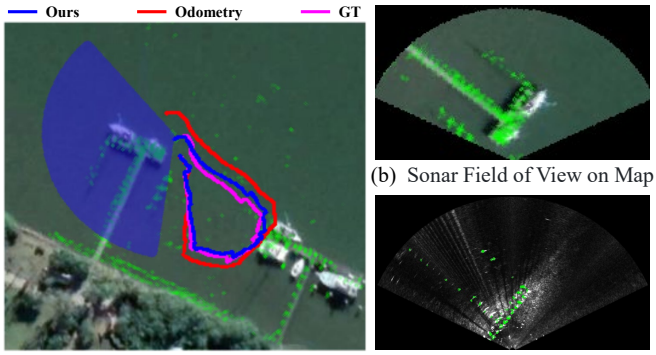
2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

DISO: Direct Imaging Sonar Odometry

Shida Xu[1,2], Kaicheng Zhang[1,2], Ziyang Hong[2], Yuanchang Liu[3], and Sen Wang[1]

***Abstract*— This paper introduces a novel sonar odometry system that estimates the relative spatial transformation between two sonar image frames. Considering the unique challenges, such as low resolution and high noise, of sonar imagery for odometry and Simultaneous Localization and Mapping (SLAM), the proposed Direct Imaging Sonar Odometry (DISO) system is designed to estimate the relative transformation between two sonar frames by minimizing the aggregated sonar intensity errors of points with high intensity gradients. Moreover, DISO is implemented to incorporate a multi-sensor window optimization technique, a data association strategy and an acoustic intensity outlier rejection algorithm for reliability and accuracy. The effectiveness of DISO is evaluated using both simulated and real-world sonar datasets, showing that it outperforms the existing geometric-only method on localization accuracy and achieves state-of-the-art sonar odometry performance. We release the source codes of the DISO implementation to the community. The source code is available at https://github.com/SenseRoboticsLab/DISO.**



(a) Trajectories and reconstructed map points

(b) Sonar Field of View on Map

(c) Sonar image and tracked points

Fig. 1. Result of the proposed DISO system on Aracati2017 dataset [8]. Blue, red and pink lines represent the trajectories from DISO, inertial odometry and the ground truth, respectively. Green points show the reconstructed map of DISO superimposed on a satellite image and a sonar view.

I. Introduction

Simultaneous Localization and Mapping (SLAM) has been extensively investigated in terrestrial environments, where approaches leveraging optical cameras and LiDAR sensors have demonstrated remarkable performance. However, underwater settings pose a unique set of challenges that substantially complicate the application of SLAM. Due to the constrained propagation characteristics of light and suboptimal illumination conditions, optical sensors exhibit limitations in their effective range underwater. This results in a significant deterioration in the performance of SLAM systems that are dependent on these types of sensors.

Imaging sonar offers a distinct advantage over optical sensors when operating in underwater environments. This is primarily because imaging sonar can function effectively over an extended range and is relatively insensitive to variation in water quality. However, the technology faces challenges, such as a low signal-to-noise ratio and limited resolution, which complicate its application in SLAM algorithms.

Prior research in imaging sonar SLAM has predominantly concentrated on feature-based and Iterative Closest Point (ICP) based approaches. For the feature-based approaches, the complications, arising from a low signal-to-noise ratio and limited resolution, pose significant challenges to effective feature matching in sonar imaging. Conventional feature extraction and matching algorithms, such as ORB[1] and SIFT[2], which have demonstrated efficacy in optical images, often underperform when being applied to sonar images. Some studies have adopted alternative strategies. For instance, Huang et al.[3] employ manually-configured features and data association. Additionally, works by Li et al.[4] and Shin et al.[5] utilized A-KAZE features, incorporating anisotropic diffusion techniques for noise reduction. [6] further detects well-constrained A-KAZE features on sonar images. These feature-based methods generally separate the feature matching process from the pose optimization process. This separation may introduce vulnerabilities, as variations in the surroundings of individual pixels attributable to low signal-to-noise ratios could adversely affect the quality of feature matching. As elucidated by Westman et al. [7], the development of a robust and generalized feature detection and matching algorithm, particularly for SLAM, remains an open challenge in the field of imaging sonar.

ICP algorithms are also employed for imaging sonar SLAM after transforming sonar imagery into point cloud representations [9]. However, it is crucial to note that the effectiveness of ICP is contingent upon a satisfactory initial pose estimate to avoid local minima. Furthermore, the SLAM performance heavily depends on the quality of point clouds converted from sonar imagery, amplifying the SLAM algorithm's sensitivity to noise due to the low quality of sonar measurements.

