2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Multi-session Underwater Pose-graph SLAM using Inter-session Opti-acoustic Two-view Factor

Hyesu Jang[1], Sungho Yoon[2] and Ayoung Kim[1,2*]

Abstract— **Concurrent mapping necessitates data association among vehicles to overcome temporal and sensor modality differences. In this work, we focus on an underwater multi-vehicle mapping scenario in which vehicles have various sensor modalities, namely sonar and camera. This inter-session sonar-optical image matching poses two main challenges. First, ensuring covisibility for the opti-acoustic pair is complex due to their projection models and field of view (FOV) difference. Second, even with secured covisible frames, feature matching over various sensor modalities is not trivial. To overcome these challenges, we complete multi-session simultaneous localization and mapping (SLAM) by introducing an opti-acoustic pairwise factor. We alleviate the covisibility requirement by introducing inter-session measurements. We achieved opti-acoustic feature matching by applying a style-transfer and integration with SuperGlue. The proposed method is validated via simulation and real underwater tank tests.**

I. Introduction and Related Works

For the multi-vehicle navigation and concurrent mapping, multi-session SLAM solves for the vehicles' consistent trajectory and mapping. Our main interest is the heterogeneously equipped underwater vehicles. A single-session underwater SLAM often combines navigational sensors (e.g., a Doppler velocity log (DVL) and fiber optic gyro (FOG)) with perceptual sensors (e.g., sonars and cameras) to overcome inevitable navigational drift. In doing so, the registration between image frames corrected navigation errors and reduced pose uncertainty [1]. As a single session, these studies mostly focused on the registration between the same type of sensors (e.g., matching between optical images [2] or sonar images [3]). Heterogeneous sensor registration is more actively studied in the terrestrial environment. For example, camera localization [4] over a LiDAR map presents a registration between different sensors. Likewise, optical camera and sonar sensor have a similar role in the underwater environment. *Opti-acoustics* provides the relation between optical images and sonar images, that can estimate the pose between two sensors. Furthermore, for underwater heterogeneous sensor fusion, we utilized *Image Style Transfer* in our previous work [5]. Developing the methods, this paper deals with *Opti-Acoustics* and *Image Style Transfer* to conduct underwater multi-session SLAM.

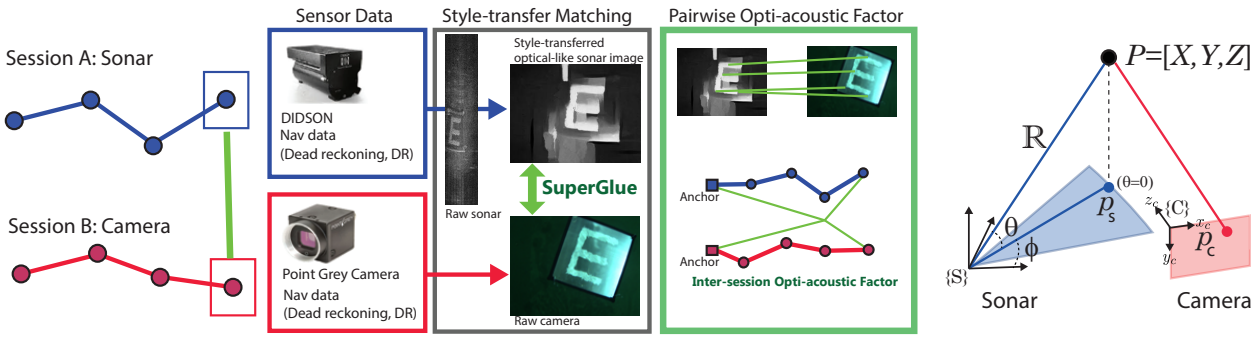
Opti-Acoustics: Underwater image registration between sonar and a camera is also known as opti-acoustic matching. A theoretical solution proposed by [6] combined sonar and camera measurements together to overcome each sensor's limitations. However, securing co-visibility in a realistic configuration and overcoming various sensor modalities present challenging issues. In their work, the opti-acoustic stereo system had a camera and sonar on each side of a rigid body to ensure covisibility. This hardware configuration could not be mounted to a moving platform or a small fleet. To overcome this issue, [7] proposed using the opti-acoustic system for non-temporal matching. Still, feature matching could not be fully solved when the authors manually cropped images to establish correspondences. Other studies examined sonar-visual-inertial SLAM [8, 9]. They resolved the visual-inertial navigation system's limitation by integrating the stereo camera with sonar range data and with inertial data.

