SONIC: Sonar Image Correspondence using Pose Supervised Learning for Imaging Sonars

Samiran Gode*, Akshay Hinduja*, and Michael Kaess

***Abstract*— In this paper, we address the challenging problem of data association for underwater SLAM through a novel method for sonar image correspondence using learned features. We introduce SONIC (SONar Image Correspondence), a pose-supervised network designed to yield robust feature correspondence capable of withstanding viewpoint variations. The inherent complexity of the underwater environment stems from the dynamic and frequently limited visibility conditions, restricting vision to a few meters of often featureless expanses. This makes camera-based systems suboptimal in most open water application scenarios. Consequently, multibeam imaging sonars emerge as the preferred choice for perception sensors. However, they too are not without their limitations. While imaging sonars offer superior long-range visibility compared to cameras, their measurements can appear different from varying viewpoints. This inherent variability presents formidable challenges in data association, particularly for feature-based methods. Our method demonstrates significantly better performance in generating correspondences for sonar images which will pave the way for more accurate loop closure constraints and sonar-based place recognition. Code as well as simulated and real-world datasets are made public on https://github.com/rpl-cmu/sonic to facilitate further development in the field.**

I. Introduction

Feature-based methods excel at localization and mapping by tracking distinct features over time. With camera frameworks, these methods utilize photometric consistency and invariance to common transformations such as scale and rotation, to produce precise feature correspondences [1].

Camera-based localization and mapping face hurdles underwater due to reduced color depth and visibility, limiting usable data. In contrast, *forward looking* or *imaging* sonars excel underwater, offering long-range visibility impervious to water particulates. While sonars are optimal for such scenarios, they currently lack robust feature descriptors.

In the camera imaging domain, feature descriptors and detection are well-researched, with notable examples being SIFT [2], ORB [3] and AKAZE [4]. The robustness of these feature descriptors stems mainly from two principles, photometric consistency and geometric invariance. Photometric consistency maintains pixel value stability against viewing angle variations and noise, while geometric resilience ensures feature recognition across varying orientations and scales. On the other hand, imaging sonars exhibit variations in intensity returns and observable shapes when viewing the same object from different angles, influenced by the object's material and geometry. These aforementioned descriptors struggle with the prevalent speckle noise and intensity variations characteristic of sonar images, especially in the polar space. An example of such a failure shown in Fig. 1, in row 3 with AKAZE and the brute force (BF) matcher. The downstream effects of this failure can result in poor, or error-prone loop closure detection and state estimation. Our work addresses this issue by using pose supervision grounded

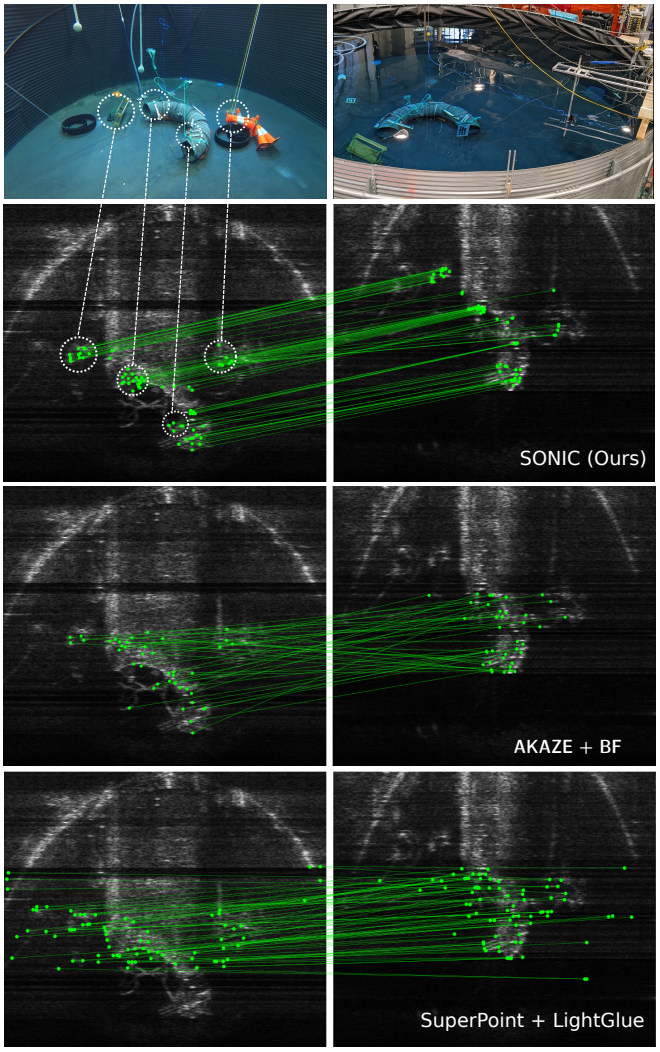


Fig. 1: Real-world matching performance: Sonar images taken from different planar positions in a test tank show our method providing significantly better matches than AKAZE with the brute force matcher, and the SuperPoint keypoints matched using LightGlue. Given the keypoints in the first frame, SONIC uses expectation matching to determine the correspondences and presents only those correspondences with high confidence.