In this paper, we propose a direct imaging sonar odometry (DISO) system which optimizes the relative transformation between sonar image frames by minimizing the overall sonar intensity errors of the points that have high intensity

[1] I-X and Department of Electrical and Electronic Engineering, Imperial College London, UK {s.xu23, k.zhang23, sen.wang}@imperial.ac.uk
[2]School of Engineering and Physical Sciences, Heriot-Watt University, UK {sx2000, kz13, zh9}@hw.ac.uk
[3]Department of Mechanical Engineering, University College London, UK yuanchang.liu@ucl.ac.uk

gradients. The underlying motivations for this work are threefold: (1) The acoustic intensity manifested in sonar imagery is a function of multiple physical attributes of the detected landmark, such as its reflection coefficient, surface normal orientation, distance from the sonar, and the presence of occlusions. Given that these factors are likely to remain relatively stable over short temporal intervals, the consistency in intensity can serve as an informative constraint for optimizing the relative transformation between two sonar images. (2) Compared to the approach of matching individual pixels in sonar imagery, which is highly susceptible to frame-to-frame variability, methods employing the overall acoustic intensity and pixel gradient are more robust to noise interference. Additionally, direct techniques obviate the need for calculating feature descriptors and performing feature matching, processes that are computationally expensive. (3) In contrast to the ICP algorithm, which relies solely on geometric information of an environment, the direct method utilizes extra intensity information. Such supplementary information holds the potential to enhance the accuracy of pose estimation.

In contrast to vision based direct methods [10], [11], [12], [13] that use images with high resolution, elevated signal-to-noise ratios, and abundant color information, sonar systems produce data that is of low resolution, laden with noise, and limited to 1-channel intensity information. Therefore, the proposed DISO has been meticulously designed to address these inherent challenges. DISO employs a multi-sensor window optimization technique and utilizes a data association scheme predicated on direct optimization methods, alongside acoustic intensity outlier rejection algorithms.

The main contributions of this work are:

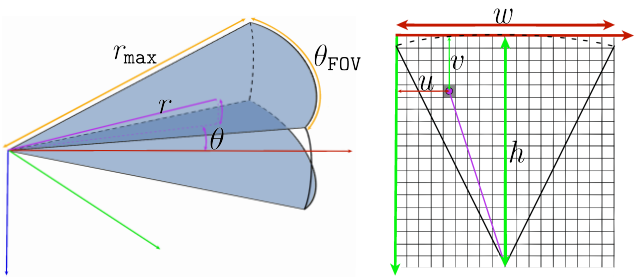
- A novel direct sonar pose optimization algorithm designed to optimize the relative transformation between two sonar frames by minimizing the aggregated sonar intensity errors of points characterized by high intensity gradients.
- Design and implementation of a full imaging sonar odometry system which addresses the unique challenges inherent in sonar imagery. It incorporates a multi-sensor window optimization technique and utilizes a data association strategy based on direct optimization.

Experimental evaluation using both simulation and real-world public datasets [8] show that the proposed DISO outperforms existing method based on ICP algorithms, and achieves state-of-the-art sonar odoemtry performance. Our source codes and simulation data will be released to the community. Given the scarcity of publicly available solutions for imaging sonar odometry or SLAM, we hope this work will benefit the research community and catalyze future research in this domain.

II. Methodology

A. Sonar Imaging Formation and Coordinates

An imaging sonar sensor transmits acoustic pulses and subsequently receives their reflected echoes. The received



(a) Imaging sonar model (b) Cartesian image frame

Fig. 2. Imaging sonar model and its Cartesian image frame.

data that can be decoded for measurements across distinct azimuth angles and ranges is often encapsulated as a Polar image whose x and y axes represent the azimuth angle θ and range r, respectively. As shown in Fig.2(a), for a given azimuth angle θ and a specified range r, the measurement corresponds to the aggregate of multiple acoustic reflections from different elevation angles ϕ along the vertical arc. Notably, the elevation information is lost during this acquisition process.

For a point ${}_{\mathtt{P}}\mathbf{p} \doteq [r, \theta]^T$ in the 2D Polar frame, its Cartesian position is defined as ${}_{\mathtt{C}}\mathbf{p} \doteq [x, y]^T = [r\cos\theta, r\sin\theta]^T$. Given the width w and height h of the Cartesian image as illustrated in Fig. 2(b), its pixel location in the Cartesian image ${}_{\mathtt{I}}\mathbf{p} \doteq [u, v]^T$ can be calculated with a 2D similarity transformation $\mathbf{S}_{\mathtt{IC}}$:

$$ {}_{\mathtt{I}}\mathbf{p} = \mathbf{S}_{\mathtt{IC}}{}_{\mathtt{C}}\mathbf{p} = \begin{bmatrix} s\ \cos\omega & -s\ \sin\omega & t_{\mathtt{x}} \\ s\ \sin\omega & s\ \cos\omega & t_{\mathtt{y}} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{1} $$

where $s \doteq \frac{h}{r}$ is the scale, ω is the rotation angle, $t_{\mathtt{x}} = \frac{w}{2}$ and $t_{\mathtt{y}} = \frac{h}{2}$ are the translation along x and y axis respectively.

