CoralSRT: Revisiting Coral Reef Semantic Segmentation by Feature Rectification via Self-supervised Guidance

Ziqiang Zheng[1†] Yuk-Kwan Wong[1] Binh-Son Hua[2] Jianbo Shi[3] Sai-Kit Yeung[1]
[1]The Hong Kong University of Science and Technology [2]Trinity College Dublin
[3]University of Pennsylvania
† corresponding author: zhengziqiang1@gmail.com; Project website: https://coralsrt.hkustvgd.com

In this supplementary file, we first comprehensively review existing coral reef analysis works, foundation models, and most similar stuff segmentation works in Sec. 1. We detail our motivation why not explicitly modeling semantic correspondences due to the biology-specific features and our problem formulation for coral reef semantic segmentation in Sec. 2. Then we provide the construction details of our CoralWorld dataset and our testing set in Sec. 3. We provide more details regarding the implementations, comparative algorithms, experimental settings, and evaluation metrics in Sec. 4, where more experimental results and corresponding ablation studies are also included. Finally, we include the discussions about the main contribution, broader impact, limitations, and future work of our work in Sec. 5

1. Related Works

1.1. Coral Reef Analysis

We first provide an explanation about corals and coral reefs to avoid potential misunderstandings. Corals are individual organisms that form the building blocks of coral reefs, which are large underwater ecosystems made up of accumulated coral skeletons from many generations of coral polyps.

Point-based coral reef analysis. The early stage of coral reef analysis [47, 62] is based on sparse points (*e.g.*, Coral Point Count with Excel extensions: CPCe [46] for short). Point annotations are manually created by experts through software tools like CPCe for calculating coverage statistics [37] and helping to assess the coral reef ecosystem [52] both spatially and temporally. However, such a point-based annotation procedure is time-consuming and labor-intensive, and it usually takes several minutes to label 100 sparse points within one image. To speed up the analysis efficiency, CoralNet [19] and ReefCloud [12] integrated deep image classification for coral reef identification [20, 39, 55] from a specified patch of image regions centered by the sparse points. The semi-supervised manner [36] is adopted to speed up coral reef analysis. However, the *patch*-based image classification cannot delineate the complicated and irregular boundary of corals. More importantly, the cropped image regions (*e.g.*, 64×64 or 224×224) may contain *multiple semantic categories* while conventional image classification algorithms [26] cannot handle the *discrepancy* between visual content and labels. Thus, in this work, we did not choose the classification as the pre-training pretext task due to such *discrepancy* and such supervised pre-training will hurt the generalization ability of optimized models.

Major limitations of point-based approaches comprise 1) significant downsampling leads to over-/under- estimation (*e.g.*, only sampling 50 points/pixels for a 4K image with $3,840 \times 3,840$ pixels with sampling ratio $0.003392‰$); 2) it cannot describe the boundary and geometry of coral reefs, failing to support dense semantic understanding or 3D reconstruction [31, 40, 48, 64, 76]. The sparse point annotations cannot indicate the growth [41] or shrinkage [45] of coral communities and cannot be used to analyze spatial distributions on a fine scale.

Coral reef semantic segmentation. Considering these drawbacks, dense coral reef semantic segmentation [56, 63], which delineates the boundary of coral reefs while yielding semantics, is gaining increasing attention from the coral reef community. Coral reefs, notoriously known for their irregular boundaries [77] and extensive taxonomic diversity [48], illustrate substantial challenges in generating precise coral reef masks. The existing coral reef segmentation works can mainly fall into two categories: 1) sparse-to-dense conversion [16–18, 58, 59] by utilizing the already available sparse point annotations, propagating labels of annotated points to neighbor regions for generating dense semantic masks; 2) data-driven coral reef semantic segmentation [43, 54, 75] through full supervision by constructing coral reef segmentation datasets and benchmarks.

Sparse-to-dense conversion involves converting sparse points into dense masks and simultaneously propagating the semantic labels of the sparse points to masks. CoralSeg [17] and Fast-MSS [56] propagated the sparse point labels based on the Superpixels [21]. However, the Superpixel-based algorithms suffered from notoriously irregular boundaries

and significant coral diversity, failing to achieve satisfactory performance under adverse conditions. PLAS [58] proposed to ensemble the outputs of different superpixel segmentation models [42, 69] to get more stable and accurate augmented dense masks. Raine *et al.* [59] proposed a human-in-the-loop procedure to determine some pivotal key points for propagating the sparse points to dense masks based on the denoised DINO features [24, 53]. HIL [59] directly used DVT [70] to perform feature denoising and conducted label propagation based on denoised DINOv2 features. Even though the involvement of human choices (*e.g.*, smart points) could help promote the semantic segmentation performance, it is not scalable to have human involvement for point annotations, and the already available sparse point annotations are usually randomly sampled. In this work, we mainly explore sparse-to-dense conversion performance based on random points. The recent foundation models (*e.g.*, SAM series [44, 61] and CoralSCOP [74]) could also be used to generate dense semantic masks by receiving the labeled sparse points. However, both overinclusive and inclusive are inevitable. Finally, we also acknowledge that both HIL and CoralSRT are performing the sparse-to-dense conversion in the feature space. Compared with HIL, our method uses feature rectification to improve features from 5 different foundation models, and further explains the effectiveness of label propagation for coral segmentation compared to promptable segmentation.

Data-driven coral reef semantic segmentation approaches optimize models with full supervision at the pixel level. Promising coral reef segmentation performance has been achieved due to revolutionizing backbones [27, 28, 67] and an increasing scale of training data [43, 74, 77]. Yet most of these are bespoke approaches designed for specific coral reef images/categories [34, 77] and are not easily adaptable to new semantic categories and datasets [19]. Furthermore, most existing coral reef segmentation algorithms are purely data-driven [71, 74] and lack domain-specific design to dissect and address the essential properties of coral reefs. More importantly, there is no connection between early sparse point based analysis and current dense coral reef segmentation, resulting in the under-utilization of the vast potential of the abundant sparse point annotations produced by the reef analysis community [12, 19].

Limitation of existing semantic coral reef segmentation algorithms. Though satisfactory segmentation performance was achieved due to the more powerful network and larger scale of training data, there are still two main limitations for existing semantic coral reef segmentation algorithms: 1) **Limited pre-defined categories**. Existing semantic coral reef segmentation performs close-set coral reef segmentation. The semantic categories are pre-defined by domain experts based on the collected visual observations in advance. The optimized model can only yield the semantic prediction belonging to the constructed label set, which heavily weakens the generalization of optimized models since the coral biologists at different sites have different label sets and their label sets may contradict with each other. 2) **Poor generalization ability**, which comes from three different factors: *limited data diversity* [63] due to the relatively small scale of training data collected at local sites; *low network capacity* that leads to poor segmentation results under some adverse conditions [72] (*e.g.*, low visibility, motion blur, dynamic lighting and under/over exposure); and the *lack of zero-shot ability* to segment unseen coral reef images.

In this work, we aim to perform coral reef semantic segmentation based on features from the powerful foundation models (FMs) optimized by a significant scale of training data without introducing any human annotations or retraining/fine-tuning the FMs. Meanwhile, our CoralSRT through rectifying the features to approach the centrality of the semantic-agnostic segment, could better model the coral reefs.

1.2. Foundation Models

We discuss the essential differences between CoralSRT and existing foundation models in detail.