Image Style Transfer: Neural network-based image enhancement is an active research topic in computer vision. Existing works on underwater imaging focused on improving image quality. For example, [10] recovered optical images from a submerged environment via a Generative Adversarial Network (GAN)-based color correction module that generated unstrained optical images. Cho et al. [11] also exploited GAN to dehaze underwater images. For imaging sonar sensors, [12] proposed crosstalk noise elimination via deep neural network (DNN). Wang et al. [13] classified sonar images with adaptive weights in a Convolutional Neural Network (CNN).

There are also reciprocal image transformations, *style transfer*, which endows different image styles while maintaining the contents. [14] proposed a CNN-based image style transfer method. They defined the representation of content and style and finally generated the designated styled images. Pix2pix [15] and CycleGAN [16] are GAN-based style transfer methods. A generator-discriminator adversarial structure creates adequate images that contain the provided information. Recently, using image localization via GAN, [17] presented impressive matching between day and night that overcame significant visual differences. In this paper, we decided to use a CNN-based style transfer since the number of datasets for opti-acoustic image pairs is limited.

Multi-Session SLAM: For multi-vehicle and multi-session SLAM, securing reliable data associations between sessions is the key challenge. Kim et al. [18] suggested integrating multiple pose graphs by introducing an anchor node. When inter-session constraints are obtained, the pose-graph solves the constraints via the *anchor* of each session. Konolige et al. [19] proposed view-based SLAM to robustly re-localize against each session by leveraging a vocabulary tree. Large-scale multi-session visual SLAM was conducted by [20], in which visual SLAM sessions were combined with

[1]H. Jang and A. Kim are with the Depart. of Civil and Env. Eng., KAIST, Daejeon, S. Korea [iriter, ayoungk]@kaist.ac.kr
[2]S. Yoon is with the Robotics Program, KAIST, Daejeon, S. Korea sungho.yoon@kaist.ac.kr



(a) Overall pipeline (b) Projection

Fig. 1: (a) Given a sonar session (blue) and a camera session (red), we complete the multi-session SLAM by introducing the inter-session pairwise opti-acoustic factor (green). The raw sonar image is style-transferred and matched against the camera raw image using SuperGlue. The matched feature correspondences are then piped into gtsam back-end to solve for the inter-session constraints. (b) The projection model of sonar and camera.

place recognition and stereo odometry-based 6-DOF pose information. For underwater application, Ozog et al. [21] proposed long-term underwater mapping using generic linear constraints (GLC)-based graph sparsification.

Unfortunately, the abovementioned methods all target the same sensor registration (mostly cameras). Unlike the existing method, we propose heterogeneous sensor registration with opti-acoustic factors. In this paper, we present multi-session SLAM by registering sonar and camera frames from different vehicles in separate sessions. The proposed method has the following key attributes:

- By further developing preliminary results in our workshop paper [5] and combining with SuperGlue [22], we complete the opti-acoustic registration via feature matching between the optical and the style-transferred sonar image.
- We formulate and implement the opti-acoustic measurement as the inter-session pairwise constraint to solve for multi-session SLAM between acoustic and optical sessions.
- We validate the proposed method using the simulation and the real in-water tank test. The proposed opt-acoustic inter-session factor allowed the optical session to register against the acoustic session via successful feature matching.

An overview of the proposed multi-session SLAM appears in Fig. 1(a). Each session includes sonar (session A, blue) and a camera (session B, red). The first session exploits navigation data for odometry and scanning sonar. The second session leverages the same navigation sensors but uses an optical camera.

II. Opti-acoustic Pairwise Factor for Multi-session SLAM

A. Opti-Acoustic Projection Model

Both camera and sonar project a 3D point onto its own image plane following a sensor-specific projection model. We denote P for the 3D point and p for the projected 2D point (Fig. 1(b)). The subscript indicates the sensor modality that describing the points, namely c for the camera, s for the sonar, and w for the world coordinate.

1) Camera Projection Model: We use the pin-hole projection model for camera which projects a 3D point P_c onto the optical image point p_c as

$$p_c = \begin{bmatrix} u \\ v \end{bmatrix} \tag{1}$$

$$s \begin{bmatrix} p_c \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \left[R_w^c | t_w^c \right] \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{2}$$

, while f_x and f_y are camera focal length, c_x and c_y are principal points. R_w^c and t_w^c are rotation matrix and translation matrix from world coordinate to camera coordinate. During this projection we lose scene depth and recovering 3D points from 2D requires the depth value as depth cannot be inferred from a monocular image.

