Elevation Angle Estimation in 2D Acoustic Images Using Pseudo Front View

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

***Abstract*—A novel method to estimate the missing dimension in 2D acoustic images for 3D reconstruction is proposed in this paper. Acoustic cameras can acquire high resolution 2D images in underwater environment insusceptible to water turbidity and light condition. However, the formulation of acoustic images leads to the missing dimension problem. Estimating the unknown elevation angle dimension is a difficult task which has recently drawn the attention of researchers. The non-bijective characteristic between 3D points and 2D pixels increases the complexity of the problem. In this paper, a novel elevation angle estimation method is proposed. The method transfers the acoustic view to pseudo front view using a deep neural network. The proposed network can estimate the missing dimension and resolve the non-bijection problem of the 2D-3D correspondence. Because of the difficulty of acquiring depth information in underwater environments, the network is trained using simulated images. To mitigate the sim-real gap, a neural style transfer method is implemented to generate a realistic image dataset for training. Simulation experiments were carried out for evaluation and real data proved the feasibility of the proposed method.**

***Index Terms*—Marine Robotics, Deep Learning for Visual Perception, Deep Learning Methods**

I. INTRODUCTION

DEPTH estimation based on monocular cameras has recently become one of the research topics that has received the most focus in computer vision. In an underwater environment, the performance of an optical camera is restricted by the visibility. The acoustic camera, a next-generation 2D forward-looking imaging sonar, has outstanding capabilities in underwater environments [1]. Acoustic images have millimeter-level resolution in the depth direction and the cameras are small in size, which makes them suitable for mounting on underwater robots such as remotely operated vehicles (ROVs) or autonomous underwater vehicles (AUVs). Acoustic cameras have already been applied in various underwater tasks, such as robot navigation, mosaicing, and mapping [2]–[4]. Despite the high performance of acoustic cameras, the formulation of acoustic images also leads to the missing dimension problem. Unlike optical cameras, in acoustic cameras, the elevation angle direction information is missing. This prevents the acquisition of the full 3D information.

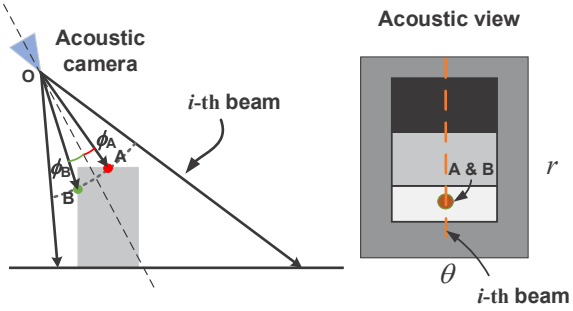


Fig. 1. Non-bijective 2D-3D correspondence: Point A and point B with different elevation angles will be projected to the same position in the acoustic image. The integration of points may lead to a higher intensity in the acoustic image. The acoustic image is a $\theta - r$ matrix in polar coordinate system.

Retrieving 3D information from acoustic images is one of the most fundamental but challenging problems in acoustic cameras. Acoustic images are described by unique imaging theories and have low signal-to-noise ratio (SNR) and complex sonar artifacts which increase the difficulty of the problem. Early research has mainly focused on sparse 3D reconstruction of acoustic images [5], [6]. The reconstructed 3D models are made up of sparse points or line frames, which are not intuitive for 3D representation. The performance of the methods is highly dependent on the performance of the feature extraction and data association processes. Most of the handcrafted features perform poorly on acoustic images [7]. Recently, researchers have focused more on dense 3D reconstruction using acoustic images. A variety of methods have been applied to this problem. They can be roughly classified into photometric stereo and multiple-view stereo methods. Photometric stereo methods utilize shadow information or model the ultrasound propagation for 3D reconstruction [8]–[10]. Such methods work in ideal conditions, but are neither robust nor general. Multiple-view stereo methods are another group of methods which achieve 3D reconstruction using multiple acoustic images [11]–[13]. They require a large number of view points for 3D reconstruction, which makes it necessary to hover around the target or employ the help of a rotator. With the rapid development of deep learning, it is now possible to achieve 3D reconstruction with a single image. Recently, a neural network-based method has been proposed to retrieve the elevation angle from a single acoustic image, with a self-supervised method

Manuscript received: October, 13, 2020; Revised December, 30, 2020; Accepted February, 2, 2021.

This paper was recommended for publication by Editor Pauline Pounds upon evaluation of the Associate Editor and Reviewers' comments.

This work was partly supported by JSPS KAKENHI, Grant Number 20K19898.

[1]Y. Wang, D. Liu, A. Yamashita, and H. Asama are with the Department of Precision Engineering, Graduate School of Engineering, University of Tokyo, Japan. {wang,liu,yamashita,asama}@robot.t.u-tokyo.ac.jp

[2]Y. Ji is with the Graduate School for Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Japan. ji-y@jaist.ac.jp

[3]H. Tsuchiya is with the Research Institute, Wakachiku Construction Co., Ltd., Japan. hiroshi.tsuchiya@wakachiku.co.jp

Digital Object Identifier (DOI): see top of this page.

based on known small camera motion to train the network [14]. However, the main problem is that the method neglects the non-bijective 2D-3D correspondence characteristic of the acoustic image, which will lead to estimation failure in many cases. As shown in Fig. 1, the red point A and the green point B will be projected to the same position in the acoustic image. This is a common occurrence in sonar imaging [8], [15]. Estimating the elevation angle of each pixel alone is insufficient for retrieving all the 3D information in the image.

