IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Acoustic-n-point for Solving 2D Forward Looking Sonar Pose Estimation

Yusheng Wang[1], Yonghoon Ji[2], Hiroshi Tsuchiya[3], Jun Ota[1], Hajime Asama[4], and Atsushi Yamashita[5]

***Abstract*—2D forward looking sonar (FLS) has gained widespread application in underwater robotics research, primarily due to its capacity to produce high-resolution images in diverse aquatic environments. This study deals with pose estimation with given 3D positions and corresponding 2D pixels, which is a fundamental problem for computer vision, denoted as acoustic-n-point (AnP) problem. It is the key part for object pose estimation, extrinsic calibration, localization, and structure from motion (SfM). We propose two methods to acquire a closed-form solution of 5 degree-of-freedom (DoF) pose. The first utilizes a non-approximated model and eliminates the cosine terms. The singular value decomposition (SVD)-based method is proposed and the nullity is discussed. The second method approximates the projection model as a linear system and conducts null space analysis for more accurate solutions. After acquiring the initial 5DoF pose, the last DoF can be acquired by constrained nonlinear least square optimization. We take advantage of both methods by selecting solutions with the smallest reprojection error. Furthermore, we explore the potential for further refinement by utilizing this solution as the initial pose with 6DoF-constrained iterative optimization. Results are evaluated on both simulation and real data.**

***Index Terms*—Marine Robotics, Localization**

I. INTRODUCTION

THE 2D forward looking sonar (FLS) stands as a vital sensor in underwater robotics, specializing in generating clear, high-resolution images unaffected by underwater turbidity or lighting circumstances. They have been already applied to various underwater robotics tasks, such as object detection, mapping, and navigation [1]–[3]. However, basic tasks such as estimating the pose with given 3D positions and corresponding 2D pixels are still open problems. The concept is similar to the perspective-n-point (PnP) of optical perspective cameras, which is one of the fundamental problems in optical-based computer vision.

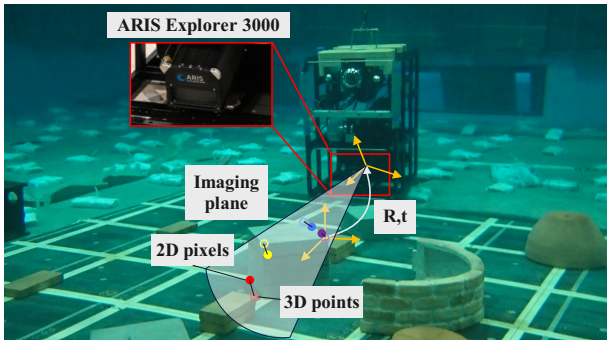


Fig. 1. Pose estimation using 2D-3D point correspondences for FLS. We name the problem as AnP.

A PnP-like algorithm for FLS is important but still absent in the underwater robotics field. We name the above problem as acoustic-n-point (AnP) in this study. Classical PnP cannot be applied to this problem considering the different geometry model. AnP is a fundamental component in various underwater applications, including tracking the pose of underwater targets, localizing underwater robots, and facilitating image registration in structure-from-motion (SfM) and simultaneous location and mapping (SLAM). Practically, artificial blocks like tetrapods [4] and remotely operated vehicles (ROVs) [5] can be exploited as underwater targets with known 3D information. With AnP, it is possible to track their poses. It is also possible to localize the FLS with respect to artificial landmarks or known 3D structures [6], [7]. As shown in Fig. 1, we can localize the ROV with given 2D-3D point correspondences. This capability is particularly crucial when establishing a controlled experimental environment, as it necessitates precise extrinsic calibration. Furthermore, it can facilitate the study of FLS-based SfM and SLAM, inspired by well-known optical camera SfM methods like COLMAP [8] and SLAM methods like ORB-SLAM2 [9], as PnP-based image registration is one of the key components.

Previous works assume the target is planar [10], [11], where the algorithm for 3D target is still an open problem. In this paper, our aim is to give a general solution to estimate the pose between the sonar coordinates and the world coordinates with given 2D-3D correspondences for both planar and 3D scenes. To the best of our knowledge, we are the first to formalize this problem. Inspired by linearization techniques for planar situations [10], [11], there are two approaches towards acquiring a closed-form solution: (i) eliminating the cosine

Manuscript received: September, 20, 2023; Revised November, 25, 2023; Accepted December, 24, 2023.

This paper was recommended for publication by Editor Giuseppe Loianno upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by JSPS KAKENHI Grant Number 23k19993 which supported the work.

[1]Y. Wang, J. Ota are with Research into Artifacts, Center for Engineering, Graduate School of Engineering, The University of Tokyo, Japan. {wang,ota}@race.t.u-tokyo.ac.jp

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

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

[4]H. Asama is with the Department of Precision Engineering, Graduate School of Engineering, The University of Tokyo, Japan. asama@robot.t.u-tokyo.ac.jp

[5]A. Yamashita is with the Department of Human and Engineered Environmental Studies, Graduate School of Frontier Sciences, The University of Tokyo, Japan. yamashita@robot.t.u-tokyo.ac.jp

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

term in the projection model using two equations from one-point correspondence, and (ii) approximating the cosine term to a constant value. Then, the linear system can be solved using linear algebra techniques. Although approach (i) uses a non-approximated model and can acquire very clean solutions, the range of the cosine term is a strong constraint that is ignored during optimization. Based on our observation, when the noise is severe, disastrous errors may occur. On the other hand, approach (ii) may introduce bias to the system, but the solution is robust to noise. This study compares the two solutions for the initial pose estimation. In practice, current imaging sonars still have much noise, and approach (ii) is more reliable in practice. For both approaches, this work utilizes singular value decomposition (SVD) to solve the linear equations. Null space analysis is carried out for both approaches to provide a robust solution. Similarly, only 5DoF can be acquired by linear systems. After acquiring a closed-form 5DoF pose, constrained univariate optimization can acquire the last DoF for both methods. To take advantage of both approaches, an empirical way is to select the solution with the minimum reprojection error, which is proven effective during experiments. Finally, a final refinement of the 6DoF pose can be carried out based on nonlinear iterative optimization using the previously computed pose as the initial value. A constrained iterative optimization is utilized for the final solution. Both simulation and real data verify our method.