2) Sonar Projection Model: Defining the sonar projection model is straightforward in spherical coordinate as a 2D point in a sonar image represents the range and the azimuth $(\Re, \theta)$. The 3D point P_s written in spherical coordinate is as below.

$$P_s = \begin{bmatrix} X_s \\ Y_s \\ Z_s \end{bmatrix} = \begin{bmatrix} \Re \cos\phi \sin\theta \\ \Re \cos\phi \cos\theta \\ \Re \sin\phi \end{bmatrix}. \tag{3}$$

For sonar, the elevation ϕ is lost in the projection as

$$\mathbf{p}_s = \begin{bmatrix} x_s \\ y_s \end{bmatrix} = \begin{bmatrix} \Re \sin\theta \\ \Re \cos\theta \end{bmatrix} \tag{4}$$

$$\begin{bmatrix} \Re \\ \theta \end{bmatrix} = \begin{bmatrix} \sqrt{X_s^2 + Y_s^2 + Z_s^2} \\ \tan^{-1}(\frac{X_s}{Y_s}) \end{bmatrix}. \tag{5}$$

3) Opti-Acoustic Stereo Model: These lost information (depth in an optical image and elevation in a sonar image), however, can be complimentarily recovered when we leverage the opti-acoustic stereo rig. Following derivation in [6], the scene depth of a camera pixel Z_c can be obtained by the

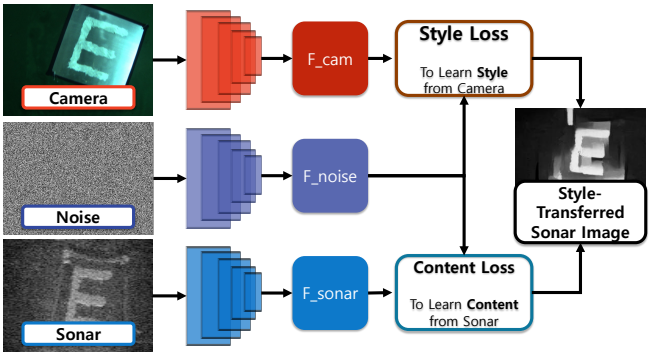


Fig. 2: The overall network for the style-transfer. Main network is VGG19[23], each feature information forms a loss for style and content of the image. Loss functions generate a synthetic image that contains both data.

Range Solution equation below.

$$\left\|\frac{\mathbf{p}}{f}\right\|^2 (Z_c)^2 + \xi_t Z_c + \left(\|t_c^s\|^2 - (\Re)^2\right) = 0 \tag{6}$$

Here, $\mathbf{p} = [x_c, y_c, f]$ is a 3D point that exists on the camera image plane. $\xi_t = \frac{2}{f}(t_c^{s\top} R_c^s \mathbf{p}_c)$ can be computed using transformation from camera to sonar. According to [6], the depth estimation in opti-acoustic epipolar geometry is unstable than other values. We will examine the effect of this depth estimation error on the opti-acoustic factor estimation in §III-B. Using the computed depth, 3D camera points are then generated from which we compute reprojection error. We briefly re-state the reprojection derivation from [6] for readibility. The projection between two sensors can be written as

$$\begin{bmatrix} x_s' \\ y_s' \end{bmatrix} = \begin{bmatrix} \frac{(Z_c/f)\mathbf{r}_1 \cdot \mathbf{p}_c + t_x}{\sqrt{1-(Z_c/\Re(\mathbf{r}_3 \cdot \mathbf{p}_c/f)+t_z/\Re)^2}} \\ \frac{(Z_c/f)\mathbf{r}_2 \cdot \mathbf{p}_c + t_y}{\sqrt{1-(Z_c/\Re(\mathbf{r}_3 \cdot \mathbf{p}_c/f)+t_z/\Re)^2}} \end{bmatrix} \tag{7}$$

and

$$\begin{bmatrix} x_c' \\ y_c' \end{bmatrix} = \frac{f}{\mathbf{r}_3' \cdot (P_s - t)} \begin{bmatrix} \mathbf{r}_1' \cdot (P_s - t) \\ \mathbf{r}_2' \cdot (P_s - t) \end{bmatrix}, \tag{8}$$

with $t = [t_x, t_y, t_z]$, and $\mathbf{r}_i$ indicates the i^{th} row of the rotation matrix R_c^s, $\mathbf{r}_i'$ for R_s^c. We combine these two reprojections into the cost to minimize and solve for the relative pose between sonar and camera in §II-C.1. Before introducing the cost, we first describe how to establish the correspondences.

B. Style-Transferred Sonar-Camera Matching

Even when observing the same object, resulting images reveals a substantial difference in terms of resolution and noise characteristic. This large visual discrepancy between sonar and optical images is the key challenge in opti-acoustic matching severely deteriorating the reliable matching. In the existing studies, manual correspondences were used in [6] or only a few successful cases were presented by using robust features [7].