A deep learning-based elevation angle estimation method is proposed in this work. A convolution neural network (CNN) is designed to transfer the acoustic view to another viewpoint called the pseudo front view, which is similar to the perspective in an optical camera. The CNN can solve the non-bijective 2D-3D correspondence problem caused by the unique imaging theory. Instead of using small motion for supervision in real applications, the labels are generated based on a simulator, and neural style transfer is used to generate a realistic dataset from the synthetic images. The contributions of this work are listed as follows:

- We form the missing dimension estimation problem into a pseudo front view depth estimation problem.
- A deep neural network is designed to estimate the front view depth from an acoustic image.
- An acoustic camera simulator is used to help realize front view depth regression.
- To mitigate the sim-real gap, neural style transfer between unpaired data is implemented to generate realistic datasets.
- The proposed methods are evaluated by simulation and real experiments. The implementation of the network, simulator and the simulation datasets are available in our GitHub[1].

The remainder of this paper is organized as follows. In Section II, related works are introduced and compared with the proposed method. Section III discusses the problem formulation. Section IV explains the proposed method for transferring the acoustic view to the front view and the method to generate a realistic dataset. The simulations and evaluations are presented in Section V, followed by the real experiments in Section VI. Finally, the conclusions and future works are presented in Section VII.

II. Related Works

As previously mentioned, sparse 3D reconstruction was first applied to acoustic images. Owing to the difficulty of autonomous feature detection and data association, manually selected features, such as corner points, were used [5], [6]. It has been later proven that AKAZE features have relatively stable performance in acoustic images. Li et al. used AKAZE features for simultaneous localization and mapping (SLAM) in a ship hull environment [2]. Westman et al. achieved SLAM with under-constrained landmarks of AKAZE features [16]. Such methods focus more on robot navigation than on 3D reconstruction. Further, Wang et al. tracked AKAZE features based on optical flow and modelled the terrain as a Gaussian process random field on a tree structure [4]. The method potentially assumes that the terrain is smooth. The computation cost increases tremendously when the feature number increases.

Dense 3D reconstruction is more intuitive for human perception. As previously mentioned, the methods can be roughly categorized into multiple-view stereo and photometric stereo methods. In multiple-view stereo methods, the use of shape-from-space-carving scheme has proved effective, owing to the fixed sensing scope. Aykin et al. applied space carving for small objects. Space that is considered empty is cut, and the remaining area is considered to constitute the object [11]. Guerneve et al. linearized the sonar projection model to an orthogonal projection and applied min-filtering to achieve a carving scheme [12]. Wang et al. used occupancy mapping, which probabilistically carves the space, and proposed an inverse sensor model to apply the method to more general scenes [13]. Although these methods provide convincing results, a large number of view points are necessary.

Photometric stereo methods are another group of methods for dense 3D reconstruction. The acoustic camera can be considered as a light source with a camera at the same position; methods like shape-from-shading have been proven to be valid for the problem. Aykin et al. modeled ultrasound propagation and sonar imaging based on a diffuse reflection assumption [8]. They proved that the physical model can be applied to 3D reconstruction with object contours. Later, Westman et al. utilized a similar scheme to the 3D reconstruction of continuous surface [10]. The method assumes that the range returns in the scene in view monotonically increase or decrease with the elevation angle. Such methods directly estimate the elevation angle of each pixel and ignore the non-bijective 2D-3D correspondence problem.

Recently, modern computer vision techniques have also been applied to this problem. DeBortoli et al. realized pixel-wise elevation angle estimation from a single acoustic image [14] using a CNN. The network is first trained by a simulator. A self-supervised scheme with small motion is then applied to train the network using transfer learning. Although impressive results have been obtained, the pixel-wise elevation angle estimation potentially assumes that one pixel corresponds to only one 3D point. Techniques based on self-supervision by small motion with warping assume photometric consistency in the acoustic image, which is not always true. Westman et al. applied non-light-of-sight (NLOS) techniques to acoustic images [17], [18]. Especially, the method on Fermat path is based on the discontinuities of the image intensity, which can be considered an effective solution to the non-bijective 2D-3D correspondence problem. However, owing to the low SNR, Fermat path detection on acoustic images is error-prone.

In this paper, we also use a CNN-based method for dense 3D reconstruction. The transfer of the acoustic view to the pseudo front view can effectively solve the non-bijective 2D-3D correspondence problem. Instead of using self-supervision by small motion, we apply a fully-supervised method to train the network. Compared to [14], our method directly do regression on front view depth instead of elevation angle map.

[1]https://github.com/sollynoay/A2FNet

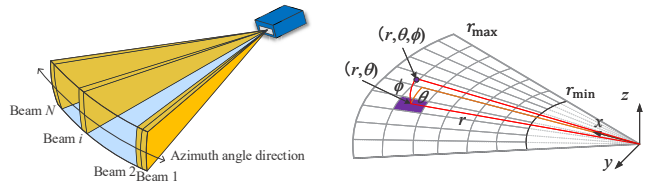


Fig. 2. Acoustic camera model. (a) As a multi-beam forward looking sonar, the sound wave can be separated into beams in azimuth angle direction. (b) A 3D point (r, θ, ϕ) in polar coordinates will be projected to (r, θ) in the imaging plane.

