Qualitative evaluation of state-of-the-art DSO and ORB-SLAM-based monocular visual SLAM algorithms for underwater applications

Juliette Drupt, Claire Dune, Andrew I. Comport, Vincent Hugel

Qualitative evaluation of state-of-the-art DSO and ORB-SLAM-based monocular visual SLAM algorithms for underwater applications

Juliette Drupt[1], Claire Dune[1], Andrew I. Comport[2], Vincent Hugel[1]

***Abstract*— Visual simultaneous localization and mapping (VSLAM) is widely investigated for airborne applications, but fewer works focus on underwater VSLAM. Previous studies of state-of-the-art VSLAM in the underwater field demonstrate that while some stereo approaches are robust to underwater visual conditions, monocular ones still lack robustness to such case. The only monocular VSLAM system able to give partial but promising results in these studies are DSO and ORB-SLAM, but these methods are still limited by tracking inconsistencies or failures from which the SLAM system fails to recover. However, recent work extend the capabilities of these approaches in place recognition and tracking failure recovery. These new developments should therefore lead to better performance in underwater conditions. This paper presents an update of previous qualitative assessments by adding recent developments of monocular DSO and ORB-SLAM. The methods are evaluated in 8 underwater scenarios, considering three criteria: the percentage of the sequence for which a localization is estimated, loop closure detection success and map and trajectory consistency. The results show the interest of multi-map approaches, namely ORB-SLAM3, in improving significantly SLAM robustness to underwater challenging visual conditions.**

I. Introduction

GNSS positioning is not available to underwater robots because of the absorption of electromagnetic waves in the first centimeters of the water column. In such GNSS-denied environments, robotic systems strongly rely on their proprioceptive sensors to estimate their location. Most underwater robots embed cameras, which are low cost, light sensors able to provide rich information about their surroundings. Visual simultaneous localization and mapping (VSLAM) can thus be a solution for underwater localization. Underwater conditions are, though, particularly challenging, because of selective color absorption, backscattering and suspended particles (Fig. 2-9). In addition, the embedded lights required for deep sea missions invalidate the lambertzian assumption of airborne VSLAM. Lastly, most airborne VSLAM systems are designed for highly structured urban or industrial environments, whereas underwater vessels often operate in natural, less structured environments.

Recent studies proposed evaluations of state-of-the-art opensource VSLAM algorithms on underwater datasets [1], [2], [3], demonstrating that while some stereo VSLAM approaches are robust to underwater visual conditions, none of the tested monocular VSLAM is able to process underwater sequences in the general case. Only two methods manage to initialize and produce partial results: ORB-SLAM [4] and DSO [5]. Their main limitations are initialization difficulties, lack of robustness in the tracking process, leading to SLAM failure, from which the system does not recover. However, these two approaches have been recently extended in ways that may improve their performances. On the one hand, LDSO [6] and DSM [7] extend DSO with loop closing functionalities, which should lead to better map point triangulation and are thus expected to reduce the risk of SLAM failure. On the other hand, Dual-SLAM [8] and ORB-SLAM3 [9] both build on ORB-SLAM and implement robust tracking loss recovery scenarios which handles relocalization failure.

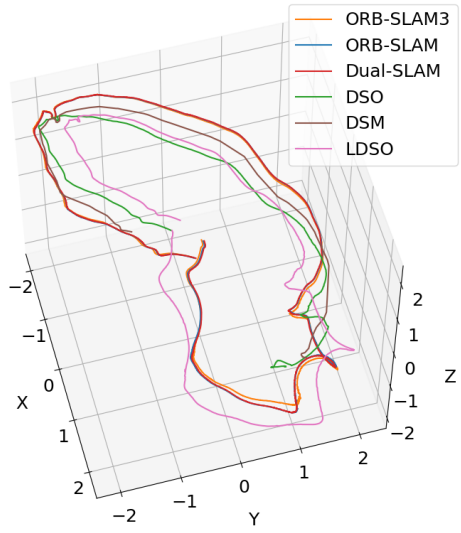


Fig. 1: Trajectories estimated by all evaluated SLAM systems on the *Aqualoc Archaeo* dataset. Trajectories are given with an unknown scale. They are aligned by a Umeyama $Sim(3)$ alignment with respect to the most complete trajectory, namely the one given by ORB-SLAM3. One can see the drift of DSO-based approaches (DSO, LDSO, DSM) compared to ORB-SLAM-based ones (ORB-SLAM, ORB-SLAM3, Dual-SLAM). Note that ORB-SLAM and Dual-SLAM estimated trajectories completely overlap.