B. Direct Sonar Pose Optimization

1) Residual Definition: Given a set of $\mathcal{N}$ keypoints $\mathcal{K} \doteq \{{}_{\mathtt{I_r}}\mathbf{p}^i\}_{i\in\mathcal{N}}$ on the reference sonar image, direct sonar pose optimization is formulated to estimate the relative transformation $\mathbf{T}_{\mathtt{S_c S_r}}$ between the reference sonar image frame $\mathtt{S_r}$ and the current sonar image frame $\mathtt{S_c}$ by minimizing the residuals of all the keypoints' acoustic intensities:

$$ E_{\mathtt{I}} \doteq \sum_{{}_{\mathtt{I}}\mathbf{p}^i \in \mathcal{K}} \| I_{\mathtt{c}}({}_{\mathtt{I_c}}\mathbf{p}^i) - I_{\mathtt{r}}({}_{\mathtt{I_r}}\mathbf{p}^i) \|^2 \tag{2} $$

where $I_{\mathtt{r}}$ and $I_{\mathtt{c}}$ are the intensities on the reference and the current sonar images at the pixel locations ${}_{\mathtt{I_r}}\mathbf{p}^i$ and ${}_{\mathtt{I_c}}\mathbf{p}^i$ respectively. ${}_{\mathtt{I_c}}\mathbf{p}^i$ denotes the pixel location on the current image that corresponds to ${}_{\mathtt{I}}\mathbf{p}^i$:

$$ {}_{\mathtt{I_c}}\mathbf{p}^i \doteq \mathbf{S}_{\mathtt{IC}}\Pi\left(\mathbf{T}_{\mathtt{S_c S_r}}\mathbf{p}^i\right) $$

where $\Pi(\cdot) : \mathbb{R}^3 \mapsto \mathbb{R}^2$ and $\Pi^{-1}(\cdot) : \mathbb{R}^2 \mapsto \mathbb{R}^3$ are project and reprojection functions for a 3D point to and from a Cartesian image assuming zero elevation, and ${}_{\mathtt{S_r}}\mathbf{p}^i \doteq \Pi^{-1}(\mathbf{S}_{\mathtt{IC}}^{-1}{}_{\mathtt{I_r}}\mathbf{p}^i)$. Because of the ambiguity of elevation, we approximate the z of the 3D point to zero. Therefore, the project function keeps the x y components of the 3D point.

2) Jacobian Definition: To solve the minimization problem of (2) with a gradient descent method, the Jacobian matrix of the residual function $\mathbf{r}_{\mathtt{I}} = I_{\mathtt{c}}({}_{\mathtt{I_c}}\mathbf{p}^i) - I_{\mathtt{r}}({}_{I_{\mathtt{r}}}\mathbf{p}^i)$ is needed. Specifically, it can be formulated with respect to the rotation $\mathbf{R}_{\mathtt{S_cS_r}}$ and translation ${}_{\mathtt{S_c}}\mathbf{p}_{\mathtt{S_cS_r}}$ of $\mathbf{T}_{\mathtt{S_cS_r}}$:

$$\mathbf{J}_{\mathtt{I}} = \frac{\partial \mathbf{r}_{\mathtt{I}}}{\partial \mathbf{T}_{\mathtt{S_cS_r}}} = [\frac{\partial \mathbf{r}_{\mathtt{I}}}{\partial \mathbf{R}_{\mathtt{S_cS_r}}}, \frac{\partial \mathbf{r}_{\mathtt{I}}}{\partial {}_{\mathtt{S_c}}\mathbf{p}_{\mathtt{S_cS_r}}}] \tag{3}$$

Since $\mathbf{R}_{\mathtt{S_cS_r}}$ is on the Special Orthogonal Group SO(3), a retraction operation is employed during the optimization process. Due to the page limit, we directly give the Jacobian matrices here:

$$\begin{aligned}\frac{\partial \mathbf{r}_{\mathtt{I}}}{\partial \mathbf{R}_{\mathtt{S_cS_r}}} &= -\frac{\partial I_{\mathtt{c}}({}_{\mathtt{I_c}}\mathbf{p})}{\partial {}_{\mathtt{I_c}}\mathbf{p}} s\ \mathbf{R}_{\mathtt{IC}} \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix} [\mathbf{R}_{\mathtt{S_cS_r}}\ {}_{\mathtt{S_r}}\mathbf{p}^i]^{\wedge} \\ \frac{\partial \mathbf{r}_{\mathtt{I}}}{\partial {}_{\mathtt{S_c}}\mathbf{p}_{\mathtt{S_cS_r}}} &= \frac{\partial I_{\mathtt{c}}({}_{\mathtt{I_c}}\mathbf{p})}{\partial {}_{\mathtt{I_c}}\mathbf{p}} \mathbf{R}_{\mathtt{IC}} \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix}\end{aligned} \tag{4}$$

where $\frac{\partial I_{\mathtt{c}}({}_{\mathtt{I_c}}\mathbf{p})}{\partial {}_{\mathtt{I_c}}\mathbf{p}}$ stands for the intensity gradient, $\mathbf{R}_{\mathtt{IC}}$ is the rotation matrix of $\mathbf{S}_{\mathtt{IC}}$, and $\wedge$ stands for the skew-matrix conversion operation.

