Exploring Real-time Semantic SLAM in Coral Reefs with MASt3R-SLAM

Jonathan Sauder
EPFL
Switzerland
jonathan.sauder@epfl.ch

Devis Tuia
EPFL
Switzerland
devis.tuia@epfl.ch

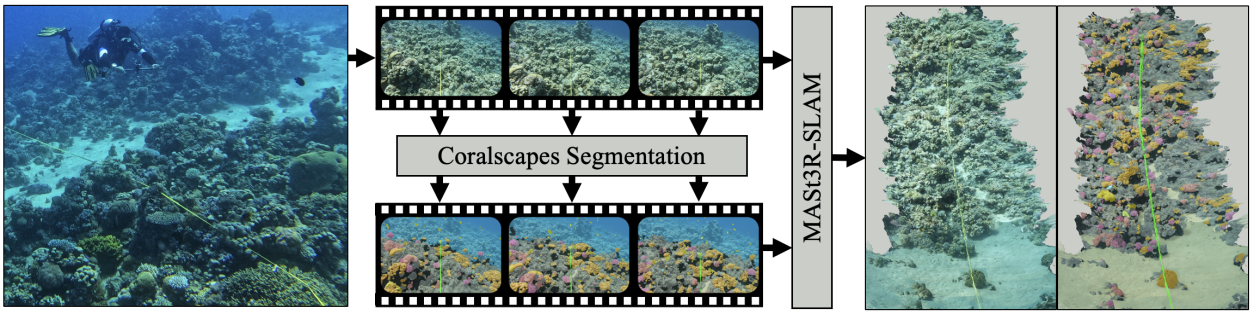

Figure 1. This paper explores the effectiveness of MASt3R-SLAM for mapping coral reefs from videos in real time with GoPro cameras. On a dataset of 24 video transects, in which a diver follows a 50m transect line (left), we segment video frames using Coralscapes (center) to filter out unwanted classes. We then assess 3D reconstructions by MASt3R-SLAM.

Abstract

Coral reefs are under unprecedented threat due to global anthropogenic climate change and local human activities, severely impacting communities relying on their services through food security or tourism income. To devise effective conservation strategies, scalable monitoring techniques are required: deep learning promises to drastically reduce the expert time used on manual analysis of coral reef survey imagery. In particular, the combination of deep monocular simultaneous localization and mapping (SLAM) with semantic segmentation of video frames offers a new paradigm for rapid semantic mapping of coral reef transects with low-cost hardware. However, many learning-based SLAM systems are trained exclusively using in-air data from urban or indoor scenes, raising questions about their generalization ability to underwater environments. In this work, we explore the synergy between recent general-purpose semantic segmentation in reefs (Coralscapes) and current state-of-the-art in deep monocular SLAM (MASt3R-SLAM). By qualitative evaluation on 24 video transects, we find that reconstructions are of very high quality locally, but that relocalization and loop closure, as well as scale-drift remain open problems, and that camera calibration remains indispensable. Our findings suggests that there is a need high quality domain-specific 3D vision datasets to train 3D pointmap matching models for underwater scenes.

1. Introduction

Coral reefs, which host 32% of marine species on less than 0.1% of the world's area [8], are existentially threatened by climate change and local human stressors [9, 11]. With current greenhouse gas emission trajectories [14], almost all coral reefs are projected to experience mass mortality, which could have catastrophic cascading effects on biodiversity [9]. The decline of reefs will severely impact more than 500 million people that rely on the services they provide (e.g. fishing or touristic income) [18] in over 100 reef nations. However, some corals exhibit signs of strong thermal resistance or adaptivity [7, 19], giving hope that some vibrant coral reefs may survive through the end of this century and beyond.

It is therefore crucial to monitor reefs at the highest possible spatial and temporal resolution to better understand the dynamics of reefs during or after heatwaves, rapidly identify & intervene on local stressors, and assess the effectiveness of conservation measures such as the establishment of marine protected areas.