The present work aims at completing the previous studies by the underwater evaluation of LDSO, DSM, Dual-SLAM and ORB-SLAM3. In line with [3], which shows that monocular VSLAM is so difficult on underwater datasets that a qualitative evaluation is sufficient as a first performance characterization step, this work only focuses on a qualitative analysis. Related works are presented in Section II, with a more detailed description of the evaluated approaches.

[1]COSMER Laboratory EA7398, Université de Toulon, France
[2]CNRS I3S Laboratory, Université Côte d'Azur, Sophia Antipolis, France

The evaluation methodology is exposed in Section III-B, including evaluation criteria and datasets. Evaluation results are presented and discussed in Section IV, leading to a conclusion in Section V.

II. Related work

The visual SLAM problem can be decomposed in two functionalities: visual primitive tracking from an image to another, leading to the estimation of an inter-frame motion, and mapping. In line with seminal work [11], these functionalities are commonly run on concurrent threads in order to conduct both localization and map building simultaneously. The common map representation consists in a graph of KeyFrames (KF) connected under covisibility criteria. VSLAM approaches can be classified according to several criteria:

- *direct* approaches, which minimize the photometric error between frames, and *indirect* ones, which use higher level features extracted from the images
- *dense* methods, which use all image points, and *sparse* ones, which rely only on a selection of them
- *loop closing* capability, where methods without this functionality are commonly denoted *visual odometries*
- the presence of *tracking failure handling* functionalities.

ORB-SLAM [4] strongly impacted VSLAM by introducing a publicly available real-time monocular, indirect, sparse VSLAM framework based on ORB features and able to perform tracking, local mapping in KF local window, relocalization in case of tracking failure and loop detection and closing, with an outbreaking accuracy. Relocalization and loop detection were performed by a BoW place recognition module based on DBoW2 [10]. Relying on relocalization as a SLAM recovery scenario is, though, limited. Depending on the visual conditions and on the system getting out of the already mapped area, relocalization may never succeed, or at least lead to important time gaps without localization estimation, which can be critical for real-life applications. In underwater VSLAM evaluations, ORB-SLAM is reported as being robust to underwater visual conditions but subject to initialization difficulties and critical relocalization failure after tracking loss [1], [2], [3]. Recent works increment ORB-SLAM with more robust SLAM failure recovery strategies. Dual-SLAM [8] and ORB-SLAM3 [9] both rely on new map creation and multi-map fusion. In case of tracking loss, Dual-SLAM initializes a new map and tries to fuse it with the previous one by running a backward SLAM. ORB-SLAM3 implements ORB-SLAM Atlas [12], which also initializes a new map in case of tracking loss to keep the SLAM running, but implements a different map fusion strategy. All old maps are stored as disconnected entities. The loop closure place recognition queries all maps. Two matched maps are then merged similarly to loop closure optimization. In addition, all maps can be used for relocalization.

Alongside with these approaches, DSO (Direct Sparse Odometry) [5] is a fully direct approach which minimizes the photometric error between a selection of pixels located along the image contour, in a sliding window. The main asset of direct methods is that they can cope with poorly textured environment and are robust to blur. DSO is however limited by the absence of place recognition functionalities for loop closing and map reuse *via* relocalization, leading respectively to map inconsistency and a lack of robustness to bad data association, and to critical SLAM failure in case of tracking loss. These observations are reported on underwater evaluations in [3]. Whereas no work addresses the problem of extending DSO with a tracking failure recovery strategy, recent works extend it with loop closure handling. LDSO (Loop Closure Direct Sparse Odometry) [6] proposes an indirect loop closure handling strategy relying on an ORB points map representation and a DBoW2 [10] place recognition, similarly to ORB-SLAM [4]. The DSO front-end is modified to include some pixels with characteristic ORB features.

DSM (Direct Sparse Mapping) [7] presents a very different photometric place loop closure strategy. Similarly to DSO and LDSO, DSM uses a window of map keyframes for current pose estimation refinement. While in DSO and LDSO this keyframe window only includes recent keyframes within a time window, denoted *temporal keyframes*, DSM tries to find in the map older keyframes with complementary viewing informations, denoted *covisible keyframe* in addition to the *temporal* ones, and computes a photometric bundle adjustment over the resulting keyframe window. *Covisible keyframe* selection is based on a guided search over the keyframe map which tracks the projection of current map points in older keyframes. Consequently, this selection strongly relies on an accurate pose prior. By querying the whole map, this process is designed to detect and handle short to long term loop closures.