II. Related Works

A. FLS Pose Estimation

FLS pose estimation can be separated into relative pose estimation (i.e., motion estimation), and absolute pose estimation. Motion estimation determines the relative pose between two viewpoints. They can be solved analytically [3] or using SLAM [12], [13]. This study aims to directly estimate absolute poses from 2D-3D point correspondences. This can be further classified into estimations based on 2D and 3D targets. Previous works mainly deal with 2D targets. Negahdaripour proposed a method to estimate extrinsic parameters and lens distortion removal simultaneously based on planar targets [14]. An initial pose is required but not discussed. Wang et al. proposed a method to linearize the projection model into weak perspective projection to acquire the initial pose [10] and apply the method to a 2D fiducial marker system [6]. In [11], a non-approximated method was proposed to estimate 6DoF pose based on a known 2D target for opti-acoustic calibration. The non-approximated method in this work is inspired from [11], but works for 3D scenes, and null space analysis is also carried out for accurate solutions. For 3D targets, Brahim et al. used a covariance matrix adaptation evolution strategy algorithm (CMA-ES) to find the FLS location based on 2D-3D correspondences [15]. However, how to acquire a robust initial pose is still an open problem. Park et al. proposed an appearance-based localization method, by minimizing the difference between the simulated image and the real image [16]. It only optimizes 3DoF pose and the computation time may take several seconds. Wang et al. proposed a similar method to estimate 6DoF pose [17]; however, it took hours to build a fine database and discretization may cause inaccuracies.

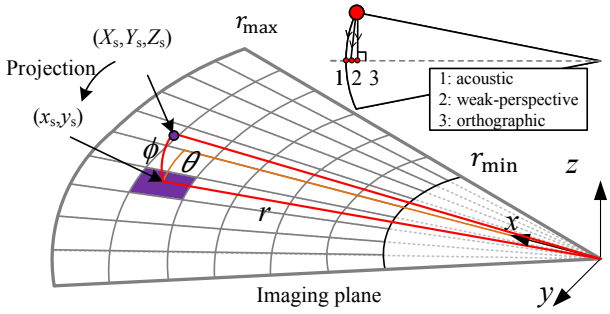


Fig. 2. Image projection model of FLS. Acoustic projection is considered to be the ideal model.

B. PnP

PnP is a long-studied problem in computer vision society. Recent techniques can be categorized into two main types when $n \geq 4$: solving polynomial equations based on cost functions or their first-order optimality conditions, and using linearization techniques in conjunction with null space analysis. The Gröbner basis method is frequently used to solve the polynomial system [18], [19]. However, for noisy data, it may face numerical stability problems. The linearization-based method is another branch, which can usually acquire faster solutions. EPnP is the first solution to provide accurate results in linear complexity [20]. They introduced control points to simplify the computation. SVD with null space analysis was carried out for robust estimations. Zhou et al. proposed a method where null space analysis is based on the rotation matrix instead of control points [21]. Our work also implements null space analysis on the rotation matrix. The benefit is that it can identify the influence of heavy noise and small point numbers according to nullity. Solving AnP may share some similarities with PnP, upon deriving the formulae, the initial 5DoF pose estimation problem can be redefined as an equation solving with linear algebra and eventually lead to a polynomial system solving challenge. On the other hand, 6DoF pose estimation considering the last DoF t_z is a unique problem for AnP.

III. Projection Model

Figure 2 shows the projection model. A 3D point in sonar coordinates can be described as (r, θ, ϕ) using polar coordinates. They can be transferred to Euclidean coordinates based on the following equation.

$$\begin{bmatrix} X_s \\ Y_s \\ Z_s \end{bmatrix} = \begin{bmatrix} r\cos\phi\cos\theta \\ r\cos\phi\sin\theta \\ r\sin\phi \end{bmatrix}. \tag{1}$$

The projected 2D point is denoted as follows.

$$\begin{bmatrix} x_s \\ y_s \end{bmatrix} = \begin{bmatrix} r\cos\theta \\ r\sin\theta \end{bmatrix}. \tag{2}$$

Based on Eqs. (1) and (2), the projection model can be described as follows.

$$\begin{bmatrix} x_s \\ y_s \end{bmatrix} = \begin{bmatrix} \cos^{-1}\phi & 0 & 0 \\ 0 & \cos^{-1}\phi & 0 \end{bmatrix} \begin{bmatrix} X_s \\ Y_s \\ Z_s \end{bmatrix}, \tag{3}$$

where $\cos^{-1}\phi$ varies from points to points. In order to simplify the problem, several works assume the $\cos^{-1}\phi$ to be a constant [10], or directly approximate to 1 [22] since the elevation aperture is usually small, e.g., for standard ARIS EXPLORER 3000 shown in Fig. 1, ϕ ranges [-7°,7°], $\cos^{-1}\phi$ ranges from 1 to 1.0075. By assuming all the $\cos^{-1}\phi$ to 1, the projection model equals the orthographic model for the optical camera. If $\cos^{-1}\phi$ is assumed to be a constant value, the projection model becomes a weak-perspective camera model. Figure 2 also shows different projection models. As a multibeam sonar, multiple fan-shape beam slices are emitted along the azimuth (θ) direction. Taking one beam slice, as shown in the top right, the red point may be projected to the imaging plane based on three paths according to different projection models. The acoustic projection model is considered to be the ideal model according to the spherical sound wave model [23].

IV. ANP PROBLEM

A. Problem Formulation

Denoting $\mathbf{p}_w = [X_w, Y_w, Z_w]^\top$ as a 3D point in world coordinates, it is projected into image coordinates based on the following equation.

$$\begin{bmatrix} x_s \\ y_s \\ 1 \end{bmatrix} = \begin{bmatrix} \cos^{-1}\phi & 0 & 0 & 0 \\ 0 & \cos^{-1}\phi & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{T}_w^s \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}, \tag{4}$$

where $\mathbf{T}_w^s$ denotes the transformation between world coordinates and sonar coordinates as follows.

$$\mathbf{T}_w^s = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}. \tag{5}$$

The aim of this research is to solve $\mathbf{R}$ and $\mathbf{t}$. The following equations can be induced from Eq. (4).

$$x_s = \cos^{-1}\phi(r_{11}X_w + r_{12}Y_w + r_{13}Z_w + t_x), \tag{6}$$

$$y_s = \cos^{-1}\phi(r_{21}X_w + r_{22}Y_w + r_{23}Z_w + t_y). \tag{7}$$

Here, $\cos^{-1}\phi$ is the major difficulty to solve the problem as a linear system. In this study, there are two approaches can be considered: (i) eliminating $\cos^{-1}\phi$ by representing $\cos^{-1}\phi$ based on Eq. (6) and substituting it into Eq. (7) and (ii) considering the range of $\cos^{-1}\phi$, assuming it to be a constant like 1. Both methods have their cons and pros. The following two subsections describe the derivation of these two approaches in detail.