Comparison with promptable segmentation foundation models. The promptable foundation models like SAM [44], SAM 2 [61], and CoralSCOP [74] could receive prompts from the users to obtain the required masks in an interactive manner. Due to their semantic-agnostic training manner, the automatically generated masks result in many false positives and false negatives. CoralSCOP was the first foundation model for coral reef segmentation, with a parallel semantic branch for coral reef segmentation. However, such pure data-driven approaches ignored the intrinsic properties of coral reefs. Unlike instance segmentation [23, 38, 49, 73], where each instance can be clearly defined by the visually consistent and standardized "structural unit" among the same semantic, coral segmentation faces challenges due to the ambiguity in instance definition. This is caused by its amorphous, self-repeating, and asymmetric characteristics. The ambiguous or even conflicting annotations within the training annotations would degrade the performance of coral segmentation and introduce ambiguities to the generated masks. In this work, we propose performing coral semantic segmentation in the feature space to enable global image understanding through feature clustering, rather than focusing on local regional analysis based on full supervision [44], where all labeled sparse points contribute to assigning labels to unlabeled points. By utilizing the readily available sparse point annotations, we could better utilize the global image information rather than *promptable regional visual understanding*, even a few sparse points provided (*e.g.*, fewer than 5 points).

Comparison with DINO series [24, 32, 53]. Due to the pre-training on the large-scale dataset, DINO and DINOv2

demonstrated a feasible ability to model the implicit semantic correspondences between different images and regions even under the self-supervised setting. However, the feature space of the DINO series may be insufficient to accurately capture the complex boundaries of coral reefs. Thus, we propose a novel feature rectifying module in the feature space to strengthen the within-segment affinity based on the supervision from humans or FMs [44, 61] optimized by full supervision. Our approach bridges the self-supervised pre-training and supervised training in the feature space without retraining the FMs.

Comparison with image-text pre-training. Unlike CLIP [57], OpenCLIP [29], and BioCLIP [65], which utilize paired image and text annotations for model optimization, we did not explicitly model the semantic correspondences based on human annotations for better preserving the flexibility to various local requirements. Furthermore, our main goal is to generate dense masks, while these image-text pre-training approaches cannot directly localize the regions of interest while delineating coral reef boundaries for further analysis.

1.3. Stuff Segmentation

The coral reef semantic segmentation could be categorized as **stuff segmentation**. COCO-Stuff [23] conducted the first attempt to do the stuff segmentation and summarized 5 key properties between "instances/things" and "stuffs": *shape*, *size*, *parts*, *instances*, and *texture*. Inspired by this work, we have also summarized the challenges of conducting coral segmentation:

- **Amorphous distribution** leading to irregular boundaries. Corals often feature non-uniform, encrusting, or intricate growth patterns that defy simple geometric descriptions, such as irregular edges.
- **Repeatability or Fractality**: The structure of corals exhibits a self-similar, fractal-like nature, where patterns or arrangements recur at varying scales.
- **Diversity**. Corals or coral reefs consist of a wide variety of components, contributing to their complex appearance.
- **Self-occlusion**: Due to clustering or overlapping elements, parts of the structure obscure others, complicating visual interpretation.
- **Asymmetry**. The whole reef structure is usually asymmetric, with amorphous shapes.

Given the high complexity and inter-heterogeneity of coral reef images characterized by the above-discussed challenges, coral reef semantic segmentation underscores a more dedicated design to model the coral reefs.

In this work, we take a further step to model the characteristics of coral reefs and how they grow, which are inherently probabilistic. We explain the key difference between segmenting the general objects (*e.g.*, fish) and corals in Figure 1. From this figure, we emphasize one key difference between segmenting coral and general objects like fish: we cannot summarize a visually consistent "structural unit" for corals, thus we cannot conduct a reasonable and consistent instance segmentation for coral reefs as demonstrated in Figure 2. We further explain the differences between instance segmentation and coral segmentation from the task formulation and requirements. First, the instances are countable based on a clear definition of a visually consistent structural unit, and corals are uncountable, where area cover is computed for corals. It is biologically challenging to define a **visually consistent structural unit** for coral reefs to obtain individual instances among various coral species. Then, existing coral reef analytical approaches focus on the **area cover computation** rather than object/instance counting. Such a domain requirement encourages the semantic/stuff segmentation of coral reefs rather than instance segmentation. The amorphous and self-repeated properties of corals also result in a higher prediction tolerance to the predicted area since we only focus on the Intersection-of-Union between the predicted masks and labeled ground truth. In detail, splitting one connected coral area two separated masks while only missing some minor areas will result in a small penalty. In contrast, splitting two dogs to three dogs will result in high prediction errors. The intrinsic self-repeatability or fractality of corals also inspires us to utilize the median or mean statistics to better model coral reefs, since there is no clearly defined structural information between different regions from the same semantic category.

2. Motivation and Problem Formulation

2.1. Biology-specific Features

In this section, we discuss the biology-specific features of corals for why not **define explicit semantic correspondences/categories** or integrate the correspondences into our model design. In our CoralSRT, the semantics are inherited from the annotated sparse points after the sparse-to-dense conversion. We summarize four reasons as follows:

- **Flexibility**. **How to define correspondences between corals** is highly subject to the domain requirement and human involvement (*e.g.* some coral biologists are separating bleached and normal corals out even if the corals are from the same species). It is not the main focus of this paper, and we advocate such flexibility to support various domain requirements. In this way, our approach demonstrated strong flexibility to various downstream coral reef analysis tasks without pre-defining the semantic label sets. Users could self-design their own label sets, and our model also illustrated a strong zero-shot ability to coral reef images from different sites all around the world, thanks to the large volume of pre-training data.
- **Reticulate pattern**[1]. Compared with other sea creatures,

[1] https : / / www . coralsoftheworld . org / page /

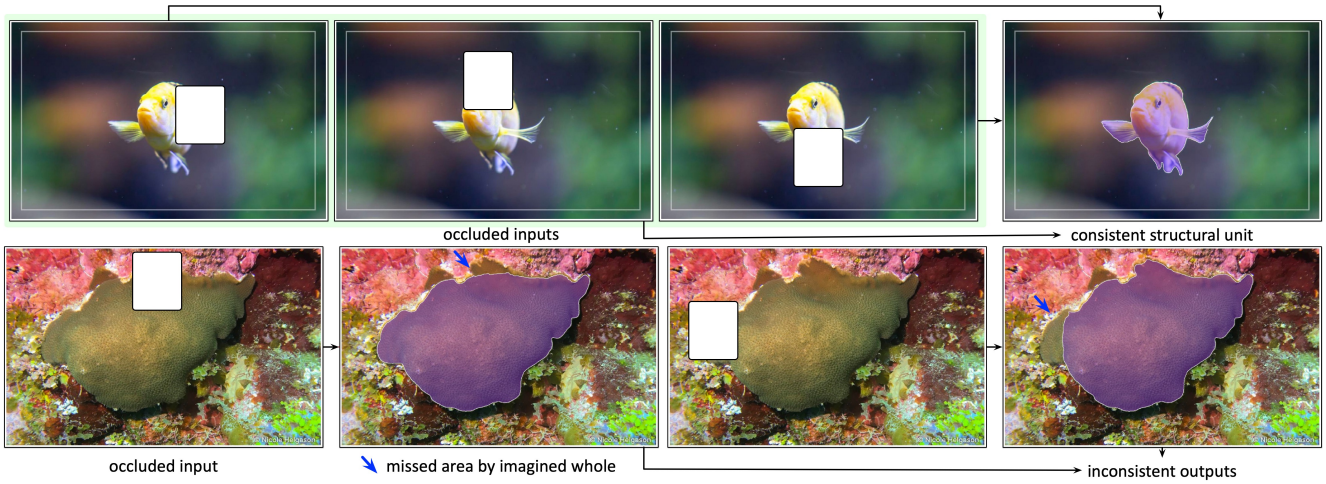

Figure 1. The key difference between segmenting the fish and the corals: the fish has a visually consistent "structural unit" while the corals do not have. No matter which part of the fish is occluded, we humans can almost imagine its boundary and shape. But for corals, we cannot imagine a consistent output from two occluded inputs with different regions occluded.