The present work focuses on the evaluation of these recent monocular ORB-SLAM and DSO based developments on underwater datasets. ORB-SLAM3 and Dual-SLAM are expected to allow the SLAM to recover more efficiently from tracking loss than ORB-SLAM, whereas LDSO and DSM are expected to produce less tracking loss than DSO due to better mapping performances through long term data association handling. Table I recaps all the methods compared and evaluated in the present work, namely ORB-SLAM, ORB-SLAM3, Dual-SLAM, DSO, LDSO and DSM.

III. Methodology

A. Datasets

Airborne VSLAM evaluation can rely on standard, public datasets recorded in different environments, featuring several sequences in the similar environments and visual conditions with various trajectories of gradual difficulty [13], [14], [15]. However, there is no equivalent in the underwater field at the time of writing, because of the important cost and resources required for acquiring such data. Previous works on VSLAM benchmark under underwater conditions released their evaluation datasets, which are composed of

TABLE I: Evaluated methods

Method	Front-end	Loop closure	Relocalization	Relocalization failure handling
ORB-SLAM [4]	indirect, sparse	DBoW2	DBoW2	No
ORB-SLAM3 [9]	indirect, sparse	DBoW2	DBoW2	New map initialization. DBoW2 for map matching and merging.
Dual-SLAM [8]	indirect, sparse	DBoW2	DBoW2	New map initialization. Backwards SLAM for map matching and merging.
DSO [5]	direct, sparse	No	No	No
LDSO [6]	direct, sparse	DBoW2 [10]	No	No
DSM [7]	direct, sparse	direct, by projecting map points according to the estimated pose	No	No

heterogeneous sequences recorded in completely different environments from one to another, with various lighting conditions and camera settings [1], [3]. Such heterogeneous datasets are particularly interesting for comparing VSLAM methods under very different conditions, but are not suitable for a detailed evaluation under specific conditions. [16] released AQUALOC, an underwater visual-inertial-pressure dataset. Similarly to standard aerial datasets, it is composed of several gradually more difficult sequences recorded in similar environments, on three different marine sites. All theses sequences, however, show quite similar conditions by featuring man-made objects lying on a planar sandy area and involving only slow camera motion As a result, the AQUALOC dataset only represents a small portion of the wide variety of underwater environments and visual conditions.

The generation of ground truth trajectories relative to underwater datasets is more difficult than for aerial datasets. Whereas airborne datasets' ground truth commonly rely on laser scans or, sometimes, motion capture systems in smaller scale indoor environments, such systems are not available in the sea. In [3], the output trajectory of a visual-inertial-SONAR-depth SLAM [17] is used as a reference for visual and visual-inertial SLAM evaluation, but this can only apply to data acquired with a very specific sensor system. In AQUALOC [16], the offline Structure-from-Motion Colmap is used to compute a reference trajectory. These two strategies assume that the use of more sensors or time and computational resources will lead to a more reliable state estimation than real-time visual-only SLAM. While the absence of a ground truth only allows a coarse comparison between VSLAM approaches, it has been shown that such qualitative evaluations are already sufficient to discriminate most monocular VSLAM works in underwater fields [3]. This is why the present work only considers a qualitative evaluation. In addition, the no-need in ground truth allows diversifying the test sequences.

The present work aims at qualitatively characterizing the performances of VSLAM methods in the underwater field in the general case. Therefore, evaluations are conducted on a selection of eight datasets chosen to have different environments and visual conditions.

The *Bus* dataset [3] (Fig. 2) is recorded in quite turbid water, by a forward facing RGB camera. The camera slowly turns around a sunken bus, hence several loop closures. The camera enters inside the bus during a small part of the trajectory (Fig. 2d), and one side of the bus is poorly illuminated (Fig. 2c). Therefore, this sequence features important visual conditions variations along the camera's trajectory, which is the main difficulty of this sequence.

The *Cave* dataset [3] (Fig. 3) is recorded in an underwater cave, with an embedded light source. It shows a natural mineral-only environment in clear water. The RGB camera is facing forward, and its motion is slow. The sequence includes several loop closures.

The *A/In* dataset [1] (Fig. 4) is recorded inside a shipwreck, with a forward facing RGB camera. Environment structure is thus closer to standard airborne indoor VSLAM evaluation datasets. The sequence mainly consists in a forward travelling and does not include any loop closure. Water is globally clear, but parts of the sequence feature suspended particles and fishes.