The authors are affiliated with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. *These authors contributed equally to this work. {ahinduja, kaess}@andrew.cmu.edu, sgode@alumni.cmu.edu

This work was partially supported by the Office of Naval Research award N00014-21-1-2482. The authors would also like to thank Eric Westman for providing code excerpts used in evaluation.

in sonar geometry. This offers a feature correspondence method for sonar images, robust against photometric and geometric variations. Recent research has improved upon the matching performance of camera images with deep neural network based approaches using ground truth correspondences [5, 6, 7]. There has also been recent semi-supervised work techniques leveraging pose-supervised learning [8]. These models, though powerful, are trained on camera data and suffer from the same problems explained above when used on sonar data.

Motivated by [8] and the availability of pose-supplemented sonar data using simulation [9], we propose a novel pose-supervised method to learn sonar image matching using a sonar epipolar-contour(see III-B) based loss function. In addition, we also enforce cyclic consistency as done in [8]. Our network learns directly from sonar images in the polar space and the loss is calculated in this space ensuring the model learns unique representations in sonar images. Specifically, our main contributions are:

- A novel semi-supervised method for sonar image correspondence, utilizing sonar-specific epipolar geometry negating the need for ground truth correspondences. This technique offers correspondences tailored for imaging sonar, producing features resilient to viewpoint changes.
- An extensive simulated dataset that comprises over 300K pairs of sonar images and their corresponding ground truth poses, collected across 10 distinct scenes. Each scene is made from randomized positioning of a specific set of objects, with repeatable sensor motion from different pose offsets.

II. Related Work

The descriptors mentioned in Section I have been utilized for sonar images in the past for applications towards acoustic structure from motion (ASFM) and feature-based SLAM [10, 11, 12, 13, 14]. These descriptors worked as long as the change in viewing angle was minimal. A couple of recent works [15, 16] give a summary of different keypoint detectors and feature extractors on sonar images and they along with [11] suggest the need for an invariant sonar specific feature descriptor.

Hand-designing a new feature descriptor for a specific type of sensor is a viable approach [17, 18], and there has been recent development on a SIFT-like descriptor made for multi-beam sonar [2]. The drawback to this process is that the descriptor parameters would need to be manually tuned for different imaging sonar models, as each sonar make has a unique elevation, bearing and range specification, and signal-to-noise profile. On the other hand, recent research on learned feature descriptors for cameras [6, 7, 19] and 3D lidars [20] have shown encouraging results when used for correspondence estimation. Similarly, methods like [21] leveraged deep neural networks for place recognition for sonar images but have not learned feature correspondences. These methods utilized deep networks to solve the problem of large variations across scenes, and multiple sensor makes. Looking specifically towards research for camera images, there have been several supervised methods such as SuperPoint [6] and LOFTR [7]. SuperPoint's network works on producing the learned feature descriptors using homographic adaptations for supervision. Follow-up work in SuperGlue, and more recently, LightGlue [5, 22] utilize the SuperPoint feature descriptors to form a robust feature matching framework. Approaches similar to SuperPoint rely heavily on abundant training data, for example from the MegaDepth dataset [23]. Other approaches, like the current state of the art, SiLK [19], depend on sub-pixel ground truth correspondences as a linear mapping during training.

These techniques, while likely to yield similar great results for sonar, are a challenging endeavor to replicate for imaging sonar. This is primarily owing to the scarcity of readily available open-access sonar data and a dearth of ground truth feature point and correspondence information. Additionally, for sonar images, the motion model for planar scenes do not reduce to a homography [24]. To address these issues, we look towards ongoing research dedicated to semi-supervised approaches that harness sensor pose data. This enables us to circumvent the necessity for precise ground truth correspondences among feature points. These methods have demonstrated success in achieving equivalent, if not better accuracy compared to their fully supervised counterparts, all while demanding a smaller volume of training data. A noteworthy example of a pose supervised method is CAPS [8]. With access to pose information relating to two images capturing the same scene, CAPS effectively employs a pair of loss functions: epipolar loss and cyclic loss. The foundation of their system rests upon the fundamental notion that a point of interest in the initial image should invariably align with the epipolar line corresponding to its counterpart in the second image. Our approach builds on the backbone of CAPS, utilizing an analog to epipolar geometry for imaging sonars which is detailed in Section IV. This gives a data and time-efficient way of obtaining trained feature descriptors for sonar images which can outperform the currently popular methods in use.