B. Non-approximated Method

If we represent $\cos^{-1}\phi$ from Eq. (6) and substituting it into Eq. (7), $\cos^{-1}\phi$ is eliminated and the system becomes linear. Noting that x_s is always larger than zero, there is no divided-by-zero problem. We may also notice that parameters like t_z are no longer observable. Hence, the problem becomes estimation of $\mathbf{r} = [r_{11}, r_{12}, r_{13}, r_{21}, r_{22}, r_{23}]^\top$ and $\mathbf{t}_s = [t_x, t_y]^\top$. Then, from one-point correspondence, the following equation can be established.

$$\begin{bmatrix} -y_s\mathbf{p}_w^\top & x_s\mathbf{p}_w^\top & -y_s & x_s \end{bmatrix} \begin{bmatrix} \mathbf{r} \\ \mathbf{t}_s \end{bmatrix} = 0. \tag{8}$$

Since there are eight unknowns but one point correspondence only provides one equation, it requires $n \geq 9$ points to solve the problem. Next, we consider decoupling $\mathbf{r}$ and $\mathbf{t}_s$ [19], [21]. Equation (8) can be further written as follows.

$$\mathbf{w}_i\mathbf{t}_s = \mathbf{h}_i\mathbf{r}, \tag{9}$$

where $\mathbf{w}_i = \begin{bmatrix} -y_s & x_s \end{bmatrix}$ and $\mathbf{h}_i = \begin{bmatrix} -y_s\mathbf{p}_w^\top & x_s\mathbf{p}_w^\top \end{bmatrix}$.With n 2D-3D point correspondences, we can build the relationship as follows.

$$\mathbf{W}\mathbf{t}_s = \mathbf{H}\mathbf{r}, \tag{10}$$

where $\mathbf{W} = [\mathbf{w}_1; \mathbf{w}_2; \ldots; \mathbf{w}_n]$ and $\mathbf{H} = [\mathbf{h}_1; \mathbf{h}_2; \ldots; \mathbf{h}_n]$. If we treat $\mathbf{r}$ as given values, the problem becomes a linear optimization of $\mathbf{t}_s$. The solution can be written in a closed form using Moore-Penrose pseudo inverse. It is able to prove that $\mathbf{W}^\top\mathbf{W}$ is invertible when n 2D points are not on the same line since the determinant of which is in Cauchy–Schwarz inequality form [24].

$$\mathbf{t}_s = (\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{H}\mathbf{r}. \tag{11}$$

Then, substituting $\mathbf{t}_s$ into Eq. (10), it is able to form the problem as follows. Theoretically, there are six unknowns, we can solve the problem with $n \geq 7$ points.

$$\mathbf{M}\mathbf{r} = \mathbf{0}, \quad \mathbf{M} = \mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{H} - \mathbf{H}. \tag{12}$$

Theoretically, the solution lies in the null space of $\mathbf{M}$ which can be expressed as follows.

$$\hat{\mathbf{r}} = \sum_{i=1}^{m} \alpha_i \mathbf{v}_i, \tag{13}$$

where $\mathbf{v}_i$ are the m right-singular vectors for the m smallest singular values from SVD of $\mathbf{M}$. The value of m can be considered as the nullity of $\mathbf{M}$. For optical cameras, m usually varies from 1 to 4. However, we discuss m from 1 to 3, since when $m = 4$, the problem is ill-conditioned in this study. We solve the coefficients α_i by the constraints of the rotation matrix.

1) Case 1: When $m = 1$, the solution is determined up to scale as $\hat{\mathbf{r}} = \alpha_1\mathbf{v}_1$. Here, $||\hat{\mathbf{r}}||_2^2 = 2$ and $||\mathbf{v}_1||_2^2 = 1$, we can directly get $\alpha_1 = \pm\sqrt{2}$. The final sign is determined by Eq. (11) that t_x should be larger than zero.

2) Case 2: When $m = 2$, the solution is $\hat{\mathbf{r}} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2$. Denoting the top two rows of $\mathbf{R}$ as $\mathbf{r}_1$=$[r_{11}, r_{12}, r_{13}]^\top$ and $\mathbf{r}_2$=$[r_{21}, r_{22}, r_{23}]^\top$, the property of the rotation matrix offers 3 constraints as follows.

$$f_1 = \hat{\mathbf{r}}_1 \cdot \hat{\mathbf{r}}_1 = 1, f_2 = \hat{\mathbf{r}}_2 \cdot \hat{\mathbf{r}}_2 = 1, f_3 = \hat{\mathbf{r}}_1 \cdot \hat{\mathbf{r}}_2 = 0. \tag{14}$$

With estimated $\mathbf{v}_1$ and $\mathbf{v}_2$, we can write the problem into $\mathbf{F}\mathbf{a} = \mathbf{c}$, where $\mathbf{a} = [\alpha_1^2, \alpha_2^2, \alpha_1\alpha_2]^\top$ and $\mathbf{c} = [1, 1, 0]^\top$. $\mathbf{F}$ belongs to

$\mathbb{R}^{3\times3}$. If we treat $\alpha_1^2, \alpha_2^2, \alpha_1\alpha_2$ as independent unknowns, the problem can be directly solved using the pseudo inverse of $\mathbf{F}$. However, since the solution may be suboptimal, to acquire a more accurate solution, we can form the problem into least square optimization [21].

$$\underset{\alpha_1,\alpha_2}{\operatorname{argmin}} M(\alpha_1,\alpha_2), \quad M(\alpha_1,\alpha_2) = ||\mathbf{Fa}-\mathbf{c}||_2^2. \tag{15}$$

To find the solution, we consider the first-order optimal, the following equations can be acquired.

$$\frac{\partial M}{\partial \alpha_1} = 0, \quad \frac{\partial M}{\partial \alpha_2} = 0. \tag{16}$$

The problem becomes solving the polynomial systems with monomials $[\alpha_1^3, \alpha_1^2\alpha_2, \alpha_1\alpha_2^2, \alpha_1, \alpha_2]$. This is similar to perspective cameras [21]. In order to solve the system, hidden variable resultant can be used [25]. Both polynomials share a common root, which necessitates that the resultant must equal zero. The problem becomes solving the following ninth-order equation.

$$k_9\alpha_2^9 + k_7\alpha_2^7 + k_5\alpha_2^5 + k_3\alpha_2^3 + k_1\alpha_2 = 0. \tag{17}$$

It is noticed that $\alpha_2 = 0$ is one of the solutions, which forms Case 2 into Case 1. We discuss the case when $\alpha_2 \neq 0$. It is obvious that by dividing α_2 on both sides and denoting $\beta = \alpha_2^2$, the problem becomes solving an equation with forth-order. In other words, a closed-form solution can be acquired.