To build the dataset, a simulator is used to generate image labels. The sim-to-real gap is mitigated using the neural style transfer method [19].

III. Problem Formulation

A. Acoustic Camera Model

As shown in Fig. 2(a), the acoustic camera is a multi-beam forward looking sonar, which contains multiple transducers in the azimuth angle direction. For each transducer, it emits 2D fan-shaped sound waves and records the backscattered intensity and time of flight information. A 3D point in the camera coordinate system can be represented as (r, θ, ϕ) using polar coordinates, as shown in Fig. 2(b). The corresponding 2D point in the image coordinate can be represented as (r, θ). This can be considered as a projection to the zero-elevation plane. For each 2D point (r, θ), it may correspond to multiple 3D points with different ϕ as shown in Fig. 1. The compression of ϕ can be considered as an integration of the intensities of all the points with the same (r, θ) coordinates [8], [12]. The intensity of a pixel on the acoustic image I_a can be represented as

$$I_a(r,\theta) = \int_{\phi_{\min}}^{\phi_{\max}} \beta(\phi) V_s(r,\theta,\phi) D_s(r,\theta,\phi) d\phi, \qquad (1)$$

where $\beta(\phi)$ models the beam pattern from each transducer, $V_s(r, \theta, \phi)$ is a measure related to the object reflectively, and $D_s(r, \theta, \phi)$ refers to the cosine of the angle between the beam direction and the surface normal of the object. For each (r, θ), discretizing ϕ as $\{\phi_1, \phi_2, \ldots, \phi_n\}$ and the i-th corresponding backscattered intensities as $I(r, \theta, \phi_i)$ gives

$$I_a(r,\theta) = \sum_{i=1}^{n} I(r,\theta,\phi_i). \qquad (2)$$

B. Acoustic Image Formulation

It is possible to generate synthetic acoustic images based on acoustic camera model. In [20], acoustic images were simulated using a GPU-based method. An echo intensity map and a pulse distance map were rendered to generate the acoustic images. By modeling D_s and V_s based on Lambertian reflectance, assuming the beam pattern β is uniformed distributed along ϕ, and setting the ray strength attenuation according to r, it is possible to acquire the echo intensity map using common ray tracing techniques in simulation environment. In this paper, we assume that the depth map and the intensity map also existed and can be estimated for real images, as shown in Fig. 3. The images are similar to the images generated from an RGB-D camera. In this paper, we refer to the view of the virtual RGB-D camera as the pseudo front view (i.e., front view) and the original view from the acoustic camera as the acoustic view. Denoting the intensity image as I_f, the depth image as D_f, and the acoustic image as I_a, this paper follows Eq. (1) and Eq. (2) based on [8], [12], using the sum of the intensity values along ϕ direction instead of using the average as in [20] to generate I_a from D_f and I_f for synthetic images. Denoting the size of D_f and I_f as $m \times n$ and the size of I_a as $m \times l$, which are determined by the image resolution and sensor scope, the minimum range as $r_{\min}$ and the range resolution as r_{res}, the acoustic image formulation process can be described in Algorithm 1.

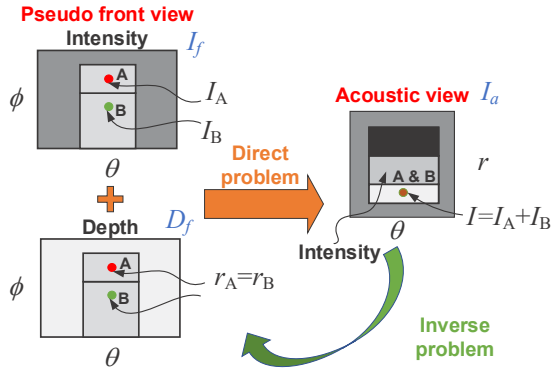


Fig. 3. Acoustic view and pseudo front view. The acoustic image can be generated from an intensity image and a depth image in the pseudo front view. We solve the inverse problem by estimating the depth image in the front view from the acoustic image.

Algorithm 1: Acoustic Image Formation Function

Input: I_f, D_f
Output: I_a

```
Initialize I_a using zero matrix O_{m×l}
for p = 0 to m − 1 do
    for q = 0 to n − 1 do
        r ← D_f(p, q), i ← I_f(p, q)
        d = ⌊(r − r_min)/r_res⌋
        if 0 ≤ d < l then
            I_a(p, d) = I_a(p, d) + i
return I_a
```

C. Inverse Problem

In this work, our aim is to solve the inverse problem of acoustic image formation, specifically, to estimate D_f from I_a. The inverse problem is highly ill-posed, and hard to be solved by model-based methods. It would be possible to solve it with additional priors or cues, such as shadow information, illuminated area position [21], physical properties of ultrasound, and the inconsistency of the intensity in the acoustic image. Instead of using a model-based method, this work directly uses

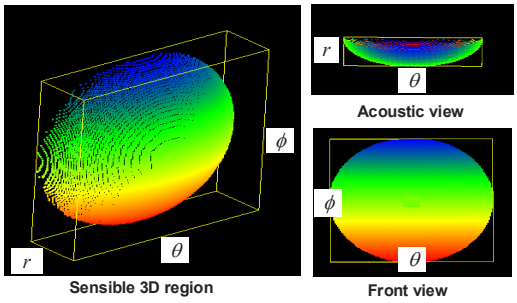


Fig. 4. The 3D model of sensible region when sensing a sphere with its acoustic view and the front view. Color refers to ϕ angle. Front view carries all the information without ambiguity.

