CoralSCOP: Segment any COral Image on this Planet – Supplementary Material –

Ziqiang Zheng[1†] Haixin Liang[1] Binh-Son Hua[2] Yue Him Wong[3]
Put ANG Jr[4] Apple Pui Yi Chui[4] Sai-Kit Yeung[1]
[1]The Hong Kong University of Science and Technology [2]Trinity College Dublin
[3]Shenzhen University [4]The Chinese University of Hong Kong
† corresponding author: zhengziqiang1@gmail.com; Project website: https://coralscop.hkustvgd.com

Abstract

In this supplementary document, we provide additional information about our constructed CoralMask dataset (Sec. 1). More qualitative and quantitative outputs are also included (Sec. 2). We also provide the detailed implementations in Sec. 3. Finally, we provide more discussions in Sec. 4.

1. CoralMask Dataset

1.1. Dataset Comparison

We first provide a comprehensive comparison between our CoralMask dataset with existing underwater/marine datasets proposed for underwater visual scene understanding in Table 1. We summarize the differences between the existing underwater and marine datasets and our CoralMask dataset from different aspects: number of categories; number of images; annotation type; whether the viewpoint of captured images changes; the image diversity; whether the dataset contains low visibility images; whether the images contain the camouflaged objects; and the research purpose (task) of these proposed datasets. Please note Eilat Fluorescence [4], Mosaics UCSD [8], CoralNet [3] and CoralVOS [23] datasets are specially proposed for coral reef research. We argue the existing coral reef datasets are mainly collected at some specific angles (mostly bird-of-view) and essentially at the same depth, making it difficult and limited for data collection. Such data collection requirements make the coral reef research limited to a specific group of people. There are also some challenging scenarios missing, such as *low visibility*, *background clutter*, *motion blur*, *occlusion between overlapping objects*, *dynamic lighting*, *color distortion*, *irregular boundaries of coral reefs*, *viewpoint variations*, *scenario variation* and *a large range of image resolutions*. These motivate us to propose our CoralSCOP dataset, which contains significant data diversity. We refer the readers to check some demo images from our CoralMask in Figure 1 for more details.

1.2. Dataset Details

The coral reef images of our CoralMask dataset contain multiple data sources: the Internet (Flickr and Google); the public data of CoralNet [3]; existing public underwa-

Table 1. A direct comparison between our CoralMask dataset with existing underwater/marine datasets. – indicates the number cannot be reported.

Datasets	Categories	Images	Annotation	Viewpoint changes	Diversity	Low Visibility	Camouflaged	Task
SUIM [11]	8	1,500	Mask	✓	Medium	✓	×	Underwater scene segmentation
MAS3K [16]	37	3,103	Mask	×	Medium	×	✓	Marine animal segmentation
Wildfish [21]	1,000	54,459	Category	✓	High	×	×	Fine-grained fish classification
Wildfish++ [22]	2,348	103,034	Category	✓	High	×	×	Fine-grained fish classification
USOD10K [10]	–	10,255	Mask	✓	High	✓	✓	Underwater salient object detection
LaRS [24]	4	4,006	Mask	×	Medium	×	×	Marine obstacle segmentation
WaterMask [17]	7	4,628	Mask	✓	High	✓	×	Underwater instance segmentation
MarineDet [9]	821	22,679	BBOX	✓	High	✓	✓	Open-marine object detection
FishNet [13]	17,357	94,532	Category/BBOX	×	High	×	×	Fine-grained fish classification and detection
Eilat Fluorescence [4]	–	142	Category	×	Low	×	×	Dense coral segmentation
Mosaic UCSD [8]	34	4,193	Mask	×	Medium	×	×	Dense coral segmentation
CoralNet [3]	191	416,512	Category	×	High	×	×	Sparse point annotation
CoralVOS [23]	–	60,456	Mask	×	Medium	✓	×	Dense coral video segmentation
CoralSCOP	–	41,297	Mask	✓	High	✓	✓	Dense coral segmentation

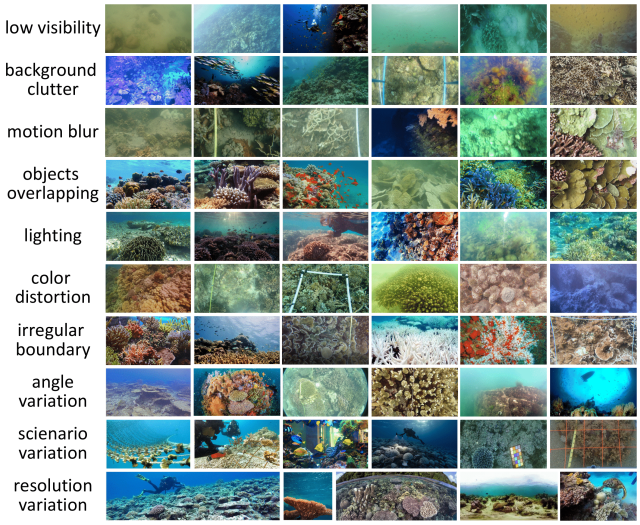

Figure 1. The example images from our CoralMask dataset. The collected coral reef images are from different environmental conditions (*e.g.*, weather, water turbidity, visibility, water depth and seasons), demonstrating a significant image diversity.

ter datasets [2, 8]; underwater surveying data contributed by coral biologists from various sites; and YouTube videos. The coral reef images are collected from various sites worldwide, with significant image diversity. The collected coral reef images have a large diversity, with large appearance, viewpoint and visibility variations. Handling complicated object shapes and boundaries is challenging because coral reef shapes can be very complex and vary greatly in size, shape, and texture. Our CoralMask dataset contains 41,297 coral reef images with 330,144 coral masks. The minimum, average, median and maximum number of coral masks for each image are *1, 8, 4 and 190*, respectively. We provide visualization of some example coral reef images (with large illumination, visibility, and diversity variations) from our CoralMask dataset in Figure 1.

1.3. Grouping Rules

All these images are labeled with dense pixel-level annotations. When performing the coral reef labeling, the annotators are asked to group the coral reefs with similar appearances to the same coral mask while separating the coral reefs with dissimilar appearances into different coral masks. In this way, we could promote the ability of the trained foundation model to yield the coral masks in a fine-grained manner and thus the generated coral masks could serve for the downstream user-defined tuning. The bleached corals have also been into consideration while the dead corals are ignored. When the coral mask annotations have been finished, the generated coral masks will be double-checked by two more different annotators to ensure the generated coral masks are accurate enough and remove the wrong coral masks.

2. Additional Results

2.1. Zero-shot Generalization Ability

We provide more qualitative coral segmentation results of coral reef images from different sites to demonstrate the strong zero-shot ability of our CoralSCOP in Figure 2. It is worth noting that all the coral masks are generated automatically without any prompts. All the testing images are downloaded from the Internet, with a significant image diversity. Please note there are some coral reefs are still missed by our CoralSCOP and our CoralSCOP also yields some false positives. We also provide more qualitative results of our CoralSCOP on the low visibility coral reef images in Figure 3, demonstrating the strong robustness of our CoralSCOP.

Comparison with SOTA. More qualitative comparisons with existing state-of-the-art semantic segmentation algorithms are provided in Figure 4. Both SAM and CoralSCOP could separate coral reefs with dissimilar appearances to different coral masks while other algorithms failed. We also provide more qualitative coral segmentation results of SAM and our CoralSCOP in Figure 5. As demonstrated, SAM heavily suffers from the over-segmentation problem and yields numerous false positives due to it cannot generate masks with semantics. In contrast, our CoralSCOP could effectively reduce the false positives and has a stronger ability to detect missed coral reefs by SAM.

2.2. Sparse-to-Dense

We provide a direct comparison between the generated coral masks by PLAS [20], SAM [14] and our CoralSCOP in Figure 6. The CPCe visualizations with sparse point annotations are also provided for reference. As illustrated, our CoralSCOP could generate more accurate and reliable coral masks based on the given sparse point annotations. The detailed and quantitative result comparisons of various algorithms for sparse-to-dense conversion are illustrated in Table 2. We report the mean average prediction error and corresponding standard deviation of the three rounds of experiments.