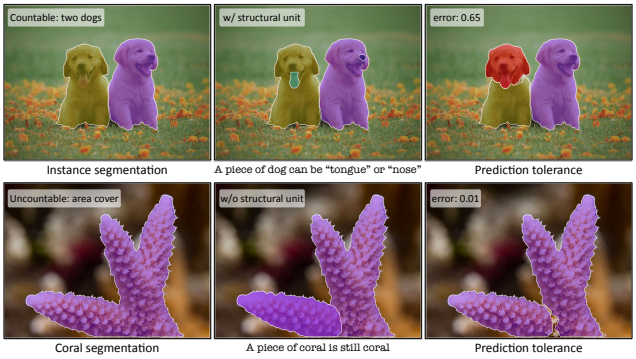

Figure 2. We differentiate instance segmentation and coral segmentation from three aspects: 1) task formulation, 2) whether it has a visually consistent structure unit, and 3) the prediction tolerance to the segmentation outputs.

corals have their specific reticulate evolution: the definition of species will change over time (*e.g.*, converging and diverging). A perfect taxonomic hierarchy of corals would be based on 1) all species being taxonomically isolated units and 2) every species being included. In the real world, these conditions can never be met. This creates an endless dilemma, for humans/experts cannot easily communicate in terms of continua. It also results in the situation that the fine-grained coral reef analysis is still closely linked to domain expertise involvement.

- **Discovering property**. From a practical perspective, the optimized model is required to discover novel coral species that do not exist in the training data. The need for redundant labeled examples for every new semantic category limits the applicability of optimized models since we do not know the semantics in advance or the data distribution of the testing data. Overfitting to the pre-defined semantic categories will weaken the generalization ability of the optimized models.
- **Inefficient visual observation**. Due to the specific underwater conditions, it is challenging to identify the semantic categories (*e.g.*, species) of the corals due to the poor visibility and indiscriminative features. Furthermore, for some very similar corals, the hierarchical taxonomy can only be reliably identified through genetic methods [66] (*e.g.*, DNA barcoding) or molecular technology, where solely visual information is far from accurate identification. Thus, clearly defining semantic categories at a fine-grained granularity will introduce noise and ambiguity to optimizing the models.

overview-of-coral-taxonomy/

2.2. Problem Formulation

In this section, we detail how we formulate the coral reef semantic segmentation. We revisit the definition of segmentation as the process of partitioning an image into **segments** with homogeneous presentations. For CRSS, we define two fundamental formulations as demonstrated in Figure 3:

- A segment is a connected region where all the pixels within it belong to the same implicit semantic category. The segments from two semantics must not be clustered into the same segment (*e.g.*, Point **A** and **B**).
- We model **within-segment** affinity (Point **B** and **D**) and **cross-segment** affinity (Segment **1** and **3**), effectively reducing intra-segment variance and enhancing inter-segment differences.

The affinity captures repeated textures, similar geometries, and biologically similar features. By modeling these two types of affinity, we enable CRSS to exhibit strong flexibility and generalization, as all CRSS tasks can be adapted to our formulation.

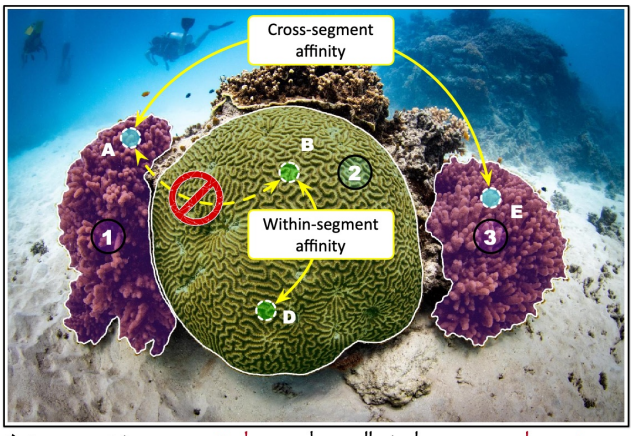

Figure 3. Our simple and fundamental problem formulation for CRSS: *segment* as the basis to model within-segment and cross-segment affinities.

3. Data Construction

3.1. Training Data Construction

The advance of the recent foundation model regarding visual understanding often comes with a greater demand for a significant scale of training data in terms of both data *diversity* and *coverage*. In this section, we discuss the detailed procedure of constructing a diverse and comprehensive coral reef dataset called **CoralWorld**.

Construction of CoralWorld dataset. To optimize a robust and effective model, it is important to build a large and diverse dataset by collecting coral reef images from a wide range of tasks, locations, and contexts. In this work, we have considered both **public** and **private** data sources to enrich the diversity and promote the quality of collected images. We illustrate the details about the composition of our constructed CoralWorld dataset in Table 1.

3.1.1. Public Data

Public coral reef images were downloaded and curated from the existing datasets, platforms, and websites.

- **CoralNet** [19]. We have downloaded all the public coral reef images from 1,050 public sources of CoralNet until the date of Nov 23, 2024. Please note that we failed to scrape some coral reef images that are a relatively small portion (estimated to be smaller than 1%) due to the unstable network link. Finally, we have obtained 555,886 coral reef images from 861 non-empty public sources (some public sources do not contain any coral reef images).
- **CoralMask** [74] dataset is the largest coral reef dataset with dense coral reef masks: 38,928 coral reef images with 299,557 coral reef masks after data cleaning. We adopt the coral reef images from the diversified CoralMask dataset to optimize our model.
- **Pangea** [60] dataset comprises 147 underwater scenes of coral reefs. Among these, 23 transects include photo quadrat images annotated by biologists with 47 labels describing habitat features and biodiversity. Following the official guidelines, we extracted 7,446 coral reef images from these 23 transect videos.
- **Moorea Labeled Corals (MLC)** [18] dataset includes 2,055 images, collected from three different habitats.
- **Benthoz15** [22] dataset consists of 407,968 expert-labeled sparse points from 9,874 geo-referenced images. In each image, up to 50 pixels were randomly annotated.
- **Shutterstock** [13]. A promising source of diverse and nearly unlimited public sources is web scrapes, such as web images from different websites. Meanwhile, while these online web sources are many orders of magnitude larger than current specially curated marine image datasets, they have significant data quality issues. We adopt the keyword filtering to remove those unrelated images based on the Alt-text information.
- **YouTube** [2]. We also downloaded coral reef videos (using different keywords to query both coral reef transect videos and casual reef videos) from YouTube. Then we manually crop the coral reef frames from the collected coral reef videos. Finally, we have obtained 12,862 coral reef images with high diversity and variation.
- **EOL** [14]. We have collected 48,139 coral reef images from 428 coral species, where each coral species has at least 10 images.
- **iNaturalist** [4]. Similarly, we have also scraped coral reef images from the iNaturalist website. In detail, 12,761 coral reef images with research-grade quality were included in our dataset. These field data captured by coral biologists have high research value and are preferred for domain analysis.
- **100 Island Challenge** [1] is a large-scale natural experiment, investigating the independent and interacting effects of oceanography, geography, and human activities in affecting the structure and growth of coral reef communities. Until now, the official website has provided the captured visual data for 48 sites, and we chose one field of data for each site to include the reef images.
- **Others**. We have also included coral reef images from other small- or medium-scale coral reef datasets such as [5–11]. For these datasets, we perform human checking and conduct down-sampling via different sampling ratios.
- **Great barrier reef**. We construct the coral reef dataset from the official website [3]. This dataset was constructed to accurately identify starfish in real-time by building an object detection model trained on underwater videos of coral reefs.