Conventional *in situ* coral reef monitoring methods are laborious, as they require coral experts to collect the data or analyze the large amounts of imagery collected. This leads

to entire countries being data-deficient, particularly countries with restricted research resources. Computer vision & machine learning methods are increasingly used to accelerate the analysis of reef imagery, but most existing methods focus on integration into existing workflows, in which sparse points within conventional photo quadrats are classified into relevant benthic substrate classes [1, 3, 16, 20]. Other solutions are specific to photogrammetric orthomosaics [27, 28], which require high-quality curated image collections as input and incur substantial computational cost [2, 5].

General-purpose segmentation of reef imagery, meaning that the input images are not restricted to a specific use-case, and that all regions of an image are segmented, including benthic classes (coral, rock, algae, sand, etc.) as well as auxiliary classes (fish, divers, water column, etc.), unlocks a wide range of use-cases. General-purpose semantic segmentation can be transferred to photogrammetric 3D reconstructions or orthomosaics, and shows synergies with underwater localization and mapping: semantic mapping of coral reefs from video allows users to rapidly conduct reef surveys under water [23], and could be deployed to marine robots, ensuring safe navigation & interaction with humans. But robust general-purpose semantic segmentation needs high-quality datasets to train deep learning models. The Coralscapes dataset [22], composed of 2075 images densely annotated with 39 classes, provides the first general-purpose segmentation dataset in coral reefs.

This paper explores the combination of segmentation using models trained on Coralscapes with a current neural SLAM system based on the pointmap matching paradigm, MASt3R-SLAM [17] (Figure 1). We assess the local and global reconstruction quality of MASt3R-SLAM, and show that semantic masking removes detrimental artifacts from the water column or unwanted objects. Our main findings are that MASt3R-SLAM produces higher local detail than existing real-time approaches for reef mapping, but is more prone to failure: tracking losses and scale drift can lead to a poor global structure.

2. Background: Deep 3D Pointmap Matching

In the last decade, deep learning has been extensively used to tackle 3D computer vision sub-tasks like feature detection [6] & matching [21], camera calibration [25], depth estimation (monocular [15]) and (multi-view [10]), and pose regression [13]. With the introduction of DUSt3R [26], the paradigm of pointmap matching from image pairs has gained popularity, combining scalable training with the ability to connect all mentioned sub-tasks, providing a clear path to foundation models in 3D vision.

The main idea is that a neural network takes two input images $I_1, I_2 \in \mathbb{R}^{H\times W\times 3}$ and, in a shared coordinate space, predicts pointmaps $X_1, X_2 \in \mathbb{R}^{H\times W\times 3}$, which are a flexible representation that combines camera intrinsics and depth estimation. In a pointmap, each 3D point corresponds to one image pixel, i.e. $I[i,j] \leftrightarrow X[i,j]$.

Given ground-truth depth maps $Z_1, Z_2 \in \mathbb{R}^{H\times W}$, camera intrinsics K, and a 6D camera pose transform between the two images $T_{1\to 2}$, the target pointmap X_1 in the coordinate space of the second image is written as $X_{1\to 2}$, and can be computed straightforwardly:

$$X_{1\to 2}[i,j] = T_{1\to 2}K^{-1}\begin{bmatrix} i \\ j \\ Z_1[i,j] \end{bmatrix}$$

. In DUSt3R, a Siamese vision transformer is trained to directly estimate the pointmaps $\hat{X}_{1\to 2}, \hat{X}_2$, from the images. This is learned on a large dataset of images with known depths, poses, and intrinsics using a regression loss:

$$\min \left(\left\| \frac{1}{\hat{n}}\hat{X}_{1\to 2} - \frac{1}{n}X_{1\to 2} \right\|_2^2 + \left\| \frac{1}{\hat{n}}\hat{X}_2 - \frac{1}{n}X_2 \right\|_2^2 \right),$$

where $\hat{n}, n$ are normalization factors. This loss is averaged over all pixels with a valid ground truth depth and re-weighted with the model's predicted confidence scores [26]. The downstream tasks (calibration, pose estimation, etc.) can be solved from the predicted pointmaps; DUSt3R showed remarkable robustness to challenging scenes, changes in viewpoint and illumination. The follower model, MASt3R [12], augments the pointmap regression objective by a matching objective, vastly improving the results. The impressive performance and robustness of MASt3R raises the question of whether it generalizes to scenes outside of the training domain, like underwater imagery.