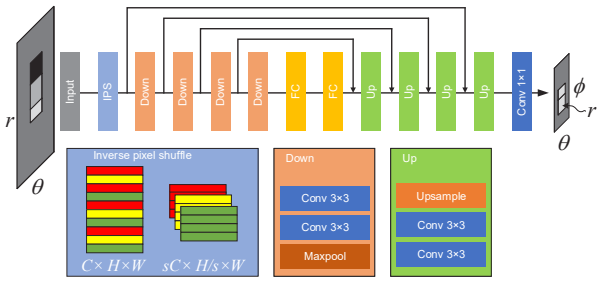


Fig. 5. A2FNet. the input is an acoustic image and the output is the pseudo front view depth map. The IPS module is used to arrange the size of the input tensor without discarding information.

a learning-based method towards the problem. By inputting I_a, we propose a CNN to learn the front view depth D_f in an end-to-end manner. The following section explains how we realize solving the problem with a CNN-based method.

IV. CNN-based Front View Depth Regression

In this study, we formulate the inverse problem as a front view depth regression problem. The ground truth of the front depth image is generated from the simulator. To mitigate the gap between the real and synthetic images, neural style transfer can be used to generate realistic images from synthetic images for training.

A. Acoustic View to Front View

1) Front view depth regression: When sensing a sphere, the sensible 3D model in polar coordinate is shown in Fig. 4. Due to the sonar imaging principle, we acquire an image in acoustic view. Directly estimating elevation angles from acoustic view would cause information loss and ambiguity problem. However, front view image contains full information without ambiguity. Estimating front view depth instead of elevation map would improve the 3D reconstruction result. An acoustic view to front view network (A2FNet) is proposed for front view depth regression.

2) Sensor Characteristics: The information quantity is usually biased in the three dimensions (r, θ, ϕ). The acoustic camera has high resolution in the depth direction r. For instance, in the 3.0 MHz mode of the ARIS EXPLORER 3000, the depth and azimuth angle resolutions are 0.003 m and $0.25°$, respectively. The aperture angle of the azimuth angle is $32°$. At effective sensing ranges between 2 m and 3.536 m, the image has the dimensions of 128 $\times$ 512 along the θ and r directions respectively. Although there is no clearly given value for the elevation angle resolution, we set the resolution to be the same as or lower than the azimuth angle direction resolution. If we set the resolution of the elevation angle to $0.4375°$, the size of the output front view image will be 128 $\times$ 32 in the θ and ϕ dimensions. It is important to arrange the size of the input tensors.

3) Network Architecture: The overall network architecture of the A2FNet is shown in Fig. 5. The first inverse pixel shuffle (IPS) block is designed to resize the image without discarding information considering sensor characteristic. This can be considered the inverse process to pixel shuffle [22]. As illustrated in Fig. 5, the IPS block rearranges the input tensor from $C \times H \times W$ to $sC \times H/s \times W$, where s denotes the scaling factor. The IPS calculation aims to downsample the depth direction of the input image while maintaining the same size with the output image. As an example, the image in Fig. 5 is deshuffled into three channels (red, yellow, and green). An encoder–decoder structure is applied to generate the front-view depth image. For CNN-based 3D reconstruction, it is common to use a 2D CNN encoder to extract the features of the 2D image and a 3D CNN decoder to generate the 3D shape. The encoder and decoder are usually connected by fully-connected (FC) layers. The extracted features are highly related to the position information in the input image. Because we are generating different views, adding fully connected layers will also improve the performance of the network. We use a 2D CNN decoder to generate the depth map. Skip connections are added to avoid the vanishing and exploding gradient problems, similar to U-Net [23]. We use ReLU as the activation function. Batch normalization is added between the activation functions and convolution layers.

4) Loss Function: In the actual application, the regression is implemented on the inverse depth instead of the depth in consideration of the infinity problem. Standard loss functions such as the L1 loss and the reverse Huber (BerHu) loss [24] can be applied for inverse depth regression with good results.

B. Synthetic Dataset Generation

An acoustic camera simulator is used to generate synthetic dataset. We also model D_s and V_s based on Lambertian reflectance and consider each transducer possessing the same beam pattern. We set the attenuation of the ray strength based on the inverse square law. I_f and D_f are generated by ray tracing. The acoustic image I_a is generated following Algorithm 1. For synthetic dataset, I_a and D_f are directly fed to the network for training. For real applications, further domain adaption process is necessary.

C. Domain Adaption for Real Applications

Although the depth label can be acquired in highly constrained experimental settings, such as in a small water tank [25], it is difficult to acquire accurate depth labels in most cases in marine environments. In this study, we train the

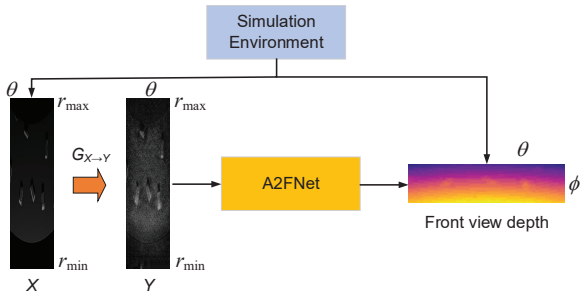


Fig. 6. Training the network with the help of the simulator and neural style transfer in the real application. The synthetic image and depth label are generated by the simulator. The generator $G_{X\to Y}$ in CycleGAN is used to transfer the synthetic image to the realistic image.

network by using the labels generated from the simulation environment, as shown in Fig. 6. To make the synthetic images closer to real images, conditional GAN (cGAN) can be used for domain adaption. In this study, CycleGAN is applied to generate a realistic dataset [19]. The advantage of CycleGAN is that it does not require strict image pairs. We first prepare two groups of images, namely, the synthetic images X and the real images Y. CycleGAN trained two generators, $G_{X\to Y}$ and $G_{Y\to X}$, and two discriminators, D_X and D_Y together. Adversarial loss and cycle consistency loss are used together to train the network. $G_{X\to Y}$ is then used to generate a realistic image from the synthetic image.

