Semantic Segmentation of Benthic Classes in Reef Environments using a Large Vision Transformer

Charlotte Sertic, Jonathan Sauder, and Devis Tuia

École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Abstract. Coral reefs are crucial for biodiversity and provide vital resources for humankind. But despite such a central role, they are confronted to increasing threats linked to climate change, pollution, and local stressors. To ensure effective conservation, efficient and scalable monitoring is key: this necessitates automated identification of benthic classes and their states on a large scale through semantic segmentation. However, segmentation of underwater videos is challenging, because of visual similarities between benthic classes, underwater distortions and limited available datasets, making it harder to create accurate and robust models. In this paper, we present a method for training a semantic segmentation model on a small dataset of video frames of coral scenes, by fine-tuning a large transformer model. Our approach uses transfer learning on the Segment Anything Model (SAM), incorporating specific training and prediction strategies. We benchmark our model against a CNN for semantic segmentation as a baseline. Our results demonstrate a substantial improvement in model performance, particularly for benthic classes that often appear as small objects and rarer classes, highlighting the potential of our approach in advancing coral reef mapping and monitoring.

1 Introduction

In the face of rising temperatures and various other stressors affecting coral ecosystems, coral monitoring is all the more important to help devise data-driven conservation strategies. Current well-established methods used to monitor corals range from divers photographing the seafloor to advanced multi spectral satellite imaging [15]. The most common approach to monitor benthic cover involves photo quadrats taken by divers, followed by expert analysis to assess coral abundance and bleaching status, which are then extrapolated to larger areas [11]. This process is labour-intensive and challenging to scale due to the need for expert evaluation. Computer vision has been applied to automate quadrat annotation [3, 18], but its efficacy hinges greatly on controlled conditions, given the complexities of underwater environments.

However, currently there are no automatic semantic segmentation systems for benthic semantic segmentation from general purpose reef images, as most systems are trained on domain-specific images such as photo quadrats [3, 5].

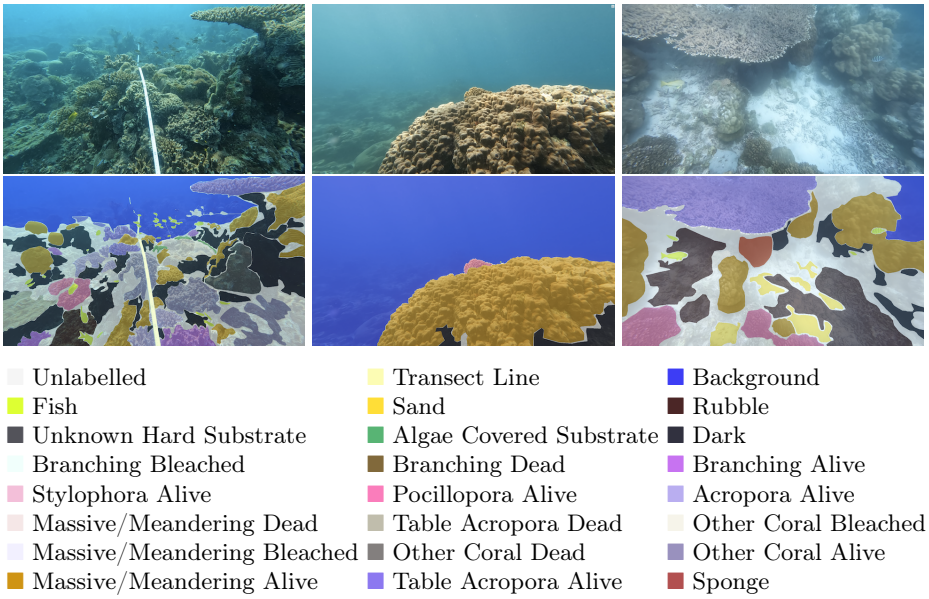


Fig. 1: Example dataset images with their labels. The labels encompass 30 classes with an additional class for unlabeled pixels.

To address these limitations, [12] proposed a coral monitoring approach for 3D mapping, integrating frame-by-frame semantic segmentation and learned Simultaneous Localization and Mapping (SLAM). This technique generates 3D point clouds with RGB data and semantic class labels, offering scalability and efficiency through video-based data acquisition. It streamlines benthic classification, eliminating the requirement for expert image analysis. This method's effectiveness relies on successful SLAM and a robust semantic segmentation algorithm.