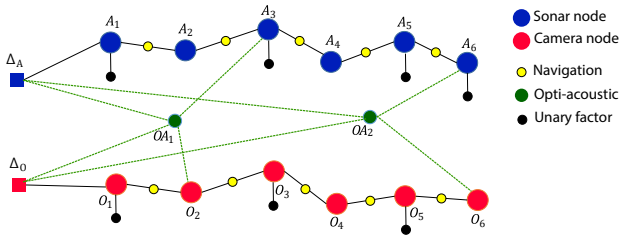


Fig. 3: Multi-session pose-graph. Sonar session (blue) and camera session (red) are constrained via pairwise opti-acoustic factor (green).

To enhance feature matching performance while generating clear image, we decided to improve noisy sonar images. Developing from our preliminary work [5], we aimed to overcome image-level discrepancy and secure robust matching. In doing so, we constructed CNN-based style transfer structure [14] preparing camera-like sonar image (Fig. 2). Unlike [17], we instead chose style-transfer to convert the style of sonar to the optical image due to the scanty underwater opti-acoustic dataset. Using the style-transferred sonar image, we apply SuperGlue to establish matching.

1) Style-Transferred Sonar Image Preparation: The base for our network is the pre-trained model VGG-19 [23] from which we extract high dimension features F. Style information named Gram matrix G_{ij}^l is defined with the inner product of feature maps i, j in layer l.

$$G_{ij}^l = \sum_k F_{ik}^l F_{jk}^l \tag{9}$$

Both style and content factors became a loss function for styled image generation. We use a noise image that has style factor S_{ij}^l and content factor C_{ij}^l. Our network minimizes style loss and content loss that generates optical-styled sonar image. For feature map number N_l, size M_l and weight factor w_l,

$$L_{style} = \sum_l \frac{w_l}{N_l^2 M_l^2} \sum_{i,j} (G_{ij}^l - S_{ij}^l)^2 \tag{10}$$

$$L_{content} = \sum_{i,j} (F_{ij}^l - C_{ij}^l)^2 \tag{11}$$

and the total loss becomes the weighted sum of these two losses.

$$L_{total} = \alpha L_{style} + \beta L_{content} \tag{12}$$

In this paper $\alpha = 1.0$ and $\beta = 10^{-6}$ were used as empirical conclusion. Using this total loss, we learn style from optic and content from the sonar image, generating an optical-like image.

2) Sonar-to-Optical Feature Matching: For the matching between optical image and styled sonar image, we adopted SuperGlue [22] using SuperPoints. As will be seen in §III-A using SuperGlue for raw sonar-camera matching is still limited without style-tranfer. Comparison to other image processing approaches will be given in §III-A.

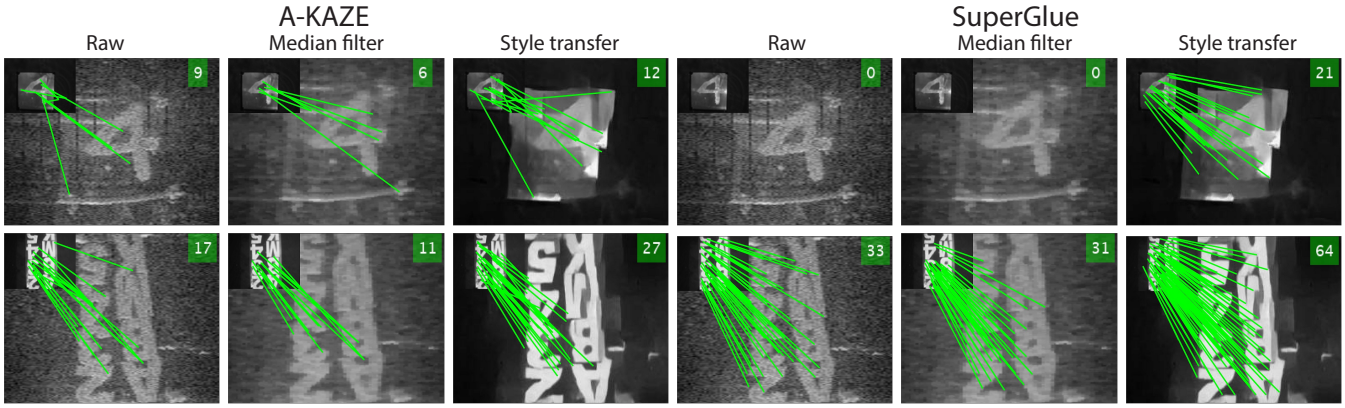


Fig. 4: Feature-matching results depending on the image-processing and matching engine. The target markers with the least matching (top) and the most matching (bottom) are presented as examples. We compared the matching results among various image-processing and feature-matching methods. The left three columns are the results from running A-KAZE as the matching algorithm. The right three columns are from SuperGlue. On the upper-left corner, corresponding optical images are provided at a reduced size to show the obtained matching. The upper-right corner shows the found number of matches. For visualization, we cropped the sonar image showing the zoomed view where the contents are concentrated.