V. Simulation Experiment

To evaluate the network, a simulation experiment was carried out. The exact ground truth can be acquired from the simulator. We compared our work with the existing state-of-the-art method.

A. Simulation Dataset

The simulator environment was built using the open-source software Blender. A torch light and a camera with the same pose are used to model the acoustic camera, which serve as the transmitter and receiver, respectively. Noting that the objects symmetric to the zero elevation plane will lead to the same image, in this study, when building datasets, we avoid the ambiguity cases caused by symmetry.

1) Water Tank Dataset: The amount of visual cues present is different in different scenes. For instance, for objects lying on the floor, there are more visual cues, such as the illuminated area or shadows, for solving the problem. We first discuss the case in which there are adequate visual cues. The dataset is comprised of two types of targets: a cylinder lying on the ground and multiple cuboid bricks with random poses on the ground. The cylinder has a radius of 0.1 m and a height of 0.25 m. The size of the cuboid is 0.1 m $\times$ 0.2 m $\times$ 0.08 m. In total, 10,000 images were generated with different poses. The images were separated into training, validation, and test datasets in the respective proportions of 80%, 10%, and 10%.

2) Floating Object Dataset: The simulation experiments with fewer visual cues focused more on simple objects. Similar to the problem setting in ElevateNet [14], the objects float or suspend in water. This may lead to the absence of visual cues such as illuminated areas and shadows. We used a cuboid, cylinder, and a sphere for evaluation. The size of the cuboid is 0.1 m $\times$ 0.1 m $\times$ 0.2 m. The radius and height of the cylinder are 0.1 m and 0.25 m, respectively. The radius of the sphere is 0.3 m. For each object, 5,000 images were generated for a total of 15,000 images. The images were also separated into training, validation, and test datasets in the proportion of 80%, 10%, and 10%.

B. Metrics

The mean average error (MAE) of the depth estimated per pixel and the chamfer distance (CD) between the ground truth and estimated point clouds were used for evaluation. We did not compare the MAE of the depth per pixel from the proposed method with the baseline method because the depth image generated by the latter is incomplete. The CD was used instead to compare the proposed method with the baseline method. The CD can evaluate both the accuracy and the completeness of the estimated results. Because the CD is influenced by the number of points, the results from the baseline method were first transferred to the front depth image with the same resolution as the proposed method. We did not evaluate the pixel-wise elevation angle in this study because one pixel in the acoustic image may correspond to multiple elevation angles. The MAE and CD metrics are given by

$$\text{MAE} = \frac{1}{HW} \times \sum_{i=1}^{H} \sum_{j=1}^{W} |\hat{D}(i,j) - D(i,j)|, \tag{3}$$

$$\text{CD} = \frac{\lambda}{S_1} \sum_{x \in S_1} \min_{y \in S_2} ||x-y||_2^2 + \frac{\lambda}{S_2} \sum_{y \in S_2} \min_{x \in S_1} ||x-y||_2^2, \tag{4}$$

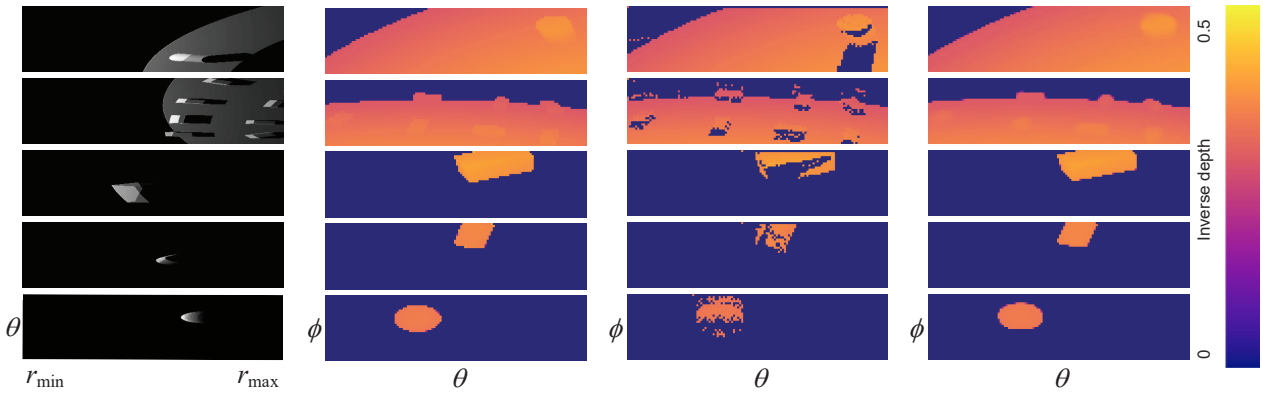
where λ was set to 500.

