CoralSCOP: Segment any COral Image on this Planet

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

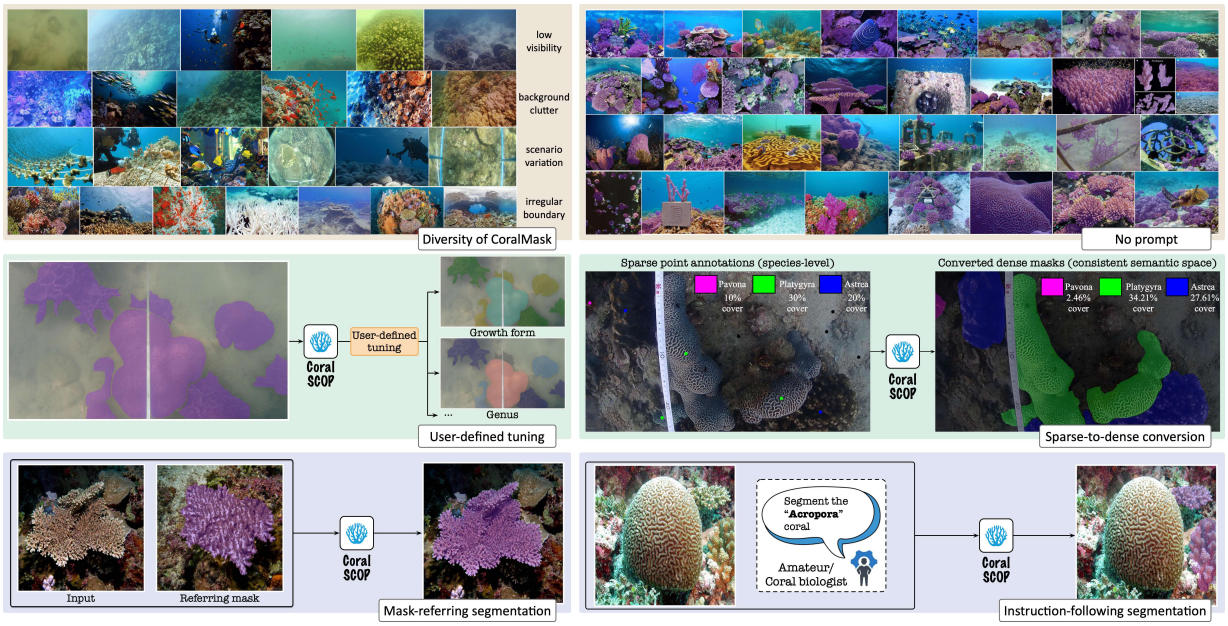

Figure 1. We present CoralSCOP, the inaugural foundation model proposed for dense coral segmentation, utilizing our newly proposed CoralMask dataset, which comprises 41,297 coral reef images and 330,144 coral masks. CoralSCOP demonstrates a strong generalization ability to unseen coral reef images, automatically generating reasonable coral masks with no prompts, while clustering similar coral reefs into the same coral mask and splitting dissimilar coral reefs into multiple coral masks. CoralSCOP could serve user-defined tuning and sparse-to-dense conversion to obtain more fine-grained, reliable and accurate coral statistics. The mask-referring segmentation and instruction-following segmentation enable both amateurs and coral biologists to generate coral masks.

Abstract

Underwater visual understanding has recently gained increasing attention within the computer vision community for studying and monitoring underwater ecosystems. Among these, coral reefs play an important and intricate role, often referred to as the rainforests of the sea, due to their rich biodiversity and crucial environmental impact. Existing coral analysis, due to its technical complexity, requires significant manual work from coral biologists, therefore hindering scalable and comprehensive studies. In this paper, we introduce ***CoralSCOP****, the first foundation model designed for the automatic dense segmentation of coral reefs. CoralSCOP is developed to accurately assign labels to different coral entities, addressing the challenges in the semantic analysis of coral imagery. Its main objective is to identify and delineate the irregular boundaries between various coral individuals across different granularities, such as coral/non-coral, growth form, and genus. This task is challenging due to the semantic agnostic nature or fixed limited semantic categories of previous generic segmentation methods, which fail to adequately capture the complex characteristics of coral structures. By introducing a novel parallel semantic branch, CoralSCOP can produce high-quality coral masks with semantics that enable a wide range of downstream coral reef analysis tasks. We demonstrate that CoralSCOP exhibits a strong zero-shot ability to segment unseen coral images. To effectively train our foundation model, we propose* ***CoralMask****, a new dataset with* ***41,297*** *densely labeled coral images and* ***330,144*** *coral masks. We have conducted comprehensive and extensive experiments to demonstrate the advantages of CoralSCOP over existing generalist segmentation algorithms and coral reef analytical approaches.*

1. Introduction

Coral reefs are vibrant and diverse ecosystems in shallow (<50m depth) and sunlit areas in tropical and subtropical seas. They are home to a vast biodiversity of marine life, including algae, invertebrates like hard and soft corals, worms, crustaceans, and echinoderms; and vertebrates like fish and marine turtles. Because of their biological importance, many measures [12, 18, 29, 33, 38] are now put under protection against human and natural threats. Underwater monitoring of reef structures, including changes in coral cover and composition, has become a routine but critical exercise to assess the health and stability of coral reef ecosystems and functions over time. To date, however, monitoring works on coral reefs remain a very time-consuming and labor-intensive task involving the physical collection of underwater photo images from the reef site and analysis of the photo images obtained (either still photos or videos). With the advancement of computer vision technology and artificial intelligence, it is envisioned that such routine monitoring of coral reef images could become more efficient with better precision and accuracy in generating data statistics that provide the community for formulating adaptive coral reef protection and conservation strategies.

Most existing coral reef analytical approaches focus on analyzing coral reef images and videos collected *in-situ* during underwater coral surveyings [15, 45, 53]. Domain expertise is required to annotate coral data, such as annotating coral growth form or genus in images and videos. To reduce complexity, existing dominant coral reef analysis algorithms are apt to involve down-sampling and adopt only some keyframes for analysis, in which a selected number of quadrats [53] are placed to determine the surveying areas. To cope with manual annotation, sparse points [3, 8, 32] are randomly sampled or specifically determined on images so that coral biologists can annotate them, facilitating the computation of coral statistics such as cover percentage, composition, and population density.

There are two main issues for these existing analytical approaches: 1) as mentioned earlier, manual annotation can be time-consuming and labor-intensive when analyzing large-scale data or conducting repeated surveys. It also imposes limitations on the spatial and temporal scales for biological surveying and is highly dependent on a group of specific coral biologists. **Scalability** has become one of the most important limiting factors that hinder more comprehensive coral reef studies. 2) The **down-sampling** [32] involved in the whole procedure (sampling quadrats areas and sampling sparse points) will inevitably introduce bias to the estimation outputs, leading to over/under estimations. Furthermore, sparse point coral analysis cannot accurately delineate *irregular boundaries* of corals in the coral reefs.