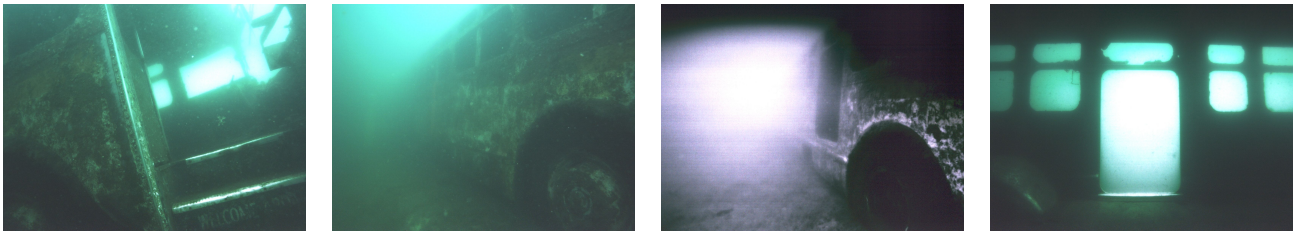
The *A/Out* dataset [1] (Fig. 5) shows a coral reef, including some mobile elements like seaweeds, fishes and suspended particles. The RGB camera is facing forward. No loop closure is included.

The *Aqualoc Harbor #01* dataset (Fig. 6) is recorded with a downwards looking grayscale fisheye camera, and shows large man-made objects lying on the sand. The sequence includes a loop closure, which is marked by an apriltag target (Fig. 6a).

The *Aqualoc Archaeo #09* dataset (Fig. 7) involves a grayscale camera which is slightly tilted downwards. The sequence features amphora hills with high texture, but also low textured sandy areas (Fig. 7c). Images show turbidity and backscattering. The sequence includes loop closures. Both *Aqualoc Harbor #01* and *Aqualoc Archaeo #09* use an embedded light source.

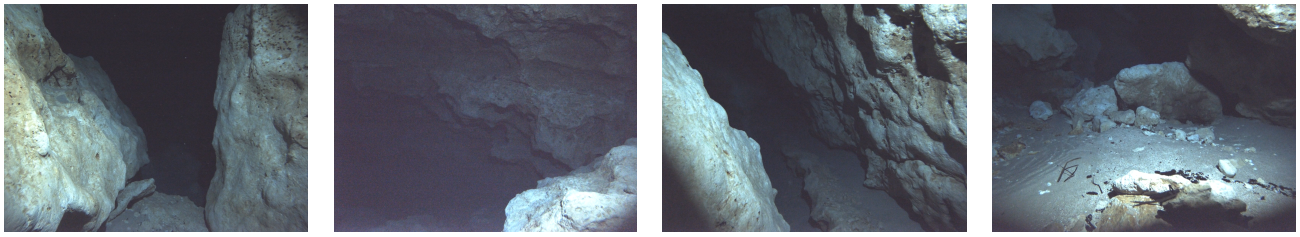
Both *Cephismer* and *Saint-Raphael* datasets are new datasets recorded by the embedded RGB camera of a BlueROV2, in challenging visual conditions. The *Cephismer* dataset (Fig. 8) is recorded in a pool. The camera is slightly tilted downwards, and a small portion of the housing appears in its field of view. The sequence features fast motion, including pure rotations, around submarine spare parts. It is recorded at a low frame rate, with the camera sometimes facing poorly textured areas. This dataset is thus particularly challenging. It also includes several loop closures.

In the *Saint-Raphael* dataset (Fig. 9), the ROV's camera is



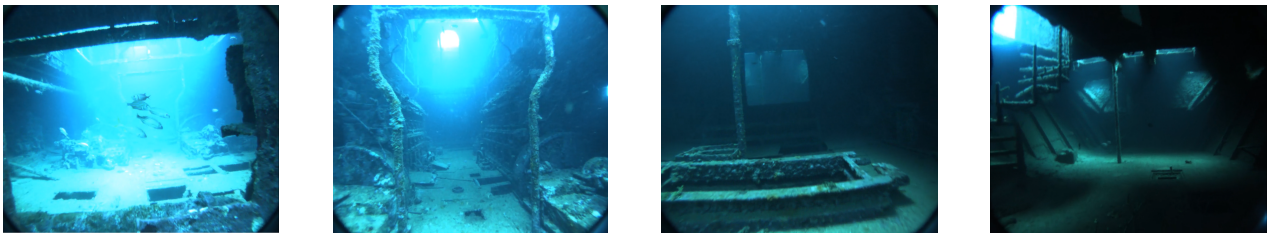
(a) (b) (c) (d)