3.1.2. Private Data

We also consider the private coral reef data from coral biologists.

Table 1. **Composition of CoralWorld dataset.** We report the list of data sources and associated splits used to construct the CoralWorld dataset, and how these data sources were included (as is, without downsampling or via rule-based downsampling). For our MSCR, we indicate the actual number of augmented images and the final number included in the final dataset. We chose to include as many data sources as possible in the pre-training coral reef data in order to cover as many sites/regions as possible.

Task/Purpose	Dataset / Split	Images	Sampled	Augmented	Final
Classification	CoralNet [19] / 1,050 sources (public)	555,886	Rule-based	MSCR	887,823
Classification	Pangea [60] / All	7,446	✘	MSCR	13,742
Classification	EOL [14] / 831 species	43,189	Human	MSCR	76,271
Classification	Benthos15 [22]	9,874	✘	MSCR	9,874
Classification	Great Barrier Reef [50]	23,501	5	MSCR	4,701
Classification	Others [5–11]	–	✘	MSCR	33,694
Segmentation	CoralMask [74]	38,928	✘	MSCR	100,352
Segmentation	Mosaics UCSD [34] / train	4,193	✘	MSCR	4,193
3D reconstruction	100islands [1] / 48 sites	157,787	✘	MSCR	630,932
3D reconstruction	Mosaics	6	Rule-based	Cropping	4,207
Transect videos	NOAA[2]	61,143	✘	MSCR	244,035
Transect videos	21 sources	14,555	✘	MSCR	31,139
Transect videos	Red sea	360	✘	MSCR	720
Web images	Shutterstock [13]	304,982	✘	MSCR	304,982
Web images	Youtube [2]	12,862	✘	MSCR	24,971
Web images	iNatualist [4]	12,761	Human	MSCR	16,011
Private data	Red sea, Indo-Ocean	204,166	Human	MSCR	204,166
Private data	Australia	3,135	Rule-based	MSCR	4,968
Private data	Pacific Ocean	4,596	Human	MSCR	31,617
Private data	India, Japan	1,867	Human	MSCR	7,468
Private data	Malaysia	1,250	Human	MSCR	4,888
Private data	Deep sea	74	Human	MSCR	365
					2,641,119

- **Routine checking**. The coral biologists deploy a series of video transects along predetermined reef sites, ensuring the camera is calibrated and securely anchored to avoid drift. As the coral biologists monitor the video footage in real-time, they coral biologists check for the clarity of the images and whether the camera's field of view captures a representative cross-section of the reef, including coral cover and fish populations. Afterward, the coral biologists review the footage for any signs of coral stress, such as bleaching, disease, or predation, and make note of any significant changes in reef health. These field data captured by the coral biologists were preferred for reef monitoring and further analysis.
- **Transect videos**. Coral reef transect videos are detailed, time-lapse recordings taken along a fixed path on the reef, capturing visual data on coral species, fish populations, and overall reef health. These videos provide a non-invasive method for monitoring changes in the ecosystem over time, allowing researchers to assess coral cover, track disturbances like bleaching or disease, and make informed decisions for conservation efforts without physically disturbing the reef. The transect videos were collected at different sites, including Hawaii, Moorea, Hong Kong, Brazil, South China, Philippines, Great Barrier Reef, Red Sea, Maldives, Indo-Ocean, Okiwala, India, Malaysia, Korea, Mexico, San Diego, and Palau.
- **Lab-setting observations of corals**. Lab-setting coral reef observations involve examining coral samples or live coral colonies in controlled environments, where factors like water temperature, light, and nutrient levels can be manipulated to study specific reef stressors or behaviors. These observations allow coral reef biologists to isolate variables, conduct detailed experiments on coral growth, and simulate potential future conditions like ocean acidification or warming. Unlike transect videos, which capture natural, real-time reef conditions, lab observations provide a deeper understanding of coral physiology and resilience under controlled stress scenarios. We also combine the

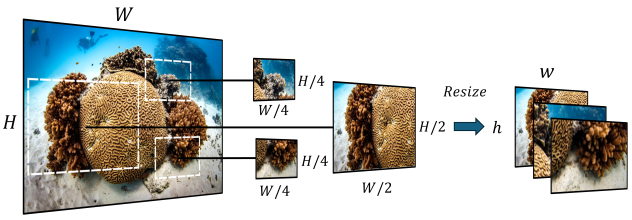

Figure 4. The illustration of multi-scale cropping to force the model to learn the scale-invariant feature representations.

laboratory observations of the corals in our CoralWorld dataset to support the fine-grained coral reef analysis. All the coral reef images were captured and contributed by local coral biologists.

3.1.3. Post-Processing and Analysis

Specific features of coral reef images. The coral reef monitoring images are usually **benthic** images (bird of view) or **quadrat** images. These images are usually high-resolution, capturing comprehensive information about the coral ecosystems. Furthermore, due to the specific underwater conditions [15, 68], visibility degradation, dynamic lighting, and color distortion are inevitable, leading to further challenges for humans or even domain experts to accurately identify the coral reefs. The motion blur within the coral reef transect videos is also a challenge. Finally, the collected coral reef images cover a large range of contexts, scenarios, and tasks.

Multi-scale cropping and resizing. Considering a specific property of the transect quadrat images (which are usually with high resolution), directly resizing one 4K or even larger image to a fixed image size (*e.g.*, 224×224 for DINO [24]) for pre-training will lead to significant information loss and the model will fail to model local regional information. To address such a limitation, we perform multi-scale cropping and resizing (**MSCR** for short) as demonstrated in Figure 4 to ensure the model can learn the scale-invariant feature representations, including both local information and global context information. Finally, we gathered the first large-scale and diverse training data of 2.64 million coral reef images to demonstrate the effectiveness of our approach.

Removing duplicated images and quality checking. In terms of data, we propose an automatic pipeline to build a dedicated, diverse, and curated coral reef image dataset instead of uncurated data from public websites and datasets. The uncurated datasets with noise lead to a significant drop in the quality of the features. The diversity and coverage of training data are important for optimizing efficient and effective foundation models. We have performed human checking for partial images from some public sources and datasets to remove those similar images to ensure the high quality of the collected pre-training data.

Diversity analysis. The whole dataset contains the reef images captured under various conditions, including close-range monitoring, transect surveying, and remote imaging. We randomly sample 400 images from the whole dataset to visualize the diversity of the collected dataset. The visualization is illustrated in Figure 5. We also provide the geographic distribution of our constructed CoralWorld dataset in Figure 6, demonstrating the diversity and coverage of the coral reef images from the whole world.