Dense segmentation [7, 10, 11, 56] provides a reasonable and satisfactory solution for coral image analysis, identifying and delineating the complex boundaries of coral colonies. It supports identifying individual coral species, thus quantifying coral reef population density/composition, and promoting 3D coral reconstruction. In computer vision, great progress has been achieved in dense segmentation of general in-air objects, led by powerful deep neural networks [13, 20] and large-scale datasets [17, 39]. Recently, the segment-anything model [31] (SAM) pre-trained on 11 million images with 1 billion masks has demonstrated a very strong generalization ability to a wide range of objects and downstream tasks [25, 26, 42, 43, 52]. It is therefore of great interest to explore how well SAM can segment corals within the coral images.

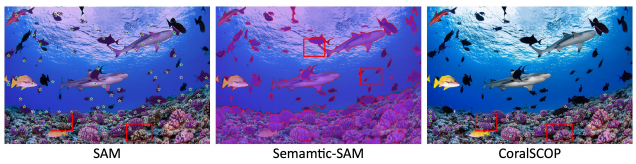

Figure 2. CoralSCOP could effectively address over-segmentation and generate more accurate coral masks than SAM and Semantic-SAM. Best viewed in color.

Unlike segmenting general in-air objects, however, precise and robust coral segmentation [1, 2, 47, 59, 64] is still a challenging problem due to *irregular coral boundaries*, *a wide range of coral appearances and forms*, *degraded visual quality of underwater images*, *complicated background with dynamic occlusions*, and *viewpoint changes* as demonstrated in Figure 1. Several important aspects for coral segmentation are lacking in SAM. First, SAM is a semantic-agnostic model, which cannot yield semantic labels for coral reef data and leads to over-segmentation as shown in Figure 2. Second, due to the fine-grained anatomical structures, uncertain and complicated boundaries, as well as the wide range of scales found in coral reefs, SAM tends to produce incomplete and inaccurate coral masks. The further Semantic-SAM [34] could generate masks from 6 implicit semantic granularities through a many-to-many matching design. However, Semantic-SAM cannot address the over-segmentation effectively either (viewed in Figure 2). To resolve these issues, a trivial approach is to directly fine-tune SAM on additional coral data. This improves the mask generation quality, but does not effectively address the over-segmentation problem since SAM is pre-trained on masks without semantics. The finetuned SAM, however, still yields numerous false positives due to the extreme ratio between pre-training data and fine-tuning data, assuming that the number of coral masks is a few thousand times fewer than billions of object masks.

In this paper, we present the first coral segmentation foundation model, CoralSCOP, trained on a significant scale of coral reef images and adaptable to a wide range of downstream coral analysis tasks. CoralSCOP could produce high-quality coral mask generation and underwater visual perception [22]. Our model is illustrated in Figure 1, demonstrating

a strong zero-shot generalization ability to segment unseen coral reef images and automatically generate reasonable coral masks with **no prompts**.

The key technical contribution of CoralSCOP is an improved network architecture over SAM by a novel **parallel semantic branch** at the decoder, which effectively addresses the semantic-agnostic issue as well as enables new training strategies for further enhanced results, *e.g.*, train with **negatives**. CoralSCOP can also receive various kinds of prompts (*e.g.*, points, boxes and texts) to guide the generation of semantic coral masks based on user intents. All these make CoralSCOP a foundation model with valuable adaptability for various downstream user-defined tasks: 1) **user-defined tuning** to support coral reef understanding from different granularities (*e.g.*, coral/non-coral, growth form and genus); 2) **sparse-to-dense** conversion by converting the already available sparse point annotations [32] to dense coral masks for more accurate coral statistics; 3) **mask-referring segmentation** and 4) **instruction-following segmentation** enable both the *amateurs* and *coral biologists* to generate coral masks, unleashing the power of textual descriptions and reusing existing coral masks.

Our CoralSCOP is trained by CoralMask, a new dataset with **41,297** coral reef images and **330,144** coral masks. Existing datasets in underwater and marine research [15, 21, 22, 24, 35, 37, 65] such as Mosaics UCSD [15] (4,193 images), WaterMask [37] (4,628 images), MAS3k [35] (3,103 images), LaRS [65] (4,006 images) and USOD10K [22] (10,255 images) have small data size with low image diversity [15] (*e.g.*, data captured at the same site) while being too generic (*e.g.*, data with a wide range of underwater objects [22, 35, 37]), which cannot directly be utilized for coral analysis. Our CoralMask dataset is specifically designed for coral analysis at a larger scale, which is about 10x larger than Mosaics UCSD [15] and 4x as USOD10K [22] in data size. Our results show strong capabilities in segmenting in-distribution and out-of-distribution coral images, either qualitatively or quantitatively.

Our contributions are summarized as follows:

- We present CoralSCOP, the first coral segmentation foundation model, which can effectively segment coral images automatically (no prompts) or semi-automatically by input prompts from the user;
- We present improved network architecture and training strategies for CoralSCOP to achieve strong zero-shot generalization ability;
- We propose CoralMask, the largest coral reef dataset to date with densely labeled coral masks to foster the development of the coral segmentation foundation model;
- We provide extensive experiments that demonstrate the SOTA results of our model. We also demonstrate the robustness of our foundation model in different application scenarios with SOTA results.

2. Related Work

Underwater Visual Analysis [21, 36, 40, 58, 63] provides an effective way to monitor and protect the underwater and marine ecosystem. To promote underwater visual recognition, various datasets and benchmarks (*e.g.*, MAS3K [35], WildFish [61], WildFish++ [62], FishNet [30], MarineDet [19], SUIM [23], *etc*) have been proposed for different purposes: fish classification [61, 62], fish counting [28, 51], marine animal segmentation [35], marine object detection [19, 21, 40, 44], salient object detection and segmentation [22, 24], semantic segmentation [15, 23, 37, 64] and name a few [36, 65]. Among these, dense segmentation [15, 64], with pixel-level analysis, can provide valuable insights into the population and biodiversity of coral reefs, and inform the development of effective management strategies to protect coral reefs.