2.3. Benthic Coral Image Segmentation

The coral reef images with the benthic view are favored by coral biologists since they can directly compute the coral coverage from the labeled sparse points or the generated dense masks. We provide more qualitative coral segmentation results of the coral reef images with the benthic view in Figure 7. As reported, our CoralSCOP could also effectively the corals from the benthic coral reef images, thus qualifying the coral cover of different coral growth forms, genera, or species. The users could directly re-define the label of the

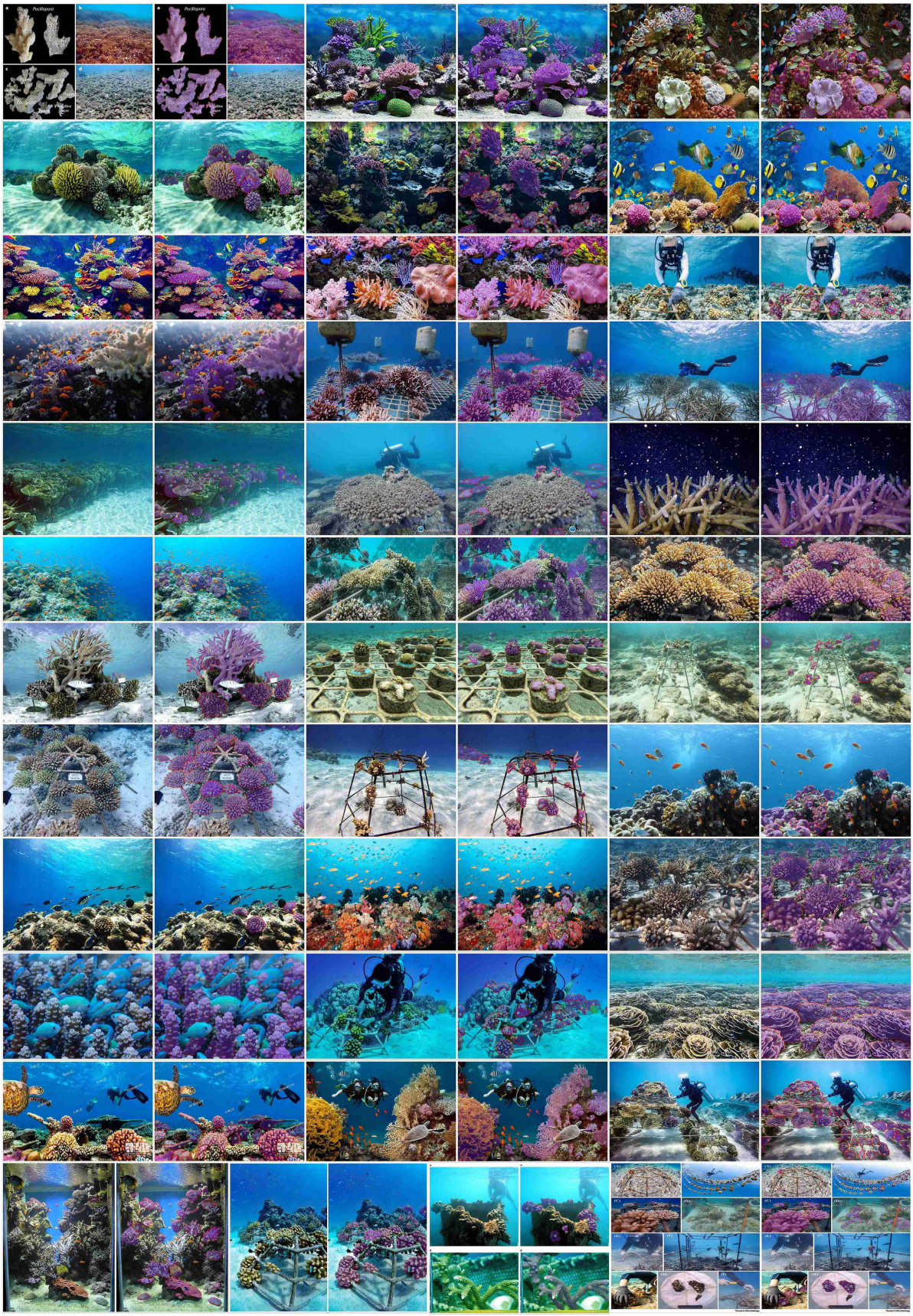

Figure 2. The zero-shot generalization ability of CoralSCOP to coral reef images from various sites. The left side is the input image while the right side illustrates the coral segmentation result of CoralSCOP.

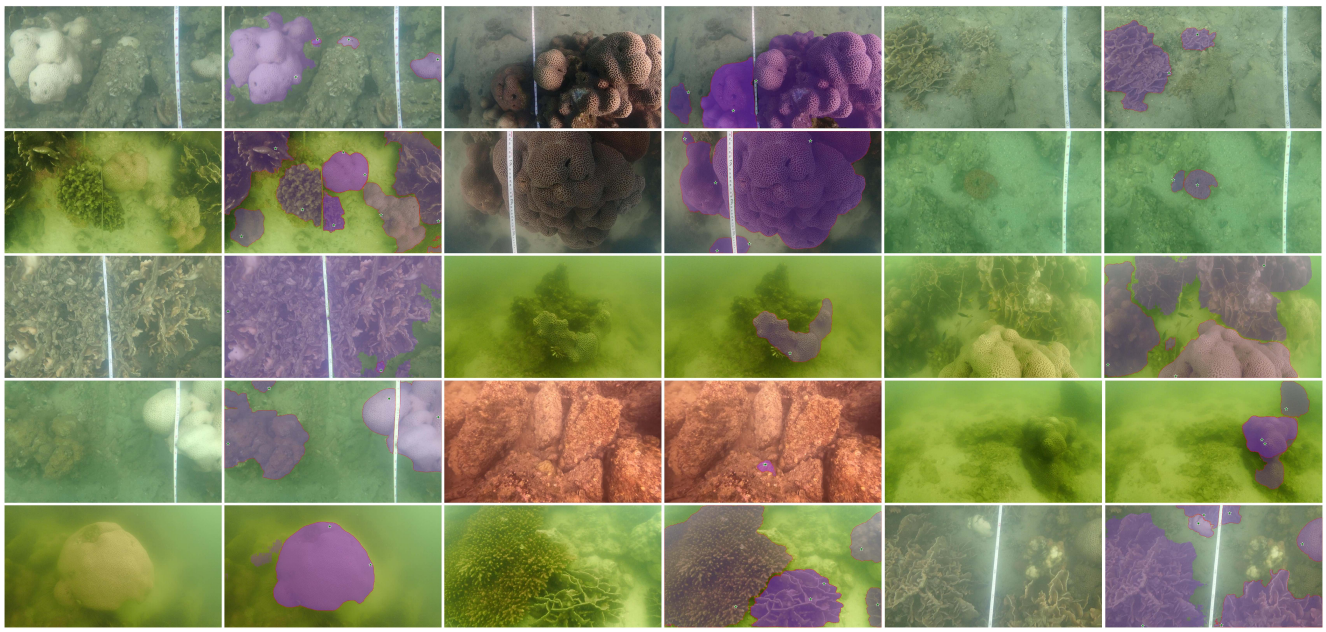

Figure 3. Our CoralSCOP demonstrates strong robustness to the low visibility coral reef images. The left side is the input image while the right side illustrates the coral segmentation result of our CoralSCOP.

Table 2. The **prediction error** under the original "sparse" (CPCe) setting and "dense" setting after sparse-to-dense conversion by SAM and CoralSCOP. We repeat the experiments 3 times to obtain the mean values and standard deviations.