3) Case 3: When $m = 3$, the solution is $\hat{\mathbf{r}} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \alpha_3\mathbf{v}_3$. Again, we write the problem into $\mathbf{Fa} = \mathbf{c}$, where $\mathbf{a} = [\alpha_1^2, \alpha_2^2, \alpha_3^2, \alpha_1\alpha_2, \alpha_2\alpha_3, \alpha_1\alpha_3]^\top$ and $\mathbf{c} = [1,1,0]^\top$. $\mathbf{F}$ belongs to $\mathbb{R}^{3\times6}$. We cannot treat the elements in $\mathbf{a}$ as independent unknowns since there are only three equations available. Considering constraints $g_1 = f_1 - f_2 = 0$ and $f_3 = 0$, we eliminate α_3 by denoting $\alpha_2 = k_1\alpha_1$ and $\alpha_3 = k_2\alpha_1$. The problem becomes $\mathbf{Gk} = \mathbf{0}$. $\mathbf{k} = [k_1^2, k_2^2, k_1k_2, k_1, k_2, 1]^\top$.

$$\underset{k_1,k_2}{\operatorname{argmin}} M(k_1,k_2), \quad M(k_1,k_2) = ||\mathbf{Gk}||_2^2. \tag{18}$$

We also use first-order optimal and the hidden variable resultant to solve the problem. The problem becomes solving an equation with one unknown but ninth degree. We use numerical solvers to acquire the solution. After solving k_1 and k_2, α_1 can be acquired based on f_1 analytically.

4) Planar: For planar case, $Z_w = 0$, $\mathbf{r} = [r_{11}, r_{12}, r_{21}, r_{22}]^\top$. We follow the same pipeline to form the problem into solving $\mathbf{Mr} = \mathbf{0}$. With three constraints f_1, f_2 and f_3, since there are two unknowns r_{13} and r_{23}, only one more variable is possible to be estimated. In other words, we can only discuss $m = 1$ for the planar case. Here, $n \geq 5$ coplanar points are required. Since the result is up to scale, the scale factor α_1 can be written as $1/\lambda$ where λ can be written in a closed form as in Appendix A [11]. For the planar case, the multiple-solution problem arises due to the sign of r_{13} and r_{23} during calculation according to f_1 and f_2. There are four combinations of signs and two can be discarded according to $\det(\hat{\mathbf{R}}) > 0$. To further determine the signs, it requires prior knowledge of the configuration, such as the sign of t_z, or the direction of normal vector of the plane. More discussions can be found in [11].

C. Approximated Method

The non-approximated method can acquire accurate solutions when there is little noise in the 2D point position and enough resolution. However, acoustic images have multi-resolution characteristics and noise is unavoidable. The $\cos^{-1}\phi$ term is a strong constraint that is eliminated for linear systems, which may lead to disastrous errors when point positions are inaccurate. To overcome this problem, we also propose an approximated method which is more robust to noise, but less accurate when there is little error in the point position.

Given the narrow range of variation for $\cos^{-1}\phi$, it is reasonable to assume $\cos^{-1}\phi = 1$. First, we can select one given 3D point as the origin of world coordinates. Denoting the corresponding 2D point as (x_{s0}, y_{s0}), we can directly acquire the following results.

$$t_x = x_{s0}, \quad t_y = y_{s0}. \tag{19}$$

Then, Eqs. (6) and (7) can be written as follows.

$$\begin{bmatrix} X_w & Y_w & Z_w & 0 & 0 & 0 \\ 0 & 0 & 0 & X_w & Y_w & Z_w \end{bmatrix} \mathbf{r} = \begin{bmatrix} x_s - t_x \\ y_s - t_y \end{bmatrix}. \tag{20}$$

With n points, we then build a $6 \times 2n$ matrix $\mathbf{A}$ and a $2n$ vector $\mathbf{b}$ based on Eq. (20). Hence, the problem is formed as

$$\mathbf{Ar} = \mathbf{b}. \tag{21}$$

There are six unknowns, and each point provides 2 equations. Here, $n \geq 3$ points (other than the origin point) are necessary to solve the problem. If $\mathbf{A}$ is full rank, the solution can be directly calculated as,

$$\mathbf{r}^* = \mathbf{A}^\dagger\mathbf{b}, \tag{22}$$

where $\mathbf{A}^\dagger$ is the pseudo inverse of $\mathbf{A}$. In practice, $\mathbf{A}$ is not always full rank and the nullity can be larger than zero. The complete solution is a particular solution adding m right-singular vectors for the m smallest singular values from SVD of $\mathbf{A}$ as follows.

$$\hat{\mathbf{r}} = \mathbf{r}^* + \sum_{i=1}^{m} \alpha_i\mathbf{v}_i, \tag{23}$$

where $\mathbf{r}^*$ is the particular solution and the second sum term is the solution of $\mathbf{Ax} = \mathbf{0}$. The singular value from largest to smallest has the property that $\sigma_1 = \sigma_2 \geq \sigma_3 = \sigma_4 \geq \sigma_5 = \sigma_6$. In other words, when σ_6 is close to zero, σ_5 is also close to zero. It is also noticed that $m = 2$ is enough.

1) Case 1: As aforementioned, if $\mathbf{A}$ is full rank, we can directly get the solution $\mathbf{r}^*$ based on Eq. (22).

2) Case 2: If the dimension of null space is not zero, since $\sigma_5 = \sigma_6$, the next case to be discussed is $m = 2$. The solution can be written as $\hat{\mathbf{r}} = \mathbf{r}^* + \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2$.

The next step is to solve α_1 and α_2 based on f_1, f_2, f_3 in Eq. (14). Similar to Case 2 in non-approximate case, we can write the problem into $\mathbf{Fa} = \mathbf{c}$, where $\mathbf{a} = [\alpha_1^2, \alpha_2^2, \alpha_1\alpha_2]^\top$, $\mathbf{c}$ is based on $\mathbf{r}^*$. We form the problem into Eq. (15), and solve the first-order optimal as Eq. (16). Since the constant terms for the two polynomials are not zero, after using the hidden variable resultant, it cannot be written in terms of Eq. (17).

This directly leads to solving an equation involving a single unknown of a ninth degree. We use numerical solvers to solve the equation.

It is worth mentioning that we did not witness $m > 2$ during experiments. In addition, the next case to be discussed is $m = 4$, there are four unknowns but only three constraints in Eq. (14) which is under-constrained.

