INVESTIGATION OF THE CHALLENGES OF UNDERWATER-VISUAL-MONOCULAR-SLAM

Michele Grimaldi[1,3], David Nakath[1,2], Mengkun She[1,2], Kevin Köser[1,2]

[1]Oceanic Machine Vision, GEOMAR Helmholtz Centre for Ocean Research Kiel, Wischhofstrasse 1-3, 24148 Kiel, Germany
[2]Marine Data Science, Department of Computer Science, Christian-Albrechts-Universität zu Kiel, 24118 Kiel, Germany
[3]Computer Vision and Robotics Research Institute (VICOROB), University of Girona, 17003 Girona, Spain

KEY WORDS: underwater, monocular, SLAM, image restoration, physically-based models, neural networks

ABSTRACT:

In this paper, we present a comprehensive investigation of the challenges of Monocular Visual Simultaneous Localization and Mapping (vSLAM) methods for underwater robots. While significant progress has been made in state estimation methods that utilize visual data in the past decade, most evaluations have been limited to controlled indoor and urban environments, where impressive performance was demonstrated. However, these techniques have not been extensively tested in extremely challenging conditions, such as underwater scenarios where factors such as water and light conditions, robot path, and depth can greatly impact algorithm performance. Hence, our evaluation is conducted in real-world AUV scenarios as well as laboratory settings which provide precise external reference. A focus is laid on understanding the impact of environmental conditions, such as optical properties of the water and illumination scenarios, on the performance of monocular vSLAM methods. To this end, we first show that all methods perform very well in air and subsequently investigate the degradation of their performance in ever more challenging underwater environments. The final goal of this study is to identify techniques that can improve accuracy and robustness of SLAM methods in such conditions. To achieve this goal, we investigate the potential of image enhancement techniques to improve the quality of input images used by the SLAM methods, specifically in low visibility and extreme lighting scenarios in scattering media. We present a first evaluation on calibration maneuvers and simple image restoration techniques to determine their ability to enable or enhance the performance of monocular SLAM methods in underwater environments.

1. INTRODUCTION

Underwater environments present unique challenges for robotic navigation. The low visibility, extreme lighting conditions, and unpredictable nature of underwater terrain make it difficult for robots to accurately perceive their surroundings and navigate effectively. Monocular vSLAM (visual Simultaneous Localization and Mapping) methods have emerged as a promising solution for underwater robot navigation, allowing robots to create a map of their surroundings while simultaneously determining their own position within that map (Durrant-Whyte and Bailey, 2006).

In monocular vSLAM, a single camera is used to capture images of the environment, which are then used to construct a map of the surrounding area. The camera's position and orientation are estimated in real-time, allowing the robot to determine its own location within the map. However, the accuracy of monocular vSLAM methods can be significantly impacted by environmental conditions, particularly low visibility and extreme lighting scenarios, which are common in underwater environments. To address this challenge, various image enhancement techniques have been proposed to improve the quality of the input images and thereby enhance the performance of monocular vSLAM methods (Song et al., 2022). These techniques comprise rather heuristic, physically-basedly-based as well as machine learning-based approaches, such as Generative Adversarial Networks (GANs).

In addition to monocular vSLAM methods, sensor fusion algorithms are crucial for accurate and robust navigation of underwater robots. Underwater environments pose unique challenges, such as low visibility and unpredictable terrain, that can significantly impact the accuracy of navigation systems. Therefore, integrating data from multiple sensors, such as Inertial Navigation Systems (INS), sonars, lasers, and cameras, can improve the overall performance of the navigation system. For example, combining data from an Inertial Navigation System (INS) and a Doppler Velocity Log (DVL) can improve the accuracy of underwater robot navigation by providing velocity measurements that are not affected by currents. Similarly, integrating data from an INS and a sonar can improve underwater localization by providing depth information that can be used to correct INS drift. Sonar sensors, such as multibeam sonars, can also be used to provide high-resolution 3D maps of underwater environments that can be used for localization and mapping (Drews Junior et al., 2016). Laser-based sensors, such as scanning laser rangefinders, can also be used to provide high-resolution 3D maps of underwater environments that can be used for localization and mapping (Palomeras et al., 2016). Incorporating data from multiple sensors can be challenging due to the different modalities and measurement noise associated with each particular sensor. However, advancements in sensor fusion algorithms, such as Extended Kalman Filters (EKF) and Unscented Kalman Filters (UKF), have made it possible to effectively combine data from multiple sensors in real-time (Yu et al., 2019), (Yang et al., 2019).

While sensor fusion using multiple modalities has shown great potential in improving underwater robot navigation and mapping, it also comes with additional cost and complexity. Therefore, in this paper, we evaluate the performance of offline and online monocular vSLAM methods for underwater robot navigation in both a real-world and a water tank in a laboratory setting. Our focus is on the impact of environmental conditions, such as water and illumination, on the performance of

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-1-W1-2023-1113-2023

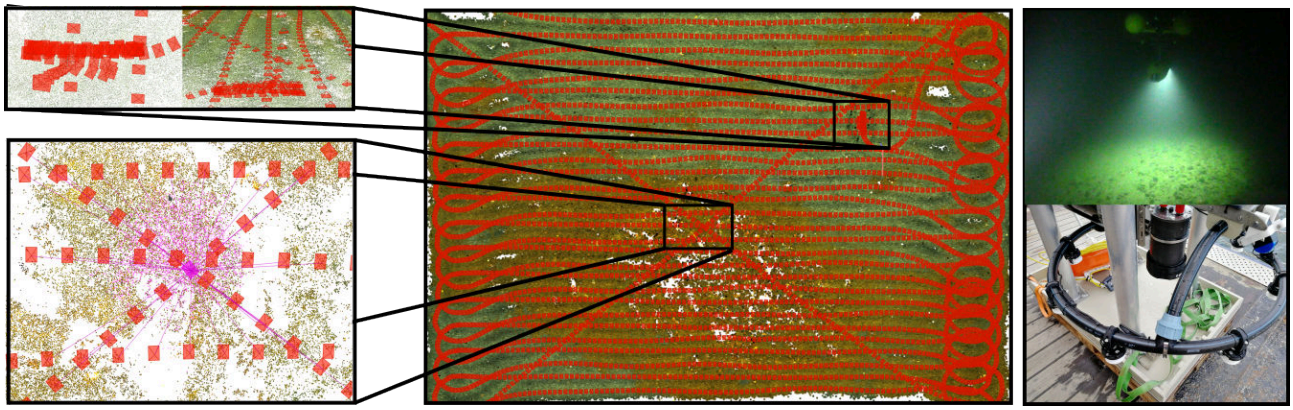

Figure 1. (Middle) Sparse Colmap reconstruction of the real Girona 500 Series AUV **A**3 surveying mission. After an initialization maneuver, a typical lawn mower patter followed by two cross tracks to support loop closing attempts are executed. The camera trajectory is drawn in red. Left lower magnification: increased number of loop closing opportunities (correspondences drawn in Pink) by executing cross-tracks, while the upper one shows a top- and a side-view of the initialization maneuver carried out to support SLAM algorithms. (Right) AUV in deep underwater mission and deployed camera-light-system.