Method	Setting	Non-coral (61.43)	Massive (22.51)	Laminar (13.06)	Branching (1.499)	Faliaceous (0.9614)	Encrusting (0.2788)	Columnar (0.2616)
CPCe [15]	10 points	$10.01_{\pm 0.2269}$	$8.903_{\pm 0.1059}$	$4.544_{\pm 0.1415}$	$0.678_{\pm 0.1804}$	$0.223_{\pm 0.0599}$	$0.480_{\pm 0.0756}$	$0.162_{\pm 0.0817}$
PLAS [20]		$\mathbf{9.886}_{\pm 0.5938}$	$8.490_{\pm 0.2171}$	$4.485_{\pm 0.3556}$	$0.675_{\pm 0.2451}$	$0.153_{\pm 0.0292}$	$0.397_{\pm 0.1358}$	$\mathbf{0.134}_{\pm 0.0932}$
SAM [14]		$16.12_{\pm 0.5343}$	$10.69_{\pm 0.0758}$	$5.846_{\pm 0.1779}$	$0.309_{\pm 0.1688}$	$0.164_{\pm 0.0888}$	$0.248_{\pm 0.0252}$	$0.135_{\pm 0.0896}$
CoralSCOP		$10.56_{\pm 0.7668}$	$\mathbf{6.371}_{\pm 0.1955}$	$\mathbf{3.677}_{\pm 0.2736}$	$\mathbf{0.287}_{\pm 0.2615}$	$\mathbf{0.034}_{\pm 0.0213}$	$\mathbf{0.211}_{\pm 0.0230}$	$0.198_{\pm 0.0899}$
CPCe [15]	20 points	$7.437_{\pm 0.7282}$	$6.218_{\pm 0.4712}$	$3.021_{\pm 0.2540}$	$0.536_{\pm 0.2408}$	$0.039_{\pm 0.0288}$	$0.308_{\pm 0.0907}$	$0.129_{\pm 0.0833}$
PLAS [20]		$6.391_{\pm 0.8405}$	$5.736_{\pm 0.2700}$	$2.952_{\pm 0.2462}$	$0.421_{\pm 0.2577}$	$0.050_{\pm 0.0058}$	$0.338_{\pm 0.0804}$	$0.141_{\pm 0.0942}$
SAM [14]		$10.71_{\pm 0.3234}$	$6.943_{\pm 0.3603}$	$3.871_{\pm 0.1936}$	$0.230_{\pm 0.0779}$	$0.249_{\pm 0.1666}$	$0.256_{\pm 0.0316}$	$\mathbf{0.055}_{\pm 0.0126}$
CoralSCOP		$\mathbf{5.334}_{\pm 0.0533}$	$\mathbf{3.791}_{\pm 0.1416}$	$\mathbf{1.555}_{\pm 0.0856}$	$\mathbf{0.119}_{\pm 0.0573}$	$\mathbf{0.049}_{\pm 0.0017}$	$\mathbf{0.225}_{\pm 0.0344}$	$0.145_{\pm 0.0930}$
CPCe [15]	50 points	$4.488_{\pm 0.3205}$	$3.569_{\pm 0.2603}$	$2.300_{\pm 0.1556}$	$0.382_{\pm 0.0135}$	$0.086_{\pm 0.0442}$	$0.217_{\pm 0.0394}$	$0.041_{\pm 0.0417}$
PLAS [20]		$3.747_{\pm 0.0866}$	$3.170_{\pm 0.0152}$	$1.671_{\pm 0.1763}$	$0.179_{\pm 0.0219}$	$0.065_{\pm 0.0177}$	$0.193_{\pm 0.0308}$	$\mathbf{0.021}_{\pm 0.0173}$
SAM [14]		$6.988_{\pm 0.4405}$	$5.194_{\pm 0.3053}$	$3.047_{\pm 0.3496}$	$0.158_{\pm 0.0367}$	$0.112_{\pm 0.0402}$	$0.545_{\pm 0.2232}$	$0.113_{\pm 0.0428}$
CoralSCOP		$\mathbf{2.832}_{\pm 0.2615}$	$\mathbf{2.657}_{\pm 0.2408}$	$\mathbf{1.396}_{\pm 0.2439}$	$\mathbf{0.042}_{\pm 0.0013}$	$\mathbf{0.022}_{\pm 0.0193}$	$\mathbf{0.154}_{\pm 0.0189}$	$0.221_{\pm 0.0031}$
CPCe [15]	100 points	$3.052_{\pm 0.2466}$	$2.648_{\pm 0.1399}$	$1.345_{\pm 0.2275}$	$0.175_{\pm 0.0180}$	$0.020_{\pm 0.0086}$	$0.159_{\pm 0.0204}$	$0.034_{\pm 0.0128}$
PLAS [20]		$2.300_{\pm 0.0757}$	$1.921_{\pm 0.0336}$	$1.253_{\pm 0.2002}$	$0.115_{\pm 0.0298}$	$0.024_{\pm 0.0198}$	$\mathbf{0.145}_{\pm 0.0320}$	$\mathbf{0.026}_{\pm 0.0165}$
SAM [14]		$7.436_{\pm 0.4764}$	$5.804_{\pm 0.4052}$	$3.378_{\pm 0.1325}$	$0.335_{\pm 0.0492}$	$0.188_{\pm 0.0538}$	$1.259_{\pm 0.0447}$	$0.098_{\pm 0.0234}$
CoralSCOP		$\mathbf{2.014}_{\pm 0.4266}$	$\mathbf{1.727}_{\pm 0.1306}$	$\mathbf{1.174}_{\pm 0.1168}$	$\mathbf{0.056}_{\pm 0.0135}$	$\mathbf{0.010}_{\pm 0.0028}$	$\mathbf{0.145}_{\pm 0.1657}$	$0.070_{\pm 0.0314}$

generated coral masks based on the required coral taxonomy to yield a more hierarchical and comprehensive biological report.

2.4. Ablation Studies

We provide a comprehensive analysis of the proposed CoralSCOP and dissect the advantage of CoralSCOP over the existing algorithms. We have included 1) SAM [14] (inference only); 2) SAM‡ (fine-tuned on our CoralMask dataset); 3) SAM-adapter [6] (fine-tuned on CoralMask with adapter design [6] while keeping $\texttt{Enc}(\cdot)$ frozen); 4) CoralSCOP^{-} (training without the negative non-coral masks); 5) CoralSCOP. All the algorithms have been conducted with three different backbones: Vit-B, Vit-L and Vit-H. Due to the constraint of computational resources, all the algorithms have been optimized for **one** epoch on our CoralMask dataset to guarantee a fair comparison. The detailed and comprehensive results of ablation studies are reported in Table 3. At the training procedure, the training prompt contains *1 random point* inside the labeled coral mask and the bounding box of the labeled coral mask or negative non-coral mask denoted as $\text{BBOX}_{\text{mask}}$. We compute results under three settings in Table 3. We could summarize such findings:

- 1) Directly fine-tuning SAM on CoralMask could promote

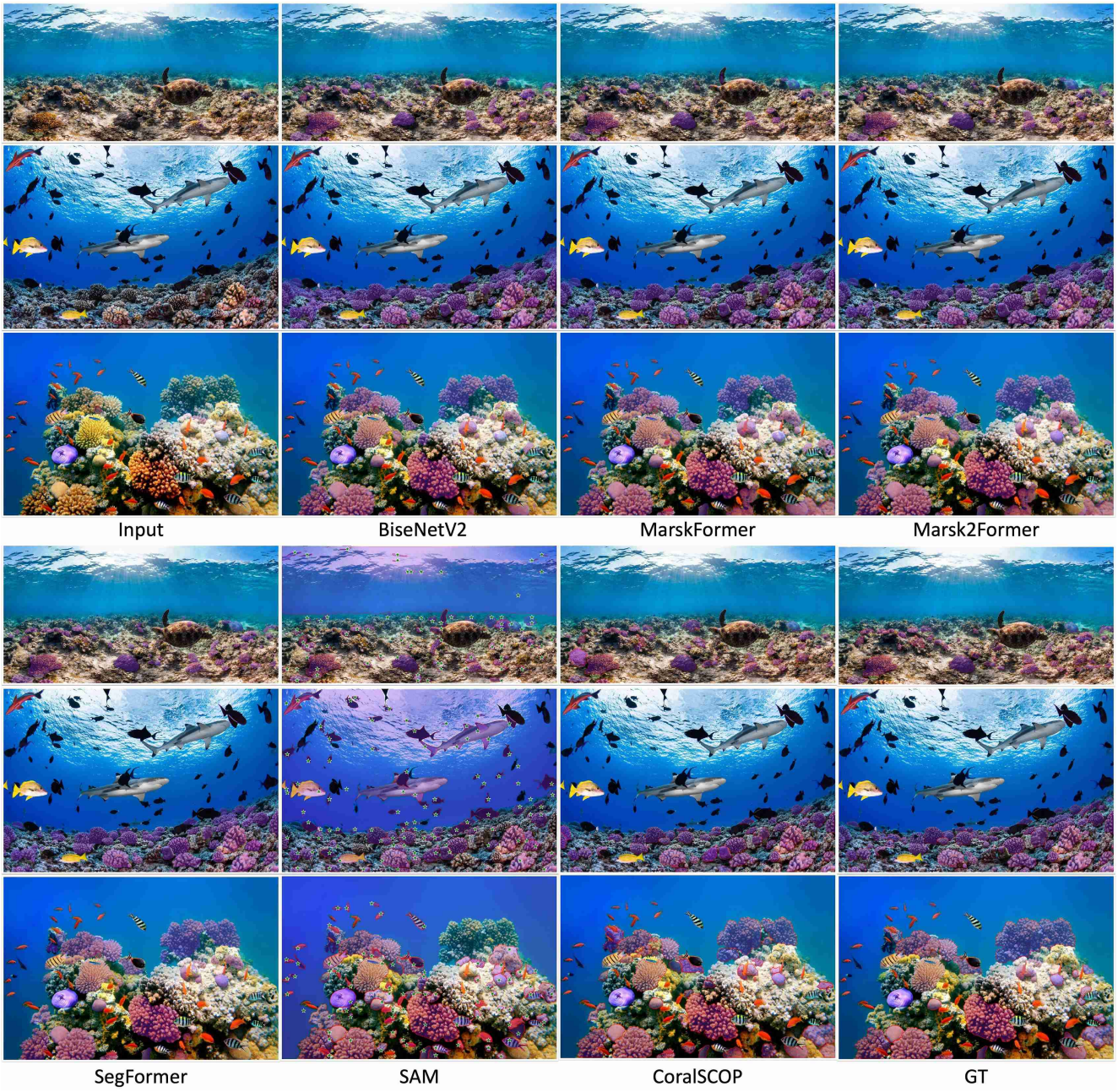

Figure 4. The qualitative zero-shot coral segmentation results of various algorithms. Note both SAM and CoralSCOP could generate multiple coral masks for further user-redefined. SAM still generates many false positives.

the ability of SAM to segment coral masks and a stronger backbone could achieve a larger performance gain, especially under the "1 point + $\text{BBOX}_{\text{mask}}^{\clubsuit}$" setting.