Fig. 2: *Bus* dataset [3]



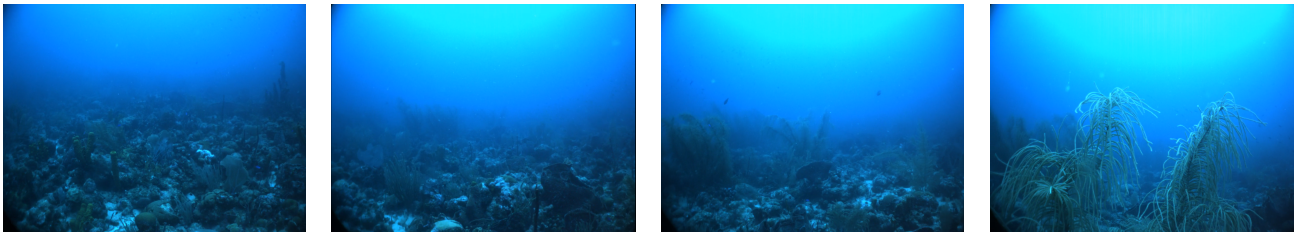
(a) (b) (c) (d)

Fig. 3: *Cave* dataset [3]



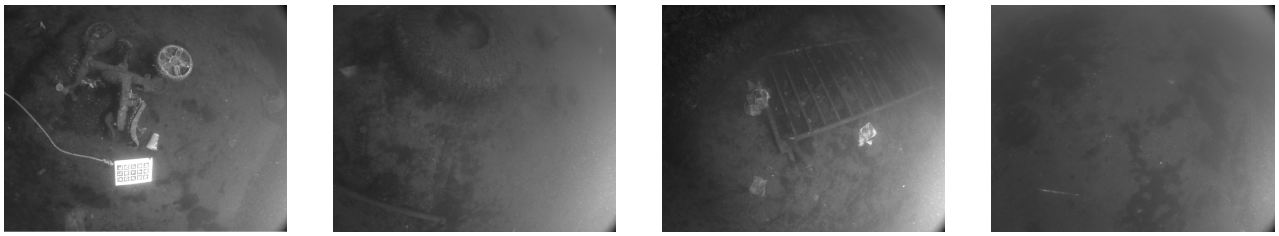
(a) (b) (c) (d)

Fig. 4: *A/In* dataset [1]



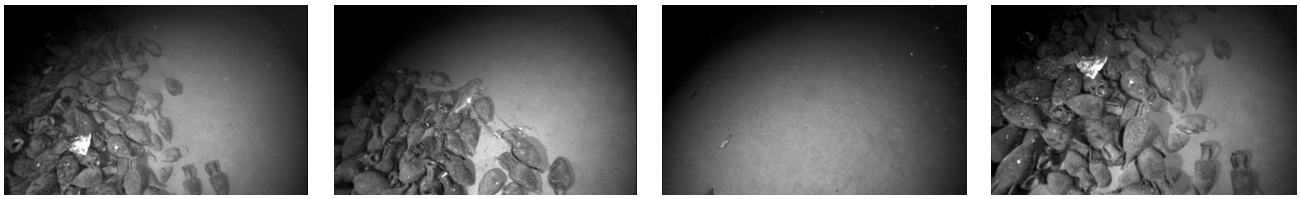
(a) (b) (c) (d)

Fig. 5: *A/Out* dataset [1]



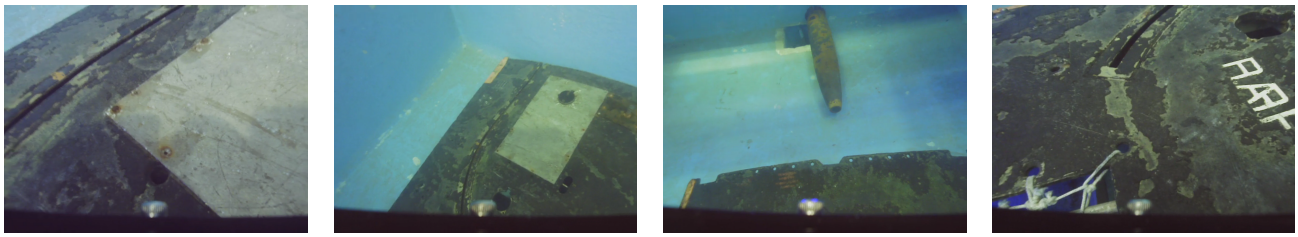
(a) (b) (c) (d)

Fig. 6: *Aqualoc Harbor (AH) #01* dataset [16]



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

Fig. 7: *Aqualoc Archaeo (AA) #09* dataset [16]



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

Fig. 8: *Cephismer* dataset

facing forward. The sequence is recorded at shallow depth, in the Mediterranean Sea, in turbid water. The sequence includes fast motions and pure rotations, and the camera sometimes happens to face plain water. Several loop closures are included. This dataset is also very challenging.