In recent literature, most existing models are limited to deep learning approaches primarily relying on convolutional neural networks (CNNs) [8,9]. Following the excellent performance of large scale transformers in other tasks, our paper explores a large vision transformer, a modified version of Segment Anything (SAM [10]) that we name SAMarine, for enhancing the semantic segmentation model. Our model can then serve either as an improved building block for [12], but also as a stand alone tool for semantic segmentation of coral reefs, as it provides insight into the distribution of the main coral types as well as their condition (whether they are bleached, dead, or alive).

A limiting factor in the research is the lack of large and densely annotated datasets. Hence, there is a lack of well-trained transformer models and limited application of transfer learning from large datasets within this domain. Challenges also arise from imaging under water, creating blur, and lack of contrast due to

color attenuation and scattering [1,2,13], which varies with depth, temperature, salinity, the particles suspended in the water, and other factors. Moreover, corals show high morphological plasticity, which can make it extremely challenging to tell some corals apart [16].

We contribute by making a set of changes in the architecture and training procedure deviating from a standard semantic segmentation setup, leading to an mIoU improvement from 0.412 to 0.513 with respect to the original work of [12]. This improvement is seen on a challenging dataset of coral reef images with fine-grained semantic segmentation labels of benthic classes.

2 Dataset

The dataset used in this paper was collected from various coral reef sites in the Red Sea in Israel, Jordan and Djibouti as part of the Transnational Red Sea Center's [1] expeditions. The segmentation dataset is comprised of annotated video frames from a dataset videos gathered by a diver swimming between 1 and 4 meters above the reef following a transect line using a GoPro Hero 10 camera at 1080×1920px. The dataset spans a diverse array of underwater environments, ranging from water-dominated scenes to those teeming with diverse coral formations, as illustrated by examples in Fig. 1. Representative video frames of the environment were selected manually and annotated by coral experts with polygons of 30 distinct classes, including one extra class dedicated to unlabeled pixels. The inclusion of this class serves to handle complex scenes where class attribution might be uncertain, while addressing challenging scenarios with difficult to outline class boundaries. These unlabeled pixels fill transitional spaces between class boundaries in such cases. The full dataset consists of 986 training images (61868 polygons) and 99 test images (7247 polygons).

The dataset has significant class imbalance both in the frequency of classes as well as their apparent sizes. This is illustrated in Fig. 2 showing the relation between the median number of pixels per polygon to the number of polygons, where for certain classes the difference in the number of annotated polygons as well as the average polygon size varies with over two orders of magnitude. The rarity of some species, like seagrass, poses challenges for the model to learn effectively, resulting in very poor scores for these classes. Consequently, in our result analysis, we differentiate between the performance of common and rare classes. Furthermore, the imbalance in size among certain classes, exemplified by fish being at least one order of magnitude times smaller than all other classes, impedes learning due to their restricted pixel representation, rendering them more challenging to grasp.

[1] http://trsc.org

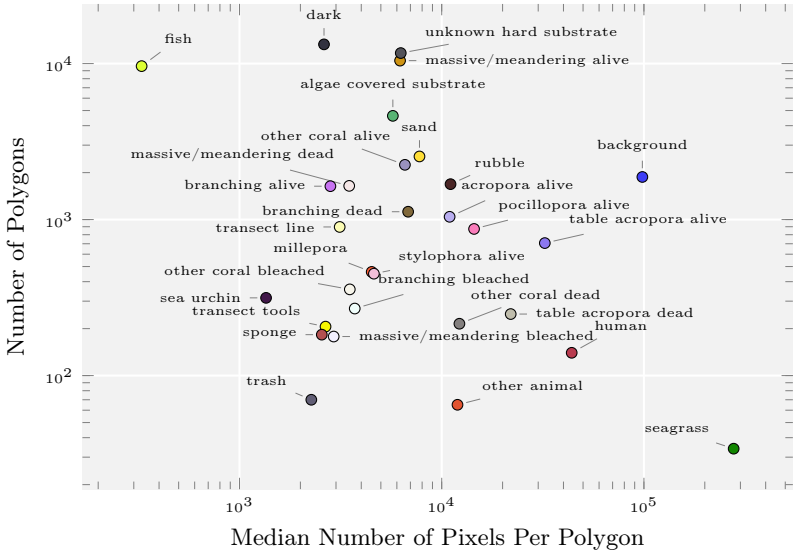


Fig. 2: Class Distribution. Relation of median number of pixels per polygon to the total number of polygons per class.