monocular vSLAM methods. To this end, we first show the good performance of the chosen SLAM methods in air and then investigate their performance degradation with respect to varying environmental conditions in a scattering medium, i.e., water. Finally, we investigate the potential of image enhancement techniques to improve accuracy and robustness in challenging underwater environments.

2. MONOCULAR VSLAM

Monocular vSLAM methods are commonly classified into three categories: feature-based methods, direct methods, and visual-inertial methods (Zhou et al., 2019). Feature-based methods, such as ORB-SLAM2 (Mur-Artal et al., 2017), rely on detecting and tracking distinctive features in the image frames to estimate camera poses and create a map of the environment. These methods have been shown to achieve accurate results in a variety of scenarios, including indoor and outdoor environments. However, they can be sensitive to changes in illumination, texture, and occlusions, which can cause feature tracking failures and affect their robustness (Mur-Artal and Tardós, 2015). Newer versions of ORB-SLAM2, ORB-SLAM3 (Campos et al., 2021) incorporate semantic information to improve the robustness and accuracy of the system. Direct methods, such as DSO (Engel et al., 2018) and LSD-SLAM (Engel et al., 2014) estimate camera motion and 3D structure directly from the intensity values of the image frames, without relying on feature detection and tracking. LSD-SLAM uses semi-dense depth maps to estimate the camera poses and create a map of the environment. This method is known for its ability to handle large-scale environments and low-texture scenes while ORB-SLAM2 uses ORB features to detect and track keypoints in the image frames, and then estimates the camera poses and creates a map of the environment based on the feature matches. BAD-SLAM (Bian et al., 2020) is another direct method that estimates the camera motion and 3D structure directly from the intensity values of the image frames. It aims to improve the accuracy and robustness of SLAM systems by directly leveraging the RGB-D information and performing real-time bundle adjustment.

These methods can provide accurate and dense reconstructions in low-texture environments, but they require significant computational power and are less robust to changes in illumination and scene geometry (Engel et al., 2017).Visual-inertial methods, such as OKVIS (Leutenegger et al., 2015) and VINS-Mono (Qin et al., 2018), fuse camera and inertial measurements to estimate camera poses and create a map of the environment. These methods can achieve accurate results in highly dynamic and challenging environments, but they require additional sensors and calibration (Forster et al., 2017). However, also these methods face challenges when operating in highly dynamic underwater or low-light environments (Köser and Frese, 2020). To address these challenges, researchers have proposed various modifications to existing methods or developed entirely new methods. For example, (Ferrera et al., 2019) proposed a new visual odometry method specifically designed to handle the challenging conditions of underwater environments without any previous image enhancement step. Howerver, in our study we focus only monocular vSLAM methods that use camera data alone and do not incorporate inertial measurements.

3. IMAGE ENHANCEMENT / RESTORATION

Underwater environments typically exhibit extreme light conditions and poor visibility, which can significantly impair the performance of monocular vSLAM methods. To mitigate this issue, researchers have proposed various techniques for enhancing underwater images, which we here broadly divide into four categories: heuristic, statistical, physically-based, and machine learning methods (see (Song et al., 2022) for a comprehensive survey).

3.1 Statistical methods

Statistical methods for underwater image restoration aim to recover the original image from its degraded underwater version. These methods use statistical models to estimate the degradation factors, such as light attenuation, scattering, and noise, and then use these estimates to restore the image. For instance, (Chiang and Chen, 2012) proposed a wavelength compensation and dehazing approach that estimates the light attenuation coefficient and compensates for color distortion caused by the water medium. In (Ancuti et al., 2012) the authors proposed an underwater image enhancement method that relies only on the degraded version of the image for input and weight measures. They used two inputs to represent color correction and contrast enhancement of the original underwater image/frame, while four weight maps are employed to enhance the visibility

of distant objects degraded by medium scattering and absorption. (Drews Jr et al., 2017) proposed a statistical approach that estimates the parameters of a degradation model and inverts the degradation process to restore the original image.

3.2 Heuristic methods

Heuristic-based methods for underwater image restoration exploit specific underwater environment characteristics. Li et al.'s underwater dark channel prior (UDCP) (Li et al., 2016) uses the dark channel prior principle to estimate transmission and restore the image. Kim and Lee's adaptive histogram equalization (AHE) (Kim and Lee, 2017) applies histogram equalization to small image regions for contrast enhancement. Pizer et al.'s contrast-limited adaptive histogram equalization (CLAHE) (Pizer et al., 1987) limits contrast enhancement to prevent overamplification of noise. In (Köser et al., 2021), the authors present a practical approach to compensating for these lighting effects on flat seafloor regions found in the Abyssal plains. The method is parameter-free and performs robust statistics-based estimates of additive and multiplicative nuisances without requiring explicit parameters for light, camera, water, or scene. Although heuristic models can produce impressive outcomes in situations that align with their underlying assumptions, their results may lack consistency in other scenarios. Moreover, they do not assure to adhere to physical principles.