In particular, the pointmap paradigm allows for flexible handling of camera models, which is attractive for underwater scenes, as virtually all underwater imagery includes radial distortions from the refraction by the change of medium (air in camera, water outside of camera). Even though DUSt3R/MASt3R are trained only on image pairs with pinhole intrinsics, it seems straightforward to extend the training dataset to include other camera models. One of the aims of this paper is to evaluate whether MASt3R – that is trained on a wide range of pinhole cameras – generalizes to distorted lenses.

In MASt3R-SLAM [17], MASt3R is incorporated in a full-fledged real-time SLAM framework, including routines for tracking, relocalization, loop-closure, & global pose graph optimization. In particular, the flexibility of MASt3R with regards to the camera model is retained by using a raymap representation instead of explicit intrinsics.

3. Method

Here, we explore the combination of general-purpose semantic segmentation (Coralscapes) with state-of-the-art

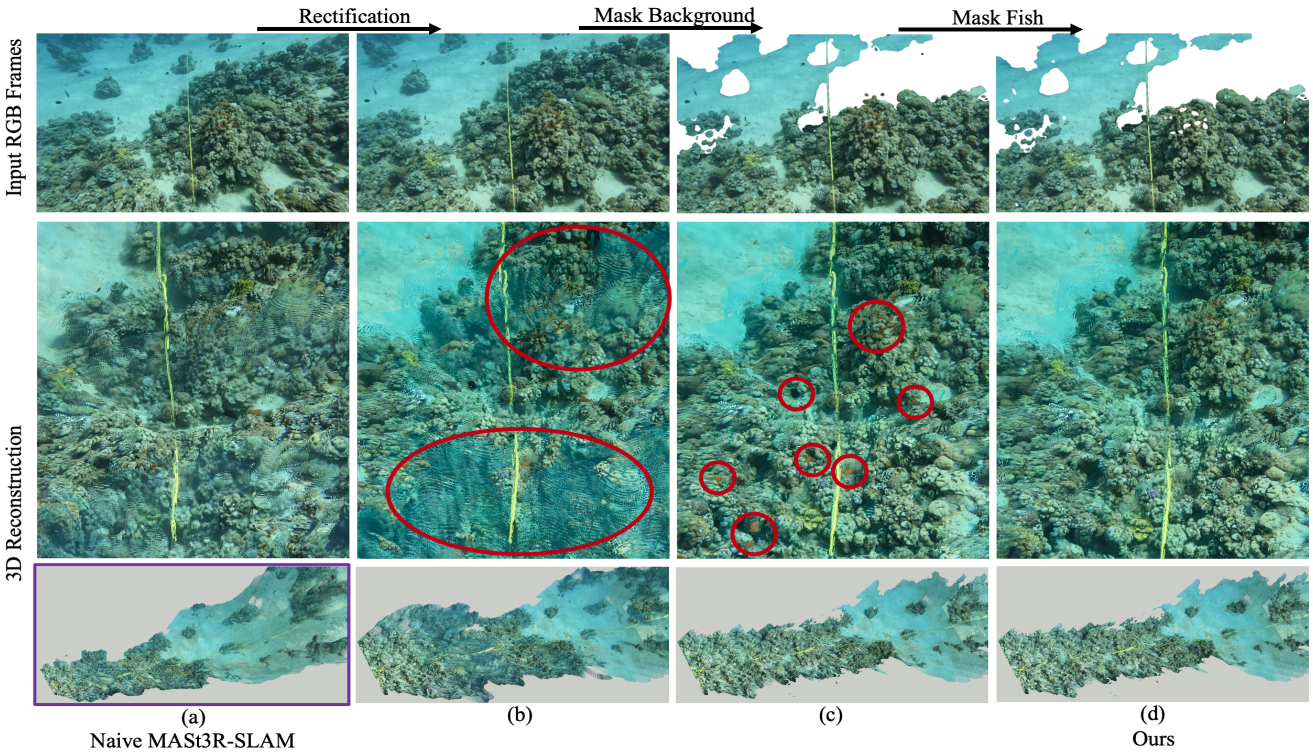

Figure 2. **Overview of the impact of rectification & masking:** When the original unrectified images are passed into MASt3R-SLAM (a), the distortion leads to a fishbowl effect, in which the reconstruction twirls upwards (purple box). While this effect is counteracted by rectifying the images, artifacts from the water column (b) and fish (c) have to be removed to obtain a clear point cloud of the reef (d).

neural SLAM (MASt3R-SLAM) for real-time semantic mapping of coral reefs with affordable consumer cameras. An overview is shown in Figure 1.

3.1. Semantic Segmentation

To enable real-time segmentation, we distill the Coralscapes SegFormer MiT-b5 model from Sauder et al. [22], which is trained to predict patches of size 1024×1024px, into a SegFormer MiT-b5 student model that segments entire frames at 1260×768px. For this, we sample random video frames from a large dataset of GoPro videos within coral reefs (62M frames) at 1920×1080px, obtain the prediction of the teacher model using four strided forward passes, and scale the image and the predictions to the resolution of the student model. The student model is trained for 100k steps with a batch size of 18 with only random horizontal flips as data augmentation using the AdamW optimizer with learning rate of $1e-5$, including 5k linear warmup and cooldown steps. Our student model surpasses the test accuracy ($82.761\% \rightarrow 83.676\%$) and mIoU ($57.800 \rightarrow 58.307\%$) of the teacher model and speeds up segmentation inference per frame from 4FPS to 18FPS on one Nvidia RTX3090 GPU.