- 2) We observed the model with backbone Vit-L achieved the largest performance gain under the "Automatic$^{\heartsuit}$" setting. We attribute this phenomenon to the reason that it requires more training time for the model with Vit-H backbone to convergence and the data scale of our CoralMask dataset cannot fully unleash the power of a very big model. The appropriate model size (Vit-L over Vit-B) will lead to the best performance.
- 3) Fine-tuning $\texttt{Enc}(\cdot)$ together could result in better automatic coral segmentation performance by comparing SAM‡ and SAM-adapter under the settings of using all the three network backbones. Especially, under the setting of using a weak network backbone (Vit-B), the coral segmen-

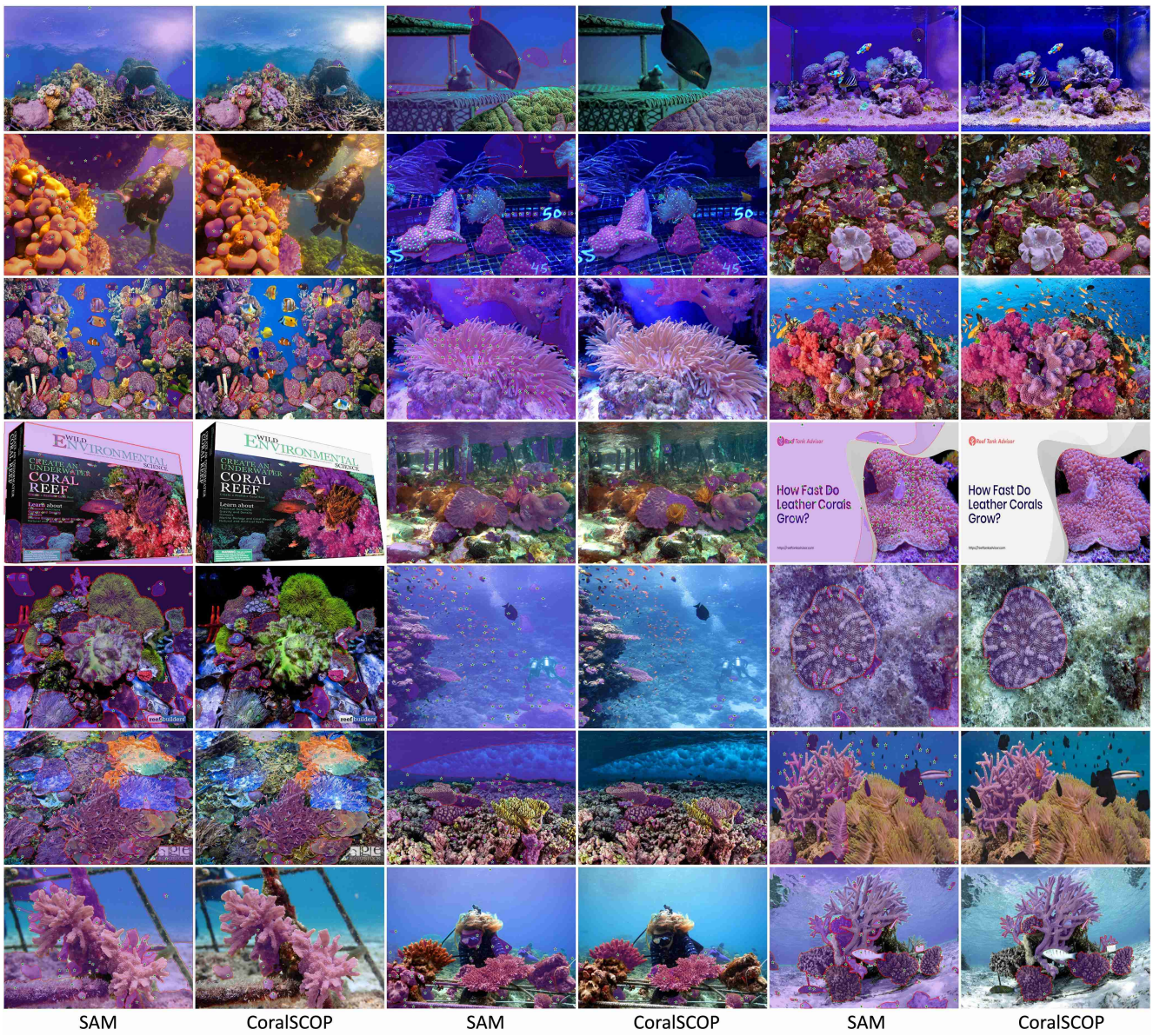

Figure 5. We provide direct comparisons between SAM and CoralSCOP. SAM still suffers from the over-segmentation problem and cannot generate accurate and complete coral masks.

tation performance of SAM-adapter drops a lot compared with SAM‡.

- 4) By comparing the coral segmentation performance of SAM‡ and our CoralSCOP, we observe that SAM‡ are with lower IoU scores while much higher MAE under almost all settings, indicating many false positives. Thus, solely fine-tuning SAM with coral masks from our CoralMask dataset cannot alleviate the over-segmentation problem well. The model has been taught what coral masks are, but not optimized by what are not coral masks. The model will tend to generate false positives on some unseen coral reef images.
- 5) By comparing the coral segmentation performance of CoralSCOP and CoralSCOP^{-}, we observe that preserving the negative non-coral masks could lead to observable IoU and pixel accuracy improvements under the "Automatic$^{\heartsuit}$" setting. With being optimized to recognize both coral masks and non-coral masks, the model could reduce the false positives. Meanwhile, the negative non-coral masks could also promote to alleviate knowledge forgetting and preserve the strong generalization ability, thus leading to a stronger model to segment coral individuals from unseen

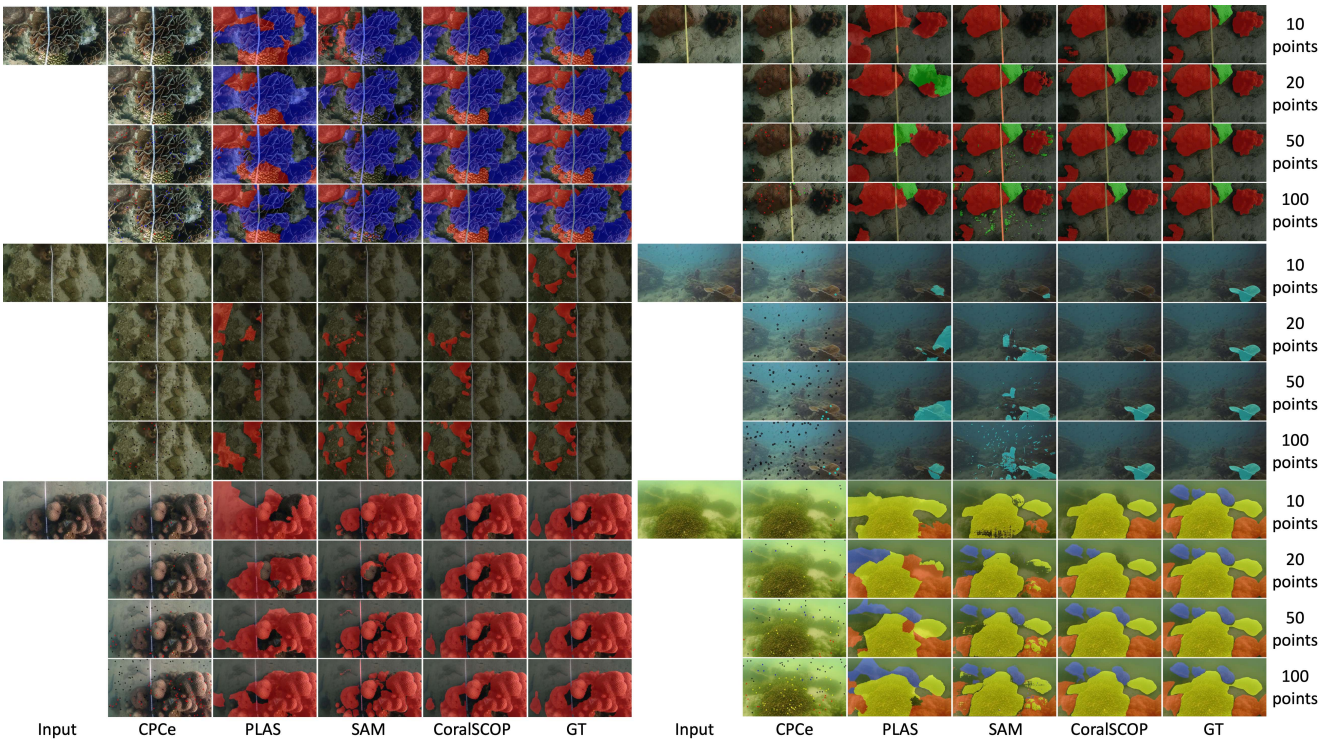

Figure 6. The sparse-to-dense conversion performance of various algorithms. The CPCe results with sparse point annotations are also provided. Best viewed in color.