3.3 Machine Learning methods

Machine learning-based methods have shown promising results in enhancing underwater images by learning from a large dataset of annotated images. One popular machine learning-based method is the Deep Underwater Image Enhancement (DUIE) framework proposed by Zhang et al. (Zhang et al., 2019). DUIE uses a convolutional neural network (CNN) to learn the mapping between the low-quality input underwater image and the high-quality output image. Generative Adversarial Networks (GANs) are a type of deep neural network that consist of a generator and a discriminator. GANs have been used to learn the mapping between degraded and enhanced underwater images. The Underwater GAN (UW-GAN) proposed by Li et al. (Li et al., 2018) aims to improve the visibility of underwater images. UW-GAN uses an underwater image dataset to train the generator to generate enhanced versions of degraded underwater images. Other GAN-based methods for underwater image enhancement include the Conditional GAN (CGAN) (Fu et al., 2018) and the Multi-Scale GAN (MSGAN) (Xu et al., 2019). Despite benefiting from the expressive capabilities of neural networks, machine learning methods are typically trained under specific conditions. However, underwater environments are often characterized by dynamic conditions and their unpredictability, which can pose challenges to these methods.

3.4 Physically-based methods

Physically-based methods for underwater image restoration aim to model the physically-based processes that cause image degradation, such as light attenuation, scattering, and absorption, and then invert these models to restore the image. In doing so, they are the only methods able to do image restoration as opposed to image enhancement. These methods typically require knowledge of the physical properties of parts of the scene and can be computationally intensive. (Garcia et al., 2017) proposed a graph-based algorithm for color correction of underwater images. This method uses a graph-based representation

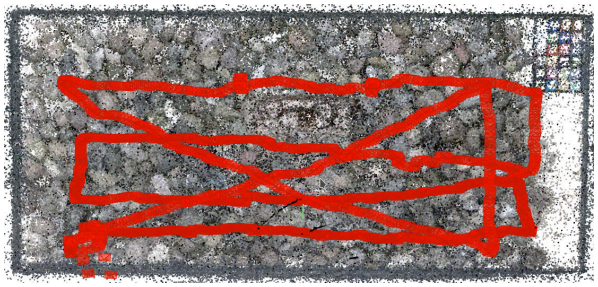

Figure 2. Top view of a sparse Colmap reconstruction of an example trajectory (in red) in the water tank: in the lower left, we exhibit an initialization maneuver (wiggle over one point), then we execute a lawn mower pattern, and finally cross the tracks, to improve the loop closing impact.

to model the relationships between different color channels in the image and to estimate the color correction factors. Another method is the Sea-thru method proposed by Akkaynak and Treibitz (Akkaynak and Treibitz, 2018) which estimates the backscatter using the dark pixels and their known range information, and then uses an estimate of the spatially varying illuminant to obtain the range-dependent attenuation coefficient. The latter method, however, is only valid for the Sunlight case, which exhibits homogeneous illumination. (Boittiaux et al., 2023) implemented multiview extensions, to improve the estimates and applied the method to datasets with artificial illumination. However, to make this approach work, the artificial illumination has to be locally homogeneous. Further approaches, which can be applied to true heterogeneous underwater artificial scenarios, i.e., with artificial illumination, are presented in (Bryson et al., 2016) and (Nakath et al., 2021). While physically-based methods have the advantage of being based on well-established principles, they can be limited by the availability and accuracy of the precise parameters needed for the models.

4. DATASETS

With Girona 500 series AUVs, we collected **A**-datasets in real waters, as well as **T**-datasets in a water tank with a precise ground truth estimate. In all settings, we carried out dedicated calibration maneuvers, which foster the initialization of SLAM approaches. Furthermore, in the tank and the AUV sets, we carried out classical lawn mower patterns with a stable flying height with subsequent cross-tracks, to support loop-closing approaches. In the tank, we additionally recorded sets, wich resemble a more free-flying scanning path. The latter brings about a lot of loop closing opportunities, which will however be impaired by big height variances, which in turn induce big changes in visual appearance in scattering media.

4.1 Water Test Tank with Ground Truth

We equipped a $2.2 \times 1 \times 0.8$m water test tank with three 50w Wasler daylight bulbs (5400k) housed in Walimex diffusors to create a homogeneous illumination setting akin to heavy atmospheric scattering. In addition, we attached two Ulanzi L2 Lite (5500k) co-moving lights, to be able to simulate active underwater light systems to a custom-build externally-tracked underwater camera (Winkel et al., 2023).

After building a small-scale test scene, we took several sets, to acquire underwater imagery with external reference as ground truth. As this is close to impossible in real waters, we equipped

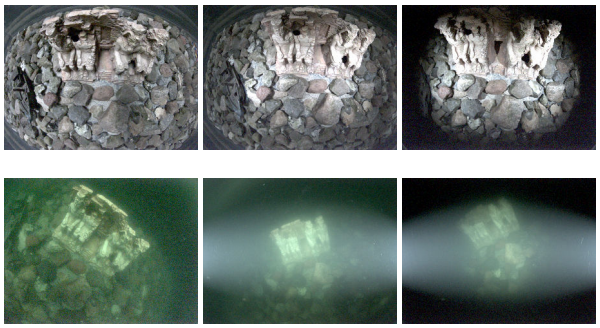

Figure 3. Left to right: Example images of tank sets: with Sunlight, Sun-and artificial light and artificial light. Upper row: in air **T**1-3, **T**4-5, and **T**6-7. Lower row: underwater **T**8-9, **T**10-11, and **T**12-13. Please note that the images are still distorted, but shown in sRGB-space for better visibility.

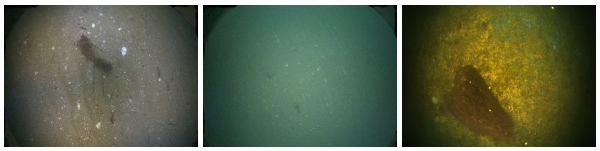

Figure 4. Left to right: Example images from **A**1, **A**2, and **A**3.

the camera with a stick and attached two Vive controllers, whose pose (position and attitude) in space can be precisely determined in air. This information can be used, to obtain a fused estimate of the pose of the underwater camera. We found the mean accuracy of the system tracking performance to be smaller 3 mm and 0.3 deg for translation and attitude, respectively (Winkel et al., 2023). For the dataset, an in-air fisheye calibration was conducted, with a residual error of 0.22px. Then, we center the camera within a dome port, to exclude refraction effects from the dataset, stemming from the traversal of interfaces of media with different optical densities (She et al., 2019, She et al., 2022). Finally, we conducted an underwater fisheye calibration of the camera, with residual reprojection error of 0.55px to capture remaining disturbances, which have not been captured by the preceding steps. Hence, we will exclusively deal with color distortion effects in those datasets. Specifically, we took homogeneously illuminated sets (**T**1-3), sets with mixed illumination (**T**4/5) and finally two sets with co-moving lights (**T**6/7) in air (see Fig. 3). Subsequently, we added water and dye for the attenuation as well as Maaloxan as the scattering agent until the working range was clearly distorted by the corresponding effects. This setup enables us to mimic the underwater conditions in Sunlight (**T**8/9), a mixed (Sun-artificial) light scenario (**T**10/11), as well as deep sea conditions, where only the artificial light is visible (**T**12/13); see Fig. 3). All former sets of those pairs execute a lawn mover pattern, while all latter sets execute free scanning trajectories with bigger depth variances.