Coral Analysis. With coral reefs facing increasing threats [45], accurately analyzing and monitoring the fragile coral reef ecosystems [2, 16, 33, 49] is more important than ever. Coral Point Count with Excel Extensions (CPCe) [32] provided the benthic coral analysis based on sparse point analysis, where sparse points are randomly sampled or determined and then annotated by coral biologists to compute the coral statistics. To reduce the human labor, the patch-based CNNs [20, 50] are integrated into CoralNet [3, 8] to automatically determine the sparse point annotation. Based on sparse point annotations, CoralSeg [1] and PLAS [47] proposed to perform sparse-to-dense conversion based on Superpixels [5] to obtain dense coral masks as pseudo ground truth for optimizing dense segmentation algorithms. Besides the sparse-to-dense conversion, some coral reef datasets [4, 15, 64] with dense coral masks have also been proposed. However, these datasets with dense masks are limited in data scale with fixed few semantic categories and image diversity which limits their generalization ability. Our CoralMask is the largest coral reef dataset with dense coral masks to date. There are also two fundamental differences between CoralSCOP and existing CPCe [32] and CoralNet [3, 8]: 1) CoralSCOP has a strong zero-shot generalization ability to unseen coral reef images without new training samples and 2) CoralSCOP supports dense segmentation, which could generate more accurate coral statistics.

Foundation Models (such as CLIP [46], ALIGN [27] and SAM [31]) have been widely favored by the whole CV community. SAM [31] optimized by vast and diverse training data, has demonstrated a strong ability to segment meaningful objects with precise object masks. SAM has been widely used for medical image segmentation [42, 43, 48, 54], camouflaged object segmentation [25, 52], challenging scenarios [9] and other applications [26, 60]. MSA [54] designed an adapter design for transferring SAM to a counterpart in segmenting medical images and SAM-adapter [9] proposed to perform camouflaged object segmentation and shadow

detection. Under both settings, domain-specific data are required for fine-tuning SAM to domain experts. However, underwater coral reef images illustrate more complicated and diversified characteristics than medical images or in-air images. Directly utilizing SAM or fine-tuning SAM for coral reef analysis cannot achieve satisfactory performance due to the over-segmentation and lack of semantics. The limited coverage of coral images in training data of SAM also results in inaccurate or incomplete coral mask generation.

3. Approach

3.1. CoralMask Dataset

We have curated a dataset of **41,297** coral reef images from different data distributions, encompassing a wide range of global coral reef sites. This collection features images sourced from the Internet, public datasets, open-source contributions, and underwater surveying images contributed by coral biologists. Most of the coral reef images are of high resolution, and we downsampled them to a resolution where the shortest side is 1,024 pixels while maintaining the aspect ratio of each image. The coral reefs have complex geometries and irregular shapes. Meanwhile, they also illustrate a wide range of textures and appearances, with different colors, patterns, and growing forms. It can be very difficult to accurately identify and segment the boundaries of individual corals from the coral colonies. To promote the adaptability of CoralSCOP to downstream tasks, we have established a series of grouping rules specifically designed to accurately separate individual corals, ensuring corals with different characteristics into separated masks. We have developed an interactive labeling tool based on SAM to promote labeling efficiency. All the coral reef images were manually labeled by coral biologists, resulting in **330,144** masks in total.

3.2. Preliminaries

Our method is built upon SAM [31], the foundation model for generic segmentation without semantic labels. SAM consists of three components, a prompt encoder $\texttt{Prop}(\cdot)$, a heavy image encoder $\texttt{Enc}(\cdot)$, and a lightweight mask decoder $\texttt{Dec}(\cdot)$. As a promotable model, SAM could be fed with an image I and a set of prompts P, including point, box, or coarse mask prompts. SAM utilizes $\texttt{Enc}(\cdot)$ to obtain image embedding, and adopts $\texttt{Prop}(\cdot)$ to encode prompts P of a length k into prompt tokens as:

$$F_I = \texttt{Enc}(I), \quad T_P = \texttt{Prop}(P), \tag{1}$$

where $F_I \in \mathbb{R}^{h\times w\times c}$ and $T_P \in \mathbb{R}^{k\times c}$, with h, w denoting the resolution of the image embedding and $c = 256$ denoting the feature dimension. F_I and T_P are then fed into mask decoder $\texttt{Dec}(\cdot)$ for mask generation. SAM constructs the input tokens of $\texttt{Dec}(\cdot)$ by concatenating the learnable mask tokens T_M generated by $\texttt{Dec}(\cdot)$ and the prompt tokens T_P for generating the mask output, formulated as:

$$M = \texttt{Dec}(F_I, \texttt{Concat}(T_M, T_P)), \tag{2}$$

where M denotes segmentation masks yielded by SAM.

SAM performs the semantic-agnostic mask generation in an automatic or interactive way through different kinds of prompts, which empower SAM with a strong generalization ability to generate precise masks for objects. However, SAM is primarily optimized by in-air images, making it less effective in segmenting underwater images, especially coral reef images. The complicated backgrounds, irregular boundaries, crowded instances, and significant diversity of corals lead to the fact that SAM still suffers from incomplete and inaccurate coral mask generation. Furthermore, SAM lacks semantics, resulting in over-segmentation and numerous false positives. Simply fine-tuned with coral masks does not address the over-segmentation problem and changes the semantic agnostic nature.

3.3. CoralSCOP

3.3.1 Coral Foundation Model Pre-training

We adopt SAM as our backbone and utilize it as a *powerful* and *stable* network initialization. The whole model is continuously pre-trained with our CoralMask dataset to better extract efficient underwater feature representations. CoralSCOP performs semantic segmentation and learns the implicit semantics from the grouping rules by the coral biologists (*e.g.*, grouping the corals with the same appearance to one coral mask while splitting dissimilar corals from crowded coral colonies). The overview of our CoralSCOP is demonstrated in Figure 3.

To address semantic-agnostic and over-segmentation problems, we have designed a **parallel semantic branch** in $\texttt{Dec}(\cdot)$. We formulate a lightweight MLP layer to project the shared feature embeddings by the mask generation branch and the semantic branch. The attention-based feature interaction is conducted between mask generation and semantic discrimination. As demonstrated in Figure 3 (we omit the IoU branch for better illustration), we couple the image embedding with prompt information to generate 4096-dimensional vectors $F_o \in \mathbb{R}^{4096\times k}$, describing the yielded coral masks. F_o could also be utilized for computing the similarity between generated coral masks, empowering CoralSCOP with the ability to perform *mask-referring segmentation*. We finally project F_o into the semantic coral space (*e.g.*, a binary semantic space discriminating the corals and non-corals in the pre-training stage or user-defined semantic space in the tuning procedure). Our design serves three purposes: training with negatives, user-defined tuning and instruction-following tuning.