C. Multi-session Pose-graph SLAM

Given the established correspondences, we estimate the relative motion between the camera and sonar. This pairwise inter-session measurement is then piped into the multi-session SLAM (Fig. 3).

1) Opti-acoustic Pairwise Measurement: Using (7) and (8), we can define the cost function by minimizing the reciprocal reprojection error and optimize for the relative transformation between sonar and camera

$$\begin{aligned} h(O_i, A_i) &= [R^*, t^*] \\ &= \underset{R,t}{\operatorname{argmin}} \sum_i \left(\left\| \begin{bmatrix} x_c \\ y_c \end{bmatrix} - \begin{bmatrix} x_c' \\ y_c' \end{bmatrix} \right\|^2 + \left\| \begin{bmatrix} x_s \\ y_s \end{bmatrix} - \begin{bmatrix} x_s' \\ y_s' \end{bmatrix} \right\|^2 \right). \end{aligned} \tag{13}$$

This relative transformation defines the relation between sonar and camera frames in each session. Because sonar and camera are rigidly mounted on vehicles, we can infer relative vehicle motion across the session.

2) Multi-session SLAM with OA Factor: For two sessions, sonar (A) and camera (O) sessions, underlying probabilistic distribution can be written as

$$P(X) \propto \prod_t f(A_{t-1}, A_t) \prod_t f(O_{t-1}, O_t) \prod_n g(O_n, A_n), \tag{14}$$

while factor $f(A_{t-1}, A_t)$ and $f(O_{t-1}, O_t)$ are odometry between time t-1 and t in sonar and camera session. The probability for opti-acoustic factor $g(O_i, A_i)$ is derived from (13) assuming Gaussian measurment noise.

$$g(O_i, A_i) \propto \exp\left(-\frac{1}{2} \left\| h(O_i, A_i) - z_i \right\|^2_{\Sigma_i} \right) \tag{15}$$

Where, z_i for measurement. For the SLAM back-end, we adopt iSAM [24] to solve the maximum *a posteriori* (MAP) estimation.

$$\begin{aligned} X^* &= \arg\max_X P(X) = \operatorname*{argmin}_X (-\log P(X)) \\ &= \operatorname*{argmin}_X \sum_i \| r_{A_i} \|^2_{\Sigma_A} + \sum_i \| r_{O_i} \|^2_{\Sigma_O} \\ &\quad + \sum_n \| h(O_n, A_n) - \chi \|^2_{\Sigma_c} \end{aligned} \tag{16}$$

r_{A_i} and r_{O_i} are residuals from acoustic and optical odometry factors.

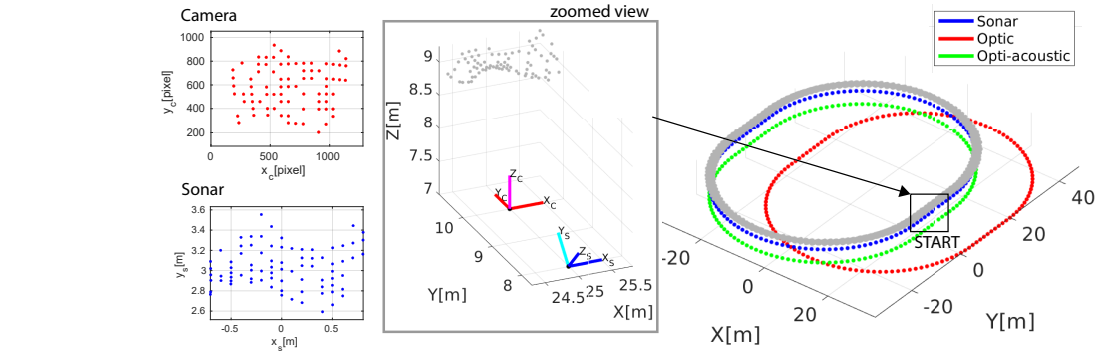
III. Experimental Results

We evaluate the proposed method using both simulation and real tank tests.