For evaluation, we undistort the images into canonical pinhole space to provide them to the SLAM algorithms. Their results are then compared to the ground truth provided by the external reference system.

4.2 AUV Datasets

We also collected three challenging real datasets with Girona 500 series AUVs in the Baltic Sea. The **A**1/2 datasets are without initialization maneuvers (see Figs. 4 a,b), while **A**3 set features an initialization maneuver tailored to SLAM approaches (see Fig. 4 c).

The AUVs have a circular-arranged active lighting system, comprised of 8 LED-compounds cast in rasin (see Fig. 1) (Sticklus et al., 2017, Song et al., 2021a). Specifically, the AUVs are equipped with a dome port camera, which was again calibrated in air with a fisheye model. Subsequently, it was centered (She et al., 2019, She et al., 2022) to avoid refraction effects. We then overtake the underwater fisheye parameters estimated by Metashapes Photoscan into an adapted Colmap (Schönberger and Frahm, 2016) version to establish ground truth with an offline reconstruction method. In the latter process, the navigation data is fused into the visual reconstruction using a prose-graph-approach (She et al., 2023). For evaluation, the images are then undistorted and provided in canonical pinhole space for the SLAM algorithms. Finally, the results from Colmap's sparse reconstruction serve as the ground truth poses.

5. EVALUATION

We evaluated the performance of four SLAM methods, ORB-SLAM2, ORB-SLAM3, LSD-SLAM, and BADSLAM. For each underwater dataset, we applied six different image enhancement methods: the UDCP algorithm, the CLAHE algorithm, UW-GAN algorithm that was trained on three different types of water and the median filter from (Köser et al., 2021). We also evaluated each method on the basic, unenhanced images. For BADSLAM and GRADSLAM, we estimated the depth maps using UW-Net (Gupta and Mitra, 2019) and UDepth (Yu et al., 2023) in underwater scenarios and Monodepth2 (Godard et al., 2019) for the in-air sets.

We categorized failures into three types: not initializing (NOT INIT), initializing but losing track (TR-Lost), and complete failure (FAILED). Not initializing means the method could not start tracking the camera pose. Initializing but losing track indicates that the method began tracking but eventually lost the camera pose without recovery. If the camera poses are lost but the map has enough keyframes, the algorithm is considered successful. Complete failure means the method provided no output (e.g., LSD-SLAM). BADSLAM may fail to start if the estimated depth map is inaccurate.

We used a solid alignment approach to match up the SLAM estimates and the ground truth. We initially temporally aligned the SLAM estimates and the ground truth interpolating the latter such that for each ground truth pose there is a corresponding estimated pose. Afterwards, we employed the $\mathbb{SIM}3$ (Allen-Blanchette et al., 2014) Umeyama (Umeyama, 1991) alignment technique, which considers scale, translation, and rotation, to accurately align the result trajectory with the ground truth in space. Specifically, we optimize

$$\min_{s,R,t} \sum_{i=1}^{k} \|\hat{x}_i - (sRx_i + t)\|_2^2, s \in \mathbb{R}; t \in \mathbb{R}^3; R \in \mathbb{SO}3 \quad (1)$$

where $\hat{x}_i$ and $x_i \in \mathbb{R}^3$ are the paired positions. This approach also entails an error-measure in the units of the ground-truth-data, i.e., position in [m] and attitude in [deg]. We used the (Grupp, 2017) Python package to perform the alignment. If the method was able to successfully initialize and track the camera pose, we used the absolute trajectory error (ATE) to compare the ground truth trajectory with the estimated trajectory. The ATE is calculated by finding the difference between the

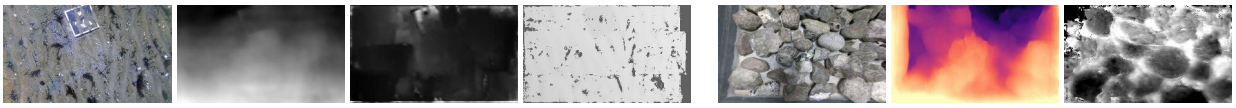

Figure 5. From left to right: base image from an **A**-set, UDepth depth map, UW-Net depth map, Colmap depth map; base image from **T**-set, Monodepth2 depth map, Colmap depth map

ground truth and estimated camera poses at each frame and then computing the root mean squared error (RMSE) of these differences. This allowed us to compare the accuracy of the SLAM methods under different conditions.
The ATE can be further broken down into the error for translation and the error for rotation. The translation error measures the difference between the estimated and ground truth position of the camera, while the rotation error measures the difference between the estimated and ground truth orientation of the camera. Both of these errors are calculated using the same approach as for the ATE, by finding the difference between the ground truth and estimated values at each frame and then computing the RMSE of these differences. The ATE errors for translation and rotation are computed as follows:

$$\begin{aligned} \text{ATE}_\text{t} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N} ||p_i - \hat{p_i}||^2} \text{ , and} \\ \text{ATE}_\text{r} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N} 2 \cdot \arccos\left(\min(|q_i^{-1} \cdot \hat{q_i}|, 1)\right) \cdot \frac{180}{\pi}}, \end{aligned} \tag{2}$$

where N is the number of poses in the trajectory, $p_i, \hat{p}_i \in \mathbb{R}^3$ and $q_i, \hat{q}_i \in \mathbb{SO}3$ are the ground truth and estimated positions and attitudes of the robot at pose i.