C. Training

We used the Pytorch framework to implement the proposed approach. An NVIDIA Geforce 1080Ti graphic card was used for training and evaluation. Training was performed on each dataset for 200 epochs. The initial learning rate of 0.001 was decreased by half with a patience of 20 epochs when a plateau was reached. We chose the final model based on the best MAE validation results. The batch size was set to 8 for training and 1 for validation. Adam was used as the optimizer [26]. Training with the floating object dataset took 7.5 hours.

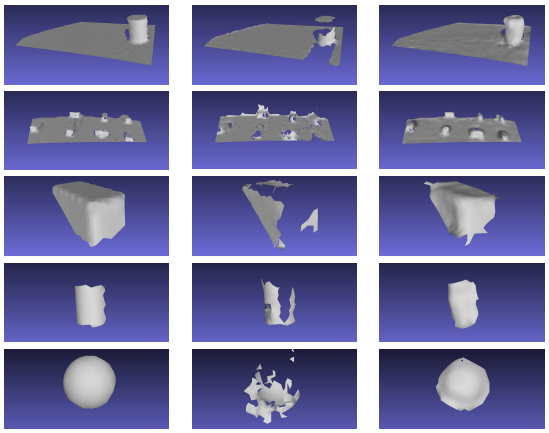
D. Results

To evaluate our method, we compared our results with the state-of-the-art ElevateNet [14]. To implement ElevateNet, we trained the U-Net [23] in a supervised manner. Instead of formulating the problem as a classification problem, we directly performed pixel-wise elevation angle regression. Table I lists the estimation results. The last three columns show the ratios (%) of CD larger than n. Lower values indicate better performance for all the results in the table.



(a) Input (b) Ground truth (c) ElevateNet [14] (d) Ours

Fig. 7. Example of simulation results. The rows correspond to the cylinder with ground, bricks with ground, cuboid, cylinder, and sphere, respectively. Depths larger than $r_{\max}$ were set to infinity, that is, their reciprocals were set to zero.



(a) Ground truth (b) ElevateNet [14] (c) Ours

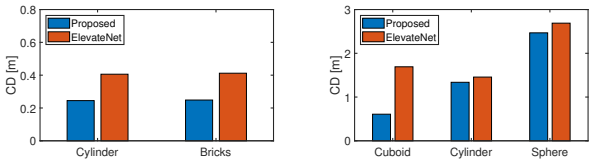
Fig. 8. Example of mesh models from the simulation results. The rows from top to bottom correspond to the cylinder with ground, bricks with ground, cuboid, cylinder, and sphere, respectively.

(a) Ground truth (b) Estimated

Fig. 9. Example of estimation error.

TABLE I
SIMULATION RESULTS

Water Tank Dataset					
Loss	MAE (m)	CD (m)	>0.25	>0.5	>0.75
L1	0.0224	**0.2468**	**10.8**	**7.5**	6.3
L2	0.0258	0.4135	82.1	15.8	**4.4**
Berhu	0.0243	0.3905	20.0	8.0	6.2
L1+no FC	**0.0212**	0.3573	15.0	8.3	6.5
L1+no IPS	0.0532	0.4120	61.7	16.6	6.1
ElevateNet [14]	–	0.4089	56.6	23.6	12.1
Floating Object Dataset					
Loss	MAE (m)	CD (m)	>1.5	>3	>4.5
L1	**0.0331**	**1.4709**	30.6	9.9	3.9
L2	0.0357	1.6346	35.7	13.3	5.7
Berhu	0.0354	1.4940	**24.2**	**8.8**	5.5
L1+no FC	0.0338	3.2626	25.5	10.4	5.4
L1+no IPS	0.0384	2.4017	64.3	24.7	9.4
ElevateNet [14]	–	1.9457	66.0	10.5	**0.3**



(a) Water tank (b) Floating object

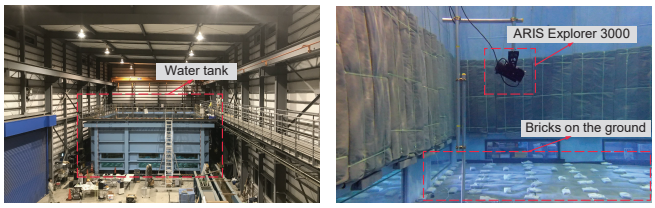
Fig. 10. Results on different targets in the datasets.

1) Loss Functions: Generally, the L1 loss largely outperforms the other losses. L2 loss may sometimes lead to large errors. The Berhu loss falls in between the L1 and L2 losses.

2) Network Architecture: Ablation tests on the network architecture were also performed. We directly downsampled the images to adjust the size of the acoustic images. The channels of the network were modified to match the adjustment. We found that the IPS blocks are extremely important to our method; the removal of these blocks may lead to poor results. The removal of the FC layers may lead to disastrous errors in the floating objects dataset.