Training with Negatives. We observed that solely relying on pre-training on domain-specific coral reef data with supervision from the coral biologists does not adequately address the over-segmentation issue. Besides, additional continuous

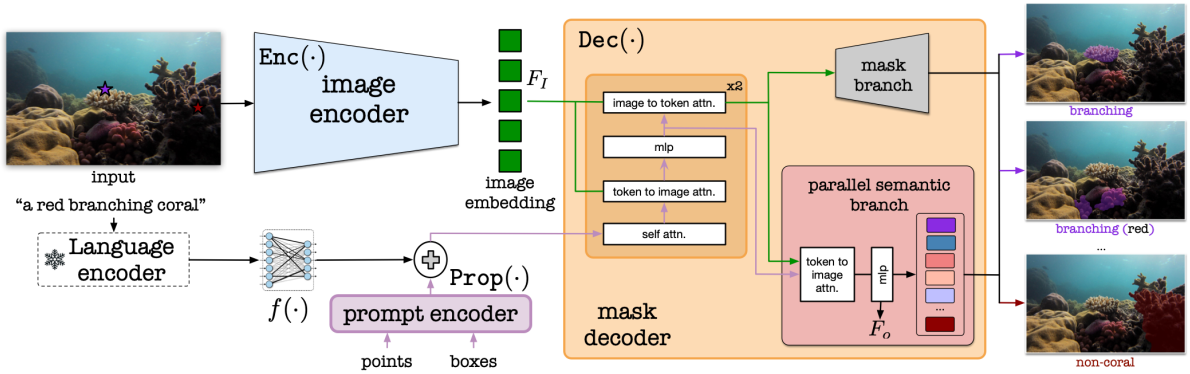

Figure 3. The proposed CoralSCOP framework. We design a parallel semantic branch inside the mask decoder, enabling the coral mask generation with semantics. The whole model is trainable in the pre-training procedure and $\texttt{Enc}(\cdot)$ is frozen in the tuning procedure.

pre-training could lead to knowledge forgetting, therefore degrading the generalization ability of CoralSCOP. To address these two issues simultaneously, we propose training-with-negatives. We generate redundant semantic-agnostic masks based on SAM with **Vit-H** backbone. Then we regard those generated masks that do **not** overlap with our labeled coral masks as masks with "non-coral" labels. The binary classification loss is computed by

$$\mathcal{L}_{cls} = -(y\log(p) + (1-y)\log(1-p)),\ p = \texttt{MLP}(F_o). \tag{3}$$

Our parallel semantic branch is optimized by both manually labeled coral masks and masks assigned with the "non-coral" label, performing binary discrimination during the pre-training procedure. By incorporating these advancements, we can preserve the strong generalization ability of CoralSCOP and alleviate the over-segmentation problem.

3.3.2 Model Tuning

After pre-training, CoralSCOP could effectively discriminate corals from the background and produce high-quality binary coral masks. In order to make it a segmentation foundation model, together with the parallel semantics branch, we designed **user-defined** tuning and **instruction-following** tuning for performing a more fine-grained coral analysis while keeping $\texttt{Enc}(\cdot)$ frozen.

User-defined Tuning. The users could self-design their own coral semantics (*e.g.*, growth form, genus, or species) and tune pre-trained CoralSCOP to downstream variants. This user-defined tuning is achieved by only modulating the parameters of the parallel semantic branch based on few-shot samples (coral masks with user-defined semantic annotations). $\mathcal{L}_{cls}^{Tun.}$ for tuning is computed as:

$$\mathcal{L}_{cls}^{Tun.} = -\sum_{c_m=1}^{N} y_{c_m}\log(p_{c_m}), \quad p_{c_m} = \texttt{MLP}(F_o) \tag{4}$$

where c_m is the semantic annotation of M while N is the number of total categories defined by users. By only tuning the parameters of the $\texttt{MLP}$ layer, we could transfer CoralSCOP to various variants aligned with the different user intents.

Instruction-following Tuning. CoralSCOP enables instruction-following segmentation. Given a textual prompt (*e.g.*, segmenting "a red branching coral"), describing corals presented in I, CoralSCOP could generate a corresponding coral mask that satisfies the user requirement. We borrow the frozen language encoder of CLIP [46] and generate the textual embedding $t \in \mathbb{R}^{768}$. We utilize the bounding box of the whole image: $\text{BBOX}_{\text{image}}$ as prompt to generate $T_P \in \mathbb{R}^{k\times c}$, where $k = 1$. A linear layer $f(\cdot)$ is designed to project t to $t^{'} = f(t)$, with the same size as T_P. $t^{'}$ is finally added to T_P element-wise:

$$M = \texttt{Dec}(F_I, \texttt{Concat}(T_M,\ T_P|\text{BBOX}_{\text{image}} + f(t))), \tag{5}$$

where M is a single mask in this case. $T_P|\text{BBOX}_{\text{image}}$ indicates that we utilize the width and height of the whole image as the box prompt. The parallel semantic branch is also utilized for discriminating whether the generated coral mask matches the given textual description. CoralSCOP will set the generated mask to "background" when the given text prompt does not match the synthesized mask. We construct both positive and negative pairs to optimize our CoralSCOP to promote its ability to generate required coral masks and alleviate the hallucination.

4. Experiments

4.1. Implementation Details

Pre-training. 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) are utilized for training our CoralSCOP. We have optimized our CoraSCOP for 5 epochs on 6 Tesla A100 GPUs and batch size per GPU is set to 1. We optimize the parameters of the whole model during the pre-training to promote $\texttt{Enc}(\cdot)$ to extract underwater visual features. The composite prompts of point prompts (1, 2, 3 or 4 random points inside the coral

Table 1. The coral mask generation quality of various algorithms under different settings.

Method	Back.	IoU ↑	Accuracy ↑	MAE ↓
DeeplabV3 [7]	R101-D8	59.39	71.76	0.1634
CGNet [55]	FCN-4xB8	59.63	79.84	0.1197
GCNet [6]	R101-D8	61.75	78.35	0.1206
BiSeNetV2 [57]	FCN-FP16	60.90	78.11	0.1129
MaskFormer [10]	R101-D32	67.21	79.46	0.0868
Mask2Former [11]	R101-D32	69.17	77.63	0.0842
SegFormer [56]	Mit-B5	71.51	84.29	0.0776
SAM$^{\heartsuit}$ [31]	Vit-H	31.16	38.05	0.5057
SAM$^{\spadesuit}$ [31]	Vit-H	44.67	52.86	0.3798
SAM$^{\clubsuit}$ [31]	Vit-H	72.80	73.92	0.1173
CoralSCOP$^{\heartsuit}$	Vit-B	59.02	79.88	0.1243
CoralSCOP$^{\spadesuit}$	Vit-B	62.90	**86.67**	0.1217
CoralSCOP$^{\clubsuit}$	Vit-B	74.00	83.73	0.0804
CoralSCOP$^{\heartsuit}$	Vit-L	65.87	78.45	0.1204
CoralSCOP$^{\spadesuit}$	Vit-L	68.23	78.34	0.0976
CoralSCOP$^{\clubsuit}$	Vit-L	**79.13**	81.11	**0.0764**

mask) and bounding box prompts are utilized for training. The final loss function is the sum of the classification loss $\mathcal{L}_{cls}$, Dice loss, Focal loss (20×) and IoU loss. **Tuning Procedure**. 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 $\texttt{Dec}(\cdot)$. For *instruction-following* tuning, we have formulated 46,610 instruction-following pairs. All the coral reef images (4,661 in total) are crawled from "Corals of the World" [14]. We optimize the whole mask decoder under this setting.