III. Preliminaries

A. Imaging Sonar Sensor Model

Imaging sonars are active acoustic sensors that measure reflection intensities of emitted sound waves. While analogous to optical cameras in projecting 3D scenes to 2D, their output images are distinctly different. We use the following coordinate frame convention, the x axis points forward from the acoustic center of the sensor, with the y axis pointed to the left and z axis pointed up, as shown in Fig. 2. Unlike the pinhole camera model, where the scene from the viewable frustum is projected onto a forward-facing image plane, for imaging sonars the scene is projected into the zero elevation plane, which is the xy plane in the sonar frame.

Modifying the notation in [25] slightly, consider a point $\mathbf{P}$ with spherical coordinates (r, θ, ϕ) - range, azimuth, and

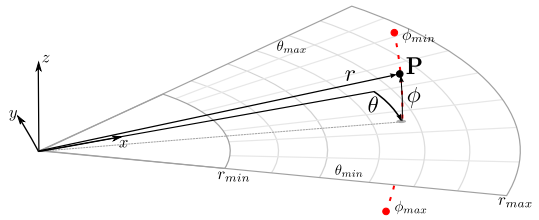


Fig. 2: The basic imaging sonar sensor model of a point feature. Each pixel provides direct measurements of the bearing / azimuth (θ) and range (r), but the elevation angle (ϕ) is lost in the projection onto the image plane - analogous to the loss of the range in the perspective projection of a camera. The imaged volume, called the frustum, is defined by the sensors limits in azimuth $[\theta_{min}, \theta_{max}]$, range $[r_{min}, r_{max}]$, and elevation $[\phi_{min}, \phi_{max}]$.

elevation, with respect to the sonar sensor. The corresponding Cartesian coordinates are then represented as:

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

In the pinhole camera model, a pixel location can indicate azimuth and elevation angles, but lacks clear range information. Any 3D point along its ray projects to the same pixel. Conversely, in the imaging sonar model, a 3D point is mapped onto the zero elevation plane as

$$\mathbf{p} = \begin{bmatrix} x_s \\ y_s \end{bmatrix} = r \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} = \frac{1}{\cos\phi} \begin{bmatrix} X_s \\ Y_s \end{bmatrix}. \tag{2}$$

In the projective camera model, each pixel has a corresponding ray that passes through the sensor origin. In contrast, the sonar sensor model has a finite elevation arc in 3D space, as seen in Fig. 2 as a red dotted line. In sonar images, a pixel conveys azimuth and range data for an arc but lacks elevation, similar to the range data being absent in the projective camera model. This nonlinear projection and limited field of view in imaging sonar sensors create complexity, making tasks that are straightforward for optical cameras seem challenging for sonar. Another complication arises from the information these pixels hold. In camera images, pixels represent light intensity from specific surface patches along corresponding rays, often resulting in unique pixel-to-patch correspondence. This similarity in pixel values across viewpoints accommodates minor camera origin shifts. In contrast, acoustic images may have pixels representing multiple surfaces along an elevation arc reflecting sensor-emitted sound. This can lead to compounded intensity in a single pixel, which may vary even with slight sensor origin changes.

B. Sonar Epipolar Geometry

Negahdaripour introduced the concept of stereo epipolar geometry in [26]. In cameras, the epipolar line is the intersection of a point's epipolar plane with the image plane. Essentially, it's the projection of the line connecting the 3D point and one camera center onto the second camera's image plane. This line represents the depth and direction from the first view. In sonar stereo, elevation arcs serve as the analog to the epipolar lines found in cameras, and their projection is

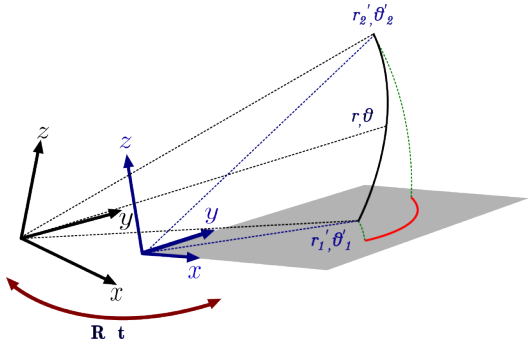


Fig. 3: Sonar Epipolar Geometry: The elevation arc of a point in the first image is transformed into the frame of the second image and then projected, which creates an epipolar contour.

an *epipolar contour*. We will describe the epipolar geometry in brief here, but refer the reader to [26] for a detailed explanation.