3.2. Camera Calibration & Rectification

We use the COLMAP [24] SfM software to obtain camera intrinsics for a given camera. This step needs to be done only once for each camera or lens setting, and can then be re-used. We found that extracting 50 frames with sufficient overlap ($80-95\%$, roughly equivalent to $7-10$FPS when swimming forward) from a video where camera translation is dominant over rotation and the reef substrate is clearly visible leads to reliable 3D reconstructions with COLMAP. We calibrate using a radial camera model with shared intrinsic parameters for all given images, which is then used to rectify images into a pinhole camera model. This process takes less than 10 minutes on a CPU-only laptop.

3.3. MASt3R-SLAM

We extend MASt3R-SLAM to handle frame-wise segmentation masks. Images and segmentation masks are scaled to an appropriate size: MASt3R-SLAM processes images with the larger image dimension scaled to 512px, while keeping the aspect ratio (as close as possible, by a multiple of 16). In the matching stage of the tracking in MASt3R-SLAM, keypoints where the segmentation prediction is 'fish', 'diver', or 'background' are masked out. If tracking fails for a new frame because of insufficient num-

ber of matches, it is re-tried once without masking any keypoints. When the final point cloud is constructed (from the images, depth and pose estimates), the semantic class is attached to each 3D point, omitting points from the three masked classes.

4. Qualitative Evaluation

4.1. Dataset

We consider a video dataset from 6 reef transect sites in Aqaba, Jordan. Each site is filmed with video transects (each video is one back & forth) four times (at different times of day), following a 50m transect line, using a GoPro Hero 10 camera in the linear lens mode with a flat-port case. Calibration is performed once on one of the videos, and re-used for all 24 videos. The videos are filmed by divers swimming at speeds of $120 - 225$ seconds per 50m pass over the transect line, and generally follow a straight line at slightly downward facing pitch angle ($10 - 45$ degrees), deliberately excluding sudden camera movements.