6. RESULTS

According to the findings presented in Annex A, the experiments conducted on the **A**1, **A**2, and **A**3 datasets revealed that the algorithms encountered challenges and exhibited poor performance due to various factors. Specifically, for the **A**1 dataset, one of the main issues identified was the lack of sufficient overlap between frames. This insufficient overlap hindered the algorithms' ability to establish robust correspondences and accurately estimate the robot's trajectory. On the other hand, for the **A**2 dataset, although there was good overlap between frames, the presence of unfavorable water and light conditions posed significant difficulties for the algorithms. These conditions, such as poor visibility, light scattering, and limited lighting, adversely affected the algorithms' ability to accurately estimate depth and track the robot's movement. Furthermore as for the case of the **A**3 dataset, the challenges were further compounded by the presence of low texture areas in addition to the water and lighting conditions. Low texture areas, which lack distinctive visual features, make it challenging for SLAM algorithms to establish reliable correspondences and accurately estimate the robot's trajectory in those regions. In fact, while the initialization trajectory of **A**3 helped for the initialization of the ORB-SLAMS and the median method performed the best on the **A**2 data-set, the ORB-SLAMS lost the track in the majority of the frames.

The experiments on the **T**8-13 dataset, which consisted of a tank with water, have shown that light conditions are critical for successful SLAM performance. The light cones produced from the artificial lights, with and without the sunlight, had a significant impact on the images, resulting in failure of the SLAM methods. This is due to the fact that the presence of water causes light to being refracted, attenuated and scattered. While we controlled for the refraction effects, by centering the camera in the dome, the two latter effects lead to changes in image appearance and the degradation of image quality. In addition, the absorption and scattering of light in water varies depending on the wavelength and scene depth, which can affect the accuracy of visual odometry and feature tracking.

Our findings indicated that LSD-SLAM was ineffective for underwater applications, as it failed to function properly on every underwater dataset we tested, consistent with previous research (Joshi et al., 2019). Furthermore, while the method was effective on the sunlight scenarios (**T**1-**T**3), it was ineffective on the in-air sets with mixed lights and in-air sets with only artificial light which is also consistent with previous research (Pascoe et al., 2017). Additionally, our findings highlight the impact of robot maneuvers, both during the trajectory and during the initialization phase, on SLAM accuracy. Specifically, we observed that low dynamic maneuvers tended to result in better accuracy for SLAM. When the robot moved in a more controlled and stable manner, the SLAM algorithm was able to more accurately estimate the robot's position and orientation. Furthermore, the initialization maneuver also had an impact on SLAM accuracy. By carefully designing and executing an appropriate initialization maneuver, we were able to improve the accuracy of the SLAM algorithm. This initialization maneuver provided the algorithm with a more accurate starting point, allowing it to establish a better understanding of the environment and subsequently improve the overall trajectory estimation.

The BADSLAM and GRADSLAM algorithms did not work with either the UDepth or UW-Net depth estimators in underwater scenarios and with Monodepth2 for the in-air sets. To ascertain whether the depth estimators were the root of the issue, we used Colmap for depth estimation and discovered that BADSLAM was capable of at least initializing itself in the **A**1/2 and **A**3 datasets and in the **T** in-air sets it succeeded alongside GRADSLAM. In regard to the UW-Net, U-Depth, and Monodepth2 depth estimators, it is important to note that they are primarily designed for forward-looking camera settings and not specifically for top-down views. Unsurprisingly, they do not perform well when applied to top-down views, such as those encountered in underwater environments (see Fig. 5).

Concerning the in-air sets, our ORB-SLAM results are aligned with results in (Song et al., 2021b) in which the authors created multiple datasets using multiple cameras, IMU and a test tank for visual inertial odometry. Finally our experiments also highlighted the importance of using appropriate image enhancement methods for different water conditions. Among the methods tested, CLAHE performed the best overall. The UW-GAN with different water types performed similarly to UDCP, and the second water type of the UW-GAN often performed as well as UDCP, possibly because the water condition used for training was similar to the water type of the dataset used for evaluation.

7. CONCLUSION

In this paper, we conducted an investigation of the challenges of underwater monocular visual SLAM. To this end, we prepared several AUV-based and controlled lab-datasets. All geometric distortions are controlled for in those sets, while they are all taken in extremely low visibility and harsh light conditions. This allows for an in-depth investigation of the impact of radiometric distortions in the underwater setting. The ground truth in the lab sets is established with a custom-build external reference system, while the AUV sets – lacking external reference – are offline-reconstructed with a modified version of Colmap. First, we showed that all selected SLAM algorithms successfully run on in air tank-dataets with good performances. Then, we evaluated several combinations of SLAM systems and preprocessing methods on all datasets. We found that no SLAM system is able to complete the real AUV-Datasets. Also from the tank datasets, only the homogeneously illuminated Sun-settings could be completed. Here, we found that the preprocessing approaches showed some initial improvements of the SLAM performance in the visually adversarial underwater environments. In addition, we can also report mild improvements, when special initialization-maneuvers are carried out. The generalization of the pre-processing methods is a direction worth to further investigate, as they seem to be heavily tuned to certain assumptions / scenarios. Furthermore, providing depth-information dependent SLAM systems with corresponding top-down-view estimates also seems to be an interesting route. However we had to resort to depth maps established in an offline fashion, as the deep learning based estimators were tuned to different use cases. Hence, in the future, we will strive to preprocess underwater imagery to mitigate radiometric distortions and at the same time improve on underwater monocular depth estimation in order to leverage the already existing big potential of in-air SLAM approaches.

8. ACKNOWLEDGMENT

This publication has been funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) Projektnummer 396311425, through the Emmy Noether Programme. Furthermore, the authors would like to thank Tim Benedikt von See, Jens Greinert, and GEOMAR's AUV team for collecting the AUV datasets within the scope of the BMBF funded CONMAR project (grant no. 03F0912A) and ships proposal MeCoMM (GPF22-1/015) and Birger Winkel for support while recording the VIVE-tracked tank-data.

REFERENCES

Akkaynak, D., Treibitz, T., 2018. Sea-thru: A method for removing water from underwater images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10), 2315–2327.

Allen-Blanchette, C., Leonardos, S., Gallier, J., 2014. Motion interpolation in sim (3).

Ancuti, C., Ancuti, C. O., Haber, T., Bekaert, P., 2012. Enhancing underwater images and videos by fusion. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 81–88.