4.2. Coral Mask Generation

We first evaluate the generated coral mask quality. We randomly sampled 941 coral reef images (non-overlapped with the training data) with significant diversity and the professional coral biologist annotators generated the ground truth coral masks. The models are required to segment the corals from the background. We have included the existing SOTA semantic segmentation algorithms DeepLabV3 [7], CGNet [55], GCNet[6], BiSeNetV2 [57], MaskFormer [10], SegFormer [56] and Mask2Former [11] for comparison. All the algorithms have been **re-trained** and **optimized** on our CoralMask dataset following the provided default setting for a fair comparison. The SAM is also utilized for inference only. We evaluate SAM and CoralSCOP under three experimental settings: "Automatic$^{\heartsuit}$" (no prompt is given); "1 point prompt$^{\spadesuit}$" (one random point inside each coral mask is given as point prompt) and "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). For the latter two settings, we repeat the mean value of 3 times of runs.

We computed IoU, pixel accuracy, and the mean absolute error (MAE) between generated coral masks and ground truth masks labeled by coral biologists in Table 1. DeepLabV3 [7], CGNet [55] and GCNet [6] models only demonstrate limited generalization ability to the unseen coral

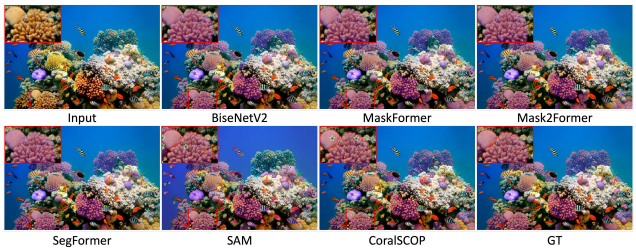

Figure 4. The coral segmentation comparisons between various algorithms. Both SAM and CoralSCOP could separate dissimilar corals into different coral masks while others failed.

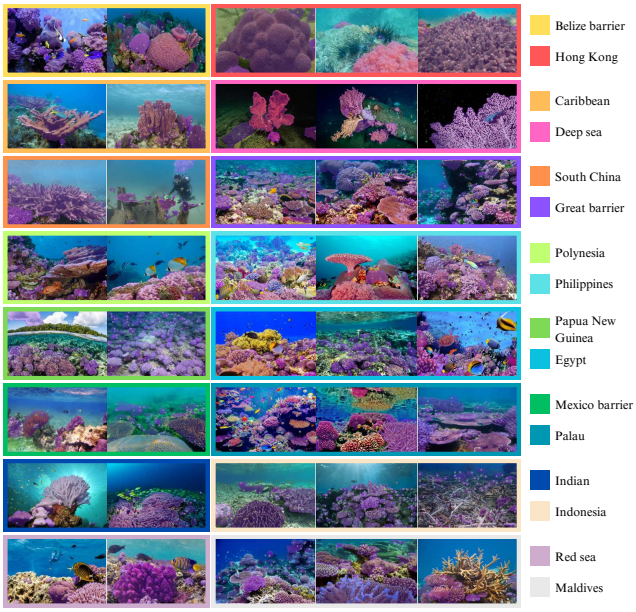

Figure 5. The automatic coral segmentation of CoralSCOP on coral reef images from different sites.

images and they failed to generate accurate coral masks. We attribute this failure to the irregular boundary and significant diversity of coral reefs. SegFormer [56], MaskFormer [10] and Mask2Former [11] illustrate a stronger ability to generate accurate coral masks. However, all these dense segmentation specialist models cannot separate corals with different appearances, growth form, genus or species representations into separated coral masks as demonstrated in Figure 4, since they are only optimized to segment corals from the background. The generated coral masks cannot be further redefined by the coral biologists aligned with user intents and the whole model is required to be re-optimized to satisfy the user requirement.

In contrast, both SAM and CoralSCOP could generate coral masks in a fine-grained manner. Figure 4 demonstrates the limitations of SAM for coral reef data. SAM suffers from the over-segmentation problem and also fails to detect some coral individuals. As demonstrated, CoralSCOP could effectively address the over-segmentation problem. CoralSCOP could achieve competitive coral mask generation performance even under the "automatic" setting. When

Table 2. The mIoU scores are computed under different settings. The numbers in blue represent the performance gains achieved by CoralSCOP.

Methods	Semantics	10-shot	20-shot	40-shot	Full
SegFormer [56]	Growth Form	32.54	36.79	38.39	59.22
CoralSCOP		52.16 (+19.62)	56.75 (+19.96)	57.27 (+18.88)	67.82 (+10.55)
SegFormer [56]	Genus	18.24	20.04	22.52	28.34
CoralSCOP		23.82 (+5.58)	26.26 (+6.22)	28.89 (+6.37)	35.53 (+7.19)

given 1 random point prompt, CoralSCOP achieves comparable performance with Mask2Former. The additional BBOX prompt further addresses the ambiguity and results in the best segmentation performance.

Spatial coral reef analysis. We provide coral segmentation results of reef images from **16** sites in Figure 5. The images have a large range of color variation, textures, visibility, perspective and biological diversity. CoralSCOP effectively recognizes the coral reef images from various sites and demonstrates a strong zero-shot generalization ability.

4.3. Semantic Coral Analysis

We then validate CoralSCOP on a wider range of biologically meaningful tasks through user-defined tuning. We offer flexible user-defined tuning experiments essential in coral reef research, including **growth form** and **genus** recognition. Growth form (500 coral images with 6 growth form annotations: 400 for training and 100 for testing) indicates the physical morphology and structure exhibited by coral reefs. We have explored the effectiveness of the few-shot tuning (*e.g.* 10, 20, and 40 training images). We randomly pick up few-shot samples from the whole training samples and we repeat experiments 3 times to report mean values. For comparison, SegFormer is conducted under the same experimental settings. The mIoU scores are computed in Table 2. CoralSCOP achieved more reliable growth form segmentation than SegFormer with nearly 20 mIoU score improvement. With the pre-trained foundation model that can effectively recognize corals, we can easily extend the foundation model to recognize the coral growth form based on few-shot samples. Similarly, we perform genus-level coral understanding (400 coral images with 14 coral genera: 300 for training and 100 for testing), which requires more fine-grained distinctions and recognizing subtle morphological differences among closely related genera (*e.g.*, Platygyra and Coelastrea). It is thus more challenging for the model to perform accurate genus-level coral segmentation. Our CoralSCOP could still achieve observable performance gains over SegFormer. It is worth noting that all the images involved in growth form and genus segmentation are with the benthic view and labeled by coral biologists.