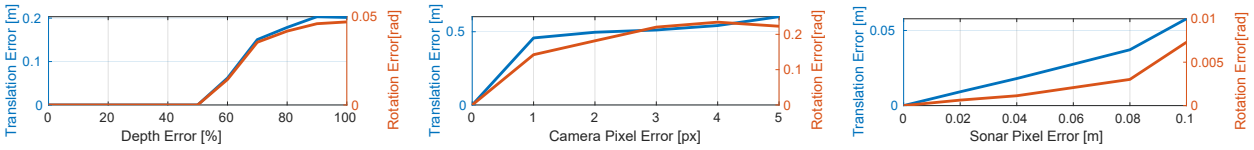
A. Opti-acoustic Feature Matching

First, we evaluate feature matching between sonar and optical images. A widely adopted solution in existing studies is to use histogram equalization for optical images [2] and median filtering for sonar images [3] in order to reduce the noise and increase the discriminability. Both image-preprocessing and feature-matching algorithms are important for robust matching between optical and sonar images. In this paper, the water condition of the experiment is clear, so that optical image processing was unnecessary. Also, we empirically verified that median filter is more appropriate than histogram equalization for sonar images. Thus, we compared the performance of sonar image processing results: raw sonar images, median filter, and our optic-style transfer. For matching, we compared SuperGlue with SuperPoint and Accelerated-KAZE (A-KAZE). A-KAZE has reliable performance for sonar images [25]. Unlike Scale Invariant Feature Transform (SIFT), A-KAZE constructs a nonlinear scale-space that preserves the contents in sonar images while reducing noise.

The feature-matching results can be seen in Fig. 4. For all of the cases, the matched number of features increased when



(a) Simulation setup and trajectory



(b) Depth error (c) Camera pixel error (d) Sonar pixel error

Fig. 5: (a) In the simulation, the vehicle was designed to follow an oval trajectory counterclockwise. On the left, the projected pixel points are illustrated for the camera (red) and the sonar (blue) sensor. The middle column shows the zoomed view of the starting point. The gray points are 3D feature points with two coordinate frames showing the camera and sonar sensor configurations. Two frames belonged to a different vehicle in a separate session. The estimated trajectories are given on the right. Initially, the camera session started from an arbitrary point but was successfully corrected after matching against the sonar session. (b)-(d) Effects of the estimated pose error when the initial depth, camera pixel, and sonar pixel contained an error, respectively.

the sonar image was style-transferred, whereas the median filter showed little improvement. SuperGlue outperformed A-KAZE by a large margin in terms of both quality and quantity. Specifically, as upper right three images in Fig. 4 depict, opti-acoustic style-transferred image overcome the limitation that SuperGlue fails to find the feature point in raw image and filtered image. The correspondences included outliers and were refined via geometric verification and confidence-level thresholding. We chose to apply SuperGlue to the style-transferred sonar images for camera-sonar matching and secured matching for all types of targets.

B. Simulated Environment

We validated the proposed pairwise inter-session constraint in a simulated environment. In the simulation, the correspondences are exactly known, in order to investigate the effects of pixel noise and the depth-estimation error on the inter-session pose estimation. In the test scenario, the sonar session was completed, which was followed by the camera mission. In an underwater environment, the camera session's starting pose cannot be known globally unless an underwater positioning system is equipped. Thus, we set the camera session's initial pose arbitrarily and leveraged opti-acoustic registration to combine two sessions. As shown Fig. 5, the two sessions were well aligned after the loop-closures were successfully registered across the sessions. Using this simulated environment, we examined the effects of pixel errors in sonar/optical images and the depth-estimation error on the final relative pose estimation. Pose-estimation error drastically increases when the depth-estimate error exceeds 50% of the true scene depth. Both optical and sonar pixel errors affected the pose estimation critically, as reported in Fig. 5(c) and Fig. 5(d).