4.2. Qualitative Results

Over all transect videos, we find that uncalibrated MASt3R-SLAM on unrectified images leads to poor global structure, with a bowl effect making the reconstruction curl upward (Fig. 2, blue box). When rectified images are used, this effect is largely counteracted. Without masking unwanted classes, artifacts from the water column or fish are visible in the reconstructions (Fig. 2, red circles). This is alleviated by the semantic masking using Coralscapes segmentation.

The results on all 24 transect scenes are summarized in Table 1. In none of the video transects, the forward and backward camera passes are correctly registered by the loop closure module: the remarkable robustness of MASt3R matching against extreme viewpoint changes does not generalize to the provided reef scenes. Tracking is successful for the entire video in 12 out of 24 scenes, and at least partially successful (more than $1/3$ of a video was tracked continuously) in 21 out of 24 scenes. In none of the scenes, MASt3R-SLAM recovered from a tracking loss via relocalization (using the opposing direction). Another challenge is scale drift: in 11 scenes with partially successful tracking, the scale drifted dramatically (with an increase/decrease factor of at least 2) over the tracked scene. Examples of the mentioned failure cases are shown in Figure 3.

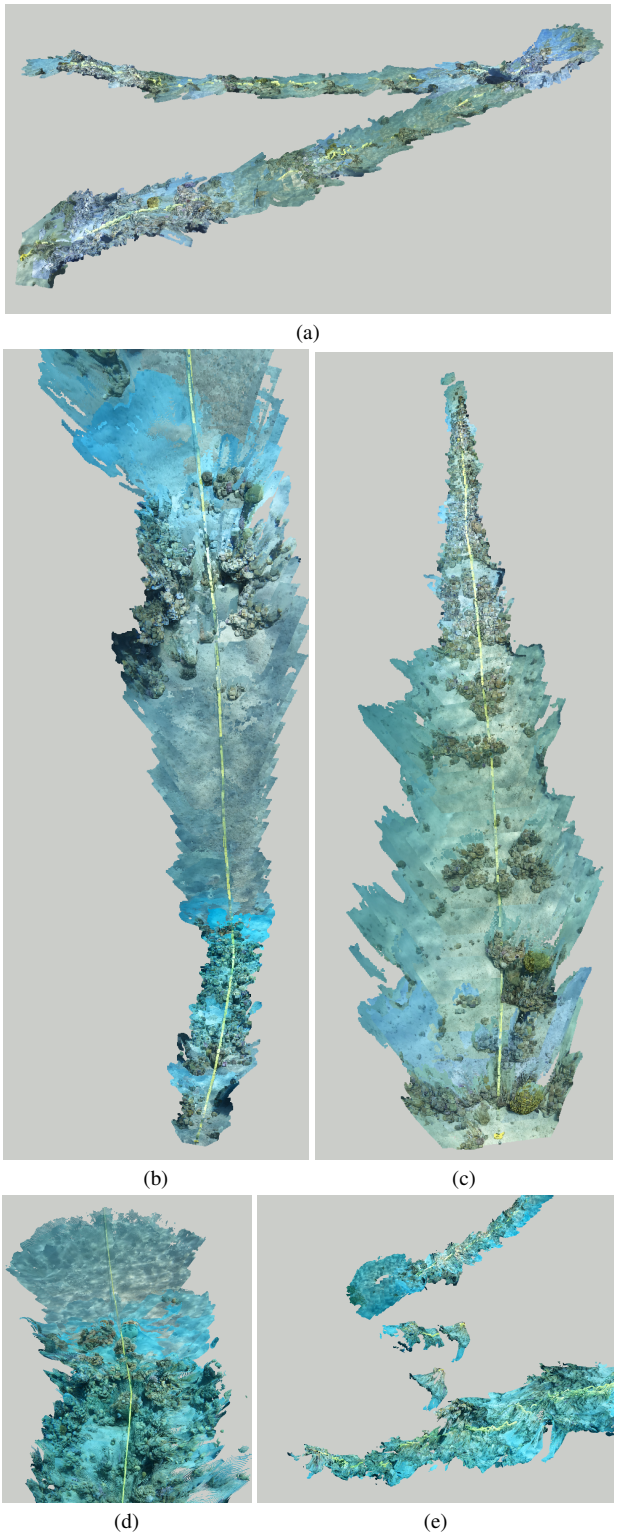

Figure 3. Failure cases: Even when tracking is successful throughout the whole video, there is no loop closure (a). Sometimes the scale explodes (b) or vanishes (c). Tracking failures (d) lead to unfinished reconstructions, and sometimes, the estimated geometry breaks down completely (e).

Loop Closure	Tracking	Stable Scale	Number of Transects
✓	✓	✓	0
×	✓	✓	5
×	✓	×	7
×	o	✓	5
×	o	×	4