Sparse-to-Dense. CoralSCOP supports sparse-to-dense conversion for more accurate coral statistics. The existing coral reef analytical approach CPCe is computing coral statistics only based on labeled sparse points: $\frac{\#\text{ of points for each growth form}}{\#\text{ of total sparse points}}$. CoralSCOP could utilize the already available sparse point annotations as prompts to generate dense semantic coral masks, where the semantics are from the point annotations. SAM [31] and PLAS [47] are chosen for comparison and 100 benthic images with 6 growth form annotations are used for testing. In Table 3, we report the average prediction error between ground truth coral composition and coral statistics computed by different algorithms after sparse-to-dense conversion: $\frac{\text{area of coral masks for each growth form}}{\text{image area}}$. The prediction error of CPCe is also reported. Please note all algorithms adopt the same sparse point annotations and are repeated 3 times to obtain the mean values. We conduct experiments using different numbers of sparse points: 10, 20, 50 and 100. In most cases, CoralSCOP achieves the lowest prediction error, demonstrating the best sparse-to-dense conversion performance. SAM even achieved worse results than CPCe, indicating the generated coral masks by SAM are wrong and then lead to higher prediction errors.

Table 3. The average **prediction error** (↓) under "sparse" (CPCe) setting and "dense" setting after sparse-to-dense conversion by PLAS, SAM and CoralSCOP. The ground truth coral composition for each growth form and non-coral composition are also provided.

Method	Setting (Points)	Non-coral (61.43)	Massive (22.51)	Laminar (13.06)	Branching (1.499)	Faliaceous (0.9614)	Encrusting (0.2788)	Columnar (0.2616)
CPCe [32]	10	10.01	8.903	4.544	0.678	0.223	0.480	0.162
PLAS [47]		**9.886**	8.490	4.485	0.675	0.153	0.397	**0.134**
SAM [31]		16.12	10.69	5.846	0.309	0.164	0.248	0.135
CoralSCOP		10.56	**6.371**	**3.677**	**0.287**	**0.034**	**0.211**	0.198
CPCe [32]	20	7.437	6.218	3.021	0.536	**0.039**	0.308	0.129
PLAS [47]		6.391	5.736	2.952	0.421	0.050	0.338	0.141
SAM [31]		10.71	6.943	3.871	0.230	0.249	0.256	**0.055**
CoralSCOP		**5.334**	**3.791**	**1.555**	**0.119**	0.049	**0.225**	0.145
CPCe [32]	50	4.488	3.569	2.300	0.382	0.086	0.217	0.041
PLAS [47]		3.747	3.170	1.671	0.179	0.065	0.193	**0.021**
SAM [31]		6.988	5.194	3.047	0.158	0.112	0.545	0.113
CoralSCOP		**2.832**	**2.657**	**1.396**	**0.042**	**0.022**	**0.154**	0.221
CPCe [32]	100	3.052	2.648	1.345	0.175	0.020	0.159	0.034
PLAS [47]		2.300	1.921	1.253	0.115	0.024	**0.145**	**0.026**
SAM [31]		7.436	5.804	3.378	0.335	0.188	1.259	0.098
CoralSCOP		**2.014**	**1.727**	**1.174**	**0.056**	**0.010**	**0.145**	0.070

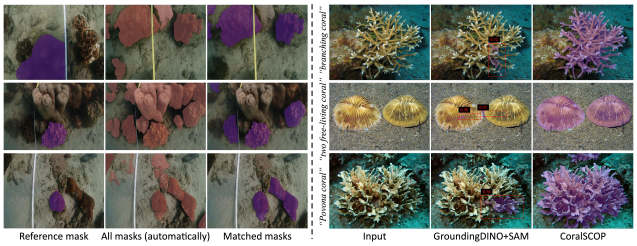

Figure 6. The coral segmentation results of instruction-following (left side) and mask-referring (right side) coral segmentation.

Mask-referring segmentation. CoralSCOP supports mask-referring coral segmentation: a reference coral mask (either from an input image or another reference image) is provided, and CoralSCOP will yield matched coral masks based on all the automatically generated coral masks. We adopt Cosine similarity to compute feature distance between masks. The masks with similarity over a tunable user-defined threshold (*e.g.*, 0.8 in our experiments) are regarded as matched masks demonstrated in the left side of Figure 6. **Instruction-following segmentation** results of CoralSCOP are illustrated in right side of Figure 6. We utilize SAM+GroundingDINO

Table 4. The coral mask generation quality of various algorithms under comprehensive settings.

Method	Back.	Automatic$^{\heartsuit}$		1 point$^{\spadesuit}$		1 point + BBOX$_{\text{mask}}^{\clubsuit}$	
		IoU ↑	MAE ↓	IoU ↑	MAE ↓	IoU ↑	MAE ↓
SAM [31]	Vit-B	23.61	0.3895	**45.76**	**0.3122**	63.03	0.1799
SAM‡		23.68	**0.2551**	32.12	0.4148	65.21	0.1676
SAM-Adapter [9]		8.821	0.3033	32.47	0.3794	62.78	0.1891
CoralSCOP^{-}		24.06	0.2915	33.52	0.3816	65.61	0.1666
CoralSCOP		**26.45**	0.2847	37.96	0.3219	**66.78**	**0.1562**
SAM [31]	Vit-L	29.83	0.4623	41.36	0.4578	57.40	0.2947
SAM‡		37.46	0.2614	44.72	0.3264	68.00	0.1470
SAM-Adapter [9]		34.38	0.3399	43.83	0.3290	67.34	0.1525
CoralSCOP^{-}		38.15	0.2601	44.87	0.3275	67.84	0.1475
CoralSCOP		**46.46**	**0.1814**	**45.65**	**0.2991**	**68.66**	**0.1408**

for comparison. The textual descriptions are fed into GroundingDINO [41] to generate the bounding box prompt (with the confidence score) for SAM and then SAM generates the corresponding masks. GroundingDINO+SAM cannot generate satisfactory coral masks.