We refer to Section III-A where we return to Eq. 1 for notation relating to the conversion of a point in polar space to the Cartesian space. Pixels are in range-bearing format since we are training sonar images. Given a point in one sonar image at some range, R and bearing θ. The point in 3D will lie along the elevation arc at the same range and bearing as defined and at some arbitrary elevation angle, φ. Since we do not know the elevation angle, the ambiguity is along the elevation arc for $\varphi_{min} <= \varphi <= \varphi_{max}$. Now, the conversion of points in 3D Cartesian space to the range bearing space is as seen in Eq. 3.

$$\boldsymbol{P} = \begin{bmatrix} R \\ \theta \\ \phi \end{bmatrix} = r \begin{bmatrix} \sqrt{x^2+y^2+z^2} \\ arctan2(y,x) \\ arctan2(x, \sqrt{x^2+y^2}) \end{bmatrix} \tag{3}$$

Assuming we have a feature point $\boldsymbol{p}$ in the first image, and $\boldsymbol{R_{1,2}}$and $\boldsymbol{t_{1,2}}$ are the known rotation and translation between the coordinate frame of image 1 and image 2. The locus of the same feature point in image 2, $\boldsymbol{p'}$ is calculated as seen in Eq. 5. This 3D feature locus can now be projected into the polar sonar image plane as in Eq. 6. Fig. 3 visually describes the process of projecting the elevation arc from the first frame onto the second image plane. The red line is the epipolar contour we use for the loss function described in the following section.

$$\boldsymbol{R_{1,2}} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix}, \boldsymbol{t_{1,2}} = \begin{bmatrix} t_x & t_y & t_z \end{bmatrix}^T \tag{4}$$

$$\boldsymbol{p'}(\phi) = \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} r_1 \cdot P + t_x \\ r_2 \cdot P + t_y \\ r_3 \cdot P + t_z \end{bmatrix} \tag{5}$$

$$\boldsymbol{p'_{proj}}(\phi) = \begin{bmatrix} r' \\ \theta' \end{bmatrix} = \begin{bmatrix} \left\| P'_p \right\|_2 \\ arctan2(y', x') \end{bmatrix} \tag{6}$$

IV. Method

In this section we describe the key aspects of our method. We first discuss obtaining keypoints, then introduce the network, and detail the loss functions and the training strategy.

A. Keypoint Detection

SONIC, like CAPS, trains feature representations and requires keypoints as an input. We use a combination of AKAZE and SuperPoint keypoints for training to ensure we have enough keypoints per frame while training. Keypoints are found in polar space as converting sonar images to euclidean spaces leads to loss of information, especially when near the sonar.

B. Network Overview

We use a CNN based encoder-decoder architecture with a differentiable matching layer and coarse-to-fine technique similar to [8]. Unlike CAPS, SONIC uses a ResNet-34 [27] base to prevent overfitting our small dataset. From here, similar to CAPS, the encoder creates a coarse representation; given the input keypoint we find the corresponding point in this representation using the differentiable matching layer. A pictorial representation of the differentiable matching layer and coarse to fine architecture is shown in Fig. 4 (a) and (b) respectively. Given two images, the network with shared weights creates the representation M_1 and M_2. For each query point x_1, the feature descriptor $M_1(x_1)$ is correlated to each point in M_2. This is used to find the probability of each point being the correspondence of x_1 in M_1 as shown in Eq. 7. A correspondence is calculated by finding the expectation of this probability in Eq. 8. Searching correspondences for all points over the image is very computationally costly, hence it makes most sense to sparsely sample the query points for supervision. This coarse to fine architecture improves on efficiency. At the coarse level, a correspondence distribution is computed over all the locations. However, at the finer level, the distribution is only computed at the highest probability location observed from the coarse map. The loss functions are imposed on both levels.

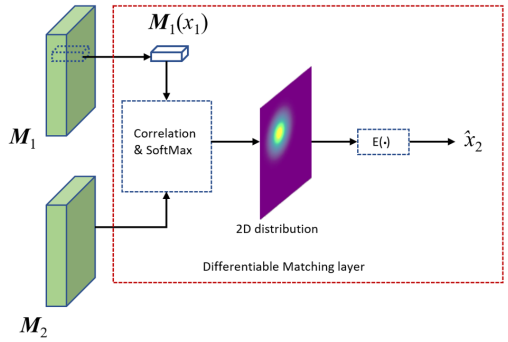
Our other modifications to the original CAPS architecture include changes to work with single channel sonar images, and condense final coarse and fine layer outputs to 64 dimensions instead of the original 128.

$$p(x|x_1, M_1, M_2) = \frac{\exp(M_1(x_1)^\top M_2(x))}{\sum_{y \in I_2} \exp(M_1(x_1)^\top M_2(y))} \tag{7}$$

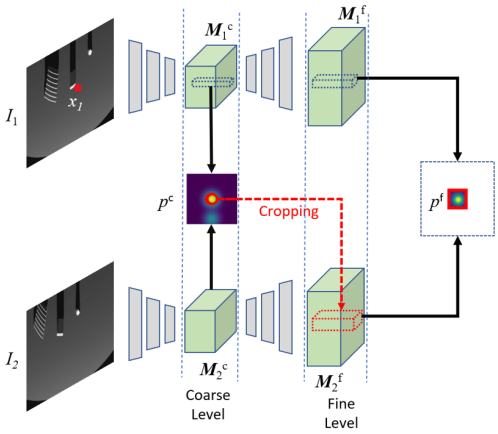
$$\hat{x}_2 = h_{1\to2}(x_1) = \sum_{x \in I_2} x \cdot p(x|x_1, M_1, M_2) \tag{8}$$