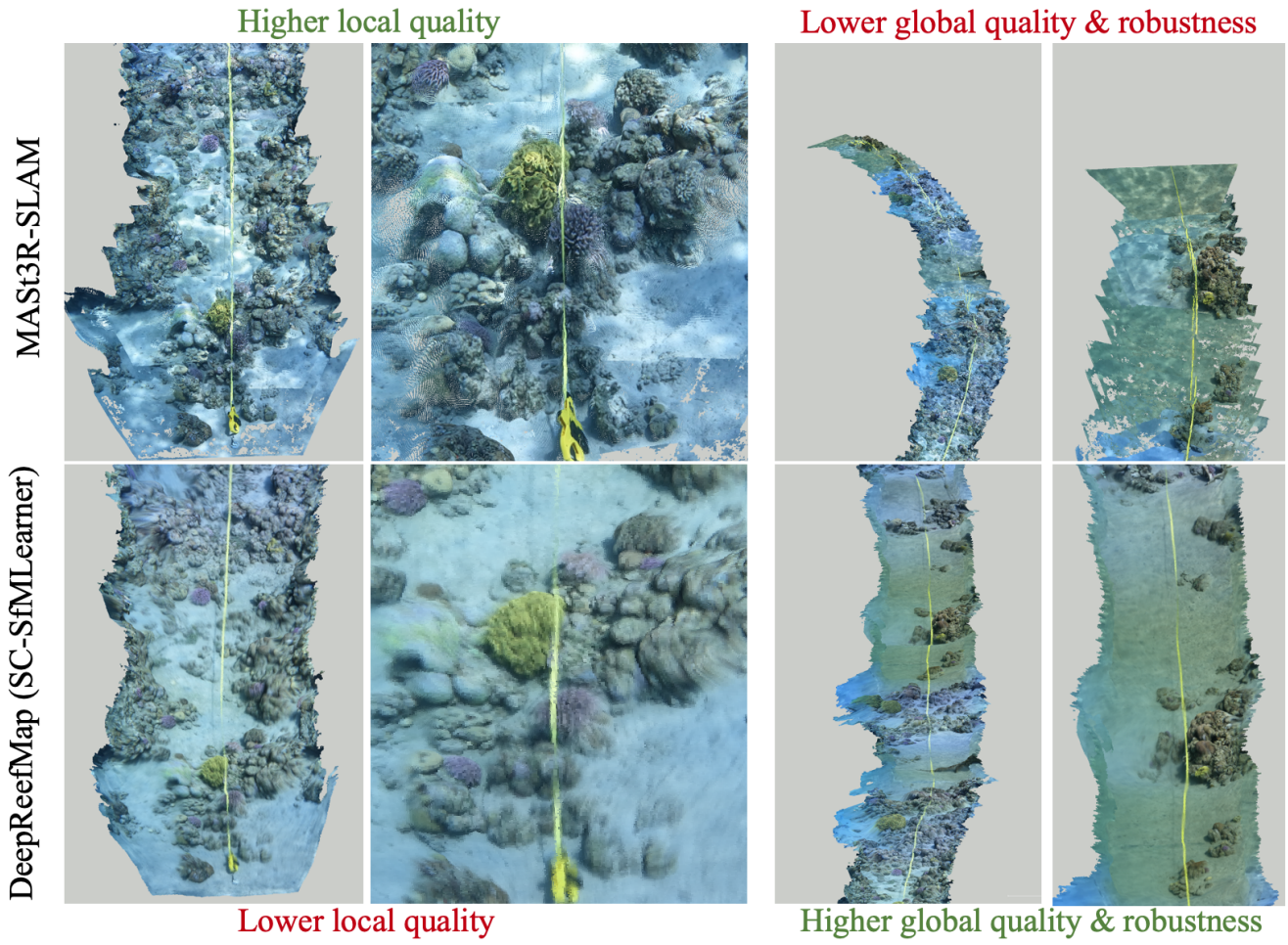

Figure 4. Qualitative comparison between reconstructions of MASt3R-SLAM (top) to DeepReefMap [23], which is currently based on the SC-SfMLearner [4] framework. Often MASt3R-SLAM produces significantly higher level of detail locally, but the global consistency remains challenging, exemplified by slight drift in pose or tracking failures (top right).

When compared directly to DeepReefMap [23], which is based on the SC-SfMLearner SLAM framework [4], we find that MASt3R-SLAM offers higher resolution and shaper details (Fig. 4): individual coral colonies, like the yellow Turbinaria, or the purple Stylophora, can be identified even from the 3D reconstruction. On the other hand, the flickering caustics from the sunlight hitting the wavy surface are propagated into the 3D model by MASt3R-SLAM.

5. Conclusion

Inspired by MASt3R-SLAM's remarkable robustness in other computer vision domains, this paper explores MASt3R-SLAM's effectiveness for mapping coral reefs. In our qualitative evaluation we find that:

- Even though MASt3R is trained on a wide variety of pinhole cameras, it does not generalize out of the box to cameras with radial distortion from refraction.
- Compared to MASt3R's demonstrated robustness to extreme pose differences, the robustness within reef scenes lags behind, reflected in the fact that MASt3R-SLAM's loop closure and relocalization modules do not work on our dataset of underwater scenes.
- Tracking works well in general, and produces higher quality reconstructions locally, when compared to DeepReefMap.