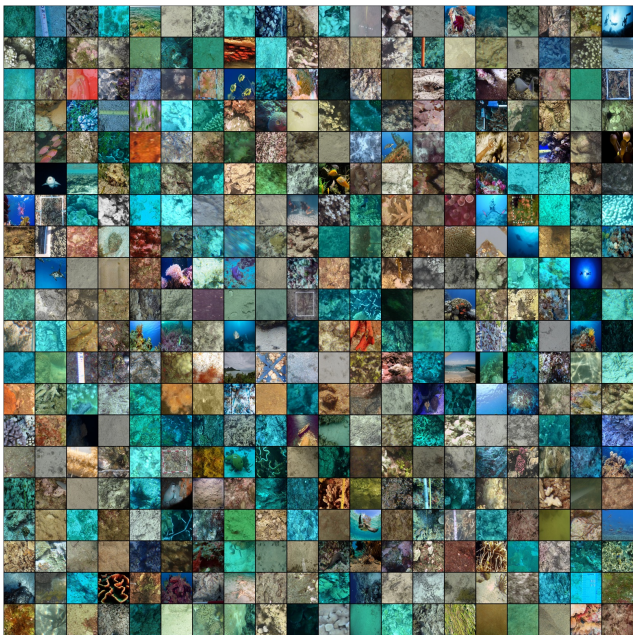

Figure 5. Randomly sampled 400 coral reef images from the constructed CoralWorld dataset.

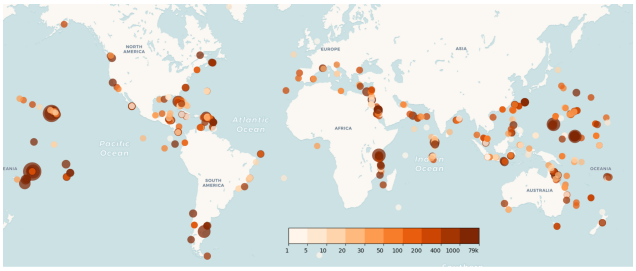

Figure 6. The geographic distribution of the collected CoralWorld dataset.

3.2. Construction of Testing Set

We provide the dataset construction details of our testing set. For annotating the coral reef images with the semantic reef masks, we utilize SAM [44] to speed up our dense semantic mask annotations. As for the data sources of our constructed evaluation set, we describe the data collection details as follows:

- **1. Deep Sea**. We curate a collection of deep-sea coral images from publicly available platforms. From this collection, we manually select 100 high-quality images in

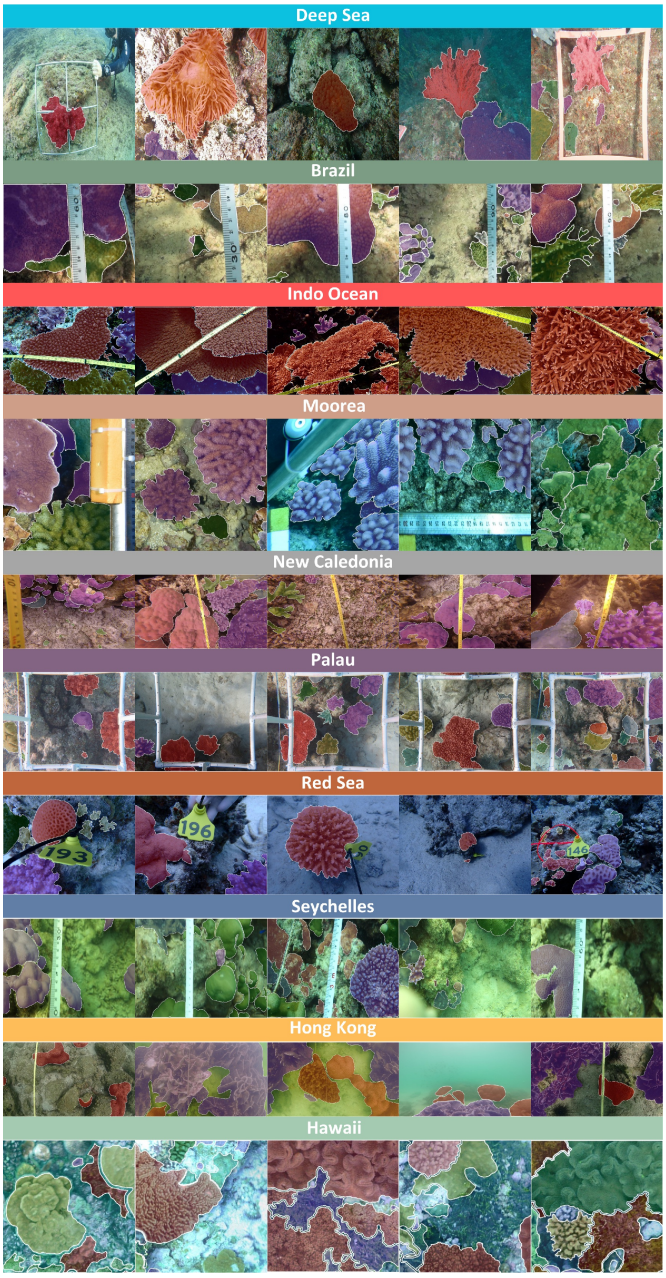

Figure 7. The example images from the 10 different subsets of our constructed testing set. We only visualize the coral reef masks for better illustration.

which the coral structures were clearly visible to construct this testing set.

- **2. Brazil**. We curate this subset from 24 coral reef transect videos captured by coral biologists. We first manually extracted 3,984 coral reef images. Then we perform the random sampling from the total images to obtain the final 100 coral reef images for human labeling.
- **3. Indo Ocean**. This testing set was developed in collaboration with a coral biologist team, who provided precise annotations for the coral regions. We select 100 representative images to construct the final testing set.
- **4. Moorea**. The local coral biologists have captured the routine field data to monitor the coral reefs. Until now, local coral biologists have collected 5,571 high-resolution coral reef quadrat images. We have randomly sampled 100 images for the local biologists to label.
- **5. New Caledonia**. Similarly, we constructed this subset based on the coral reef transect videos, which were contributed by the local coral biologists. We extract one image frame every 3 seconds from the 21 collected reef videos. We finally obtained 3,256 coral reef images and randomly sampled 100 coral reef images for human labeling.
- **6. Palau**. We construct this subset based on the 547 coral reef images contributed by the local biologists. All the coral reef images are benthic quadrat images. We randomly sampled 100 coral reef images and invited the local coral biologists to do the semantic labeling according to their requirements.
- **7. Red Sea**. We curate this subset from the collection of visual observations of the field data. The coral biologists have captured 856 diverse visual images from comprehensive conditions. We sample 100 images for annotating.
- **8. Seychelles**. We construct this subset from 14 transect videos captured by the local biologists. Similarly, we extracted one frame every three seconds and finally obtained 2,845 coral reef images. We randomly sampled 100 frames for expert annotations.
- **9. Hong Kong**. We curate this subset from the existing HKCoral dataset [77], which contains 6 growth form annotations, including Encrusting, Massive, Faliceous, Laminar, Branching, and Columnar.
- **10. Hawaii**. We curate this subset from the existing Mosaics UCSD dataset [34], which contains 34 semantic categories, including both algae and other non-coral organisms. We randomly sample 100 uncorrupted images from the testing set.