The main characteristics of these eight datasets are recapped in Tab. II.

B. Evaluation method

Sequences are processed with all evaluated VSLAM approaches, namely ORB-SLAM [4], ORB-SLAM3 [9], Dual-SLAM [8], DSO [5], LDSO [6] and DSM [7]. Since ORB-SLAM, Dual-SLAM and DSM do not support fisheye camera models, these three approaches are not evaluated on the *Aqualoc Harbor #01* dataset. When possible, evaluations are conducted in real-time conditions, using ROS middleware. The only approaches that do not implement real-time processing are LDSO and DSM. The evaluation of these two methods are therefore non-real-time in the present work.

The parameters of each VSLAM method are tuned manually for each dataset, following all available documentation provided by the authors. For each VSLAM approach, the following criteria are evaluated:

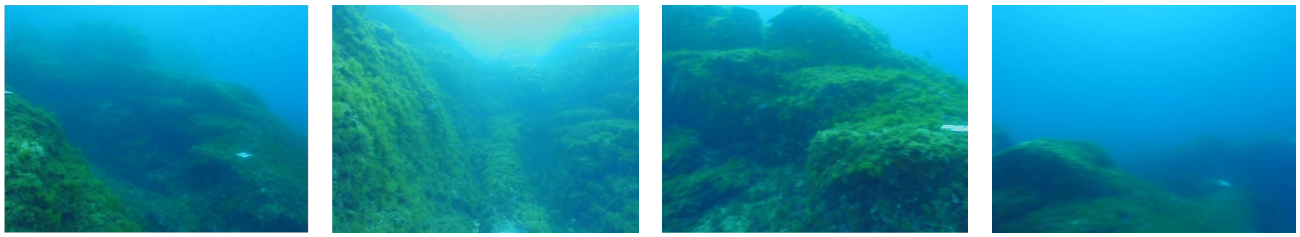
- *Ability to track the complete trajectory*, evaluated by the percentage of the sequence duration for which a localization is computed. One can notice that this does not take into account the reliability of the estimated pose.
- *Loop closure detection and handling capability*, evaluated qualitatively by a color mark which indicates wether the SLAM system manages to detect and process the main sequence's loops (*green*), only a few of them (*yellow*), or none of them (*red*).
- *Localization and mapping consistency*, also evaluated qualitatively by a color mark. A *green* one means that the SLAM outputs consistent trajectory and map on almost all the sequence duration. A *yellow* one indicates consistent trajectory and map on more than half of the sequence, and an *orange* one corresponds to consistent trajectory and map during less than half of the sequence. Lastly, a *red* mark indicates that the SLAM is able to initialize or produce completely inconsistent outputs.

IV. Results

Evaluations are carried out in real-time on a computer with an Intel i7-10610U CPU @ 1.80GHz × 8, 16 GB RAM, running Ubuntu 18.04 and ROS Melodic. In order to take into account the non-deterministic behavior of multithreaded applications, the reported observations are based on the median out of 3 runs per SLAM for each dataset. Results are reported in Tab. III, including the percentage of the sequence for which a localization is computed, and qualitative loop closure capability and consistency marks as defined in Section III-B. ORB-SLAM3 being a multimap SLAM system, the final number of disconnected maps is also indicated.

First, one can notice that none of the tested methods manages to completely process all sequences with a qualitatively fair accuracy. In addition, DSO-based approaches give particularly poor performances compared to ORB-SLAM-based ones.

One can see from Tab. III that ORB-SLAM fails to process important parts of the test sequences. This is caused by important initialization delays (ORB-SLAM even fails to initialize on the *A/In* sequence) and tracking failure recovery disability, hence the need for SLAM recovery strategies. Whereas Dual-SLAM's failure recovery strategy seems inefficient in the test datasets, leading