Table 3. The coral mask generation quality of various algorithms under comprehensive settings.

Method	Backbone	Automatic$^{\heartsuit}$			1 point$^{\spadesuit}$			1 point + BBOX$_{\text{mask}}$$^{\clubsuit}$		
		IoU ↑	Accuracy ↑	MAE ↓	IoU ↑	Accuracy ↑	MAE ↓	IoU ↑	Accuracy ↑	MAE ↓
SAM [14]	Vit-B	23.61	44.18	0.3895	$45.76_{\pm 0.1946}$	$59.14_{\pm 0.0718}$	$0.3122_{\pm 0.0026}$	$63.03_{\pm 0.1007}$	$64.89_{\pm 0.1415}$	$0.1799_{\pm 0.0006}$
SAM‡		23.68	60.34	0.2551	$32.12_{\pm 0.2812}$	$53.77_{\pm 0.3630}$	$0.4148_{\pm 0.0007}$	$65.21_{\pm 0.1032}$	$67.34_{\pm 0.1098}$	$0.1676_{\pm 0.0001}$
SAM-Adapter [6]		8.821	32.34	0.3033	$32.47_{\pm 0.4317}$	$54.29_{\pm 0.2397}$	$0.3794_{\pm 0.0013}$	$62.78_{\pm 0.0836}$	$64.78_{\pm 0.0745}$	$0.1891_{\pm 0.0006}$
CoralSCOP^{-}		24.06	53.62	0.2915	$33.52_{\pm 0.4328}$	$48.52_{\pm 0.2461}$	$0.3816_{\pm 0.0019}$	$65.61_{\pm 0.1097}$	$67.26_{\pm 0.0987}$	$0.1666_{\pm 0.0006}$
CoralSCOP		26.45	56.45	0.2847	$37.96_{\pm 0.4348}$	$41.44_{\pm 0.2713}$	$0.3219_{\pm 0.0023}$	$66.78_{\pm 0.1303}$	$69.07_{\pm 0.1028}$	$0.1562_{\pm 0.0003}$
SAM [14]	Vit-L	29.83	39.52	0.4623	$41.36_{\pm 0.3996}$	$48.99_{\pm 0.1087}$	$0.4578_{\pm 0.0031}$	$57.40_{\pm 0.1888}$	$58.07_{\pm 0.1954}$	$0.2947_{\pm 0.0010}$
SAM‡		37.46	52.65	0.2614	$44.72_{\pm 0.5179}$	$56.95_{\pm 0.1279}$	$0.3264_{\pm 0.0020}$	$68.00_{\pm 0.0065}$	$70.16_{\pm 0.0690}$	$0.1470_{\pm 0.0001}$
SAM-Adapter [6]		34.38	46.19	0.3399	$43.83_{\pm 0.3061}$	$56.55_{\pm 0.1263}$	$0.3290_{\pm 0.0026}$	$67.34_{\pm 0.0673}$	$69.47_{\pm 0.0672}$	$0.1525_{\pm 0.0005}$
CoralSCOP^{-}		38.15	53.64	0.2601	$44.87_{\pm 0.5484}$	$56.87_{\pm 0.1325}$	$0.3275_{\pm 0.0062}$	$67.84_{\pm 0.0785}$	$69.12_{\pm 0.0773}$	$0.1475_{\pm 0.0003}$
CoralSCOP		46.46	75.62	0.1814	$45.65_{\pm 0.9050}$	$58.80_{\pm 0.1449}$	$0.2991_{\pm 0.0057}$	$68.66_{\pm 0.0754}$	$70.79_{\pm 0.0862}$	$0.1408_{\pm 0.0001}$
SAM [14]	Vit-H	31.16	38.05	0.5057	$44.67_{\pm 0.2025}$	$52.86_{\pm 0.0907}$	$0.3798_{\pm 0.0009}$	$72.80_{\pm 0.0060}$	$73.92_{\pm 0.0541}$	$0.1173_{\pm 0.0002}$
SAM‡		32.89	48.75	0.2900	$48.02_{\pm 0.3630}$	$58.23_{\pm 0.0839}$	$0.3200_{\pm 0.0035}$	$75.17_{\pm 0.0624}$	$77.21_{\pm 0.0234}$	$0.0994_{\pm 0.0001}$
SAM-Adapter [6]		37.55	44.60	0.3736	$46.92_{\pm 0.4367}$	$56.03_{\pm 0.0897}$	$0.3285_{\pm 0.0025}$	$75.14_{\pm 0.1001}$	$76.67_{\pm 0.1023}$	$0.1019_{\pm 0.0004}$
CoralSCOP^{-}		37.13	58.42	0.2743	$49.15_{\pm 0.2371}$	$62.11_{\pm 0.0975}$	$0.2462_{\pm 0.0018}$	$75.09_{\pm 0.0915}$	$76.37_{\pm 0.0891}$	$0.1024_{\pm 0.0004}$
CoralSCOP		37.42	68.82	0.2385	$52.92_{\pm 0.1266}$	$68.62_{\pm 0.0946}$	$0.2276_{\pm 0.0014}$	$76.62_{\pm 0.0616}$	$79.15_{\pm 0.0868}$	$0.0873_{\pm 0.0003}$

coral reef images.

2.5. Failure Cases

Our CoralSCOP is not without limitations. We found that CoralSCOP only demonstrates limited automatic coral mask generation performance under crowded scenarios with occlusions from dynamic objects when visibility is heavily limited. We provide some failure cases of our CoralSCOP in Figure 8 to inspire the whole coral reef research community on how to further promote the coral segmentation performance.

3. Implementation Details

Pre-training procedure. During the pre-training procedure, 1.3 million masks (including 330,144 coral masks labeled by coral biologists and 978,968 non-coral masks generated by SAM with Vit-H backbone) are utilized for training our CoralSCOP. We have optimized our CoraSCOP for 5 epochs