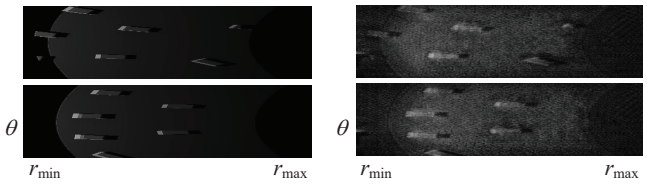
3) Comparison with baseline: The proposed method outperformed the baseline method on both datasets. To visualize the results, we show some examples of the front depth estimation and 3D mesh model results in Fig. 7 and 8. The mesh models were generated using MeshLab. We first estimated the normals of the point cloud and then used the ball pivoting method to generate meshes. The proposed method performs better because it generates more complete results. The CD metric considers both the accuracy and the intersection over union (IoU). The non-bijective projection problem is a common occurrence in acoustic cameras and heavily affects the baseline method. As shown in Fig. 8, our method can generate a more complete 3D model from a single image. In Fig 10, we show the estimation results for each type of target in the datasets. It is proved that the proposed method outperforms the baseline method in 3D reconstruction. In other words, front view depth regression works better than acoustic view elevation angle regression here.

4) Computation Cost: It took approximately 4 ms to estimate one image with the image size of 128 × 512 on a



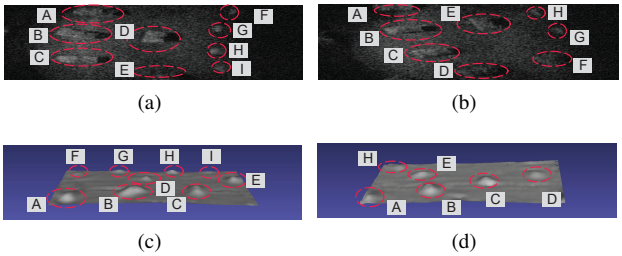
(a) (b)

Fig. 11. Real experiment environment: (a) the water tank at Wakachiku Construction Co., Ltd. (b) ARIS Explorer 3000 and the bricks on the ground.



(a) (b)

Fig. 12. Examples of domain adaption results: (a) input simulation images and (b) output images after domain adaption.



(a) (b) (c) (d)

Fig. 13. Real experiment results: (a) and (b) are the input acoustic images, (c) and (d) are the mesh models generated by the proposed method.

NVIDIA GeForce 1080Ti GPU. Because the frame rate of the acoustic camera is approximately 10 Hz, the proposed method can work in real time.

E. Discussion

For datasets that include the ground, the error is smaller, which is partially because the CD is influenced by the number of points. For floating objects, the point numbers are smaller, which may lead to larger errors. The larger errors are also due to the presence of fewer visual cues, which degrades the estimation accuracy. As shown in Fig. 9, although the shape of the sphere is complete, the sphere appears at a different position along the elevation direction, which may lead to a large error based on our metrics. Such problems rarely occur if there are sufficient visual cues.

VI. Real Experiment

The real experiment was carried out in a water tank at Wakachiku Construction Co., Ltd., as shown in Fig. 11(a). The ARIS Explorer 3000 was operated in 3.0 MHz mode with a resolution of 0.003 m in the range direction. The receiver gain of the sensor was set to 8 dB. The acoustic camera was mounted on a rotator to adjust the pose for data collection as shown in Fig. 11(b). Bricks with random poses were placed on the ground. The size of each brick is approximately 0.1 m $\times$ 0.2 m $\times$ 0.04 m. The minimum and maximum ranges of the acoustic images were set to 2.8 m to 4.312 m. The image size is 128 $\times$ 512.

A. Dataset Generation

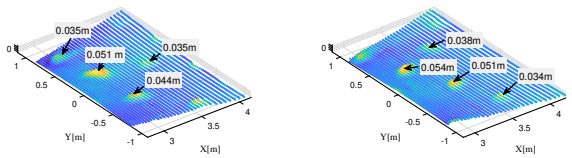
Depth labels can be collected in a constrained underwater environment [25]. However, in a larger-scale environment, the collection of depth labels is much more difficult. It is possible to model a similar scene and train the network based on the synthesized images. However, the intensity images I_f from the simulator are not sufficiently realistic. In this study, CycleGAN was used to transfer the domain to generate realistic images for training. We first collected data in the water tank and modeled a similar scene in the simulator. We chose 293 real images and 239 synthetic images with the same sizes and scales to train the network. The vanilla CycleGAN was applied in this study [19]. We trained the networks for 150 epochs. The learning rate was set to 0.0002 and reduced every 50 epochs. The batch size was set to 1. The synthetic images X were generated by placing brick models with random poses and changing the camera pose. We generated 3,500 images with depth labels and transferred the synthetic images X to realistic images using $G_{X\to Y}$. Examples of the images $G_{X\to Y}(X)$ after domain adaption are shown in Fig. 12. It can be seen that the generated images are very close to the real images. We roughly estimate the SNR as $\bar{y}/\sigma$, where $\bar{y}$ and σ refer to the mean value and standard deviation of the images, respectively. The average SNRs of the 293 real images, the 3,500 synthetic images before domain adaption, and the synthetic images after domain adaption are 1.83, 1.11, and 1.90, respectively. These results indicate that the noise levels after domain adaptation are similar to those in the real images. In the future, more systematic studies will be performed on the image quality of the images after domain adaption. After generating the dataset, we used the same parameters in the simulation experiment to train the A2FNet.