C. Real Underwater Tank Test

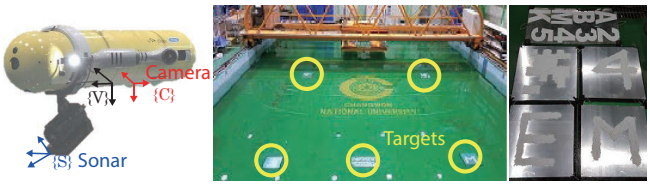
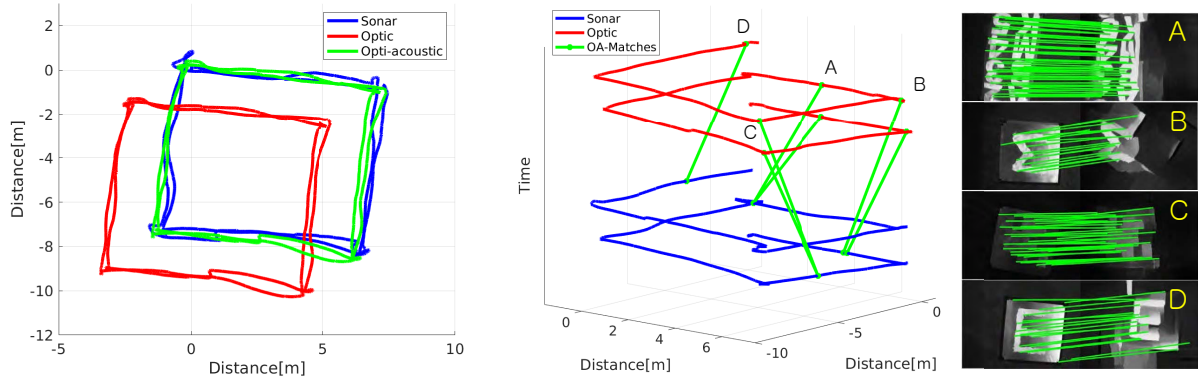


Fig. 6: The water tank test. The vehicle (left) is equipped with sonar and camera. The water tank test includes five markers on each corner and on a track. Marker deployment (middle) and sample markers (right).

The last validation was a real underwater tank test. The vehicle trajectory was designed to follow a 7 m by 7 m square route with a marker placed on each corner of the square. The sensor configuration, test setup, and markers are shown in Fig. 6. A forward-looking Dual frequency IDentification SONar (DIDSON) was installed in the front of the vehicle in a tilted configuration. The angle of the sonar sensor was 30° from the water surface. Sonar azimuth FOV is 14.4°, and elevation FOV is 7°. A down-pointing Point



(a) Top view trajectory

(b) Time elevation graph

Fig. 7: (a) Real in-water tank test. The first sonar session (blue) was followed by the second camera session (red). The initial pose of the camera session was set arbitrarily, showing an offset from the sonar session. The opti-acoustic factor connected two sessions, with the inter-session factors (the green lines in (b)) aligning the corrected camera session (green) to the sonar session (blue). (b) The time-elevation graph with the z-axis representing the mission time. The inter-session measurements are with in green lines, with the matching samples on the right.

Grey camera was mounted behind the vehicle center. Camera focal length $[f_x, f_y] = [1717.61, 1722.09]$, with image pixels $[1380, 1024]$.

In the test, the vehicle traveled the square path four times with both sensors mounted. To test the multi-session scenario with sonar-camera matching, we separated the mission into two sub-missions. We assumed the first half solely relied on sonar, and the remaining mission only exploited the camera. In a word, we obtained full data for all rounds but utilized only half for each mission. Although the vehicle possessed full data, we eliminated camera data for the first two rounds, and sonar data for the latter part. Instead, full data were utilized only for error evaluation. To evaluate the separate multi-session case, we assumed that no exact start pose was known for the camera session by initiating the pose graph at an arbitrary pose, similarly as in the simulation. Thus, the camera session initially showed an offset from the sonar session without an opti-acoustic factor.

The corresponding SLAM results are depicted in Fig. 7. Although the optical session started at a different point from the sonar session, the opti-acoustic factor adjusted the camera session trajectory with respect to the sonar session. Successfully matched feature points are provided in Fig. 7 for each opti-acoustic factor. Due to the large FOV difference, the associated vehicle poses for matched pair reveals horizontal displacement. Specifically, the loop-closures in the time elevation graph (i.e., green lines in Fig. 7) is slanted rather than vertical even registered over the same object. This slanted loop-closures indicate the FOV difference between camera and sonar. Overall, the constrained two sessions retained consistent trajectories.

Although no ground truth trajectory was given in the test, we can evaluate the effects of the opti-acoustic factor quantitatively by using the original sonar-based trajectory as the baseline. Because the two sessions originally belonged to the same mission as mentioned, we can evaluate the camera session's accuracy by using the original trajectory as the baseline. As listed in Table I, the error in the camera trajectory was large because the initialization occurred at an arbitrary point. The opti-acoustic factor stitched the optic session consistently with the sonar session, yielding a smaller error from the baseline trajectory.

	w/o OA factors	w/ OA factors
RMSE [m]	2.0161	0.2917

TABLE I: RMSE to the baseline trajectory. Using opti-acoustic factors, the error significantly reduced and consistent to the sonar session.