3 Method

3.1 SAMarine

Model Architecture. Our model, **SAMarine** (Fig. 3), is an adaptation of the pre-trained Segment Anything Model (SAM) [10] to the domain of semantic segmentation of corals. It is based on the architecture proposed by [20] for medical semantic segmentation. For our specific semantic segmentation task, we aim at classifying each pixel into one of the 30 benthic classes, deviating from the binary foreground-background segmentation in SAM. To do so, we design SAMarine to output 30 binary segmentation masks, one per benthic class. In these experiments, we do not actively design prompts, and use the default prompt from the original SAM.

Training. We use a hybrid loss function that combines cross-entropy loss ($Loss_{CE}$) and dice loss ($Loss_{Dice}$) [14], weighted by parameter λ as shown in Eq. (1). This approach allows us to strike a balance between distribution-based and geometry-based optimisation.

$$Loss_{Combined} = \lambda \cdot Loss_{CE} + (1 - \lambda) \cdot Loss_{Dice} \tag{1}$$

Moreover, we perform a number of augmentations (center / random cropping and resizing, rotation of up to 15°, jittering, horizontal flips) and vary their strength. To tackle the size of the pre-trained backbones, we use LoRa [7] and identify the best rank through ablations.

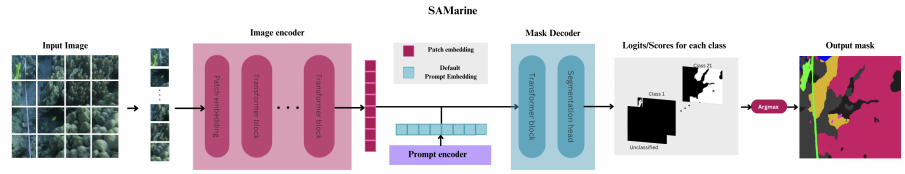


Fig. 3: SAMarine Architecture. Overview of our SAMarine approach. On the left side, the input image undergoes a patch-based embedding process before being forwarded to the transformer encoder. The resulting image embedding is then queried by a default prompt embedding in the mask decoder. The mask decoder generates a class-specific score mask, which leads to the output segmentation mask on the right side of the figure. Illustration inspired by [20].

Inference. To handle the loss of resolution that would be necessary if we fed an entire image (see next section Sec. 3.2), we feed the large input image to SAMarine in smaller patches and then reconstruct the full resolution map, where the softmax probabilities are averaged in the overlapping parts.

3.2 Baseline

The baseline method is a standard semantic segmentation setup. A DeeplabV3+ [4] model with a *resnext50_32x4d* backbone pre-trained on ImageNet [6], is trained using cross-entropy loss, excluding the unlabelled class, with classes weighted inversely to the square root of the pixel count to address class imbalance. We use random resized crops, horizontal flips, rotations and colour jitter as augmentations. For handling large images at inference, we resize the input image from 1080×1920px to 352×608px, and then upsample the prediction to the original size, as is standard practice in many semantic segmentation setups [17, 19].

4 Results

4.1 Implementation Details

We assess the baseline method using both the resizing and the patch-based prediction (as in SAMarine) with a patch size of 1024. Subsequently, we evaluate the SAMarine model, with the 'Vit-H' backbone with a LoRA rank of 128. The model is assessed using patch-based prediction with single patches of size 512. Unless specified otherwise, $\lambda = 1$ (a pure cross-entropy loss) is used with medium-strength augmentations.
To assess our models, we use mean Intersection over Union (mIoU) as the evaluation metric, considering the class imbalance. We categorise the results into common and rare classes, focusing on the former as the latter exhibits high volatility, making their outcomes less reliable.

4.2 Results

Results are shown in Tab. 1, where we observe an overall performance increase for the baseline from the resizing to the patching approach (0.412 to 0.431), mostly related to an improved performance in the *Common* classes (increase from 0.587 to 0.632). This shows that for underwater coral scenes, due to the abundance of details, the loss of resolution due to the resizing approach leads to a loss in performance.