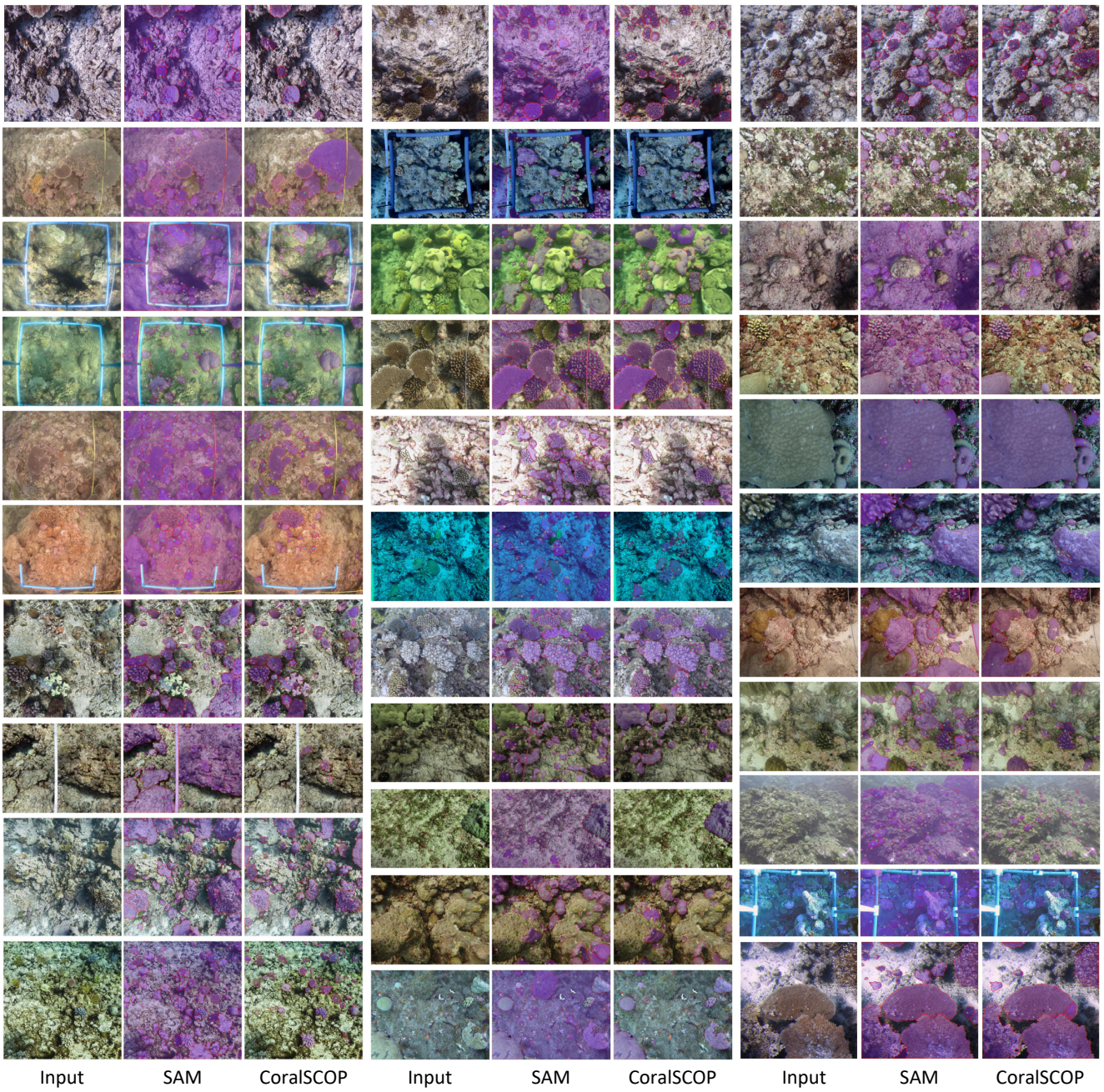

Figure 7. Coral segmentation performance comparison between SAM and CoralSCOP on benthic coral reef images. Compared with SAM, CoralSCOP could yield more accurate coral mask generation and reduce the false positives, thus leading to more reliable and accurate coral coverage computation.

on 6 Tesla A100 GPUs and the batch size per GPU is set to 1. It requires 17.2, 20.7 and 25.8 GPU hours to optimize the model with Vit-B, Vit-L and Vit-H backbones for only **one** epoch. We optimize the parameters of the whole model during the pre-training procedure to promote $\texttt{Enc}(\cdot)$ to extract underwater visual features as demonstrated in Figure 9. The composite prompts of point prompts (1, 2, 3 or 4 random points inside the coral mask) and bounding box prompts are utilized for training. The coarse mask prompts are not involved in our experiments. The final loss function is the sum of the classification loss $\mathcal{L}_{cls}$, Dice loss, Focal loss ($20\times$) and IoU loss. The latter three loss functions are the default loss functions from the original SAM implementation.

Tuning procedure. We have designed two tuning proce-

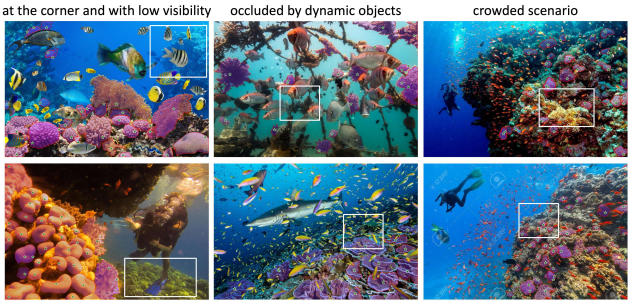

Figure 8. The failure cases of CoralSCOP. Best viewed in color.

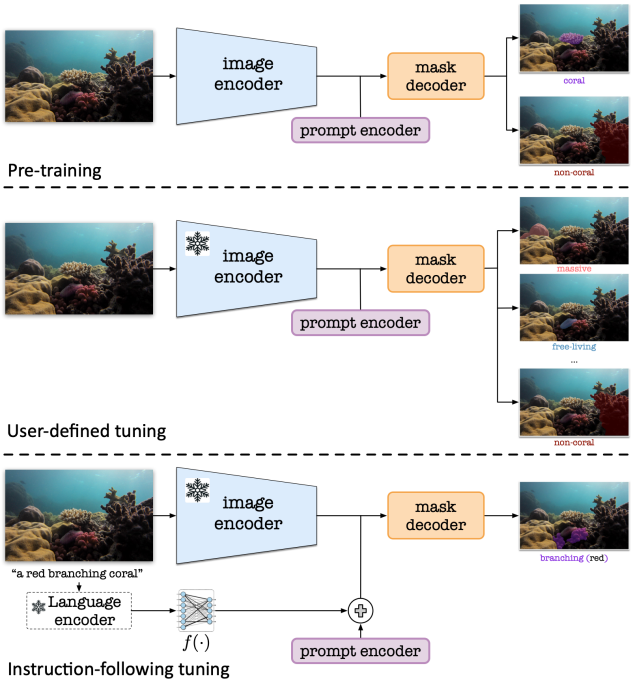

Figure 9. We provide the detailed overview frameworks of our CoralSCOP under the pre-training procedure; user-defined tuning procedure; and instruction-following tuning procedure.

dures: *user-defined* tuning and *instruction-following* tuning as illustrated in Figure 9. Please note the heavy image encoder $\texttt{Enc}(\cdot)$ is kept frozen during the tuning procedures.

For the *user-defined* tuning, the coral masks with corresponding user-defined semantic annotations are fed into CoralSCOP and we only optimize the MLP layer in $\texttt{Dec}(\cdot)$. The users could self-design the label of generated coral masks, ensuring the flexibility of the proposed CoralSCOP for downstream coral reef analysis tasks.

For the *instruction-following* tuning, we borrow the language decoder of CLIP [19] to generate the textual embedding to empower CoralSCOP with the ability to understand the user intents. The CLIP model has been optimized by a large scale of image-text pairs and possesses a strong ability to recognize and understand visual concepts, object shapes, texture, and color, by associating visual images with corresponding textual descriptions. Under this setting, we have formulated 46,610 instruction-following pairs. 4,661 coral images with 831 different coral species are crawled from "Corals of the World" [7]. For coral mask generation, we only label the corals belonging to the given coral species when there are multiple coral species. The generated sentences are paired with the labeled images to formulate textual input and mask output pairs. There are 831 coral species from 142 coral genera in total. Based on the coral species names, we asked ChatGPT-3.5 [18] to generate 5 sentences to describe the distinctive appearances of such coral species. Then we pair such 5 sentences with the images from the corresponding coral species. We adopt a simple prompt template to formulate the final textual input: "this image contains <*coral species name*>. <*description generated by ChatGPT-3.5*>". For example "this image contains *Acropora aspera*; *Acropora aspera* is *a branching-like coral and it exhibits a vibrant and intricate structure*". The texts in *Italic* will be replaced by other counterparts based on the corresponding annotations.

We regard these formulated textual inputs and coral mask outputs as positive pairs. We have also constructed the negative pairs to alleviate the hallucination, which means the generated coral mask does not match the given textual description. We randomly sample the textual descriptions from other coral genera to formulate the negative pairs considering the coral reef images from different coral genera share various appearance representations. There are in total 46,610 (4,661 $\times$ 5 $\times$ 2) pairs for performing the instruction-following tuning. The generated sentences are paired with the labeled images to formulate textual input and mask output pairs. We optimize the whole mask decoder under this setting.

Coral mask generation. We have designed three different experimental settings for coral mask generation:

- "Automatic$^{\heartsuit}$" (no prompt is given). We follow the automatic mask generation pipeline of SAM [14] and generate the grid points (32 $\times$ 32) as point prompts for automatic coral mask generation. The IoU threshold and stability threshold are set to 0.82 to remove the automatically generated low-quality coral masks.
- "1 point prompt$^{\spadesuit}$" where one random point inside each coral mask is given as point prompt. The point prompt is randomly sampled from the whole coral reef mask.
- "1 point prompt and $\text{BBOX}_{\text{mask}}{}^{\clubsuit}$" (one random point inside the coral mask and BBOX of the coral mask are provided together as prompts). Besides, $\text{BBOX}_{\text{image}}$ indicates that we utilize the width and height of the whole image as the box prompt.

3.1. User-defined Tuning

In this section, we provide more details about the training data involved in the user-defined tuning procedure.

Growth form. We adopt 500 coral reef images with 6 growth form definitions:

- **Massive** corals grow in a spherical or hemispherical manner, being solid. They are resistant to strong water currents and are therefore commonly found in shallow and mid-depth waters.
- **Encrusting** corals (also called crustose corals) are highly tolerant of strong water currents. It is a growth form in which the coral colony forms a thin, flat layer that adheres tightly to the substrate.
- **Foliaceous** corals, such as the pagoda coral Tubinaria mesenterina, are scroll-like in their appearance. They form horizontally flattened, unifacial plates or lobes that are attached to the reef substrate from the basal (ventral) surface.
- **Columnar** corals, such as the catch bowl coral (Isopora palifera), are pillar or finger-like corals that form. Columnar corals do not have the secondary branches seen in the branching coral growth type.
- **Laminar** corals have a flat upper surface which gives them a table-like structure. A growth form in which the coral colony forms flat, leaf-like structures that are attached to the substrate.
- **Branching** corals, such as the thin birds-nest coral (Seriatopra hystrix), often found in areas of high wave action, are antler or staghorn-like in their appearance.