This suggests that MASt3R would substantially benefit from training or fine-tuning on in-domain data, motivating the creation of diverse high-quality 3D datasets from the coral reef domain or other ecological domains. Currently, for coral reef monitoring & conservation practitioners that use video data to map coral reefs, established systems still provide more robustness compared to MASt3R-SLAM. Improving camera-agnostic RGB SLAM systems like MASt3R-SLAM holds enormous potential for ecological monitoring in coral reefs and beyond, especially given the synergy with semantic segmentation.

References

[1] AIMS. Reefcloud.ai. https://reefcloud.ai/, 2023. 2

[2] Daniel TI Bayley and Andrew OM Mogg. A protocol for the large-scale analysis of reefs using structure from motion photogrammetry. *Methods in Ecology and Evolution*, 11(11): 1410–1420, 2020. 2

[3] Oscar Beijbom, Peter J Edmunds, David I Kline, B Greg Mitchell, and David Kriegman. Automated annotation of coral reef survey images. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1170–1177. IEEE, 2012. 2

[4] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Zhichao Li, Le Zhang, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth learning from video. *International Journal of Computer Vision*, 129(9):2548–2564, 2021. 5

[5] JHR Burns, D Delparte, RD Gates, and M Takabayashi. Integrating structure-from-motion photogrammetry with geospatial software as a novel technique for quantifying 3d ecological characteristics of coral reefs. *PeerJ*, 3:e1077, 2015. 2