3) Planar Case: For planar case, our group previously proposed a method approximating the projection model into a weak perspective [10]. This study tries to form the problem into orthographic projection directly.

$$\begin{bmatrix} X_w & Y_w & 0 & 0 \\ 0 & 0 & X_w & Y_w \end{bmatrix} \mathbf{r} = \begin{bmatrix} x_s - t_x \\ y_s - t_y \end{bmatrix}, \tag{24}$$

where $\mathbf{r} = [r_{11}, r_{12}, r_{21}, r_{22}]^\top$. The singular value from largest to smallest has the property that $\sigma_1 = \sigma_2 \geq \sigma_3 = \sigma_4$. Since the constraints are not enough to solve the problem when $\mathbf{A}$ is not full rank, we do not carry out further discussions. The solution can be acquired by pseudo inverse as Eq. (22). Two more points other than the world coordinates origin are required to solve the problem.

D. Rotation matrix and t_z

After acquiring $\hat{\mathbf{r}}_1$ and $\hat{\mathbf{r}}_2$, $\hat{\mathbf{r}}_3$ can be acquired by $\mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2$. In case the estimated rotation matrix $\hat{\mathbf{R}}$ is not orthogonal, orthogonalization can be carried out using SVD [24]. Assuming $\hat{\mathbf{R}} = \mathbf{U}\Sigma\mathbf{V}^\top$, the closest $\mathbf{R} \in SO(3)$ is $\mathbf{U}\mathbf{V}^\top$. Both approaches eliminate t_z during the estimation process. In order to estimate t_z, one possible method is to utilize the 3D world coordinate origin and the corresponding 2D pixel (x_{s0}, y_{s0}). This is feasible under low-noise and high-resolution conditions. For the approximated method, since we assume Eq. (19), t_z will always be zero. For a more robust solution, we build a least square cost function and use sequential least squares quadratic programming (SLSQP) to solve the optimization problem [26]. For FLS, the aperture angle in the elevation direction is a strong constraint that cannot be ignored. We use constrained iterative optimization to make sure all the 3D points are within the sonar scope.

$$\begin{aligned} &\underset{t_z}{\operatorname{argmin}} \, ||proj(\mathbf{P}, \hat{\mathbf{R}}, \hat{\mathbf{t}}_s, t_z) - \mathbf{S}||_F^2, \\ &\text{s.t.} \quad \phi_{\min} \leq \phi_i \leq \phi_{\max}, \quad \phi_i \in \{\phi_1, \ldots, \phi_n\}, \end{aligned} \tag{25}$$

where F denotes the Frobenius norm, $\mathbf{S} \in \mathbb{R}^{n\times 2}$ denotes the measured 2D pixel positions and ϕ_i are the elevation angles of 3D points in sonar coordinates. $proj(.)$ represents the transform process in Eq. (4).

E. Initial Solution Selection

As mentioned above, there are two approaches to acquiring the initial 5DoF pose. Considering the non-coplanar configuration, for the non-approximated approach, there are three cases, and for the approximated approach, there are two cases. Selecting the proper case as the solution is ad-hoc, for instance, it requires a threshold for singular values to identify the rank of $\mathbf{M}$. Furthermore, selection between approximated and non-approximated methods is also difficult. We noticed that the reprojection error is usually informative. By selecting the solution with the smallest reprojection error, similar to [20], for most situations the best solution among the $3 + 2$ cases can be selected. Denoting the three cases non-approximated cases 1, 2, and 3 as c_{n1}, c_{n2}, and c_{n3}, the two approximated cases as c_{a1} and c_{a2}, the solution selection can be described as follows. Here, $e(.)$ refers to the reprojection error, and $\tilde{c}$ is the selected solution.

$$\tilde{c} = \underset{c \in c_{n1}, c_{n2}, c_{n3}, c_{a1}, c_{a2}}{\operatorname{argmin}} \; e(c). \tag{26}$$

F. 6DoF Constrained Iterative Optimization

The previous steps can acquire an initial estimate of 6DoF pose, denoted as $\hat{\mathbf{R}}_{\text{init}}$ and $\hat{\mathbf{t}}_{\text{init}}$. To acquire more accurate results, a final iterative optimization can be carried out. We use quaternions to represent the rotation matrix for optimization. we transfer $\hat{\mathbf{R}}_{\text{init}}$ into $\hat{\mathbf{q}}_{\text{init}}$ as the initial value for optimization. We also use SLSQP to solve the refined 6DoF pose $\hat{\mathbf{R}}_{\text{fine}}, \hat{\mathbf{t}}_{\text{fine}}$ as follows.

$$\begin{aligned} &\underset{\mathbf{q},\mathbf{t}}{\operatorname{argmin}} \, ||proj(\mathbf{P}, \mathbf{q}, \mathbf{t}) - \mathbf{S}||_F^2, \\ &\text{s.t.} \quad \phi_{\min} \leq \phi_i \leq \phi_{\max}, \quad \phi_i \in \{\phi_1, \ldots, \phi_n\}. \end{aligned} \tag{27}$$

G. Summarization

For clarity, we give the summarization of the computation process in Algorithm 1.

Algorithm 1: AnP algorithm

Input : n 2D points and 3D positions
Output: Pose $\hat{\mathbf{R}}_{\text{fine}}, \hat{\mathbf{t}}_{\text{fine}}$

1 Computing initial 5DoF for the non-approximated cases c_{n1}, c_{n2}, and c_{n3}.
2 Estimating t_z for the non-approximated cases c_{n1}, c_{n2}, and c_{n3} using Eq. (25).
3 Computing initial 5DoF for the approximated cases c_{a1} and c_{a2}.
4 Estimating t_z for the approximated cases c_{a1} and c_{a2} using Eq. (25).
5 Selecting the case with the minimum reprojection error $\tilde{c}$ based on Eq. (26) to get $\hat{\mathbf{R}}_{\text{init}}, \hat{\mathbf{t}}_{\text{init}}$.
6 Refining the 6DoF solution based on Eq. (27) to acquire $\hat{\mathbf{R}}_{\text{fine}}, \hat{\mathbf{t}}_{\text{fine}}$.
7 **return** $\hat{\mathbf{R}}_{\text{fine}}, \hat{\mathbf{t}}_{\text{fine}}$

V. Experiment

A. Simulated Data

We assume the FLS is ARIS EXPLORER 3000. 3D points were randomly generated within a $[1.6, 2.8] \times [-0.6, 0.6] \times [-0.3, 0.3]$ box in the sonar coordinates. We set the first point as the world coordinates origin as $\mathbf{t}^{gt}$ and randomly rotated the points [18]. The process of generating points is completed once n 3D points have been generated within the scope of the sonar. We carried out 300 independent trials for each experiment. We measure the absolute error in degrees