C. Supervision

Like CAPS, we propose two loss functions, modifying it for sonar images. We introduce the sonar-epipolar and sonar-cyclic loss. Given the relative pose between two image frames $I_1$and I_2, we use Eq. 5 and 6 to determine the epipolar contour for a given point P in the I_1. A correctly predicted point in the I_2 would lie on this contour. We use the distance between this predicted point and contour as a metric to optimize over. The epipolar loss term, $L_{epipolar}$ for a given point x_1 in I_1 in Eq. 9 is defined as the shortest distance between the predicted correspondence of $\hat{x}_2$ and the epipolar contour of x_1 in the second image I_2. In our



(a) Differentiable Matching Layer



(b) Coarse to Fine Module

Fig. 4: Network architecture highlights: a) For each qeury point x_1, its corresponding location $\hat{x}_2$(Eq. 8) is represented as the expectation of a distribution computed from the correlation between the feature descriptors. The associated uncertainty also helps in reweighting training loss. During training, keypoints serve as queries (b) Searching correspondence across the entire image is costly. The location of the correspondence p^c at the coarse level is used to ascertain a local window at the fine level, p^f is found in this window using differentiable matching.

implementation, we sample points along the elevation arc of the first point in the first image and then transform and project them on to the second image to create a discrete epipolar contour of the sampled points. In the polar frame, the loss is thus the minimum of the distance between the predicted point and each point on the arc as seen in Eq. 13. The epipolar loss only checks for the predicted match to lie on the estimated contour. To further constrain the system, a cyclic consistency loss, L_{cyclic}, is utilized which aims to keep the forward-backward mapping of the point consistent. The weighted sum of the losses $L_{epipolar}$ and L_{cyclic} is our final loss function as seen in Eq. 13. A point to note is that the distances found for both the losses are in the range-bearing space. Due to nature of sonar images, it is important for the network to learn this distinction, and the combined loss terms in this representation support this. A graphical representation of the losses is seen in Fig. 5.

$$L_{epipolar}(x_1) = dist(h_{1\to2}(x_1), ep\_contour) \tag{9}$$

$$L_{cyclic}(x_1) = \|h_{2\to1}(h_{1\to2}(x_1) - x_1)\|_2 \tag{10}$$

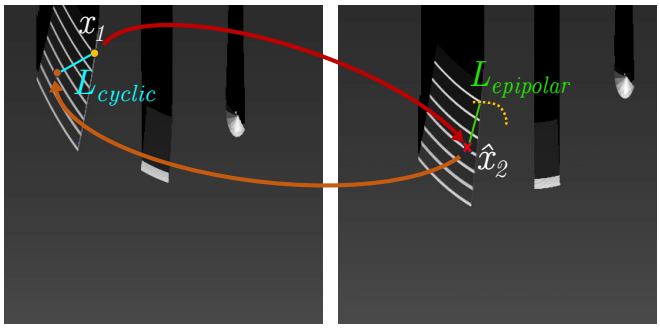


Fig. 5: Loss functions: The yellow point x_1 represents a queried keypoint in the first image. The red cross $\hat{x_2}$is the predicted point. The yellow dotted line represents the sampled points on the epipolar contour of point x_1. $L_{epipolar}$ is the shortest distance to the epipolar contour, or simply the epipolar loss. L_{cyclic} is the cyclic loss to assert that the mapping of the feature point is close to its original position.

$$L_{(I_1,I_2)} = \sum_{i=1}^{n} [L_{epipolar}(x_1^i) + \lambda L_{cyclic}(x_1^i)] \tag{11}$$

$$L_{cyclic}^{1,2} = r_1^2 + r_2^2 - 2r_1 r_2 (cos(\theta_1 - \theta_2)) \tag{12}$$

$$L_{epipolar}(p'_{proj}(\varphi), r_2, \theta_2) = \min_{\phi}(r'(\varphi)^2 + r_2^2 - 2r'(\varphi) r_2 (cos(\theta'(\varphi) - \theta_2))) \tag{13}$$

D. Data and Training

Acquiring numerous sonar images with ground truth pose is challenging, and real-world data often lacks feature-rich frames for network training. While there are sonar datasets like those by Singh et al. [28] and Malios et al. [29] for object classification or SLAM, they either lack pose information or do not use wide elevation imaging sonars. Thus, simulated environments, particularly the HoloOcean simulator [9, 30], supplemented by custom worlds, offer an enhanced solution for data generation.

For our primary application, seafloor mapping, sonar images were gathered from 1-4m above the floor at a downward pitch of 10-20°. Multiple trajectories with varying rotation and translation offsets produced image pairs. Our dataset comprises approximately 300K training and 30K validation pairs. We configure our simulated images on the Blueprint Subsea M1200d's [31] low frequency mode. In this mode, the images have a maximum azimuthal field of view of 130° and elevation of 20°. The maximum range is set to be 10m. The image is comprised of 512×512 range and bearing bins. Image noise parameters are marginally varied around the values provided in HoloOcean's example scenarios.