[6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2

[7] Maoz Fine, Hezi Gildor, and Amatzia Genin. A coral reef refuge in the red sea. *Global change biology*, 19(12):3640–3647, 2013. 1

[8] Rebecca Fisher, Rebecca A O'Leary, Samantha Low-Choy, Kerrie Mengersen, Nancy Knowlton, Russell E Brainard, and M Julian Caley. Species richness on coral reefs and the pursuit of convergent global estimates. *Current Biology*, 25 (4):500–505, 2015. 1

[9] Ove Hoegh-Guldberg. Climate change, coral bleaching and the future of the world's coral reefs. *Marine and freshwater research*, 50(8):839–866, 1999. 1

[10] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2821–2830, 2018. 2

[11] Terry P Hughes, Andrew H Baird, David R Bellwood, Margaret Card, Sean R Connolly, Carl Folke, Richard Grosberg, Ove Hoegh-Guldberg, Jeremy BC Jackson, Janice Kleypas, et al. Climate change, human impacts, and the resilience of coral reefs. *science*, 301(5635):929–933, 2003. 1

[12] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 2

[13] Siddharth Mahendran, Haider Ali, and René Vidal. 3d pose regression using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 2174–2182, 2017. 2

[14] Valérie Masson-Delmotte, Panmao Zhai, Hans-Otto Pörtner, Debra Roberts, Jim Skea, Priyadarshi R Shukla, Anna Pirani, Wilfran Moufouma-Okia, Clotilde Péan, Roz Pidcock, et al. Global warming of 1.5 c. *An IPCC Special Report on the impacts of global warming of*, 1(5), 2018. 1

[15] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 2

[16] MERMAID. Marine ecological research management aid (mermaid). https://datamermaid.org/, 2023. 2

[17] Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16695–16705, 2025. 2

[18] NOAA. Fact sheet: Coral reefs. `https://www.coast.noaa.gov/states/fast-facts/coral-reefs.html`, 2022. Accessed: 2022-09-09. 1

[19] Eslam O Osman, David J Smith, Maren Ziegler, Benjamin Kürten, Constanze Conrad, Khaled M El-Haddad, Christian R Voolstra, and David J Suggett. Thermal refugia against coral bleaching throughout the northern red sea. *Global change biology*, 24(2):e474–e484, 2018. 1

[20] Gaia Pavoni, Massimiliano Corsini, Federico Ponchio, Alessandro Muntoni, Clinton Edwards, Nicole Pedersen, Stuart Sandin, and Paolo Cignoni. Taglab: Ai-assisted annotation for the fast and accurate semantic segmentation of coral reef orthoimages. *Journal of Field Robotics*, 39(3):246 – 262, 2022. 2

[21] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2

[22] Jonathan Sauder, Viktor Domazetoski, Guilhem Banc-Prandi, Gabriela Perna, Anders Meibom, and Devis Tuia. The coralscapes dataset: Semantic scene understanding in coral reefs. *arXiv preprint arXiv:2503.20000*, 2025. 2, 3

[23] J. Sauder, G. Banc-Praudi, A. Meibom, and **D. Tuia**. Scalable semantic 3d mapping of coral reefs with deep learning. *Methods Ecol. Evol.*, in press. 2, 5

[24] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[25] Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. Geocalib: Learning single-image calibration with geometric optimization. In *European Conference on Computer Vision*, pages 1–20. Springer, 2024. 2

[26] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2

[27] Matan Yuval, Iñigo Alonso, Gal Eyal, Dan Tchernov, Yossi Loya, Ana C Murillo, and Tali Treibitz. Repeatable semantic reef-mapping through photogrammetry and label-augmentation. *Remote Sensing*, 13(4):659, 2021. 2

[28] Matan Yuval, Iñigo Alonso, Gal Eyal, Dan Tchernov, Yossi Loya, Ana C Murillo, and Tali Treibitz. Repeatable semantic reef-mapping through photogrammetry and label-augmentation. *Remote Sensing*, 13(4):659, 2021. 2