C. Window Optimization

In window optimization, data association is assumed established (see Section II-D.5 on data association). Instead of minimizing the acoustic intensity error, re-projection error is minimized in the window optimization. Specifically, given a set of 3D landmark position $\mathcal{L} \doteq \{{}_{\mathtt{S_0}}\mathbf{l}^i\}_{i\in\mathcal{N}}$ with respect to the first sonar frame $\mathtt{S}_0$, corresponding 2D pixel locations $\mathcal{K} \doteq \{{}_{\mathtt{I_j}}\mathbf{p}^i\}_{i\in\mathcal{N},\mathtt{j}\in\mathcal{M}}$ in the set of frames $\mathcal{M}$, and the frame poses $\mathcal{F} \doteq \{\mathbf{T}_{\mathtt{S_0S_j}}\}_{\mathtt{j}\in\mathcal{M}}$, the window optimization is to minimize the below reprojection error

$$E_{\mathtt{P}} \doteq \sum_{i\in\mathcal{N},\mathtt{j}\in\mathcal{M}} \|\mathbf{S}_{\mathtt{IC}}\Pi(\mathbf{T}^{-1}_{\mathtt{S_0S_j}}{}_{\mathtt{S_0}}\mathbf{l}^i) - {}_{\mathtt{I_j}}\mathbf{p}^i\|^2 \tag{5}$$

Notably, the initial pose within the window, along with its corresponding landmark, is held fixed as a reference frame. Subsequent poses and landmarks are then optimized through computational algorithms.

When odometry measurement $\hat{\mathbf{T}}_{\mathtt{S_jS_{j-1}}}$ is available, it can be also added into the optimization with an odometry rotation residual defined as:

$$E_{\mathtt{O_R}} \doteq \sum_{\mathtt{j}\in\mathcal{M}} \|\mathbf{R}_{\mathtt{S_jS_{j-1}}} \boxminus \hat{\mathbf{R}}_{\mathtt{S_jS_{j-1}}}\|^2 \tag{6}$$

where $\boxminus : \mathrm{SO}(3) \times \mathrm{SO}(3) \to \mathfrak{so}(3)$ denotes the subtraction operation in SO(3). The logarithmic map is subsequently employed to map the result into $\mathbb{R}^3$, serving as a quantification of rotation error.

Its translation residual is obtained as

$$E_{\mathtt{O_P}} \doteq \sum_{\mathtt{j}\in\mathcal{M}} \|{}_{\mathtt{S_j}}\mathbf{p}_{\mathtt{S_jS_{j-1}}} - \mathbf{R}_{\mathtt{S_jS_{j-1}}}\hat{\mathbf{R}}^T_{\mathtt{S_jS_{j-1}}}{}_{\mathtt{S_j}}\hat{\mathbf{p}}_{\mathtt{S_jS_{j-1}}}\|^2 \tag{7}$$

D. System Implementation

We introduce the implementation of the proposed DSIO system. After providing an overview, we will focus on the main technical implementations related to frame-to-frame tracking, frame-to-window tracking and window optimization.

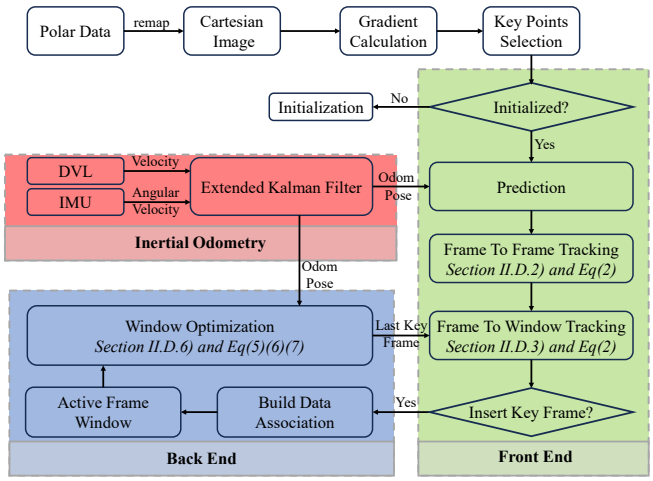


Fig. 3. System overview

*1) **Overview**:* The system overview is illustrated in Fig. 3. It can be divided into 4 main parts: data pre-processing, inertial odometry, front-end and back-end: (a) In the data pre-processing, the system initially transforms raw Polar data to a Cartesian image representation, as delineated in Section II-A. Subsequently, the Sobel operator is employed to rapidly compute the intensity gradient across the sonar image. Locations whose gradient values exceed a pre-defined threshold are identified as preliminary candidates for keypoints. According to their respective gradient magnitudes, the top N candidates are kept. To ensure a balanced distribution of keypoints across the image, a grid-based filter is implemented. Within each grid cell, the point with the highest gradient magnitude is selected as a keypoint. The system initializes the initial set of landmarks using the first sonar frame. (b) The inertial odometry estimates relative motion $\hat{\mathbf{T}}_{\mathtt{S_jS_{j-1}}}$ between two consecutive sonar frames by fusing measurements of a Doppler Velocity Log (DVL) and an Inertial Measurement Unit (IMU) using an extended Kalman filter [14]. This odometry measurement is synchronized with the sonar image by timestamp. (c) In the front-end, the odometry measurements serve as an initial approximation for the direct sonar optimization in the Frame-to-Frame Tracking. Moreover, for instances where the sonar image quality is compromised, evidenced by an insufficient gradient leading to a paucity of keypoints, the front-end employs the odometry prediction to continue pose tracking. (d) In the back-end stage, window optimization is performed across the keyframes and the landmarks with the odometry measurements added as pose constraints between consecutive sonar frames. We implement a multi-threaded architecture for the front-end and the back-end, similar to visual SLAM systems.