Method	**mIoU**		
	Overall	*Common*	*Rare*
Baseline (Resizing)	0.412	0.587	0.271
Baseline (Patching)	0.431	0.632	0.281
SAMarine (Patching, $\lambda = 1$)	0.492	0.685	0.337
+ Balanced combined loss ($\lambda = 0.5$)	0.494	0.688	0.336
+ Stronger augmentations	**0.513**	**0.693**	**0.369**

Table 1: Numerical comparison between our SAMarine and the DeepLabV3+ baseline.

Most importantly, there is a significant performance leap observed from the baseline to our proposed SAMarine, visible in both rare and common classes. Our experiments revealed that employing a balanced combined loss with $\lambda = 0.5$, alongside more severe data augmentation strategies, resulted in improved performance, leading to an overall increase from 0.431 (baseline) to 0.492 (for the base SAmarine) and 0.513 (for the final model). The common classes enjoy an improvement of 6%, while the rare classes of 8%. The increased performance for smaller classes is attributed partly to the use of patch-based prediction, which preserves resolution and maintains accuracy in predicting these compared to resized prediction method. Additionally, the overall performance boost across all classes could stem from the efficacy of the transformer architecture employed or the benefits of transfer learning from pre-training on a large segmentation dataset.

Ablation studies about the different model architectures and the hyper-parameter λ are reported in Fig. 4. As illustrated in Fig. 4a, superior performance was attained using the larger 'Vit-H' backbone compared to 'Vit-B', along with a larger rank of 128. Additionally, in Fig. 4b, a $\lambda = 0.5$ yielded the best results, contrasting sharply with the inferior performance observed when using only the dice loss ($\lambda = 0$).

The visual analysis in Fig. 5 reveals insights into model performance on small classes. Compared to the baseline resizing approach, the patch-based method shows a marked improvement in predicting fish instances. Notably, the SAMarine model surpasses others by predicting more small fish instances with greater boundary delineation accuracy. This underscores SAMarine's effectiveness in capturing fine-grained details.

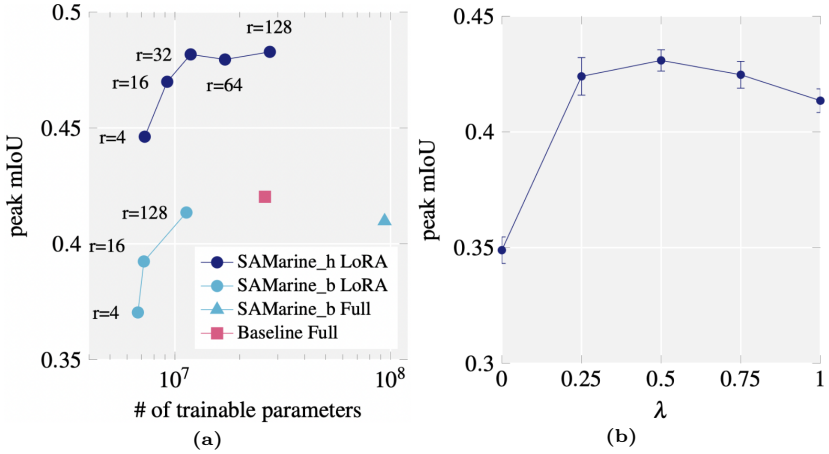
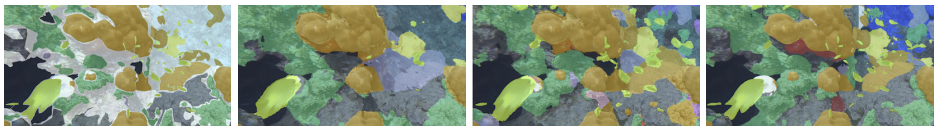


Fig. 4: Ablation Analysis. Figure (a) illustrates the ablations conducted to determine the best backbone architecture for SAMarine and the best rank, found to be using 'Vit-H' and a LoRA rank of 128. Figure (b) shows the best value for the loss function hyper-parameter λ as 0.5.



(a) Label (b) Baseline (Resiz.) (c) Baseline (Patch.) (d) SAMarine

Fig. 5: Qualitative Analysis. Example predictions using various methods. The fish class, highlighted in yellow, demonstrates superior performance and detail capture with the SAMarine approach.