Epipolar and cyclic losses are given 0.7 and 0.3 weights, respectively. We train the network for 32 epochs with a batch size of 14 pairs, using a Nvidia GeForce 4090 RTX with 24GB memory.

V. Evaluation

We assessed matching performance by comparing SONIC, AKAZE, and the camera-trained LightGlue model. As highlighted in Section II AKAZE is the preferred keypoint and descriptor for imaging sonar applications. LightGlue [22] is a recent improvement on SuperGlue, with a performance equivalent to LoFTR [7] and MatchFormer [32] which are considered the current state of the art for camera image correspondence. We compare SONIC with LightGlue since both use SuperPoint keypoints and require a keypoint detector, unlike other detector-free matchers.

Assuming a planar scene we evaluate the number of matched keypoints that are considered to be inliers for each of AKAZE, LightGlue and SONIC. Inliers are classified by projecting the keypoints from the reference image using the ground truth relative pose information as described in Section III-B onto the query image. The threshold selected for simulation images is 12 pixels, which corresponds to 0.23m in range or 3° in bearing. We increase this threshold for real images to be 20 pixels, which corresponds to 0.2m and 5° in range and bearing respectively. To show the benefit of our method towards downstream tasks such as feature-based SLAM, we evaluate performance in relative sensor pose recovery using the two-view acoustic bundle adjustment framework to recover sensor pose information as presented by Westman et al. [14]. It is typical for hovering underwater vehicles to be equipped with pressure sensors for depth estimation, and to be limited to planar motion [33]. Most graph-based solutions for underwater SLAM would thus model *Z*, *pitch* and *roll* as a separate unary factor, independent to planar motions in *X*, *Y*, and *Yaw* [34]. Thus we focus mainly on evaluating the average absolute error in planar translation (*xy*) and rotation (*yaw*). Prior pose estimates are derived from corrupted ground truth data. These poses are used to filter matches. We perform a Z-test on the distance between the corresponding projected and matched query keypoints. We prune matches not falling under a 2σ threshold for all three methods. The same noisy pose prior is provided to the acoustic bundle adjustment optimizer along with matched keypoints from the reference and query images, passed as their bearing and range values. In SLAM application, the data association could be improved by using RANSAC [35] or JCBB [14].

We present and discuss results for our simulated datasets first, and then for the real world data from a test tank.

A. Simulation

We sample sonar image pairs, unused for SONIC's training, with varied pose differences including minor roll, pitch, and z variations, and broader x, y, and yaw differences. These pairs are categorized into two groups: small and large variation. The small variation group contains pairs with *yaw* $< \pm 5°$ and *x,y* $< \pm 1.5$m. The large variation group contains pairs with up to $\pm 40°$ variation in rotation and ± 7 meter in translation. Evaluation images had the same parameters for range, bearing and elevation as used in training. We first look at the inlier ratio in I, where both small and large variation groups show significantly more ground truth pose agreeable matches than the other two methods. We then look at Table II to see the average absolute error and standard deviation for planar translation(meters) and rotation(radians). Our method outperforms the other two in all situations.

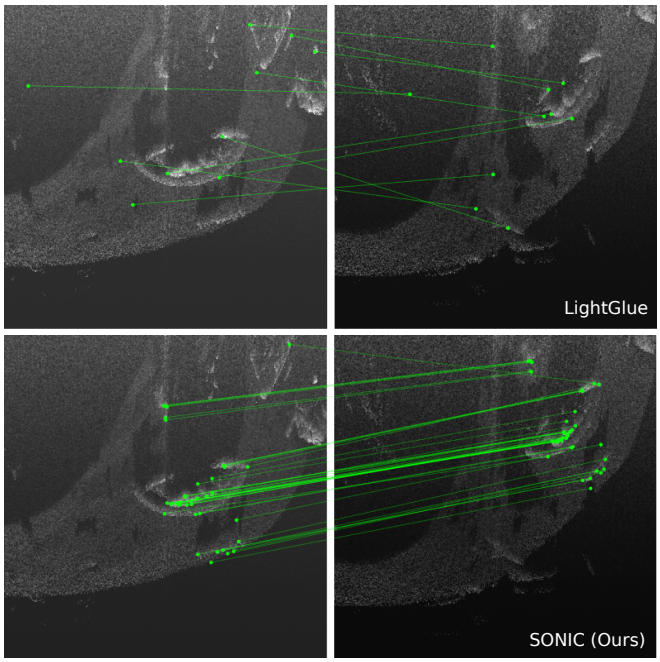


Fig. 6: Simulation Matching Performance: LightGlue is unable to match features when the structures' shape warps under sensor motion, a common phenomenon in polar images.