IV. Conclusion

In this work, we presented an underwater multi-session SLAM method that can overcome different sensor modalities among vehicles. In doing so, we introduced a style-transferred sonar image and an opti-acoustic pairwise factor between sessions registering optical images over sonar images. In future work, we plan to extend the style-transfer to less artificial objects and consider a general underwater application.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (2020R1C1C1006620)

References

[1] F. S. Hover, R. M. Eustice, A. Kim, B. Englot, H. Johannsson, M. Kaess, and J. J. Leonard, "Advanced perception, navigation and planning for autonomous in-water ship hull inspection," *International Journal of Robotics Research*, vol. 31, no. 12, pp. 1445–1464, Oct. 2012.

[2] A. Kim and R. M. Eustice, "Real-time visual SLAM for autonomous underwater hull inspection using visual saliency," *IEEE Transactions on Robotics*, vol. 29, no. 3, pp. 719–733, Jun. 2013.

[3] H. Johannsson, M. Kaess, B. Englot, F. Hover, and J. Leonard, "Imaging sonar-aided navigation for autonomous underwater harbor surveillance," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 4396–4403.

[4] J. J. Youngji Kim and A. Kim, "Stereo camera localization in 3D LiDAR maps," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2018, pp. 1–9.

[5] H. Jang, G. Kim, Y. Lee, and A. Kim, "CNN-based approach for opti-acoustic reciprocal feature matching," in *ICRA Workshop on Underwater Robotics Perception*, May. 2019.

[6] S. Negahdaripour, H. Sekkati, and H. Pirsiavash, "Opti-acoustic stereo imaging: On system calibration and 3-d target reconstruction," *IEEE Transactions on Image Processing*, vol. 18, no. 6, pp. 1203–1214, 2009.

[7] D.-H. Gwon, Y.-S. Shin, Y. Kim, A. Kim, Y. Lee, and H.-T. Choi, "Nontemporal relative pose estimation for opti-acoustic bundle adjustment," in *Proceedings of the IEEE/MTS OCEANS Conference and Exhibition*, 2016, pp. 1–5.

[8] S. Rahman, A. Q. Li, and I. Rekleitis, "Sonar visual inertial slam of underwater structures," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2018, pp. 1–7.

[9] ——, "Svin2: An underwater slam system using sonar, visual, inertial, and depth sensor," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019, pp. 1861–1868.

[10] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson, "Watergan: unsupervised generative network to enable real-time color correction of monocular underwater images," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 387–394, 2018.

[11] Y. Cho, H. Jang, R. Malav, G. Pandey, and A. Kim, "Underwater image dehazing via unpaired image-to-image translation," *International Journal of Control, Automation and Systems*, vol. 18, no. 3, pp. 605–614, 2020.

[12] M. Sung, H. Cho, T. Kim, H. Joe, and S.-C. Yu, "Crosstalk removal in forward scan sonar image using deep learning for object detection," *IEEE Sensors Journal*, vol. 19, no. 21, pp. 9929–9944, 2019.

[13] X. Wang, J. Jiao, J. Yin, W. Zhao, X. Han, and B. Sun, "Underwater sonar image classification using adaptive weights convolutional neural network," *Applied Acoustics*, vol. 146, pp. 145–154, 2019.

[14] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[16] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[17] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. V. Gool, "Night-to-day image translation for retrieval-based localization," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2019, pp. 5958–5964.

[18] B. Kim, M. Kaess, L. Fletcher, J. Leonard, A. Bachrach, N. Roy, and S. Teller, "Multiple relative pose graphs for robust cooperative mapping," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2010, pp. 3185–3192.

[19] K. Konolige, J. Bowman, J. Chen, P. Mihelich, M. Calonder, V. Lepetit, and P. Fua, "View-based maps," *International Journal of Robotics Research*, vol. 29, no. 8, pp. 941–957, 2010.

[20] J. McDonald, M. Kaess, C. Cadena, J. Neira, and J. J. Leonard, "Real-time 6-dof multi-session visual slam over large-scale environments," *Robotics and Autonomous Systems*, vol. 61, no. 10, pp. 1144–1158, 2013.

[21] P. Ozog, N. Carlevaris-Bianco, A. Kim, and R. M. Eustice, "Long-term mapping techniques for ship hull inspection and surveillance using an autonomous underwater vehicle," *Journal of Field Robotics*, vol. 33, no. 3, pp. 265–289, 2016.

[22] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4938–4947.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] M. Kaess, A. Ranganathan, and F. Dellaert, "isam: Incremental smoothing and mapping," *IEEE Transactions on Robotics*, vol. 24, no. 6, pp. 1365–1378, 2008.

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