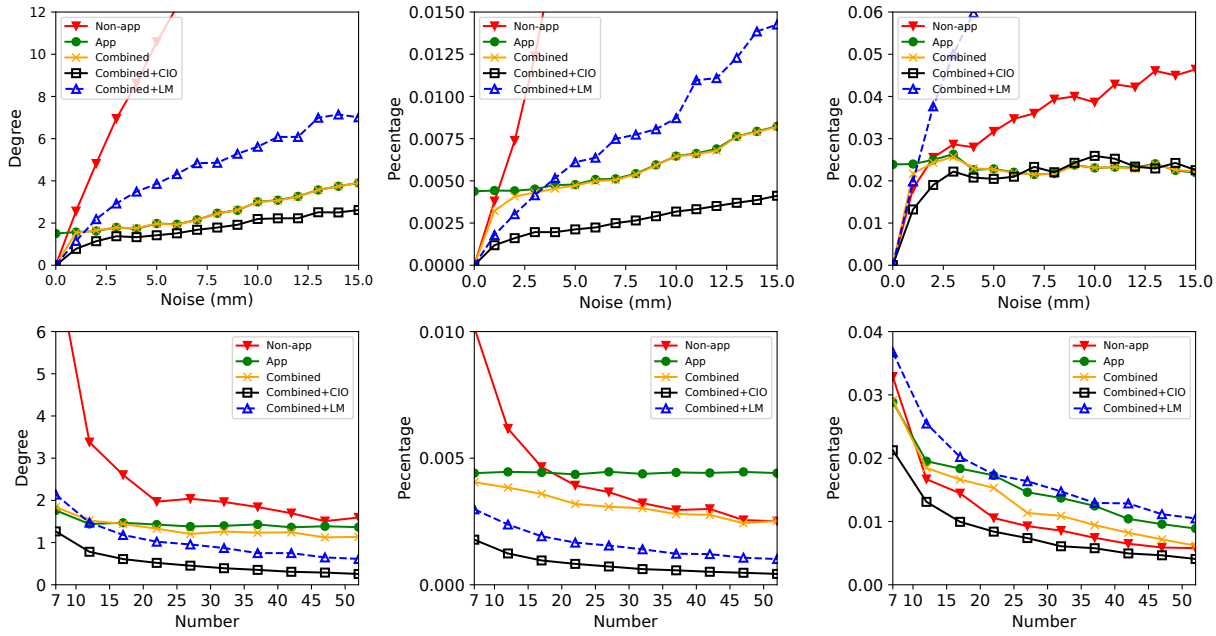


Fig. 3. Non-planar case. From left to right: rotation error, $\mathbf{t}_s$ error and t_z error. The first row: influence of Gaussian noise with 10 points for pose estimation. The second row: influence of point number under noise level 0.002 mm.

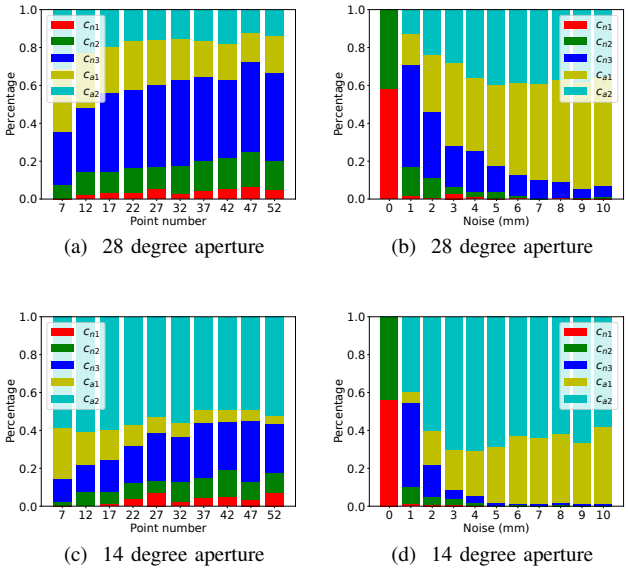
between ground truth $\mathbf{R}^{gt}$ and the estimated rotation matrix $\hat{\mathbf{R}}$, defined as $\max_{k=1}^{3} \arccos(dot(\mathbf{r}_k^{gt}, \hat{\mathbf{r}}_k))/\pi \times 180$. We separate the evaluation of $\hat{\mathbf{t}}_s = [\hat{t}_x, \hat{t}_y]^\top$ and $\hat{t}_z$. The translation errors are defined as $||\hat{\mathbf{t}}_s - \mathbf{t}_s^{gt}||/||\mathbf{t}^{gt}||$ and $|\hat{t}_z - t_z^{gt}|/||\mathbf{t}^{gt}||$. For planar case, we directly evaluate $\hat{\mathbf{t}}$ by $||\hat{\mathbf{t}} - \mathbf{t}^{gt}||/||\mathbf{t}^{gt}||$. We find the roots of the ninth-degree polynomials using NumPy and implement LM and SLSQP using SciPy.

1) Noise and Point Number: We discuss Gaussian noise and point number influences. We first set the elevation aperture angle to standard 14 degrees. Simulation results for non-planar situations are shown in Fig 3. Again, we select solutions according to the reprojection error, Non-app refers to the solution selected among c_{n1}, c_{n2}, and c_{n3}. Similarly, App refers to the solution selected from c_{a1} and c_{a2}, and Combined refers to the solution from all five cases, i.e., $\tilde{c}$. CIO refers to 6DoF-constrained iterative optimization. We first set the point number to 10 and check the influence of noise. The non-approximated method outperforms the approximated methods when the noise level is close to zero, but the error increases drastically when the noise increases. The results for the approximated method are more stable. Solution selection based on reprojection error is effective here, which can take advantage of both methods. For the influence of point number, we checked the low noise case where the standard deviation is 2 mm. It is noticed that the non-approximated method exceeds the approximated method for low-noise (2 mm) and large-point-number cases ($\geq$20) on $\mathbf{t}_s$ estimation. In practice, considering the 3mm for ARIS EXPLORER 3000, noise under 2mm is hard to realize, requiring sub-pixel refinement. The error in t_z is still a problem [6], [10]. When noise increases, the observability is low. Additional information may be required for accurate measurement [6].

2) Elevation Angle Constraints: An ablation test was carried out on the 6DoF CIO, we replaced it with levenberg-marquardt (LM) method as the unconstrained optimization, which was used in [11] for refinement. As shown in Fig 3, obviously, direct LM can lead to poor results even worse than the initial solution when the noise increases. The estimation of t_z may be disastrous. For low-noise situations, CIO also performs much better. This proves the importance of the elevation angle constraints.