*2) **Frame-to-Frame Tracking**:* Frame-to-Frame Tracking estimates the relative transformation between consecutive sonar frames through the direct sonar pose optimization in Section II-B. Since direct method tends to be sensitive to low sonar frame rates, a multi-scale pyramid strategy is introduced, transitioning from coarse to fine scales, and the

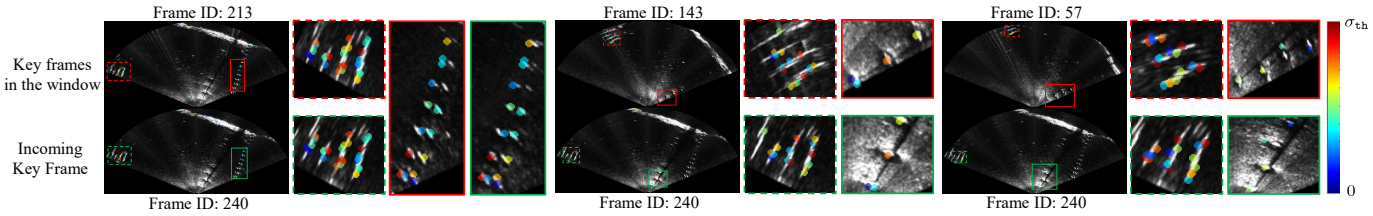


Fig. 4. Data association between the incoming keyframe and existing landmarks in the active window. Points of the same color in both the incoming and existing keyframes indicate that a landmark is being tracked, and a data association is established. The color of each point represents the error in acoustic intensity. The threshold for this error is denoted by σ_{th}.

inertial odometry is incorporated as an initial guess. The outcome of the Frame-to-Frame Tracking serves as an initial estimate for the Frame-to-Window tracking to refine.

*3) **Frame-to-Window Tracking**:* Pose estimates from the Frame-to-Frame Tracking can drift quickly due to its frame-to-frame nature. Therefore, a Frame-to-Window Tracking is introduced to estimate the relative transformation between the current frame and last keyframe within an active optimization window using the direct sonar pose optimization in Section II-B. Meanwhile, the results derived from the Frame-to-Frame Tracking are employed as the initial values for the Frame-to-Window Tracking optimization. Additionally, outlier landmarks are rejected by limiting the maximum intensity error defined in Eq(2). This signifies that the inlier landmarks have been successfully tracked in the current frame.

*4) **Keyframe Selection**:* The keyframe selection decision is made upon the below two primary criteria:

- The temporal interval to last keyframe exceeds a predetermined time threshold;
- The inlier quantity fails below a specified threshold.

The first criterion is designed to ensure a homogeneous distribution of keyframes within the active window, while also mitigating excessive drift in odometry measurements. Conversely, the second is established to ensure that the selected keyframe is adequately informative for the optimization process. This measure simultaneously augments the robustness of the tracking mechanism by mitigating the risk of utilizing an insufficient number of points for direct optimization which lead the optimization dominated by noise.

*5) **Data Association**:* Once a frame is selected as a keyframe, its points are associated with the existing landmarks that are visible in the nearby keyframes. As shown in Fig. 4, the keyframes within the active window exhibit substantial photometric differences and significant relative motions between them. This is challenging for direct method because it is prone to converging to local minima. To mitigate this issue, we utilize the pose estimates from the Frame-to-Window Tracking algorithm as initial values to perform the direct optimization between the incoming keyframe and all the keyframes in the active window to establish data association based on acoustic intensity error. The point pairs whose intensity errors fall below a threshold are considered successfully associated and will be used in the following

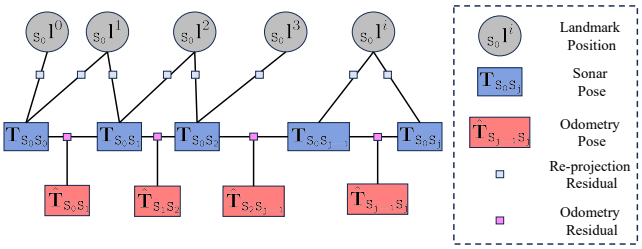
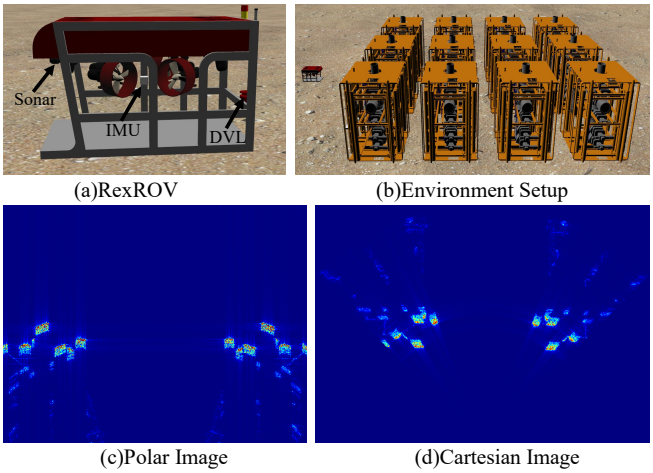


Fig. 5. Factor graph of window optimization.