B. Real Experiment Results

After training the network, we input real images from the water tank into the network to evaluate its performance. Figs. 13(a) and 13(b) show examples of the input images, and Figs. 13(c) and 13(d) their corresponding generated meshes. The letters in the images indicate the corresponding bricks. The results prove that the network trained by the dataset works on real images. Most of the complete bricks were successfully reconstructed. On the other hand, some of the bricks on the upper boundaries, such as F and G in Fig. 13(b), were not reconstructed. This is an acceptable result because the bricks on the upper boundaries are incomplete and blurred. To evaluate the estimation results, we compared the heights of the bricks with the ground truth. We manually chose the height at one point on each complete brick, as shown in Fig. 14. The average error of the heights is 0.007 $\pm$ 0.004 m.

VII. Conclusions

In this work, we proposed a novel solution for the missing dimension problem in acoustic cameras by estimating



(a) (b)

Fig. 14. Real experiment results. The color indicates the height information. We manually labeled the heights of the complete bricks.

the depth in the pseudo front view. The proposed method can extract all the information from the acoustic images by considering the non-bijective 2D-3D correspondence. A novel network was proposed for depth regression in the front view. Considering the difficulty of collecting depth labels in an underwater environment, neural style transfer was employed to generate a realistic dataset for training. Both simulation and real experiments were performed to evaluate the proposed method. Our simulator and synthetic dataset are made open source for comparison.

Future works may include more systematic evaluations of complex scenes in real experiments and mounting the camera to an underwater vehicle. The method can be integrated into SLAM and other autonomous systems. We used CNN-based regression to solve the problem in this study, it would be also interesting to test model-based method in the future. Other problems such as ambiguity caused by imaging symmetry would also be considered. In this study, we used neural style transfer to build the dataset. This required both the collection of real images and the generation of synthetic images from a simulator, which may not be sufficiently efficient for field applications, especially for unstructured environments such as seabed. Currently, there is no open-source dataset available. It would take much effort to collect real acoustic images, model similar scene in the simulator, generating synthetic images, and train the domain adaption network. Training the depth estimation network in a self-supervised manner would be desirable for robotics applications. Future work may also include training the proposed method in a self-supervised manner. Test on generalization performance of the network will also be carried out in the future.

REFERENCES

[1] E. Belcher, W. Hanot, and J. Burch, "Dual-frequency identification sonar (didson)," *Proc. IEEE Int. Symp. Underwater Technol.*, pp. 187–192, Apr. 2002.

[2] J. Li, M. Kaess, R. M. Eustice, and M. Johnson-Roberson, "Pose-graph slam using forward-looking sonar," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2330–2337, Jul. 2018.

[3] 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," *J. Field Robot.*, vol. 32, no. 1, pp. 123–151, 2015.

[4] J. Wang, T. Shan, and B. Englot, "Underwater terrain reconstruction from forward-looking sonar imagery," *Proc. IEEE Int. Conf. Robot. Autom.*, pp. 3471–3477, May 2019.

[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] T. A. Huang and M. Kaess, "Incremental data association for acoustic structure from motion," *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 1334–1341, Oct. 2016.

[7] E. Westman, A. Hinduja, and M. Kaess, "Feature-based slam for imaging sonar with under-constrained landmarks," *Proc. IEEE Int. Conf. Robot. Autom.*, pp. 3629–3636, May 2018.

[8] 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, 2016.

[9] H. Cho, B. Kim, and S. Yu, "Auv-based underwater 3-d point cloud generation using acoustic lens-based multibeam sonar," *IEEE J. Ocean. Eng.*, vol. 43, no. 4, pp. 856–872, Oct. 2018.

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

[11] 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.

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

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

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

[15] Y. Ji, S. Kwak, A. Yamashita, and H. Asama, "Acoustic camera-based 3d measurement of underwater objects through automated extraction and association of feature points," *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst.*, pp. 224–230, Sep. 2016.

[16] E. Westman, A. Hinduja, and M. Kaess, "Feature-based slam for imaging sonar with under-constrained landmarks," *Proc. IEEE Int. Conf. Robot. Autom.*, pp. 3629–3636, May 2018.

[17] E. Westman, I. Gkioulekas, and M. Kaess, "A volumetric albedo framework for 3d imaging sonar reconstruction," *Proc. IEEE Int. Conf. Robot. Autom.*, 2020.

[18] E. Westman, I. Gkioulekas, and M. Kaess, "A theory of fermat paths for 3d imaging sonar reconstruction," *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020.

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

[20] R. Cerqueira, T. Trocoli, G. Neves, S. Joyeux, J. Albiez, and L. Oliveira, "A novel gpu-based sonar simulator for real-time applications," *Comput. Graph.*, vol. 68, pp. 66–76, 2017.

[21] Y. Wang, Y. Ji, H. Woo, Y. Tamura, H. Tsuchiya, A. Yamashita, and H. Asama, "Rotation estimation of acoustic camera based on illuminated area in acoustic image," *IFAC-PapersOnLine*, vol. 52, no. 21, pp. 163–168, Sep. 2019.

[22] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1874–1883, 2016.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Med. Image Comput. Comput. Assist. Interv.*, pp. 234–241, 2015.

[24] Z. Fang, X. Chen, Y. Chen, and L. V. Gool, "Towards good practice for cnn-based monocular depth estimation," *Proc. IEEE/CVF Winter Conf. on Appl. Comput. Vis.*, Mar. 2020.

[25] M. Sung, H. Cho, J. Kim, and S. Yu, "Sonar image translation using generative adversarial network for underwater object recognition," *Proc. IEEE Int. Symp. Underwater Technol.*, pp. 1–6, 2019.

[26] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proc. Int. Conf. Learn. Represent.*, Dec. 2014.