4.4. Further Analysis

Ablation studies. We compare CoralSCOP (two variants: with negatives as CoralSCOP and without negatives as CoralSCOP^{-}) and SAM variants: SAM (inference only), SAM‡ (fine-tuned on CoralMask) and SAM-adapter [9] (fine-tuned on CoralMask with adapter design [9] while keeping $\texttt{Enc}(\cdot)$ frozen). The training prompt contains *1 random point* inside the labeled coral mask and BBOX$_{\text{mask}}$. 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. We compute results under three settings in Table 4. We could conclude such findings: 1) directly fine-tuning SAM on CoralMask mask could promote the ability of SAM to segment coral masks and a stronger backbone could achieve larger performance gain; 2) when optimized by prompt with 1 random point and BBOX$_{\text{mask}}$, the ability of mask decoder to generate precise coral masks based on only 1 point prompt has been weakened while it could more precise coral mask generation with additional BBOX$_{\text{mask}}$ prompt; 3) fine-tuning $\texttt{Enc}(\cdot)$ together could result in better automatic coral segmentation performance; 4) with only coral masks for fine-tuning cannot address the over-segmentation well by comparing SAM‡ and CoralSCOP: SAM‡ are with lower IoU score while much higher MAE under almost every setting, indicating many false positives; 5) the negative masks could promote IoU score and reduce false positives (lower MAE) under the automatic setting by comparing CoralSCOP and CoralSCOP^{-}.

Promoting underwater visual analysis. We then demonstrate the heavy image encoder of CoralSCOP is more effective than the counterpart of SAM on underwater visual analysis (*e.g.*, underwater salient object segmentation on USOD10K dataset [22]). $\texttt{Enc}(\cdot)$ of both SAM and CoralSCOP are frozen and we only optimize the mask decoder. The BBOX$_{\text{image}}$ is used as the prompt for generating the salient object prediction. We follow [22] and compute results in Table 5. We observe the frozen $\texttt{Enc}(\cdot)$ of CoralSCOP could result in better performance than that of SAM, indicating better underwater feature extraction ability.

Table 5. The underwater salient object segmentation results on the USOD10K dataset [22].

Method	Back.	S_m ↑	E_e^{max} ↑	max$F$ ↑	MAE ↓
SAM [31]	Vit-B	0.8695	0.9199	0.8445	0.0387
CoralSCOP		0.8739 (+0.0044)	0.9250 (+0.0051)	0.8512 (+0.0067)	0.0353 (-0.0044)
SAM [31]	Vit-L	0.8843	0.9279	0.8658	0.0336
CoralSCOP		0.8884 (+0.0041)	0.9338 (+0.0059)	0.8707 (+0.0049)	0.0316 (-0.0037)

4.5. Discussions

Our development of CoralMask dataset contributes toward standardizing and aggregating expertly curated labeled mask annotations, defining how to split the individual coral from a mix of coral colonies. CoralSCOP can continue to grow from sparse-to-dense conversion and incorporate more labeled data from the coral community, accelerating the processing of visual data critical for developing conservation strategies to achieve a healthy and sustainable global coral reef.

Compared with the hitherto commonly used CPCe to generate coral cover or other coral parameters, our CoralSCOP can generate statistics with greater accuracy (less error) within a much shorter time (around 348 seconds for similar work, *e.g.*, 100 images in CPCe that take hours). Our CoralSCOP should therefore find extensive application (*e.g.*, coral bleaching and biodiversity statistics) in coral monitoring works around the world. The generated coral masks can be used for promoting 3D coral scene understanding (*e.g.*, rugosity computation), removing noisy backgrounds to better model the geometry information of coral colonies. This should lead to, for example, better differentiation of coral vs non-coral objects from coral images, which is very important in assessing coral cover in any new sites under assessment for its conservation value.

5. Conclusion

CoralSCOP represents a significant step forward in the field of underwater coral reef analysis, as the first segmentation foundation model in marine research. It could support scalable and efficient dense coral segmentation, providing wide users with a seamless and efficient tool for analyzing coral reef images/videos. We hope that our attempts can provide the community with some insights into the future development of introducing the large foundation model in domain-specific research. We will make CoralMask and CoralSCOP publicly available, which hopefully benefit researchers and coral biologists in their own research.

Acknowledgement. This work was supported by the Innovation and Technology Support Programme of the Innovation and Technology Fund (Ref: ITS/200/20FP) and the Marine Conservation Enhancement Fund (MCEF20107 and MCEF23EG01) and an internal grant from HKUST (R9429).

References

[1] 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

[2] Oscar Beijbom, Peter J Edmunds, David I Kline, B Greg Mitchell, and David Kriegman. Automated annotation of coral reef survey images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1170–1177. IEEE, 2012. 2, 3

[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. 2, 3

[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. 3

[5] Michael Van den Bergh, Xavier Boix, Gemma Roig, Benjamin de Capitani, and Luc Van Gool. Seeds: Superpixels extracted via energy-driven sampling. In *European Conference on Computer Vision (ECCV)*, pages 13–26. Springer, 2012. 3

[6] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 0–0, 2019. 6

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4):834–848, 2017. 2, 6

[8] 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. 2, 3

[9] 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. 3, 8

[10] Bowen Cheng, Alexander G Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *arXiv preprint arXiv:2107.06278*, 2021. 2, 6

[11] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022. 2, 6

[12] Joshua E Cinner, Cindy Huchery, M Aaron MacNeil, Nicholas AJ Graham, Tim R McClanahan, Joseph Maina, Eva Maire, John N Kittinger, Christina C Hicks, Camilo Mora, et al. Bright spots among the world's coral reefs. *Nature*, 535 (7612):416–419, 2016. 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[14] Turak E. and DeVantier L.M. Corals of the world. http://coralsoftheworld.org/v0.01(Beta), 2016. 6

[15] 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. 2, 3

[16] Jose Nuno Gomes-Pereira, Vincent Auger, Kolja Beisiegel, Robert Benjamin, Melanie Bergmann, David Bowden, Pal Buhl-Mortensen, Fabio C De Leo, Gisela Dionísio, Jennifer M Durden, et al. Current and future trends in marine image annotation software. *Progress in Oceanography*, 149: 106–120, 2016. 3

[17] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5356–5364, 2019. 2

[18] Andreas F Haas, Mohamed FM Fairoz, Linda W Kelly, Craig E Nelson, Elizabeth A Dinsdale, Robert A Edwards, Steve Giles, Mark Hatay, Nao Hisakawa, Ben Knowles, et al. Global microbialization of coral reefs. *Nature microbiology*, 1(6):1–7, 2016. 2

[19] Liang Haixin, Zheng Ziqiang, Ma Zeyu, and Sai-Kit Yeung. Marinedet: Towards open-marine object detection. *arXiv preprint arXiv:2310.01931*, 2023. 3