5 Conclusion

We present a visual transformer-based model for automatic semantic segmentation of benthic classes in reef scenes on a dataset of expert-annotated video frames. This methodology incorporates transfer learning from a large transformer model (Segment Anything), strong augmentations, a combined loss function and patch-based prediction. Our model SAMarine outperforms the baseline, particularly in smaller classes, and is able to accurately predict boundaries around these. This is significant, as there are a lot of small benthic classes, which are challenging to discern in these unprocessed large underwater scenes. These findings also highlight the need for larger datasets of densely labeled images of benthic habitats in coral reef scenes, with a special focus on rare or naturally small organisms. Overall, improved performance in automatic semantic segmentation marks a solid step toward implementing coral monitoring at larger scale.

References

1. Akkaynak, D., Treibitz, T.: A revised underwater image formation model. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6723–6732 (2018) 3
2. Akkaynak, D., Treibitz, T.: Sea-thru: A method for removing water from underwater images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1682–1691 (2019) 3
3. Beijbom, O., Edmunds, P.J., Kline, D.I., Mitchell, B.G., Kriegman, D.: Automated annotation of coral reef survey images. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1170–1177 (2012). https://doi.org/10.1109/CVPR.2012.6247798 1
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018) 5
5. Chen, Q., Beijbom, O., Chan, S., Bouwmeester, J., Kriegman, D.: A new deep learning engine for coralnet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 3693–3702 (October 2021) 1
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848 5
7. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9 4
8. King, A., Bhandarkar, S., Hopkinson, B.: A comparison of deep learning methods for semantic segmentation of coral reef survey images. pp. 1475–14758 (06 2018). https://doi.org/10.1109/CVPRW.2018.00188 2
9. King, A., M.Bhandarkar, S., Hopkinson, B.M.: Deep learning for semantic segmentation of coral reef images using multi-view information. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019) 2
10. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (October 2023) 2, 4
11. Obura, D.O., Aeby, G., Amornthammarong, N., Appeltans, W., Bax, N., Bishop, J., Brainard, R.E., Chan, S., Fletcher, P., Gordon, T.A.C., Gramer, L., Gudka, M., Halas, J., Hendee, J., Hodgson, G., Huang, D., Jankulak, M., Jones, A., Kimura, T., Levy, J., Miloslavich, P., Chou, L.M., Muller-Karger, F., Osuka, K., Samoilys, M., Simpson, S.D., Tun, K., Wongbusarakum, S.: Coral reef monitoring, reef assessment technologies, and ecosystem-based management. Frontiers in Marine Science **6** (2019). https://doi.org/10.3389/fmars.2019.00580, https://www.frontiersin.org/articles/10.3389/fmars.2019.00580 1
12. Sauder, J., Banc-Prandi, G., Meibom, A., Tuia, D.: Scalable semantic 3d mapping of coral reefs with deep learning (2023) 2, 3
13. Sauder, J., Tuia, D.: Self-supervised underwater caustics removal and descattering via deep monocular slam. In: European Conference on Computer Vision (ECCV) (2024) 3

14. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3. pp. 240–248. Springer (2017) 4
15. Teague, J., Megson-Smith, D.A., Allen, M.J., Day, J.C., Scott, T.B.: A review of current and new optical techniques for coral monitoring. Oceans **3**(1), 30–45 (2022). https://doi.org/10.3390/oceans3010003, https://www.mdpi.com/2673-1924/3/1/3 1
16. Todd, P.: Todd p. a. — morphological plasticity in scleractinian corals. biological reviews. Biological reviews of the Cambridge Philosophical Society **83**, 315–37 (09 2008). https://doi.org/10.1111/j.1469-185X.2008.00045.x 3
17. Wang, J., Liu, B., Xu, K.: Semantic segmentation of high-resolution images. Science China Information Sciences **60**, 1–6 (2017) 5
18. Williams, I., Couch, C., Beijbom, O., Oliver, T., Vargas-Angel, B., Schumacher, B., Brainard, R.: Leveraging automated image analysis tools to transform our capacity to assess status and trends of coral reefs. Frontiers in Marine Science **6** (04 2019). https://doi.org/10.3389/fmars.2019.00222 1
19. Wittich, D., Rottensteiner, F.: Appearance based deep domain adaptation for the classification of aerial images. ISPRS Journal of Photogrammetry and Remote Sensing **180**, 82–102 (2021). https://doi.org/https://doi.org/10.1016/j.isprsjprs.2021.08.004, https://www.sciencedirect.com/science/article/pii/S0924271621002045 5
20. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv e-prints pp. arXiv–2304 (2023) 4, 5