Bian, J., Wang, S., Li, C., Li, M., 2020. Bad-slam: Bundle adjusted direct rgb-d slam. *IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 10540–10546.

Boittiaux, C., Marxer, R., Dune, C., Arnaubec, A., Ferrera, M., Hugel, V., 2023. SUCRe: Leveraging Scene Structure for Underwater Color Restoration. *arXiv preprint arXiv:2212.09129*.

Bryson, M., Johnson-Roberson, M., Pizarro, O., Williams, S. B., 2016. True Color Correction of Autonomous Underwater Vehicle Imagery. *Journal of Field Robotics*, 33(6), 853-874. https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21638.

Campos, C., Elvira, R., Gómez, J. J., Montiel, J. M. M., Tardós, J. D., 2021. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial and Multi-Map SLAM. *IEEE Transactions on Robotics*, 37(6), 1874-1890.

Chiang, J.-H., Chen, K.-C., 2012. Wavelength compensation and dehazing for underwater image enhancement. *Optics Express*, 20(24), 25011–25021.

Drews Jr, P., Barsi, Á., Czúni, L., 2017. A statistical approach to underwater image restoration. *Journal of Visual Communication and Image Representation*, 48, 207–218.

Drews Junior, P., Bezerra, L., Medeiros, A., Campos, M., 2016. Underwater navigation using multibeam sonar and ins sensors. *OCEANS 2016 MTS/IEEE Monterey*, IEEE, 1–5.

Durrant-Whyte, H., Bailey, T., 2006. Simultaneous localization and mapping: part I. *IEEE Robotics & Automation Magazine*, 13(2), 99-110.

Engel, J., Koltun, V., Cremers, D., 2018. Direct sparse odometry. *IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2172–2180.

Engel, J., Schöps, T., Cremers, D., 2014. Lsd-slam: Large-scale direct monocular slam. *European Conference on Computer Vision (ECCV)*, Springer, 834–849.

Engel, J., Usenko, V., Cremers, D., 2017. Direct sparse odometry with rolling shutter. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 1251–1260.

Ferrera, M., Moras, J., Trouvé-Peloux, P., Creuze, V., 2019. Real-Time Monocular Visual Odometry for Turbid and Dynamic Underwater Environments. *Sensors*, 19(3). https://www.mdpi.com/1424-8220/19/3/687.

Forster, C., Carlone, L., Dellaert, F., Scaramuzza, D., 2017. Manifold-based optimization for end-to-end visual-inertial Odometry. *IEEE Transactions on Robotics*, 33(1), 1–20.

Fu, X., Huang, Y., Ding, X., Liao, Y., Paisley, J., 2018. Removing the water from the unseen underwater image. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3207–3215.

Garcia, R., Gasull, A., Trias, J., Palou, A., Travé-Massuyes, L., 2017. Color correction of underwater images using a graph-based algorithm. *Journal of Imaging*, 3(4), 52.

Godard, C., Mac Aodha, O., Firman, M., Brostow, G. J., 2019. Digging into Self-Supervised Monocular Depth Prediction.

Grupp, M., 2017. evo: Python package for the evaluation of odometry and slam. `https://github.com/MichaelGrupp/evo`.

Gupta, H., Mitra, K., 2019. Unsupervised Single Image Underwater Depth Estimation. *arXiv preprint arXiv:1905.10595*.

Joshi, B., Rahman, S., Kalaitzakis, M., Cain, B., Johnson, J., Xanthidis, M., Karapetyan, N., Hernandez, A., Li, A. Q., Vitzilaios, N., Rekleitis, I., 2019. Experimental comparison of open source visual-inertial-based state estimation algorithms in the underwater domain. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 7227–7233.

Kim, H.-J., Lee, S.-H., 2017. Underwater image enhancement using an adaptive histogram equalization algorithm. *International Conference on Control, Automation and Information Sciences (ICCAIS)*, IEEE, 449–452.

Köser, K., Frese, U., 2020. Challenges in underwater visual navigation and SLAM. *AI technology for underwater robots*, 125–135.

Köser, K., Song, Y., Petersen, L., Wenzlaff, E., Woelk, F., 2021. Robustly Removing Deep Sea Lighting Effects for Visual Mapping of Abyssal Plains. *arXiv preprint arXiv:2110.00480*. https://arxiv.org/abs/2110.00480.

Leutenegger, S., Lynen, S., Bosse, M., Siegwart, R., Furgale, P., 2015. Keyframe-based visual-inertial slam using nonlinear optimization. *IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 1565–1572.

Li, C., Anwar, H., Sung, M.-C., Hsu, W., 2018. WaterGAN: Unsupervised Generative Network to Enable Real-time Color Correction of Monocular Underwater Images. *IEEE Robotics and Automation Letters*, 3(4), 3282–3289.

Li, C., Anwar, S., Tan, R. T., 2016. Underwater image enhancement by removing backscatter with the underwater dark channel prior. *IEEE Transactions on Image Processing*, 25(6), 2809–2823.

Mur-Artal, R., Montiel, J. M. M., Tardós, J. D., 2017. Orb-slam2: an open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 1655–1661.

Mur-Artal, R., Tardós, J. D., 2015. ORB-SLAM: a Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, 31(5), 1147–1163.

Nakath, D., She, M., Song, Y., Köser, K., 2021. In-situ joint light and medium estimation for underwater color restoration. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3731–3740.

Palomeras, N., Ridao, P., Carreras, M., 2016. Underwater SLAM with sampling sonar and scanning laser rangefinder. *Robotics and Autonomous Systems*, 79, 38–50.

Pascoe, G., Maddern, W., Tanner, M., Piniés, P., Newman, P., 2017. Nid-slam: Robust monocular slam using normalised information distance. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1435–1444.

Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., Romeny, B., Zimmerman, J. B., 1987. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3), 355–368.

Qin, T., Li, P., Shen, S., 2018. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4), 1004–1020.

Schönberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

She, M., Nakath, D., Song, Y., Köser, K., 2022. Refractive geometry for underwater domes. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183, 525–540.

She, M., Song, Y., Mohrmann, J., Köser, K., 2019. Adjustment and calibration of dome port camera systems for underwater vision. *Pattern Recognition: 41st DAGM German Conference, DAGM GCPR 2019, Dortmund, Germany, September 10–13, 2019, Proceedings 41*, Springer, 79–92.