Our constructed evaluation set considered the geometric and species diversity of the coral reefs. We have built the first medium-scale, diverse, and comprehensive semantic coral reef segmentation benchmark to better measure the ability of coral reef segmentation algorithms. Our evaluation set could also serve as a valuable benchmark to conduct cross-site and open-set coral reef segmentation. We will release our evaluation set to the coral reef analysis community to speed up the automatic coral reef analysis. The constructed evaluation set supports the semantic understanding of coral reefs from different levels to align with the domain requirements, such as the growth form, genus, or even species.

To better illustrate the diversity and coverage of the constructed testing set, we visualize the randomly sampled testing images with corresponding semantic mask annotations

for each sub-set in Figure 7.

4. Experiments

In this section, we provide more details about the experimental settings, the dataset used, the implementations, the evaluation metrics employed, the comparative algorithms, more experimental results, and corresponding ablation studies to demonstrate the effectiveness of the proposed algorithm.

4.1. Experimental Settings

Datasets. We first provide the details of the datasets used in our paper. **Mosaics UCSD** [34] dataset contains 23 taxonomic and 8 functional groups with dense ground truth masks [34]. This dataset is the only publicly available dataset that supports coral genus segmentation with dense ground truth masks. It contains 4,193 training images and 729 test images with 34 semantic classes and the background class. All the images are with 512×512 size. We only choose the testing set of this set for evaluation only and removed the corrupted testing images (696 images left after data cleaning). **HKCoral** [77] dataset contains coral reef images collected in the wild and corresponding semantic mask annotations from the growth form levels. It contains 2,515 images annotated by 6 various growth forms and the background class. We follow the training/validation/testing split with 1500/500/515, and we only report the experimental results of the testing set. The exact dataset usage in our experiments is illustrated in Table 2. **Seaview** dataset [37] encompasses over one million standardized downward-facing "photo-quadrat" images (covering approximately one square meter of the sea floor) with 55 million sparse point annotations, which were collected between 2012 and 2018 at 860 transect locations around the world, including the Caribbean and Bermuda, the Indian Ocean (Maldives, Chagos Archipelago), the Coral Triangle (Indonesia, Philippines, Timor-Leste, Solomon Islands), the Great Barrier Reef, Taiwan, and Hawaii. This dataset contains more than 1 million coral reef images in total. The images from the Seaview dataset do not cover the coral reef images with various viewpoints and fields of view. Even though the collected quadrat images of the Seaview dataset are high quality, the restricted diversity and coverage also poses challenges for models pre-trained on the Seaview dataset when handling the coral reef images with random viewpoints and field of views. **BenthicNet** [51] dataset is a global compilation of seafloor images. We download all the unlabeled images from the provided link[3] and there are around 1.45 million images in total, including both labeled and unlabeled images. Please also note that the BenthicNet dataset contains multiple sources with images sampled from the videos (*e.g.*, Pangea [60]) with a high sampling rate. There is a high potential to contain images with strikingly similar images in the whole dataset. We only adopt the Seaview and BenthicNet datasets for pre-training.

[3] https://www.frdr-dfdr.ca/repo/files/9/published/publication_609/submitted_data/01_BenthicNet/images

Table 2. Exact dataset usage in our experiments. **S2D**: Sparse-to-dense conversion.

Datasets	CoralWorld (2.64M images)	UCSD Mosaics (test: 696 images)	HKCoral (train/test: 1,500/515 images)
Usage	Self-supervised learning	Zero-shot **S2D** (testing only)	**S2D** for pseudo labels to optimize semantic segmentation models
Labels	No labels	dense masks as GT for performance comparison; random point sampling from GT masks for **S2D**	

Evaluation metrics. In this work, we adopt **mIoU** and **mPA** as the main evaluation metrics. mIoU is widely regarded as the standard metric for segmentation tasks. mPA calculates the ratio of correctly classified pixels for each class to the total number of pixels within that class, providing an average accuracy measure across all classes. Different from existing approaches that compute the class-level mIoU and mPA, we choose to compute the **image-level** mIoU and mPA considering two main reasons: 1) the distribution of semantic categories is highly imbalanced and the semantic category sets between various sites are different. Reporting the mIoU among the whole testing set with 10 different subsets is challenging and cannot effectively measure the ability of various models to conduct the sparse-to-dense conversion. Thus, we report the mIoU and mPA for every individual image and compute the average score for all the testing images. 2) Computing the class-level mIoU and mPA scores for promptable segmentation models is challenging due to overlapping regions between the masks generated by sparse points with different semantic categories. To resolve this, we calculate the IoU and PA for each semantic category and report the image-level mIoU and mPA to more effectively evaluate the sparse-to-dense conversion performance. In other words, each semantic category in an image has its own independent semantic mask, with the label inherited from the provided sparse points.

4.2. Implementation Details

Optimization of CoralSRT and CoralSRT-. To optimize both CoralSRT and CoralSRT-, we adopted the network architecture DVT [70] and removed the positional embedding. For optimizing CoralSRT-, we adopt the coral reef masks provided by the CoralMask dataset as the human-annotated masks. As suggested by CoralSCOP [74], we also generate the non-coral masks based on SAM 2 as the negative masks. We utilize both positive and negative masks to force CoralSRT- to strengthen the within-segment affinity in the feature space constructed from various models and backbones. To optimize the CoralSRT, we utilize the model-generated masks from SAM 2 as the supervision and remove the masks with area values smaller than 1,024. For running SAM 2, we modified the stability threshold to 0.85 and kept all the hyperparameters default. For both CoralSRT and

CoralSRT-, we set the batch size to 32 and the number of training iterations to 50,000. We conduct the training of both CoralSRT and CoralSRT- on a single GTX 3090 GPU.

DINO pre-training. We conduct the pre-training experiments on 6 Nvidia H800 GPUs with a batch size of 224 per GPU. The image resolution is set to 224×224 and the number of training epochs is set to 100. We adopt the ViT-B/16 as the network backbone and follow the official hyperparameters provided by DINO [24] to conduct the pre-training.

Sparse-to-dense conversion. We adopt the official codes[4] of Fast-MSS to perform the sparse-to-dense conversion. For evaluating the SAM series and CoralSCOP under the prompt-based setting♠, there are two settings for running the point-based algorithms: 1) one point by one point and 2) combining all the sparse points from the same semantic class. Under both settings, we utilize all the points from other semantic classes as negative points. We empirically found that the former setting will lead to better sparse-to-dense conversion due to grouping two geometrically separated regions into one mask will lead to visible artifacts. **Point sampling**: during the evaluation procedure, sparse points are randomly sampled from dense masks within the whole image.

For evaluating various foundation models (SAM, SAM 2, DINO, DINOv2, and CoralSCOP) under the feature-based setting♣, we adopt the features from the last layer (11_{th} layer) of those foundation models for a fair comparison. We then conduct the KNN clustering following the setting of HIL [59] to perform the label propagation. We included DVT [70] and FeatUp [35] for comparing the rectified features. Please note that we conduct the feature clustering (KNN and $K = 1$ as suggested in [59]) based on the same labeled sparse points for all algorithms under the feature-based setting. We did not provide the ablation studies of using various values of K since [59] already demonstrated that increasing K would not increase performance gains, and our experimental results are also aligned with this observation.

Optimization of semantic segmentation models. The three semantic segmentation models (DeeplabV3 [25], SegFormer [67], and Mask2Former [28]) were optimized by 80,000 iterations following the official configurations on GTX 3090 GPUs. We adopt ResNet101-D8, MiT-B5 and Swin Transformer (Base) network backbones for DeeplabV3, SegFormer and Mask2Former, respectively.