3) Solution Selection: Since we refer to reprojection error for solution selection, it is interesting to figure out which case has the smallest reprojection error before refinement. The elevation angle aperture for ARIS can also be modified to 28 degrees using a special lens, we also discussed the aperture angle size. As shown in Fig. 4, we also checked the influence of noise on a 10-point situation. For 28 degrees, if the noise is not severe, cases from the non-approximated method may occupy a substantial portion. However, for the 14-degree case, if noise increases over one pixel, the approximated method may most likely be selected. We also checked low-noise situations when the noise is 2 mm, when the point number increases, the portion of the non-approximated method significantly increases.

4) Planar Case: Results on planar cases are shown in Fig. 5. We set the aperture angle to 14 degrees. The non-approximated method performs well in low-noise and large-point-number situations. We also compared to planar-AnP [10], it is interesting that the weak-perspective-based method is quite robust with a small number of points and a high noise level. It is worth noting that in situations with low noise and a high number of points, there are instances where the non-approximated method yields solutions that are closer to the ground truth, even though the reprojection error may be larger.



(a) 28 degree aperture (b) 28 degree aperture

(c) 14 degree aperture (d) 14 degree aperture

Fig. 4. Case selection. For (a) and (c), the point number is 10. For (b) and (d), the noise level is 0.002 mm.

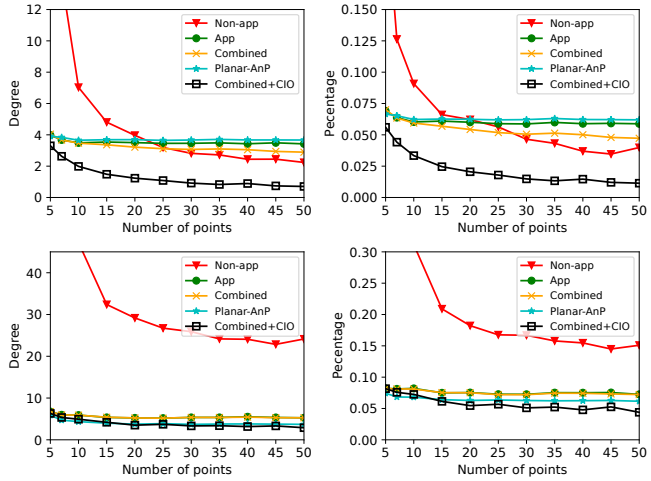


Fig. 5. Planar case. Top row: noise levels are 0.002 mm. Second row: noise levels are 0.01 mm. Left column: rotation error. Right column: translation error.

During evaluation, the sign of t_z is considered to be given considering multi-solutions for planar case.

5) Time Cost: To test on-board performance, we employ our program on Raspberry Pi Model B with a CPU Broadcom BCM2711 @1.5GHz. The computation time is shown in Fig. 6. The current method can ensure 10 Hz performance with less than 300 points. There is still space to improve computation efficiency since we carry out iterative optimization for t_z for all the cases which is time-consuming.

B. Real Data

We also tested the proposed method on real data [17], [27]. We used ARIS Explorer 3000 under 3.0MHz mode with a resolution of 3 mm. The scene setup is the same as Fig. 1. We also compared the result in [17]. Note that the method in [17]

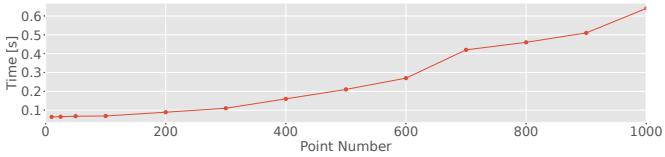


Fig. 6. On-board computation time.

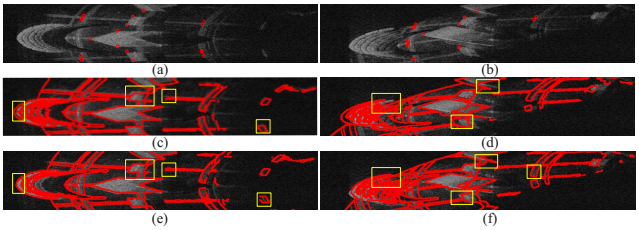


Fig. 7. Real data result. (a) and (b) display the real images, with red circles highlighting corner points for pose estimation. (c) to (f) show contours generated by extracting edges from synthetic images produced by our simulator, based on the estimated poses. (c) illustrates the pose estimated by [17], while (d) represents the pose obtained by rotating the estimate in (c). (e) and (f) depict results from our proposed method. The contours generated by our method exhibit a closer match to the real image, indicating enhanced accuracy in pose estimation.

built a large database with synthetic images and corresponding poses. The pose for the real image was found by finding the closest synthetic image in the database based on edge information. For the fine database, it took 7 hours for computation. To test the proposed algorithm, we manually selected the 2D pixels and acquired the corresponding 3D positions from CAD model. In our experimental environment, numerous readily identifiable corner points, such as line intersections and cube vertices, are present. Figures 7(a) and (b) show the corner points we selected. We calculated the poses based on the proposed method and rendered synthetic images based on them. We used the simulator with multi-path reflection support which can generate acoustic images with correct geometric information [28]. To evaluate the result, we extracted the edge of the synthetic image using Canny detection and overlaid it on the real image. The result in [17] is shown in Fig. 7(c). Figure 7(d) was generated by rotating from the pose estimate in Fig. 7(c). The results from the proposed method are shown in Figs. 7(d) and (f). Qualitatively, the contours from the estimated poses better match the real images. We tested the overall PSNR [29] between the synthetic and real images, as shown in Table I, for method [17] and the proposed method. The estimated poses are also written in Table I for reference. With the poses estimated by the proposed method, there is an increase of +0.11 dB and + 0.67 dB compared to the ones from poses generated with the previous method. Our proposed method demonstrates superior pose estimation results, as evidenced by more accurate contour matches and higher similarity scores. It is worth mentioning that since the corner points were manually selected, it cannot reach sub-pixel level accuracy. The approximated method has a much smaller reprojection error and is chosen as the initial solution. During real experiment, we noticed that if one 2D-3D correspondence

TABLE I
RESULTS IN FIG. 7: PSNR AND ABSOLUTE POSES

	PSNR	Translation [m]			Rotation [deg.]		
	(dB)	x	y	z	$roll$	$pitch$	yaw
(c)	20.50	2.664	0.077	2.005	0.13	51.62	84.79
(d)	20.68	2.638	0.072	2.027	-0.06	51.62	87.17
(e)	19.74	2.664	0.077	2.005	44.91	-28.68	110.65
(f)	20.41	2.622	0.028	2.054	43.11	-24.18	110.87

is wrongly input, the error may be disastrous. Keypoints from smooth objects without obvious corners are also ignored in this study. These insights motivate us towards the valuable future direction of developing automatic and robust methods for associating 2D and 3D data [30].