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

Fig. 9: *St-Raphael* dataset

TABLE II: Main characteristics of the evaluation datasets

Dataset	fps (Hz)	Resolution (pixels)	Duration (s)	Embedded light	Loop closure(s)	Depth (m)	Camera
Bus [3]	12.5	1200x1600	584	no	yes	20	RGB
Cave [3]	12.5	1200x1600	709	yes	yes	20	RGB
A/In [1]	15	640x776	88	no	no	unknown	RGB
A/Out [1]	4	640x776	53	no	no	unknown	RGB
AH #01 [16]	20	512x640	289	yes	yes	3	grayscale
AA #09 [16]	20	608x968	349	yes	yes	380	grayscale
Cephismer	5	480x640	156	no	yes	1.5	RGB
St-Raphael	20	480x640	472	no	yes	20	RGB

TABLE III: Results

		ORB-SLAM	ORB-SLAM3	Dual-SLAM	DSO	LDSO	DSM
Bus	% localized	38.08	64.03 (1 map)	33.83	15.74	21.25	14.65
	Loop closure				x		
	Consistency						
Cave	% localized	79.81	99.96 (1 map)	99.78	7.89	2.97	46.26
	Loop closure				x		
	Consistency						
A/In	% localized	91.23	99.62 (1 map)	99.62	99.03	99.00	99.60
	Loop closure	x	x	x	x	x	x
	Consistency						
A/Out	% localized	0.0	86.07 (1 map)	0.0	92.55	99.0	99.0
	Loop closure	x	x	x	x	x	x
	Consistency						
AH	% localized	x	99.71 (1 map)	x	77.73	79.08	x
	Loop closure	x		x	x		x
	Consistency	x		x			x
AA	% localized	99.79	99.51 (1 map)	99.72	88.07	99.0	82.50
	Loop closure				x		
	Consistency						
Cephismer	% localized	3.88	66.72 (5 maps)	3.88	28.50	42.15	41.03
	Loop closure				x		
	Consistency						
St-Raphael	% localized	6.81	64.46 (12 maps)	3.43	9.51	9.56	0.05
	Loop closure				x		
	Consistency						

to similar performances than ORB-SLAM, one can see that ORB-SLAM3 significantly improves the localization capabilities from ORB-SLAM. Indeed, ORB-SLAM3 outputs a localization on longer sequence portions, due to faster initialization and new map creation in case of SLAM failure what allows keeping the SLAM running. ORB-SLAM3's initialization's algorithm is the same as ORB-SLAM but with a different library, which may lead to faster computation and explain the improved initialization capabilities of ORB-SLAM3. ORB-SLAM3's tracking failure handling strategy with new map initialization is however particularly interesting. In the *Bus* dataset, this strategy allows covering a more important portion of the sequence than ORB-SLAM, since the system does not have to wait to reach an already mapped area to keep running. In difficult sequences leading to repetitive tracking failures, like in the *Cephismer* and *St-Raphael* datasets, this multimap approach results in disconnected trajectory parts and submaps, which cover an important portion of the full video sequence. Finally, all ORB-SLAM-based approaches show the same good loop closure detection and handling capabilities, which appears to be robust to underwater conditions. However, this same place recognition module fails to detect most map overlaps when running

ORB-SLAM3 on the *Cephismer* and *St-Raphael* scenarios, failing to fuse these maps into a global one in these particularly difficult sequences.

Whereas ORB-SLAM based approaches, and namely ORB-SLAM3, manage to consistently process the majority of the duration of all test sequences, DSO-based methods prove to be far less robust to underwater visual conditions. Similarly to [3], we observe that DSO is able to process at least partially some of the sequences and produce a quite realistic map of the environment during this time interval. The best DSO results are observed for the most structured environments, and in particular for the *A/In*, *Aqualoc Harbor* and *Aqualoc Archaeo* datasets which feature man-made objects. Images from these datasets show quite clear object contours, which are more adequate for DSO's tracking, which relies on close to contours pixel patches. DSO also suffers from local tracking inconsistencies, drift, and the incapacity to recover from tracking failure that may happen soon after initialization like in the *Cave*, *Cephismer* and *St-Raphael* datasets. These limits enlighten the interest of extending DSO with loop closing capabilities in order to reduce or correct these local tracking inconsistencies and lead to a more reliable mapping of the environment, for more robustness. However, both LDSO and DSM's loop closing implementations on a DSO basis show important limitations in the underwater field. LDSO fails to detect loops, and shows similar performances than DSO in terms of delay before failure and SLAM consistency. It is also very slow compared to DSO, and requires up to several seconds to process a single image. LDSO's loop detection process is very close to the one of ORB-SLAM, which manages to detect and process most loop closures. This difference in loop detection success with very close detection methods might be caused by the bad triangulation of map points used for loop detection, which prevents any geometric consistency validation. Finally, DSM also fails to give better performances than DSO on the evaluated underwater scenarios. It is also extremely slow, up to dozens of seconds per frame. In addition, DSM's loop detection strongly relies on a good pose prior, resulting in loop detection failure in all sequences because of localization drift. In the *Cave* dataset, bad pose prior even leads to false loop detections, which decrease SLAM performances. On the other hand, in the *Aqualoc Archaeo* dataset, DSM manages to detect accurate covisibilities between recent but first disconnected keyframes, hence the yellow loop closing mark. Successfully including older covisible keyframes with complementary points of view in the tracking window leads to a less important trajectory drift than DSO and LDSO on this sequence, as represented in Fig. 1.