4.3. More Results

In this section, we provide more experimental results to provide insights for coral reef analysis.

4.3.1. Dissecting Promptable Segmentation Models

In this section, we dissect why the promptable segmentation models cannot achieve satisfactory sparse-to-dense conversion performance compared with performing feature clustering in the feature space. We attribute this failure due to the intrinsic properties of coral reefs: the distribution of corals is irregular, amorphous, and self-replicating. We cannot have a visually consistent structural unit to separate consistent instance masks [49, 73] for segmenting the corals. As demonstrated in Figure 8, the annotators have their own preferences for annotating the coral masks. This lead to inconsistent mask annotations within the training annotations, therefore weakening the ability of models to segment the corals. Furthermore, we also notice that the promptable segmentation models are sensitive to the spatial choices of sparse points as demonstrated in Figure 9. It is very tricky to find out the pivot points within each mask to generate consistent and precise dense masks after the sparse-to-dense conversion. More importantly, the annotated sparse points are usually randomly sampled and annotated by the coral reef biologists. Such spatial sensitivity of the promptable segmentation models is not favored by the coral reef biologists.

We acknowledge the automatic ability of the promptable segmentation models to segment the coral reefs based on the

Figure 8. Since there is no visually consistent structural unit to separate the corals, the users have their own preferences for annotating the masks, which leads to the inconsistency between the annotated masks from different annotators. Please note the color is only used for illustrating different masks and is without semantics.

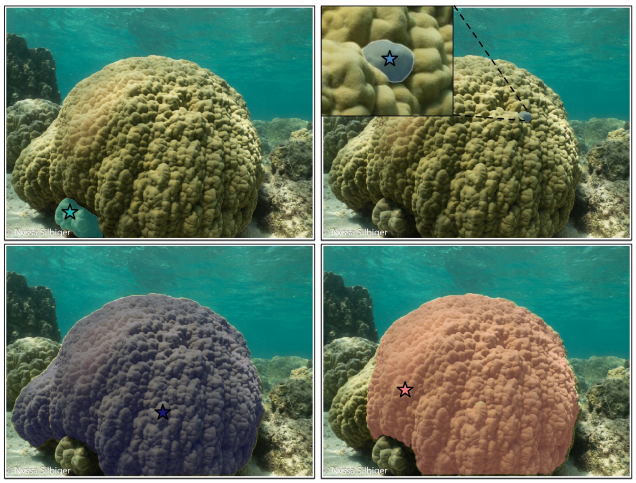

Figure 9. The promptable segmentation models are sensitive to the spatial choices of the sparse points (indicated by the stars). Various point prompts will lead to totally different mask outputs. Please note the color is only used for illustrating different masks and is without semantics.

[4] https://github.com/JordanMakesMaps/Fast-Multilevel-Superpixel-Segmentation

grid point prompts, especially CoralSCOP with the parallel semantic branch to address the over-segmentation issue. However, it is not the optimal solution to utilize the promptable segmentation models to perform the sparse-to-dense conversion based on randomly labeled sparse points due to the irregular and amorphous distribution of coral reefs. There is no visually consistent structural unit to separate the different coral masks as the instance segmentation, making promptable segmentation models less efficient on coral reef segmentation.

4.3.2. Model-Agnostic

The proposed CoralSRT is model-agnostic and could be combined with various foundation models, effectively enhancing the within-segment affinity. We conduct experiments based on DINO, DINOv2, SAM, SAM 2 and CoralSCOP and provide the detailed qualitative comparison in Figure 10. As demonstrated, CoralSRT could yield more effective and consistent features to better serve the downstream semantic segmentation task. Especially, CoralSRT could heavily alleviate the grid artifacts of SAM and SAM 2.

4.3.3. Generalization Ability of CoralSRT

Generalization ability to random coral reef images from Internet. We provide more qualitative results of our CoralSRT to the random coral reef images scraped from the Internet based on DINO and DINOv2. The PCA visualization (first 3 principal components) of original features and rectified features by CoralSRT were illustrated in Figure 11. Our method demonstrates a strong generalization ability to distinguish the coral reefs. The learning in the feature space enables the model to recognize the coral reef features extracted by powerful FMs and promote the understanding ability of how to cluster the coral reefs with implicit semantics.

4.3.4. Zero-shot Sparse-to-Dense Conversion

One of the biggest contributions of our CoralSRT to the whole reef analysis community is that our method provides an efficient and effective way to convert the redundant sparse point annotations from the whole reef analysis community to the dense semantic masks in a zero-shot manner. The generated dense masks are valuable for the 3D semantic reconstruction of the reef community and also serve for more reliable and accurate coral cover computation. Besides, the converted dense masks could also be utilized as the pseudo ground truth for optimizing dense segmentation models for local sites without any additional human annotations. Our CoralSRT does not require any training or fine-tuning. Considering the Seaview dataset [37] has provided redundant sparse point annotations, we conduct the zero-shot sparse-to-dense conversion on the Seaview dataset and provide the qualitative results in Figure 12. As illustrated in Figure 12, most of the converted dense masks are reasonable. The proposed method has the potential to push the reef analysis to the era of dense masks without introducing additional human annotations or further retraining/finetuning the foundation models.

4.4. Ablation Studies

In this section, we provide more experimental results to dissect the effectiveness of each component of the proposed method.

4.4.1. Comparison with FeatUp and DVT

Since FeatUp and DVT were not optimized by coral reef images, to make a fair comparison with our algorithm, we train both FeatUp and DVT on the CoralMask dataset and evaluate the performance of models on the testing set of Mosaics UCSD dataset in a zero-shot manner. Please note that we do **not** use any coral reef mask labels from the CoralMask dataset and we utilize SAM 2 to generate the mask supervision to ensure there is no any additional labels introduced during the whole training procedure. We follow the official instructions of DVT and Featup to conduct experiments on the CoralMask dataset until the convergence. We report the quantitative result comparison with these two algorithms in Table 3. Please note that all the algorithms are using the same labeled sparse points. As demonstrated, optimizing DVT and FeatUp on the CoralMask dataset does not lead to the performance gains than training them on the ImageNet or COCO-Stuff datasets since they are mainly focusing on the geometric correspondences between various data augmentations (FeatUp) and the internal feature artifacts (DVT) within the ViT architecture. Compared with these two models, our method demonstrates a more powerful ability to conduct the coral reef segmentation since we introduced the mask supervision in the feature space to strengthen the within-segment affinity for better modeling the coral reefs.

4.4.2. Sensitivity to Spatial Choice of Sparse Points

We also conduct experiments on the testing set of the Mosaics UCSD dataset in a zero-shot manner. We conduct the experiments under various settings by 5 times and report the mean scores and corresponding standard deviation. Please note that all the algorithms use the same labeled sparse points to make a fair comparison. All the quantitative results are reported in Table 4. As reported, the proposed method could achieve the best sparse-to-dense conversion performance under the setting of multiple trials. In this experiments, we report the quantitative results of CoralSRT-  to report the upper bound of our method.