All these coral reef images are with a benthic view.

Genus. There are 14 different coral genera involved in genus-level coral recognition: **Goniopora**; **Lithophyllon**; **Plesiastrea**; **Pavona**; **Platygyra**; **Dipastrea**; **Echinophyllia**; **Porites**; **Leptastrea**; **Favites**; **Cyphastrea**; **Coscinaraea**; **Galaxea**; **Acropora**. We have included 400 coral reef images for training and testing. Similarly, all the coral reef images are with the benthic view. Genus-level coral recognition is more challenging than recognizing coral growth forms due to several factors: the finer distinctions required, the diversity within genera, and the subtle morphological differences among closely related genera. We provide the example images with semantic masks (both growth form and genus) in Figure 10 and Figure 11, respectively. In Figure 11, to avoid potential misleading and provide better readability, we only visualize one coral mask belonging to the given coral genus for better illustration. The images are captured from different islands and sites in Hong Kong with significant diversity.

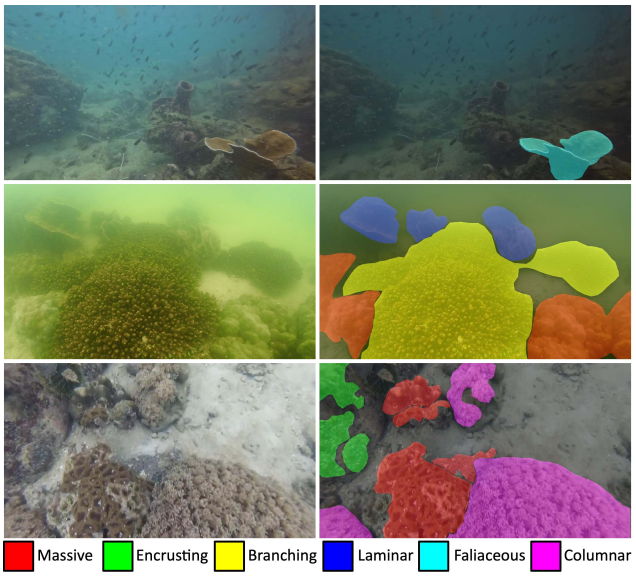

Figure 10. Example coral masks from the selected 6 growth forms.

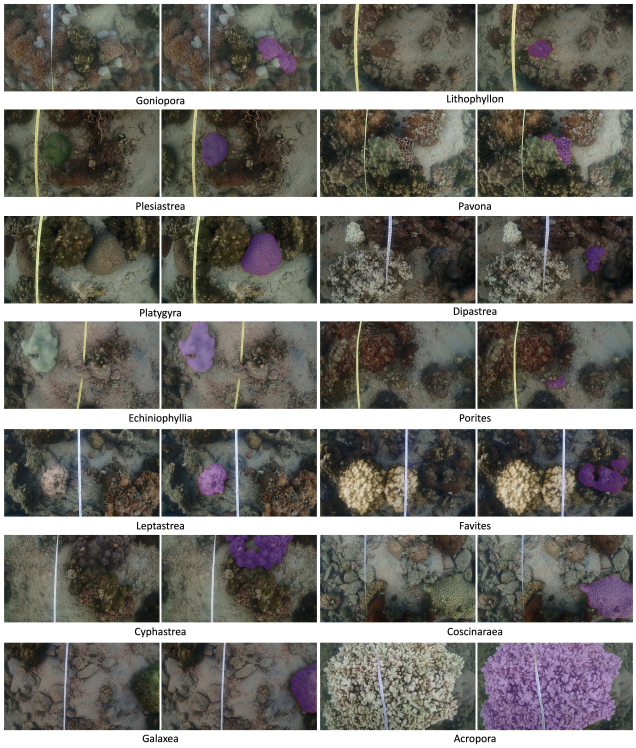

Figure 11. Example coral masks from the selected 14 coral genera.

4. Discussions

4.1. Comparison with Existing Coral Analysis

There are two essential differences between CoralSCOP and existing CPCe [15], ReefCloud [1], and CoralNet [3, 5]:

- **Zero-shot generalization ability**. For unseen coral images, CoralNet requires few-shot training samples (*e.g.*, 20 labeled training samples) to perform model fine-tuning for discriminating unseen coral images. The users are required to do the sparse point based annotation based on random or determined points (usually 100 points), making the model segment unseen reef images. The proposed CoralSCOP could perform zero-shot dense coral segmentation without new training samples.
- **Sparse vs. Dense**. CoralSCOP yields dense pixel-wise segmentation outputs, which could generate more precise and accurate coral statistics compared with the CPCe and CoralNet. The sparse point based analysis may lead to over/under estimations and cannot reveal coral boundaries. Furthermore, the dense segmentation could serve for downstream 3D reconstruction and video monitoring. CoralSCOP provides advanced analytical capabilities required for in-depth spatial analysis, statistical modeling, and dense segmentation.

CoralNet (semi-automatic) and CPCe rely on manual identification and counting of coral points, which introduces a degree of subjectivity and potential bias. More importantly, the manual process can be time-consuming, especially when analyzing large datasets or conducting repeated surveys. These challenges require a coral foundation model with a stronger capability for accurate and robust dense coral mask generation.

4.2. Potential Impact

Foundation model for coral reefs. CoralSCOP provides an advanced interactive way for both amateurs and experts to obtain required coral masks. CoralSCOP could unleash the power of textual descriptions, providing an interactive way to segment corals for both amateurs and experts. We envision our attempt as one of the first steps towards scientific discovery in the coral reef domain assisted by the foundation model. The pre-trained powerful coral foundation model will further promote the understanding of coral reefs, and pave the foundation for future discoveries.

Coral statistics. The coral biologists could utilize our CoralSCOP to compute the coral statistics for the coral reef images with the benthic view. The coral coverage statistics could be the ratio of the generated dense coral masks over the total image area. The abundance, coverage, composition, and distribution of coral reefs play a very important role in marine ecosystems. With the increasing attention and more advanced equipment in collecting underwater surveying/diving videos, huge coral reef images/videos have been collected for different purposes. CoralSCOP contributes towards efficient coral analysis for these collected coral reef images/videos, yielding coral coverage, composition and population estimation.

4.3. Application

3D coral reconstruction. The generated coral masks could be used for promoting 3D coral scene understanding. We perform structure-from-motion (Pix4D [1] is used in our experiments) for 3D reconstruction to better model the structure and geometry information of coral colonies following the experimental setup of [23]. The segmented coral masks are utilized as binary masks to remove the noisy background and preserve the 3D coral scene. The corresponding 3D reconstruction results with the orthographic view are illustrated in Figure 12. With the coral masks generated by CoralSCOP, we could significantly reconstruct more **accurate**, **robust**, and **detailed** coral colonies without geometry distortions. Meanwhile, we could also remove the background of the 3D model for better monitoring of the coral ecosystems.

Coral reef rendering. Similarly, the coral masks could also promote the coral rendering performance [12]. We adopt the official codes [2] of 3D Gaussian Splatting to perform experiments and report the corresponding coral rendering performance in Figure 13 under the two settings: without and with the coral masks generated by CoralSCOP. With the help of our CoralSCOP, we could obtain better rendering results with better visual quality.

Bleached coral analysis. CoralSCOP could also be utilized for bleached coral analysis, identifying the bleached coral areas and computing the relative bleached coral ratio. We first utilize CoralSCOP to automatically segment the coral masks and then generate the bleached coral areas as follows:

- **Coral mask generation**. We utilize CoralSCOP to generate the coral mask and we only preserve these detected coral areas considering that the bleached coral could only come from these coral areas.
- **RGB→Greyscale**. We convert the RGB images to greyscale images to alleviate the influence of color intensity changes.
- **Thresholding**. We convert the grayscale images into binary mask images based on a user-defined bleaching threshold. The value over the given bleaching threshold is set to 1 and otherwise 0.

Through these, we could roughly obtain the bleached coral masks. We provide some results by using different values of the bleaching threshold in Figure 14. The coral biologists could adjust the bleaching threshold according to their requirements. The identification of bleached coral needs calibration from coral biologists, so we leave it open for the user to define the scale of bleaching.

