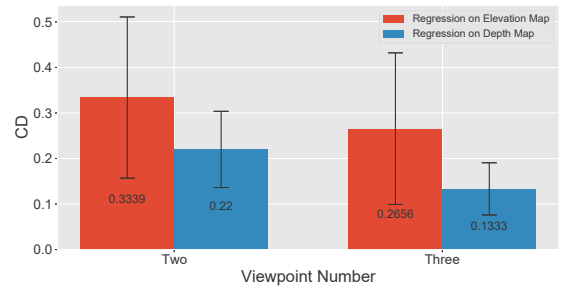
3) Ablation on Elevation Plane Warping: An ablation test was carried out on the elevation plane warping block when there were more viewpoints (three). By removing the warping process and simply adding the cost volumes from the reference frame and the source frames, the final result may be even worse than the case using two viewpoints with warping as shown in Table IV. In other words, the elevation plane warping successfully fuse the information from multiple frames with known motion information.

4) Comparison with Occupancy Mapping: We also compared the proposed method with a previous method using occupancy mapping [2]. We used two frames, one reference frame and one source frame, with known relative poses to

TABLE IV

ABLATION TEST ON ELEVATION PLANE WARPING

	Terrain	
	w / o warping (three)	w warping (three)
CD [m]	0.2734±0.0819	**0.1333±0.0573**
MAE [m]	16.26 ± 4.59	**9.13±2.70**

generate the local map. With only two frames with a small roll angle difference, the results of occupancy mapping did not converge. On terrain dataset, the CD error for the test set was 46.83 ± 19.28.

VI. CONCLUSIONS

In this study, we proposed a novel method for retrieving the missing 3D information in the acoustic images. By utilizing elevation plane warping to fuse the information from multiple frames and learning the pseudo front depth, our proposed method achieved state-of-the-art performance on both simulation and real datasets. It can also deal with severe ambiguous cases such as submerging sphere in water bodies. Our simulation datasets were made open source for future comparisons. The methods can be applied to underwater vehicles with two forward-looking sonars or a single forward-looking sonar with multiple views.

Future work may include self-supervised network training and attempting various motions. Collecting datasets in real underwater environments is extremely difficult. In this study, the datasets were collected in a structured water tank where depth labels could be generated from our simulator. However, this is difficult to transfer to oceanic environments. Training the network without depth labels is worth investigating. Other methods, such as generative adversarial networks (GANs) may be valuable for mitigating the sim-to-real gap so that we can train the network based on synthetic data [26]. In this work, the motions between the reference frame and the source frames were pure roll rotation; other motions may also be tested in the future.

REFERENCES

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

[2] Y. Wang, Y. Ji, H. Woo, *et al.*, "Acoustic camera-based pose graph slam for dense 3-d mapping in underwater environments," *IEEE Journal of Oceanic Engineering*, vol. 46, no. 3, pp. 829–847, 2021.

[3] 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.

[4] S. Negahdaripour, "On 3-d motion estimation from feature tracks in 2-d fs sonar video," *IEEE Transactions on Robotics*, vol. 29, no. 4, pp. 1016–1030, Aug. 2013.

[5] N. T. Mai, H. Woo, Y. Ji, Y. Tamura, A. Yamashita, and H. Asama, "3-d reconstruction of underwater object based on extended kalman filter by using acoustic camera images," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 1043–1049, 2017.

[6] N. T. Mai, H. Woo, Y. Ji, Y. Tamura, A. Yamashita, and H. Asama, "3-d reconstruction of line features using multi-view acoustic images in underwater environment," *Proceedings of the 2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI2017)*, pp. 312–317, Nov. 2017.

[7] T. A. Huang and M. Kaess, "Towards acoustic structure from motion for imaging sonar," *Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2015)*, pp. 758–765, Sep. 2015.

[8] 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, Jul. 2017.

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

[10] 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.

[11] E. Westman and M. Kaess, "Wide aperture imaging sonar reconstruction using generative models," *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2019)*, pp. 8067–8074, Nov. 2019.

[12] R. DeBortoli, F. Li, and G. A. Hollinger, "Elevatenet: A convolutional neural network for estimating the missing dimension in 2d underwater sonar images," *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2019)*, pp. 8040–8047, 2019.

[13] S. Negahdaripour, "Application of forward-scan sonar stereo for 3-d scene reconstruction," *IEEE Journal of Oceanic Engineering*, vol. 45, no. 2, pp. 547–562, 2020.

[14] J. McConnell, J. D. Martin, and B. Englot, "Fusing concurrent orthogonal wide-aperture sonar images for dense underwater 3d reconstruction," *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2020)*, pp. 1653–1660, 2020.

[15] J. Li, M. Kaess, R. M. Eustice, and M. Johnson-Roberson, "Pose-graph slam using forward-looking sonar," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2330–2337, 2018.

[16] J. Wang, T. Shan, and B. Englot, "Underwater terrain reconstruction from forward-looking sonar imagery," *Proceedings of the 2019 IEEE International Conference on Robotics and Automation (ICRA2019)*, pp. 3471–3477, 2019.

[17] E. Westman, I. Gkioulekas, and M. Kaess, "A theory of fermat paths for 3d imaging sonar reconstruction," *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2020)*, pp. 5082–5088, 2020.

[18] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *Journal of Machine Learning Research*, vol. 17, pp. 1–32, 2016.

[19] N. Mayer, E. Ilg, P. Hausser, *et al.*, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR2016)*, pp. 4040–4048, 2016.

[20] A. Kendall, H. Martirosyan, S. Dasgupta, *et al.*, "End-to-end learning of geometry and context for deep stereo regression," *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV2017)*, pp. 66–75, 2017.

[21] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," *Proceedings of the 2018 European Conference on Computer Vision (ECCV2018)*, pp. 767–783, 2018.

[22] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, "Deepmvs: Learning multi-view stereopsis," *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR2018)*, pp. 2821–2830, 2018.

[23] R. Collins, "A space-sweep approach to true multi-image matching," *Proceedings of the 1996 IEEE Conference on Computer Vision and Pattern Recognition (CVPR1996)*, pp. 358–363, 1996.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI2015)*, pp. 234–241, 2015.

[25] Y. Wang, Y. Ji, D. Liu, H. Tsuchiya, A. Yamashita, and H. Asama, "Simulator-aided edge-based acoustic camera pose estimation," *OCEANS 2022 - Chennai*, pp. 1–4, 2022.

[26] D. Liu, Y. Wang, Y. Ji, H. Tsuchiya, A. Yamashita, and H. Asama, "Simulator-aided edge-based acoustic camera pose estimation," *Advanced Robotics*, vol. 35, no. 3-4, pp. 242–254, 2021.