Window Optimization step for reprojection error calculation. Three examples are given in Fig. 4. Thanks to the introduced data association strategy and the high-quality initial estimates from the front-end, good data association is achieved, even in the presence of substantial spatial disparities between the keyframes.

*6) **Window Optimization**:* Similar to bundle adjustment in visual odometry/SLAM algorithms, the window optimization in Section II-C jointly optimizes the landmark locations and the sonar frame poses through minimizing the acoustic re-projection errors in (5). Its factor graph is shown in Fig. 5.

The choice of re-projection error over intensity error is primarily informed by two considerations. Firstly, the intensity values in sonar imagery are significantly influenced by several factors, not limited to the physical characteristics of the landmark such as reflection coefficients and normal orientation. The amplitude of the acoustic signal experiences attenuation as a function of the spatial distance between the sonar system and the designated landmark, leading to variations in the power of the returned acoustic echoes. Given that the window may span a considerable temporal duration and encompass substantial motion, the intensity associated with a particular landmark may undergo significant fluctuations. Consequently, the consistency of landmark intensity cannot be reliably maintained within the optimization window. Secondly, utilizing re-projection error obviates the need for resampling the image during the optimization process. This computational efficiency is particularly advantageous when optimizations are conducted across multiple frames. Consequently, it enables the incorporation of a greater number of keyframes into the window. This, in turn, allows us to leverage additional information, thereby reducing drifts.



(a)RexROV (b)Environment Setup

(c)Polar Image (d)Cartesian Image

Fig. 6. Simulation setup and simulated sonar data

TABLE I
PERFORMANCE COMPARISON IN SIMULATION

	Translation (%)			Rotation (Degree Per Meter)		
	Ours DISO	BlueROV SLAM*	Odometry	Ours DISO	BlueROV SLAM*	Odometry
Seq1	**3.40**	10.00	6.15	**0.276**	0.44	0.43
Seq2	**4.29**	12.18	14.48	**0.41**	1.29	1.29
Seq3	**4.18**	9.37	8.47	**0.488**	0.89	0.68

III. EXPERIMENTAL EVALUATION

We employ the metrics and the toolset proposed in [15] to evaluate the proposed sonar odometry. BlueROV SLAM [9], one of the very few open-source implementations of imaging sonar SLAM, is selected as a competing method. It employs the ICP algorithm and pose graph optimization. Given that our proposed technique focuses solely on odometry without a loop-closure module, the loop-closure functionality of the BlueROV SLAM is deactivated for a fair comparative evaluation.

A. Simulation Evaluation

1) Simulation Setup: We perform simulation validation first since it provides accurate underwater ground truth for evaluation. The DAVE Aquatic Virtual Environment (DAVE) [16], an open-source platform for underwater simulation, is used. A ray model based multi-beam sonar simulation [17], a DVl, and a IMU are integrated into the RexROV model as shown in Fig. 6(a). A number of underwater structures is span on the sea bottom. The RexROV is controlled to move around the structure for collecting three sequences of the sonar, DVL and IMU data. An example of a raw Polar and its converted Cartesian images are given in Fig. 6(c) and (d). In addition, for the fairness of comparison, we use the same odometry data for the proposed method and the BlueROV SLAM[9]. The odometry data is generated by using an open-source EKF implementation [18]. The EKF odometry takes input from DVL and IMU and outputs pose estimates at 30Hz. Notably, BlueROV SLAM [9] requires Occulus sonar message data. We convert the sonar simulation data to the Occulus sonar message data type for it.

TABLE II
PERFORMANCE COMPARISON ON ARACATI2017 REAL DATA

	Translation (%)			Rotation (Degree Per Meter)		
	Ours DISO	BlueROV SLAM*	Odometry	Ours DISO	BlueROV SLAM*	Odometry
Seq1	**5.91**	11.64	17.97	**0.17**	0.19	NaN
Seq2	**9.08**	10.77	16.63	**0.19**	0.46	NaN
Seq3	**7.28**	19.37	13.83	**0.16**	0.22	NaN
Overall	**8.69**	16.25	17.69	**0.25**	0.32	NaN

The root-mean-square error (RMSE) of relative translation and rotation error [19] is delineated in Table I. Comparative evaluation reveals that the proposed method exhibits superior performance across all sequences and outperforms the BlueROV SLAM approach, with marked reductions in both translational and rotational error metrics. This demonstrates the benefits of the proposed direct sonar method against the ICP based technique. Fig. 7 shows the trajectories corresponding to the three sequences. It is evident that the trajectories estimated by the proposed DISO most closely approximate the groundtruth.