SONIC is specifically trained to identify features which are more likely to be invariant to viewpoint changes, as well as learn the geometric variance of features across these changes. The inductive bias of convolutional neural nets also ensures that it looks at a area and structure around the keypoint and finds matches closest to that in the other image. AKAZE and other traditional descriptors try to describe features such as as corners, edges and blobs which do not stay consistent in sonar images. LightGlue, and other learned matchers also look at the image as a whole and thus perform much better than AKAZE. However, since LightGlue is trained on camera images it cannot predict accurate matches when the structures or keypoints change over significant translation or rotational variation due to the change in object geometries occurring in the polar image-space as seen in Fig. 6.

B. Test Tank Evaluation

Transferring simulated performance to real-world scenarios is challenging. As SONIC was solely trained on simulated data, it's crucial to assess if it learned the geometry of view-invariant features in actual sonar data. Accurately modeling sonar noise is challenging, making real-image performance evaluation vital. We conducted experiments in a test tank with a sliding gantry system with a configurable 6 degrees of freedom sensor pose as seen in Fig. 1. We used a Leica Total Station 16 [36] with a tracking prism for ground truth estimation. Here, we show two sample frames, taken at differing positions with a change of 1.153m and 1.310m in $\mathbf{x}$ and $\mathbf{y}$ and a yaw rotation of $-15°$. The other degrees of freedom remained constant. We observe our method is able to provide the right matches, whereas previously used methods like AKAZE with symmetric nearest-neighbour brute force matching and SuperPoint descriptors with LightGlue are unable to do so. The matching results of the three methods in the tank can also been seen in the same figure. While we trained our images on a maximum range setting of 10 meter, the real world parameters was fixed to 6m due to the small size of the tank (7m diameter). We also had to limit the keypoint detectors for all three methods from detecting points beyond the general position of the objects due to significant acoustic reflections between the metallic surfaces of the tank wall and pipe placed inside the tank. Due to the limited change in pose, and concentrated features, the two-view acoustic bundle adjustment framework was unable to resolve the relative pose for most image pairs. However, we can see from the inlier ratio in Table I that SONIC is once again more likely to provide better correspondences.

TABLE I: Percentage of Inliers

Method	Simulation - Small Variation	Simulation - Large Variation	Test Tank
AKAZE	24.23%	10.22%	40.08%
LightGlue	39.13%	11.18%	51.92%
SONIC	**49.43%**	**23.56%**	**74.53%**

TABLE II: Planar Translation and Rotational Accuracy

Method	Simulated data			
	Small Variation		Large Variation	
	Translation (m)	Rotation (rad)	Translation (m)	Rotation (rad)
AKAZE	μ: 4.24, σ: 2.40	μ: 1.56, σ: 1.22	μ: 5.06, σ: 2.35	μ: 1.44, σ: 1.01
LightGlue	μ: 3.62, σ: 4.52	μ: 0.97, σ: 1.12	μ: 3.49, σ: 3.53	μ: 0.97, σ: 1.00
SONIC	μ: **0.88**, σ: **2.01**	μ: **0.25**, σ: **0.61**	μ: **2.23**, σ: **3.35**	μ: **0.60**, σ: **0.82**

VI. Conclusions and Future Work

We propose SONIC, a pose-supervised model that solves the challenging problem of feature correspondence in sonar images. Our results demonstrate that SONIC excels over existing techniques in sonar image feature extraction, addressing the need for sonar-centric feature correspondence. Our method achieves this through a novel sonar epipolar loss and a cyclic consistency loss. The epipolar loss utlizes the epipolar contour projected by an arc onto the second image using relative pose information. This guides the predicted correspondence to align with this contour. Simultaneously, our consistency loss ensures cyclic congruence between the predicted and the initial keypoints. Our method, trained solely on simulated data, demonstrates promising real-world feature correspondence due to its understanding of underlying geometry. However, more thorough open water tests are needed, along with supplementing real data towards training to close the gap between simulation and real-world results. Future applications of this work include the extension of the method to incorporate different sonar frequency modes and parameters under one model. Recent work on image matching using attention [37] opens the door to enable cross-sonar feature matching and improving performance, which could potentially help cross-platform localization and mapping.

REFERENCES

[1] M. Mur-Artal, Raúl, J. M. M., and J. D. Tardós, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[2] W. Zhang, T. Zhou, C. Xu, and M. Liu, "A SIFT-like feature detector and descriptor for multibeam sonar imaging," Jul 2021.

[3] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 International conference on computer vision*, pp. 2564–2571, 2011.

[4] P. F. Alcantarilla and T. Solutions, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," *IEEE Trans. Patt. Anal. Mach. Intell*, vol. 34, no. 7, pp. 1281–1298, 2011.

[5] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4938–4947, 2020.

[6] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 224–236, 2018.

[7] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8922–8931, 2021.