She, M., Song, Y., Nakath, D., Köser, K., 2023. Efficient large-scale auv-based visual seafloor mapping.

Song, Y., Nakath, D., She, M., Elibol, F., Köser, K., 2021a. Deep sea robotic imaging simulator. *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*, Springer, 375–389.

Song, Y., Nakath, D., She, M., Köser, K., 2022. Optical imaging and image restoration techniques for deep ocean mapping: A comprehensive survey. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 90(3), 243–267.

Song, Y., Qian, J., Miao, R., Xue, W., Ying, R., Liu, P., 2021b. Haud: A high-accuracy underwater dataset for visual-inertial odometry. *2021 IEEE Sensors*, 1–4.

Sticklus, J., Kwasnitschka, T., Hoeher, P. A., 2017. Method and device for potting an led luminaire potted in a potting compound, and led luminaire. US Patent App. 15/533,130.

Umeyama, S., 1991. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4), 376-380.

Winkel, B., Nakath, D., Woelk, F., Köser, K., 2023. Design, implementation and evaluation of an external pose-tracking system for underwater cameras.

Xu, J., Wen, L., Li, J., Wang, Z., 2019. Msgan: Multi-scale gradient aggregation with adversarial networks for enhancing underwater imagery. *Proceedings of the IEEE International Conference on Computer Vision*, 4464–4473.

Yang, Z., Zhou, Y., Dong, Y., 2019. An integrated navigation system for underwater robots using an improved UKF algorithm. *Sensors*, 19(6), 1366.

Yu, B., Wu, J., Islam, M. J., 2023. Udepth: Fast monocular depth estimation for visually-guided underwater robots. *Accepted at the IEEE International Conference on Robotics and Automation (ICRA)*, IEEE.

Yu, W., Qiu, B., Zhang, H., 2019. A survey of underwater robot navigation. *Journal of Marine Science and Engineering*, 7(6), 183.

Zhang, C., Li, X., Li, C., Fu, K., 2019. Deep Underwater Image Enhancement. *IEEE Transactions on Image Processing*, 28(1), 206–219.

Zhou, Y., Xiao, N., Wen, J., Zhang, Y., 2019. Visual Monocular SLAM for Autonomous Underwater Vehicle: Recent Advances and Future Challenges. *IEEE Access*, 7, 16279–16297.

A. ANNEX ALGORITHM PERFORMANCES

A1, Real mission w/ Girona 500 AUV

	Base	CLAHE	UDCP	UW-GAN Water 1	UW-GAN Water 2	UW-GAN Water 3	Median
ORB-SLAM2	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost
ORB-SLAM3	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost
BADSLAM (UDepth)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
BADSLAM (UW-Net)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
BADSLAM (Colmap)	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost
LSD-SLAM	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UDepth)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UW-Net)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(Colmap)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED

A2, Real mission w/ Girona 500 AUV

	Base	CLAHE	UDCP	UW-GAN Water 1	UW-GAN Water 2	UW-GAN Water 3	Median
ORB-SLAM2	NOT INIT	TR-Lost	TR-Lost	NOT INIT	NOT INIT	NOT INIT	TR-Lost
ORB-SLAM3	NOT INIT	TR-Lost	NOT INIT	NOT INIT	NOT INIT	NOT INIT	TR-Lost
BADSLAM (UDepth)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
BADSLAM (UW-Net)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
BADSLAM (Colmap)	NOT INIT	TR-Lost	TR-Lost	NOT INIT	NOT INIT	NOT INIT	TR-Lost
LSD-SLAM	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UDepth)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UW-Net)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(Colmap)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED

A3, Real mission w/ Girona 500 AUV

	Base	CLAHE	UDCP	UW-GAN Water 1	UW-GAN Water 2	UW-GAN Water 3	Median
ORB-SLAM2	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost
ORB-SLAM3	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost
BADSLAM (UDepth)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
BADSLAM (UW-Net)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
BADSLAM (Colmap)	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost	TR-Lost
LSD-SLAM	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UDepth)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UW-Net)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(Colmap)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED

T8, Water Tank, homogenous illumination, lawn mower

	Base	CLAHE	UDCP	UW-GAN Water 1	UW-GAN Water 2	UW-GAN Water 3	Median
ORB-SLAM2	3.5°|1.42	**2.2°**|1.28	3.12°|1.41	4.0°|1.52	4.1°|1.53	3.9°|1.61	NOT INIT
ORB-SLAM3	4.5°|1.61	3°|**1.19**	2.6°|1.51	3.3°|1.54	3.8°|1.50	3.3°|1.59	NOT INIT
BADSLAM (UDepth)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
BADSLAM (UW-Net)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
LSD-SLAM	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UDepth)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UW-Net)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED

T9, Water Tank, homogenous illumination, scanning trajectory

	Base	CLAHE	UDCP	UW-GAN Water 1	UW-GAN Water 2	UW-GAN Water 3	Median
ORB-SLAM2	2.7°|1.34	2.2°|1.26	2.2°|1.52	2.7°|1.47	2.8°|1.46	2.9°|1.61	NOT INIT
ORB-SLAM3	2.9°|1.53	**1.8°**|**1.24**	2.6°|1.51	2.62°|1.49	3.32°|1.59	2.9°|1.47	NOT INIT
BADSLAM (UDepth)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
BADSLAM (UW-Net)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
LSD-SLAM	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UDepth)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UW-Net)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED

T10, Water Tank, mixed illumination, lawn mower

	Base	CLAHE	UDCP	UW-GAN Water 1	UW-GAN Water 2	UW-GAN Water 3	Median
ORB-SLAM2	NOT INIT	TR-Lost	NOT INIT	NOT INIT	NOT INIT	NOT INIT	NOT INIT
ORB-SLAM3	NOT INIT	TR-Lost	NOT INIT	NOT INIT	NOT INIT	NOT INIT	NOT INIT
BADSLAM (UDepth)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	NOT INIT
BADSLAM (UW-Net)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
LSD-SLAM	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UDepth)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UW-Net)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED

T11, Water Tank, mixed illumination, scanning trajectory

	Base	CLAHE	UDCP	UW-GAN Water 1	UW-GAN Water 2	UW-GAN Water 3	Median
ORB-SLAM2	NOT INIT	TR-Lost	NOT INIT	NOT INIT	NOT INIT	NOT INIT	NOT INIT
ORB-SLAM3	NOT INIT	TR-Lost	NOT INIT	NOT INIT	NOT INIT	NOT INIT	NOT INIT
BADSLAM (UDepth)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	NOT INIT
BADSLAM (UW-Net)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
LSD-SLAM	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UDepth)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UW-Net)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED

T12, Water Tank, artificial illumination, lawn mower

	Base	CLAHE	UDCP	UW-GAN Water 1	UW-GAN Water 2	UW-GAN Water 3	Median
ORB-SLAM2	NOT INIT	TR-Lost	NOT INIT	NOT INIT	NOT INIT	NOT INIT	NOT INIT
ORB-SLAM3	NOT INIT	TR-Lost	NOT INIT	NOT INIT	NOT INIT	NOT INIT	NOT INIT
BADSLAM (UDepth)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
BADSLAM (UW-Net)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
LSD-SLAM	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UDepth)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UW-Net)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED

T13, Water Tank, artificial illumination, scanning trajectory

	Base	CLAHE	UDCP	UW-GAN Water 1	UW-GAN Water 2	UW-GAN Water 3	Median
ORB-SLAM2	NOT INIT	TR-Lost	NOT INIT	NOT INIT	NOT INIT	NOT INIT	NOT INIT
ORB-SLAM3	NOT INIT	TR-Lost	NOT INIT	NOT INIT	NOT INIT	NOT INIT	NOT INIT
BADSLAM (UDepth)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
BADSLAM (UW-Net)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
LSD-SLAM	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UDepth)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM(UW-Net)	FAILED	NA	FAILED	FAILED	FAILED	FAILED	FAILED

T1-7, Tank in-air

	T1	T2	T3	T4	T5	T6	T7
ORB-SLAM2	1.68°|0.09	1.31°|0.08	2.12°|0.12	1.68°|0.10	2.47°|0.15	2.23°|0.21	2.54°|0.16
ORB-SLAM3	1.60°|0.09	1.23°|**0.07**	2.94°|0.17	1.92°|0.12	2.27°|0.14	2.22°|0.30	2.23°|0.29
LSD-SLAM	1.89°|0.10	1.68°|0.09	1.98°|0.20	2.20°|0.24	1.87°|0.20	FAILED	FAILED
BADSLAM (Monodepth2)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
BADSLAM (Colmap)	**1.38°**|0.08	1.47°|0.11	1.87°|0.12	1.94°|0.20	2.09°|0.25	2.07°|0.19	2.17°|0.23
GRADSLAM (Monodepth2)	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED
GRADSLAM (Colmap)	1.45°|0.12	1.42°|0.13	2.10°|0.15	2.22°|0.17	2.3°|0.27	2.28°|0.24	2.41°|0.28

B. ANNEX ORIGINAL AND PREPROCESSED IMAGES

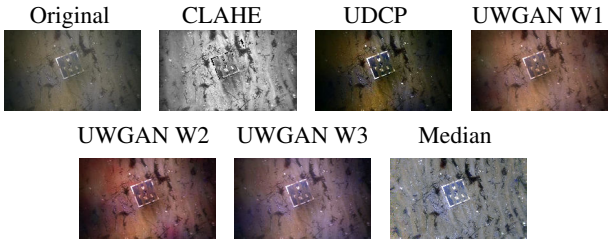

Figure 6. **A**1

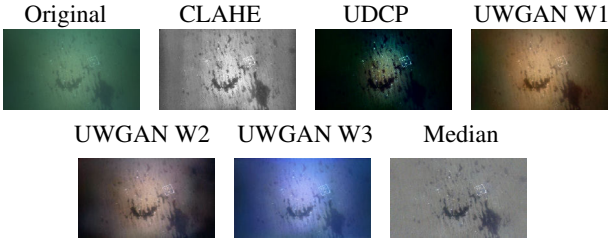

Figure 7. **A**2

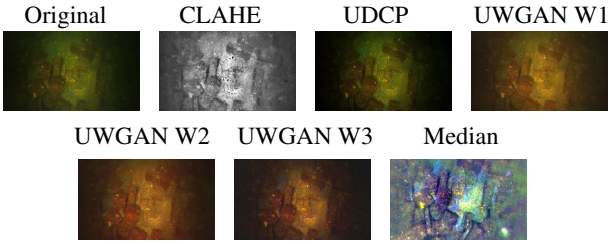

Figure 8. **A**3

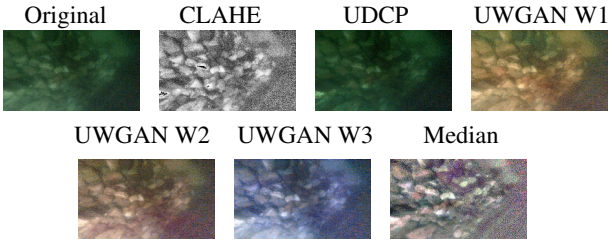

Figure 9. **T**8-9: sunlight

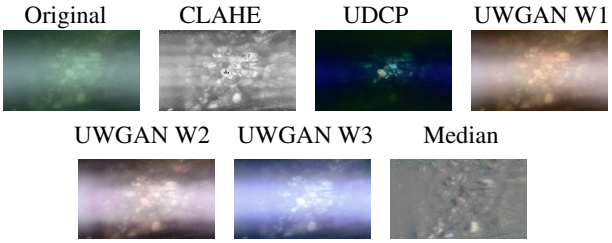

Figure 10. **T**10-11: sunlight and artificial lights

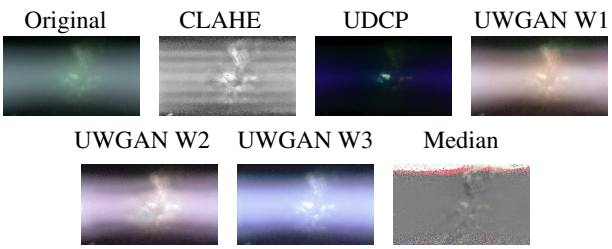

Figure 11. **T**12-13: artificial lights