References

[1] Reefcloud.ai. https://reefcloud.ai/. 10

[1] https://www.pix4d.com/

[2] https://github.com/graphdeco-inria/gaussian-splatting

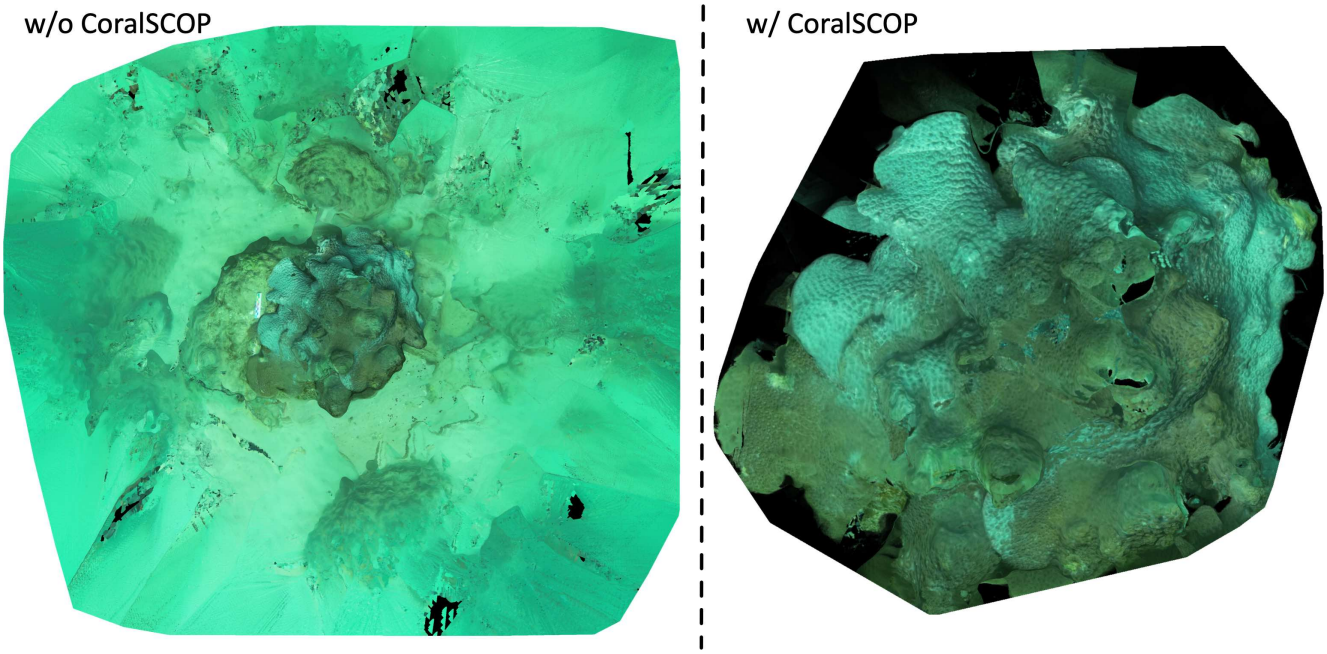

Figure 12. The 3D coral reconstruction under the two settings: without and with our CoralSCOP. The generated coral masks could be utilized for removing the non-coral backgrounds and alleviating the geometric distortions, enabling the coral scene understanding in a 3D fashion.

Figure 13. The radiance field rendering under two settings: without and with our CoralSCOP. The generated coral masks could be utilized for promoting the rendering performance. Please zoom in to check more details.

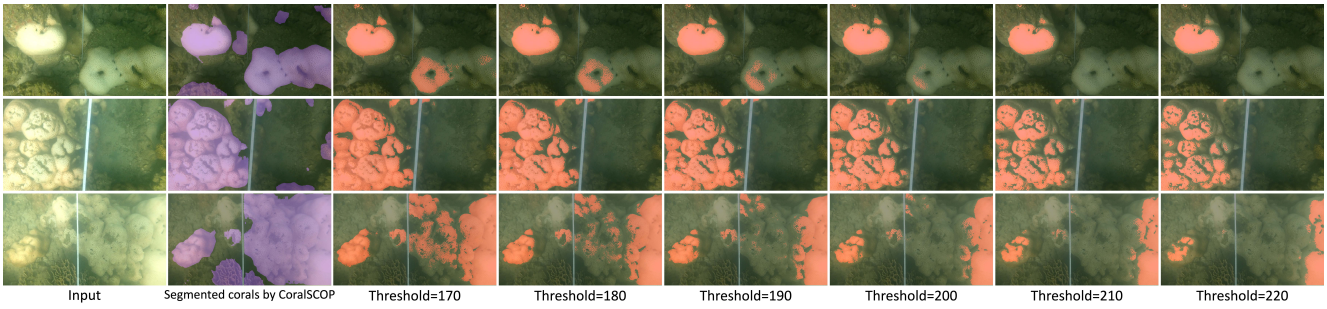

Figure 14. The bleached coral analysis by our CoralSCOP of using different values of the bleaching threshold.

[2] Inigo Alonso, Matan Yuval, Gal Eyal, Tali Treibitz, and Ana C Murillo. Coralseg: Learning coral segmentation from sparse annotations. *Journal of Field Robotics*, 36(8):1456–1477, 2019. 2

[3] Oscar Beijbom, Peter J Edmunds, Chris Roelfsema, Jennifer Smith, David I Kline, Benjamin P Neal, Matthew J Dunlap, Vincent Moriarty, Tung-Yung Fan, Chih-Jui Tan, et al. Towards automated annotation of benthic survey images: Variability of human experts and operational modes of automation. *PloS one*, 10(7):e0130312, 2015. 1, 10

[4] Oscar Beijbom, Tali Treibitz, David I Kline, Gal Eyal, Adi Khen, Benjamin Neal, Yossi Loya, B Greg Mitchell, and David Kriegman. Improving automated annotation of benthic survey images using wide-band fluorescence. *Scientific reports*, 6(1):1–11, 2016. 1

[5] Qimin Chen, Oscar Beijbom, Stephen Chan, Jessica Bouwmeester, and David Kriegman. A new deep learning engine for coralnet. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3686–3695, 2021. 10

[6] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Shangzhan Zhang, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. *arXiv preprint arXiv:2304.09148*, 2023. 4, 7

[7] Turak E. and DeVantier L.M. Corals of the world. `http://coralsoftheworld.org/v0.01(Beta)`, 2016. 9

[8] Clinton B Edwards, Yoan Eynaud, Gareth J Williams, Nicole E Pedersen, Brian J Zgliczynski, Arthur CR Gleason, Jennifer E Smith, and Stuart A Sandin. Large-area imaging reveals biologically driven non-random spatial patterns of corals at a remote reef. *Coral Reefs*, 36(4):1291–1305, 2017. 1, 2

[9] Liang Haixin, Zheng Ziqiang, Ma Zeyu, and Sai-Kit Yeung. Marinedet: Towards open-marine object detection. *arXiv preprint arXiv:2310.01931*, 2023. 1

[10] Lin Hong, Xin Wang, Gan Zhang, and Ming Zhao. Usod10k: a new benchmark dataset for underwater salient object detection. *IEEE Transactions on Image Processing (TIP)*, 2023. 1

[11] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1769–1776. IEEE, 2020. 1

[12] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4): 1–14, 2023. 11

[13] Faizan Farooq Khan, Xiang Li, Andrew J Temple, and Mohamed Elhoseiny. Fishnet: A large-scale dataset and benchmark for fish recognition, detection, and functional trait prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20496–20506, 2023. 1

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 4, 7, 9

[15] Kevin E Kohler and Shaun M Gill. Coral point count with excel extensions (cpce): A visual basic program for the determination of coral and substrate coverage using random point count methodology. *Computers & geosciences*, 32(9): 1259–1269, 2006. 4, 10

[16] Lin Li, Eric Rigall, Junyu Dong, and Geng Chen. Mas3k: An open dataset for marine animal segmentation. In *International Symposium on Benchmarking, Measuring and Optimization*, pages 194–212. Springer, 2020. 1

[17] Shijie Lian, Hua Li, Runmin Cong, Suqi Li, Wei Zhang, and Sam Kwong. Watermask: Instance segmentation for underwater imagery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1305–1315, 2023. 1

[18] OpenAI. Introducing chatgpt. 2022. 9

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 9

[20] Scarlett Raine, Ross Marchant, Brano Kusy, Frederic Maire, and Tobias Fischer. Point label aware superpixels for multi-species segmentation of underwater imagery. *IEEE Robotics and Automation Letters (RA-L)*, 2022. 2, 4

[21] Peiqin Zhuang, Yali Wang, and Yu Qiao. Wildfish: A large benchmark for fish recognition in the wild. In *ACM international conference on Multimedia (ACM MM)*, pages 1301–1309, 2018. 1

[22] Peiqin Zhuang, Yali Wang, and Yu Qiao. Wildfish++: A comprehensive fish benchmark for multimedia research. *IEEE Transactions on Multimedia (TMM)*, 23:3603–3617, 2020. 1

[23] Zheng Ziqiang, Xie Yaofeng, Liang Haixin, Yu Zhibin, and Sai-Kit Yeung. Coralvos: Dataset and benchmark for coral video segmentation. *arXiv preprint arXiv:2310.01946*, 2023. 1, 11

[24] Lojze Žust, Janez Perš, and Matej Kristan. Lars: A diverse panoptic maritime obstacle detection dataset and benchmark. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1