[8] Q. Wang, X. Zhou, B. Hariharan, and N. Snavely, "Learning feature descriptors using camera pose supervision," in *European Conference on Computer Vision*, pp. 757–774, Springer, 2020.

[9] E. Potokar, S. Ashford, M. Kaess, and J. G. Mangelson, "HoloOcean: An underwater robotics simulator," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 3040–3046, IEEE, 2022.

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

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

[12] Y. S. Shin, Y. Lee, H. T. Choi, and A. Kim, "Bundle adjustment from sonar images and SLAM application for seafloor mapping," in *Proc. of the IEEE/MTS OCEANS Conf. and Exhibition*, pp. 1–6, Oct. 2015.

[13] J. Li, M. Kaess, R. 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.

[14] E. Westman and M. Kaess, "Degeneracy-aware imaging sonar simultaneous localization and mapping," *IEEE J. of Oceanic Engineering*, 2019.

[15] A. Oliveira, B. Ferreira, and N. Cruz, "A performance analysis of feature extraction algorithms for acoustic image-based underwater navigation," *J. Mar. Sci. Eng.*, vol. 9, p. 361, 2021.

[16] P. Tueller, R. Kastner, and R. Diamant, "A comparison of feature detectors for underwater sonar imagery," in *OCEANS 2018 MTS/IEEE Charleston*, pp. 1–6, 2018.

[17] P. Hansen, P. Corke, W. Boles, and K. Daniilidis, "Scale invariant feature matching with wide angle images," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1689–1694, IEEE, 2007.

[18] J. Yang, Z. Cao, and Q. Zhang, "A fast and robust local descriptor for 3D point cloud registration," *Information Sciences*, vol. 346-347, pp. 163–179, 2016.

[19] P. Gleize, W. Wang, and M. Feiszli, "Silk – simple learned keypoints," in *ICCV*, 2023.

[20] A. Dewan, T. Caselitz, and W. Burgard, "Learning a local feature descriptor for 3D LiDAR scans," pp. 4774–4780, 10 2018.

[21] P. O. C. S. Ribeiro, M. M. dos Santos, P. L. J. Drews, S. S. C. Botelho, L. M. Longaray, G. G. Giacomo, and M. R. Pias, "Underwater place recognition in unknown environments with triplet based acoustic image retrieval," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 524–529, 2018.

[22] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "LightGlue: Local Feature Matching at Light Speed," in *ICCV*, 2023.

[23] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[24] S. Negahdaripour, "On 3-D motion estimation from feature tracks in 2-D FS sonar video," *IEEE Trans. Robotics*, vol. 29, pp. 1016–1030, Aug. 2013.

[25] N. Hurtós, D. Ribas, X. Cufí, Y. Petillot, and J. Salvi, "Fourier-based registration for robust forward-looking sonar mosaicing in low-visibility underwater environments," *J. of Field Robotics*, vol. 32, no. 1, pp. 123–151, 2014.

[26] S. Negahdaripour, "Analyzing epipolar geometry of 2-D forward-scan sonar stereo for matching and 3-D reconstruction," in *Proc. of the IEEE/MTS OCEANS Conf. and Exhibition*, pp. 1–10, 2018.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[28] D. Singh and M. Valdenegro-Toro, "The marine debris dataset for forward-looking sonar semantic segmentation," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, (Los Alamitos, CA, USA), pp. 3734–3742, IEEE Computer Society, oct 2021.

[29] A. Mallios, E. Vidal, R. Campos, and M. Carreras, "Underwater caves sonar data set," *The International Journal of Robotics Research*, vol. 36, no. 12, pp. 1247–1251, 2017.

[30] E. Potokar, K. Lay, K. Norman, D. Benham, T. B. Neilsen, M. Kaess, and J. G. Mangelson, "HoloOcean: Realistic sonar simulation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8450–8456, 2022.

[31] Blueprint Subsea, "Blueprint subsea oculus M1200d." `https://blueprintsubsea.com/oculus/oculus-m-series`.

[32] Q. Wang, J. Zhang, K. Yang, K. Peng, and R. Stiefelhagen, "Matchformer: Interleaving attention in transformers for feature matching," 2022.

[33] M. Kaess, H. Johannsson, B. Englot, F. Hover, and J. Leonard, "Towards autonomous ship hull inspection using the Bluefin HAUV," in *Ninth International Symposium on Technology and the Mine Problem*, May 2010.

[34] P. Teixeira, M. Kaess, F. Hover, and J. Leonard, "Underwater inspection using sonar-based volumetric submaps," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, (Daejeon, Korea), pp. 4288–4295, Oct. 2016.

[35] 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*, pp. 1–6, 2015.

[36] Leica Geosystems, "Leica TS16: Robotic Total Station." `https://leica-geosystems.com/en-us/products/total-stations/robotic-total-stations/leica-ts16`.

[37] O. Wiles, S. Ehrhardt, and A. Zisserman, "Co-attention for conditioned image matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15920–15929, 2021.