B. Real-World Evaluation

The real-world evaluation was carried out using the Aracati 2017 dataset [8]. This dataset was acquired in the marina of the Yacht Club of Rio Grande, Brazil, utilizing a LBV 300-5 ROV. The ROV was outfitted with a Blue View p900-130 sonar for data collection. To facilitate the ROV's navigation in the area, it was tethered to a floating platform equipped with a Differential Global Positioning System (DGPS) and a magnetic compass. Ground truth data was obtained through the DGPS and compass. Additionally, the vehicle's odometry was computed utilizing velocity commands in conjunction with data from the magnetic compass. Notably, this dataset only provides cropped Cartesian images which cause ambiguity of the scale s described in (1). So a similarity transformation alignment provided by [15] is conducted to the estimated trajectory to align the trajectory with the ground truth. BlueROV SLAM [9] requires Occulus sonar ping data as input. However, the Aracati2017 dataset only provides Cartesian images. So we convert the Cartesian images to point clouds and provide them to BlueROV SLAM. For the purpose of easy visualization, the entire sequence with a total duration of 44 minutes was partitioned into three sub-sequences, each of which has an approximately equivalent duration.

As shown in Table II, the proposed method outperforms others in all the sequences in terms of both translation and rotation errors. Notably, we do not evaluate the rotation errors of odometry because the groundtruth and odometry rotations share the identical readings from from the compass. Fig. 8 shows the estimated trajectories of the three sequences. Similar to the simulation outcome, the trajectories of the

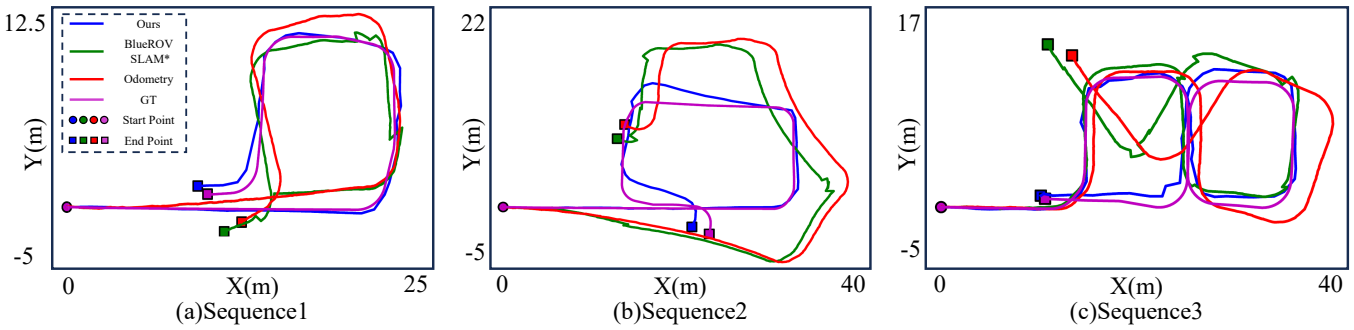


Fig. 7. Trajectories of the three simulation sequences.

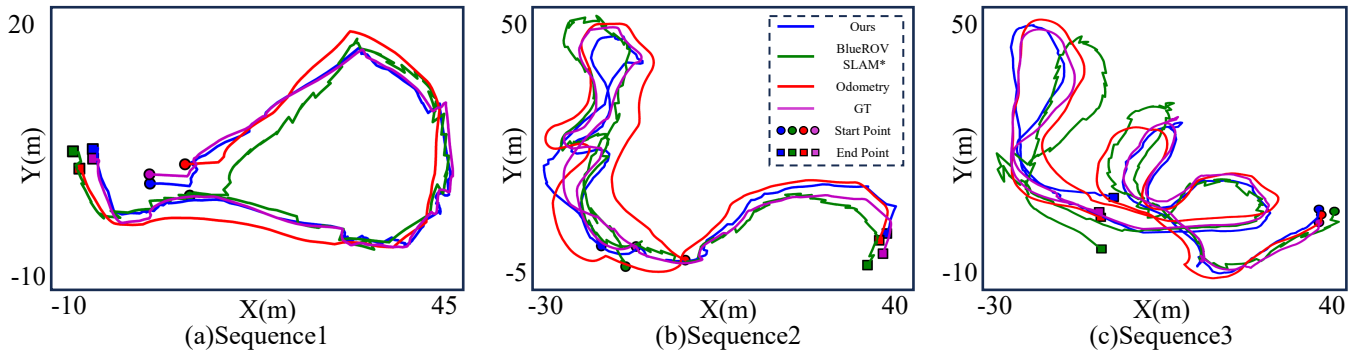


Fig. 8. Trajectories on real-world dataset Aracati2017 [8].