V. Conclusion and future works

This work provides a qualitative underwater evaluation of recent monocular developments on DSO [5] and ORB-SLAM [4]. The main underwater VSLAM challenges appear to be the handling of tracking loss and place recognition. Whereas DSO's recent loop closure extensions DSM [7] and LDSO [6] are not robust enough to tracking failure, the recent multimap extension ORB-SLAM3 [9] of ORB-SLAM seems promising, and allows keeping computing a localization and a map even after a tracking loss with relocalization failure. The other ORB-SLAM multimap SLAM recovery extension, Dual-SLAM, has been evaluated, but does not demonstrate important robustness improvement compared to ORB-SLAM in the scenarios evaluated. Lastly, if ORB-SLAM3 appears to be already quite robust to underwater challenging visual conditions, it could be improved by investigating a more robust place recognition algorithm for map fusion.

References

[1] A. Q. Li, A. Coskun, S. M. Doherty, S. Ghasemlou, A. S. Jagtap, M. Modasshir, S. Rahman, A. Singh, M. Xanthidis, J. M. O'Kane, and I. Rekleitis, "Experimental comparison of open source vision-based state estimation algorithms," in *Springer Proceedings in Advanced Robotics*. Springer International Publishing, 2017, pp. 775–786.

[2] F. Hidalgo, C. Kahlefendt, and T. Braunl, "Monocular ORB-SLAM application in underwater scenarios," in *OCEANS*. Kobe, Japan: MTS/IEEE, May 2018.

[3] B. Joshi, N. Vitzilaios, I. Rekleitis, S. Rahman, M. Kalaitzakis, B. Cain, J. Johnson, M. Xanthidis, N. Karapetyan, A. Hernandez, and A. Q. Li, "Experimental comparison of open source visual-inertial-based state estimation algorithms in the underwater domain," in *IROS*. IEEE/RSJ, Nov. 2019.

[4] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[5] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, Mar. 2018.

[6] X. Gao, R. Wang, N. Demmel, and D. Cremers, "LDSO: Direct sparse odometry with loop closure," in *IROS*. IEEE/RSJ, Oct. 2018.

[7] J. Zubizarreta, I. Aguinaga, and J. M. M. Montiel, "Direct sparse mapping," *Transactions on Robotics*, vol. 36, no. 4, pp. 1363–1370, Aug. 2020.

[8] H. Huang, W.-Y. Lin, S. Liu, D. Zhang, and S.-K. Yeung, "Dual-SLAM: A framework for robust single camera navigation," in *IROS*. IEEE/RSJ, Oct. 2020.

[9] C. Campos, R. Elvira, J. J. G. Rodriguez, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM," *Transactions on Robotics*, Dec. 2021.

[10] D. Galvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, Oct. 2012.

[11] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *2007 6th IEEE and ACM Int. Symposium on Mixed and Augmented Reality*. IEEE, Nov. 2007.

[12] R. Elvira, J. D. Tardos, and J. Montiel, "ORBSLAM-atlas: a robust and accurate multi-map system," in *IROS*. IEEE/RSJ, Nov. 2019.

[13] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, Aug. 2013.

[14] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, Jan. 2016.

[15] D. Schubert, T. Goll, N. Demmel, V. Usenko, J. Stuckler, and D. Cremers, "The TUM VI benchmark for evaluating visual-inertial odometry," in *IROS*. IEEE/RSJ, Oct. 2018.

[16] M. Ferrera, V. Creuze, J. Moras, and P. Trouvé-Peloux, "AQUALOC: An underwater dataset for visual–inertial–pressure localization," *The International Journal of Robotics Research*, vol. 38, no. 14, pp. 1549–1559, Oct. 2019.

[17] S. Rahman, A. Q. Li, and I. M. Rekleitis, "Svin2: An underwater slam system using sonar, visual, inertial, and depth sensor," in *IROS*. IEEE/RSJ, 2019.