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 3

[21] Jungseok Hong, Michael Fulton, and Junaed Sattar. Trashcan: A semantically-segmented dataset towards visual detection of marine debris. *arXiv preprint arXiv:2007.08097*, 2020. 3

[22] 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. 2, 3, 8

[23] 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. 3

[24] Md Jahidul Islam, Ruobing Wang, and Junaed Sattar. Svam: saliency-guided visual attention modeling by autonomous underwater robots. *arXiv preprint arXiv:2011.06252*, 2020. 3

[25] Ge-Peng Ji, Deng-Ping Fan, Peng Xu, Ming-Ming Cheng, Bowen Zhou, and Luc Van Gool. Sam struggles in concealed

scenes–empirical study on segment anything. *arXiv preprint arXiv:2304.06022*, 2023. 2, 3

[26] Wei Ji, Jingjing Li, Qi Bi, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. *arXiv preprint arXiv:2304.05750*, 2023. 2, 3

[27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, pages 4904–4916. PMLR, 2021. 3

[28] Justin Kay, Peter Kulits, Suzanne Stathatos, Siqi Deng, Erik Young, Sara Beery, Grant Van Horn, and Pietro Perona. The caltech fish counting dataset: A benchmark for multiple-object tracking and counting. In *European Conference on Computer Vision (ECCV)*, pages 290–311. Springer, 2022. 3

[29] Linda Wegley Kelly, Craig E Nelson, Andreas F Haas, Douglas S Naliboff, Sandi Calhoun, Craig A Carlson, Robert A Edwards, Michael D Fox, Mark Hatay, Maggie D Johnson, et al. Diel population and functional synchrony of microbial communities on coral reefs. *Nature communications*, 10(1): 1691, 2019. 2

[30] 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. 3

[31] 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, 3, 4, 6, 7, 8

[32] 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. 2, 3, 7

[33] Natalie Levy, Ofer Berman, Matan Yuval, Yossi Loya, Tali Treibitz, Ezri Tarazi, and Oren Levy. Emerging 3d technologies for future reformation of coral reefs: Enhancing biodiversity using biomimetic structures based on designs by nature. *Science of The Total Environment*, 830:154749, 2022. 2, 3

[34] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023. 2

[35] 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. 3

[36] Lin Li, Bo Dong, Eric Rigall, Tao Zhou, Junyu Dong, and Geng Chen. Marine animal segmentation. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 32 (4):2303–2314, 2021. 3

[37] 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. 3

[38] Bing Lin, Yiwen Zeng, Gregory P Asner, and David S Wilcove. Coral reefs and coastal tourism in hawaii. *Nature Sustainability*, 6(3):254–258, 2023. 2

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 2

[40] Xiaomin Lin, Nitin J Sanket, Nare Karapetyan, and Yiannis Aloimonos. Oysternet: Enhanced oyster detection using simulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5170–5176. IEEE, 2023. 3

[41] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 8

[42] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 2, 3

[43] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *arXiv preprint arXiv:2304.10517*, 2023. 2, 3

[44] Md Modasshir and Ioannis Rekleitis. Enhancing coral reef monitoring utilizing a deep semi-supervised learning approach. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1874–1880. IEEE, 2020. 3

[45] Benjamin Paul Neal, Adi Khen, Tali Treibitz, Oscar Beijbom, Grace O'Connor, Mary Alice Coffroth, Nancy Knowlton, David Kriegman, B Greg Mitchell, and David I Kline. Caribbean massive corals not recovering from repeated thermal stress events during 2005–2013. *Ecology and Evolution*, 7(5):1339–1353, 2017. 2, 3

[46] 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. 3, 5

[47] 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, 3, 7

[48] Saikat Roy, Tassilo Wald, Gregor Koehler, Maximilian R Rokuss, Nico Disch, Julius Holzschuh, David Zimmerer, and Klaus H Maier-Hein. Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. *arXiv preprint arXiv:2304.05396*, 2023. 3

[49] Stuart A Sandin, Esmeralda Alcantar, Randy Clark, Ramón de León, Faisal Dilrosun, Clinton B Edwards, Andrew J Estep, Yoan Eynaud, Beverly J French, Michael D Fox, et al. Benthic assemblages are more predictable than fish assemblages at an island scale. *Coral reefs*, 41(4):1031–1043, 2022. 3

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[51] Guolei Sun, Zhaochong An, Yun Liu, Ce Liu, Christos Sakaridis, Deng-Ping Fan, and Luc Van Gool. Indiscernible object counting in underwater scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13791–13801, 2023. 3

[52] Lv Tang, Haoke Xiao, and Bo Li. Can sam segment anything? when sam meets camouflaged object detection. *arXiv preprint arXiv:2304.04709*, 2023. 2, 3

[53] Vasilis Trygonis and Maria Sini. photoquad: a dedicated seabed image processing software, and a comparative error analysis of four photoquadrat methods. *Journal of experimental marine biology and ecology*, 424:99–108, 2012. 2

[54] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. 3

[55] Tianyi Wu, Sheng Tang, Rui Zhang, Juan Cao, and Yongdong Zhang. Cgnet: A light-weight context guided network for semantic segmentation. *IEEE Transactions on Image Processing (TIP)*, 30:1169–1179, 2020. 6

[56] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems (Neurips)*, 34:12077–12090, 2021. 2, 6, 7

[57] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision (IJCV)*, 129(11): 3051–3068, 2021. 6

[58] Ziqiang Zheng, Jipeng Zhang, Tuan-Anh Vu, Shizhe Diao, Yue Him Wong Tim, and Sai-Kit Yeung. Marinegpt: Unlocking secrets of ocean to the public. *arXiv preprint arXiv:2310.13596*, 2023. 3

[59] Jiageng Zhong, Ming Li, Hanqi Zhang, and Jiangying Qin. Combining photogrammetric computer vision and semantic segmentation for fine-grained understanding of coral reef growth under climate change. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 186–195, 2023. 2

[60] Tao Zhou, Yizhe Zhang, Yi Zhou, Ye Wu, and Chen Gong. Can sam segment polyps? *arXiv preprint arXiv:2304.07583*, 2023. 3

[61] 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. 3

[62] Peiqin Zhuang, Yali Wang, and Yu Qiao. Wildfish++: A comprehensive fish benchmark for multimedia research. *IEEE Transactions on Multimedia (TMM)*, 23:3603–3617, 2020. 3

[63] Zheng Ziqiang, Ha Tan-Sang, Chen Yingshu, Liang Haixin, Chui Apple Pui-Yi, Wong Yue-Him, and Yeung Sai-Kit. Marine video cloud: A cloud-based video analytics platform for collaborative marine research. 2023. 3

[64] 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. 2, 3

[65] 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. 3