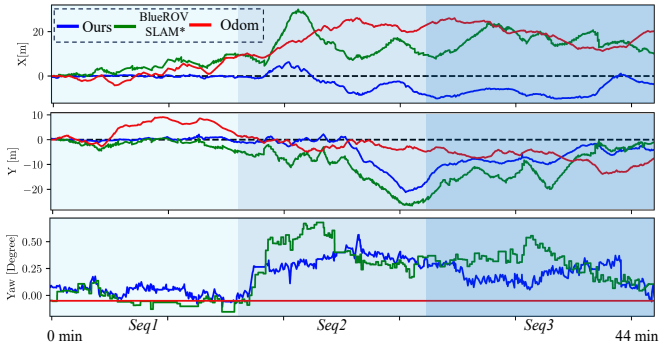


Fig. 9. Localization errors along time on Aracati2017 dataset.

proposed method is the most accurate to the groundtruth. Fig. 9 shows the error change along timestamp on x, y and yaw on the overall Aracati2017 dataset. The proposed method shows better performance on all axis. The Odometry error on yaw is always very close to 0 because the odometry and the groundtruth orientations are generated by the compass.

C. Discussions

The noted superiority of the proposed method over the ICP-based method, such as BlueROV SLAM, can be attributed to two primary factors:

- Sensitivity to Initial Conditions: ICP algorithms are known to be highly sensitive to initial values. In scenarios characterized by elevated levels of point cloud extraction on sonar imagery, the initial values provided are often inadequate, leading the ICP algorithm to converge to local minima. Instead, the proposed method leverages the intensity gradient information and the coarse-to-fine optimization on the pyramid images to improve its robustness compared with ICP.
- Robustness to Outliers: The ICP algorithm is notably susceptible to the influence of outliers, whereas the proposed DISO manifests heightened robustness in handling such anomalous data points. DISO introduces a new intensity-based inlier selection scheme, which is robust to the outliers.

IV. Conclusions

We present a novel algorithm for direct sonar optimization, specifically designed to optimize the relative pose between two distinct sonar frames by minimizing the error in acoustic intensity. The algorithm is incorporated into a meticulously designed sonar odometry system, which takes into account the unique challenges associated with sonar data namely, its low signal-to-noise ratio and low resolution.

Experiments are conducted in both simulated environments and real-world aquatic settings. The results demonstrate our proposed method consistently demonstrated superior performance in minimizing both translational and rotational errors, as compared to the ICP-based approach.

One primary constraint of the present study lies in the incapacity of the proposed algorithm to address loop closure. Another constraint is the assumption of zero elevation limits the proposed method to only a 2D planar scenario. Future research endeavours will explore the feasibility of integrating loop closure capabilities and extending to a 3D environment.

REFERENCES

[1] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 2564–2571.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

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

[4] 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.

[5] Y.-S. Shin, Y. Lee, H.-T. Choi, and A. Kim, "Bundle adjustment from sonar images and slam application for seafloor mapping," in *OCEANS 2015-MTS/IEEE Washington*. IEEE, 2015, pp. 1–6.

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

[7] E. Westman and M. Kaess, "Degeneracy-aware imaging sonar simultaneous localization and mapping," *IEEE Journal of Oceanic Engineering*, vol. 45, no. 4, pp. 1280–1294, 2019.

[8] Aracati 2017. [Online]. Available: https://github.com/matheusbg8/aracati2017

[9] W. Jinkun, C. Fanfei, H. Yewei, M. John, S. Tixiao, and E. Brendan, "Virtual maps for autonomous exploration of cluttered underwater environments," in *IEEE Journal of Oceanic Engineering*. IEEE, 2022.

[10] C. Forster, M. Pizzoli, and D. Scaramuzza, "Svo: Fast semi-direct monocular visual odometry," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 15–22.

[11] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[12] R. Wang, M. Schwörer, and D. Cremers, "Stereo dso: Large-scale direct sparse visual odometry with stereo cameras," in *2017 International Conference on Computer Vision (ICCV)*, Venice, Italy, October 2017.

[13] D. Luo, Y. Zhuang, and S. Wang, "Hybrid sparse monocular visual odometry with online photometric calibration," *International Journal of Robotics Research*, vol. 41, no. 11-12, pp. 993–1021, 2022.

[14] S. Xu, T. Luczynski, J. S. Willners, Z. Hong, K. Zhang, Y. R. Petillot, and S. Wang, "Underwater visual acoustic slam with extrinsic calibration," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 7647–7652.

[15] Z. Zhang and D. Scaramuzza, "A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[16] F. R. Lab. Project dave. [Online]. Available: https://field-robotics-lab.github.io/dave.doc/

[17] W.-S. Choi, D. R. Olson, D. Davis, M. Zhang, A. Racson, B. Bingham, M. McCarrin, C. Vogt, and J. Herman, "Physics-based modelling and simulation of multibeam echosounder perception for autonomous underwater manipulation," *Frontiers in Robotics and AI*, vol. 8, p. 706646, 2021.

[18] robot localization. [Online]. Available: https://github.com/cra-ros-pkg/robot_localization

[19] Z. Zhang and D. Scaramuzza, "A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry," in *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2018.