VI. CONCLUSIONS

In this work, we raise the AnP problem and offer two closed-form solutions for the initial 5DoF pose. It is noticed that though the non-approximated is unbiased, it is sensitive to noise. On the other hand, the approximated method will get biased results when there is no noise but is robust when the noise level increases. After finding the last DoF based on the iterative method, we also found that it is possible to take advantage of both methods by selecting the solution with the smallest reprojection error. Further refinements with nonlinear optimization considering the sonar scope can acquire more accurate results. Future work may include improving the accuracy and the computation efficiency of t_z, automatic and robust 2D-3D data association using learning-based method [30], and applying AnP to SfM and SLAM.

APPENDIX

A. Scale for Non-approximated Method in Planar Case

Based on f_1, f_2, f_3, the following solution can be acquired. The sign of λ is determined by $t_x > 0$.

$$\begin{aligned} \lambda^2 &= \frac{(\hat{r}_{11}^2 + \hat{r}_{12}^2 + \hat{r}_{21}^2 + \hat{r}_{22}^2) + \delta_1\delta_2}{2}, \\ \delta_1 &= ((\hat{r}_{11} - \hat{r}_{22})^2 + (\hat{r}_{12} + \hat{r}_{21})^2)^{\frac{1}{2}}, \\ \delta_2 &= ((\hat{r}_{11} + \hat{r}_{22})^2 + (\hat{r}_{12} - \hat{r}_{21})^2)^{\frac{1}{2}}. \end{aligned} \tag{28}$$

REFERENCES

[1] K. Xie, J. Yang, and K. Qiu, "A dataset with multibeam forward looking sonar for underwater object detection," *Scientific Data*, vol. 9, no. 739, pp. 1–8, 2022.

[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] 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, 2013.

[4] F. Dentale, G. Donnarumma, and E. P. Carratelli, "Numerical wave interaction with tetrapods breakwater," *International Journal of Naval Architecture and Ocean Engineering*, vol. 6, no. 4, pp. 800–812, 2014.

[5] P. Santos, M. Santos, P. Trslic, E. Omerdic, D. Toal, and G. Dooly, "Autonomous tracking system of a moving target for underwater operations of work-class rovs," *Proceedings of the MTS/IEEE conference OCEANS 2021: San Diego – Porto*, pp. 1–6, 2021.

[6] Y. Wang, Y. Ji, D. Liu, *et al.*, "Acmarker: Acoustic camera-based fiducial marker system in underwater environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5018–5025, 2020.

[7] J. Pyo, H. Cho, and S. Yu, "Beam slice-based recognition method for acoustic landmark with multi-beam forward looking sonar," *IEEE Sensors Journal*, vol. 17, no. 21, pp. 7074–7085, 2017.

[8] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," *Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[10] Y. Wang, Y. Ji, H. Woo, *et al.*, "Planar anp: A solution to acoustic-n-point problem on planar target," *Proceedings of the MTS/IEEE conference OCEANS 2020 Singapore*, 2020.

[11] D. Yang, B. He, M. Zhu, and J. Liu, "An extrinsic calibration method with closed-form solution for underwater opti-acoustic imaging system," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 9, pp. 6828–6842, 2020.

[12] 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.

[13] E. Westman, A. Hinduja, and M. Kaess, "Feature-based slam for imaging sonar with under-constrained landmarks," *Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3629–3636, 2018.

[14] S. Negahdaripour, "Calibration of didson forward-scan acoustic video camera," *Proceedings of the MTS/IEEE conference OCEANS 2005*, vol. 2, pp. 1287–1294, 2005.

[15] N. Brahim, D. Guériot, S. Daniel, and B. Solaiman, "3d reconstruction of underwater scenes using didson acoustic sonar image sequences through evolutionary algorithms," *Proceedings of the IEEE conference OCEANS 2011-Spain*, pp. 1–6, 2011.

[16] J. Park and J. Kim, "Robust underwater localization using acoustic image alignment for autonomous intervention systems," *IEEE Access*, vol. 10, pp. 58 447–58 457, 2022.

[17] Y. Wang, Y. Ji, D. Liu, H. Tsuchiya, A. Yamashita, and H. Asama, "Simulator-aided edge-based acoustic camera pose estimation," *Proceedings of the MTS/IEEE conference OCEANS 2022-Chennai*, pp. 1–4, 2022.

[18] Y. Zheng, Y. Kuang, S. Sugimoto, K. Astrom, and M. Okutomi, "Revisiting the pnp problem: A fast, general and optimal solution," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2344–2351, 2013.

[19] L. Kneip, H. Li, and Y. Seo, "Upnp: An optimal o(n) solution to the absolute pose problem with universal applicability," *Proceedings of the 2014 European Conference on Computer Vision (ECCV)*, pp. 127–142, 2014.

[20] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o(n) solution to the pnp problem," *International Journal of Computer Vision*, vol. 81, pp. 155–166, 2009.

[21] L. Zhou and M. Kaess, "An efficient and accurate algorithm for the perspecitve-n-point problem," *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6245–6252, 2019.

[22] 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.

[23] R. J. Urick, *Principles of Underwater Sound*. Peninsula Pub, 1996.

[24] W. Ford, *Numerical Linear Algebra with Applications*. Elsevier, 2015.

[25] D. A. Cox, J. Little, and D. O'Shea, *Using Algebraic Geometry*. Springer, 2015.

[26] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer, 2006.

[27] Y. Wang, Y. Ji, H. Tsuchiya, H. Asama, and A. Yamashita, "Learning pseudo front depth for 2d forward-looking sonar-based multi-view stereo," *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8730–8737, 2022.

[28] Y. Wang, C. Wu, Y. Ji, H. Tsuchiya, H. Asama, and A. Yamashita, "2d forward looking sonar simulation with ground echo modeling," *Proceedings of the 20th International Conference on Ubiquitous Robots (UR)*, pp. 724–729, 2023.

[29] A. Horé and D. Ziou, "Image quality metrics: Psnr vs. ssim," *Proceedings of the 2010 International Conference on Pattern Recognition (ICPR)*, pp. 2366–2369, 2010.

[30] D. Campbell, L. Liu, and S. Gould, "Solving the blind perspective-n-point problem end-to-end with robust differentiable geometric optimization," *Proceedings of the 2020 European Conference on Computer Vision (ECCV)*, pp. 244–261, 2020.