4.4.3. Investigating Pre-training Datasets

We provide the zero-shot sparse-to-dense conversion performance of DINO features (ViT-B/16) pre-trained on different datasets in Table 5 on the testing set of the Mosaics UCSD dataset. The corresponding results of rectified features by

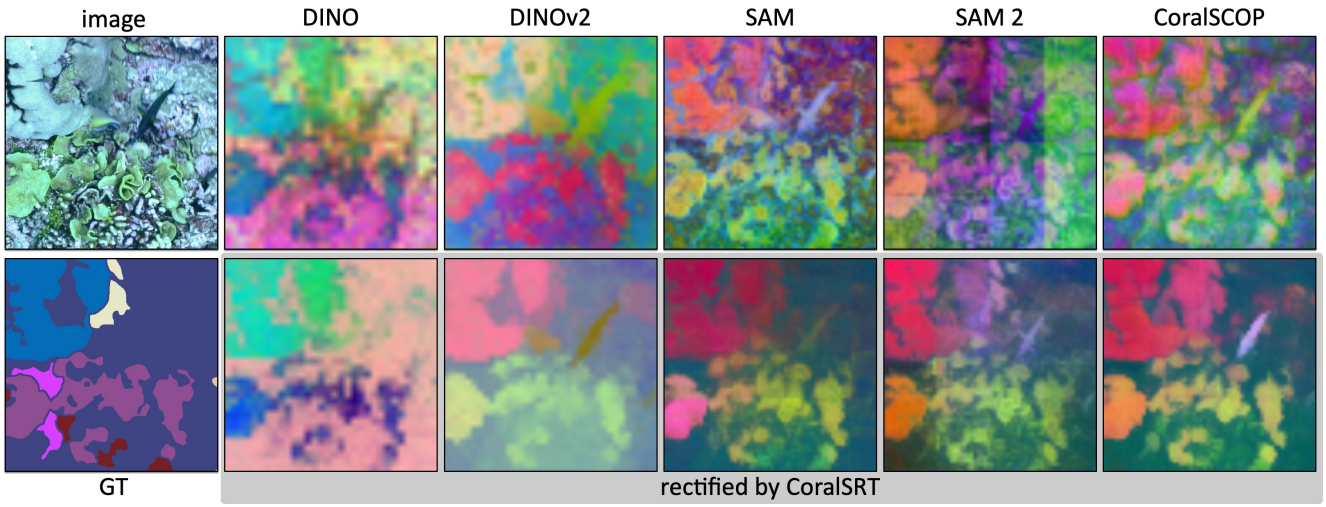

Figure 10. PCA visualization (first 3 components) of both original features and rectified features by our CoralSRT from various foundation models.

Table 3. Quantitative zero-shot comparisons between DVT [70], FeatUp [35] and CoralSRT on Mosaics UCSD dataset [34]. Both DVT and FeatUp were optimized on the CoralMask dataset [74] for a fair comparison.

Methods	5 points		10 points		20 points		50 points		100 points	
	mIoU	mPA	mIoU	mPA	mIoU	mPA	mIoU	mPA	mIoU	mPA
DVT [70]	16.85	18.82	24.37	28.64	31.68	37.21	43.27	51.12	51.01	61.03
FeatUp [35]	15.86	18.01	23.57	27.63	31.45	37.11	43.48	51.12	52.03	62.14
CoralSRT	18.15	20.98	26.45	30.67	33.27	40.01	44.66	53.03	52.18	62.19

our CoralSRT were also reported in Table 5. To further investigate the features from various models, we illustrate the PCA visualization in Figure 13. The rectified features produced by our CoralSRT exhibit higher within-segment affinity and offer more efficient feature representations, leading to improved performance in semantic segmentation.

Furthermore, as reported in Table 5, with more pre-training images, the DINO models could generate more efficient features with fewer holes or artifacts, demonstrating a stronger ability to cluster similar regions. We also notice that diversity and quality are important factors for constructing a more efficient feature space. Even if the BenthicNet dataset contains more pre-training data than the Seaview and CoralWorld-1M datasets, the generated features from the DINO model pre-trained on the BenthicNet dataset are still worse than the DINO models pre-trained on the Seaview and CoralWorld-1M datasets. Our proposed CoralSRT could effectively promote the sparse-to-dense conversion performance under all the settings, demonstrating the effectiveness of the proposed method.

5. Discussions

5.1. Contribution Claim

We first discuss our main contributions over existing works. We have three fundamental contributions to coral reef segmentation:

- **Problem revisiting**. We have revisited coral reef semantic segmentation. We have comprehensively dissected the key difference between segmenting general objects and corals: whether there is a visually consistent structural unit. Based on the intrinsic properties of coral reefs, we have defined the segment as the basis for performing coral reef segmentation to model *within-segment* and *between-segments* affinities, where the intrinsic properties of coral reefs, and the domain requirements were considered. Our approach provides a novel perspective for performing coral reef segmentation.
- **Largest coral reef pre-training dataset construction**. We gathered the biggest and most diverse CoralWorld dataset with 2.6 million coral reef images to validate our approach. We have also comprehensively dissected the relationships between the pre-training data and the constructed feature space. Our experimental results demonstrated that the model could learn transferable features from the natural images, alleviating the efforts for collecting domain-specific data. Meanwhile, the diversity, quality, and coverage are also important factors for constructing efficient feature space.
- **Bridging point-based analysis and mask-based analysis**. Our work has bridges the sparse point based analytical approaches and coral reef semantic segmentation in the

Table 4. Quantitative zero-shot comparisons with specialist algorithms on Mosaics UCSD dataset [34]. All the experiments are repeated with **5 times** to obtain the mean values and standard deviations with the same sparse points.

Methods	5 points		10 points		20 points		50 points		100 points	
	mIoU_{std}	mPA_{std}	mIoU_{std}	mPA_{std}	mIoU_{std}	mPA_{std}	mIoU_{std}	mPA_{std}	mIoU_{std}	mPA_{std}
Fast-MSS [56]	$1.40_{0.077}$	$9.90_{0.353}$	$2.47_{0.054}$	$11.90_{0.205}$	$4.17_{0.063}$	$13.49_{0.222}$	$7.47_{0.083}$	$15.29_{0.142}$	$9.68_{0.064}$	$15.99_{0.150}$
PLAS [58]	$12.78_{0.260}$	$14.69_{0.303}$	$17.67_{0.229}$	$21.03_{0.358}$	$23.99_{0.210}$	$29.06_{0.202}$	$36.38_{0.185}$	$43.23_{0.207}$	$46.35_{0.285}$	$53.30_{0.285}$
HIL [59]	$16.69_{0.399}$	$18.99_{0.411}$	$23.69_{0.253}$	$27.78_{0.345}$	$31.89_{0.107}$	$38.12_{0.138}$	$43.20_{0.151}$	$52.21_{0.202}$	$51.10_{0.230}$	$61.29_{0.280}$
FeatUp [35] (DINO)	$15.85_{0.391}$	$17.89_{0.461}$	$23.03_{0.292}$	$26.73_{0.382}$	$31.69_{0.128}$	$37.27_{0.175}$	$43.97_{0.288}$	$51.85_{0.263}$	$52.54_{0.314}$	$61.56_{0.360}$
FeatUp [35] (DINOv2)	$15.90_{0.383}$	$17.99_{0.386}$	$23.10_{0.292}$	$26.94_{0.360}$	$31.99_{0.144}$	$37.85_{0.144}$	$44.50_{0.237}$	$52.74_{0.288}$	$53.28_{0.315}$	$62.60_{0.313}$
CoralSRT-🔥	$19.44_{0.494}$	$21.34_{0.536}$	$27.11_{0.368}$	$30.77_{0.508}$	$35.50_{0.122}$	$41.39_{0.171}$	$46.11_{0.123}$	$55.09_{0.184}$	$52.89_{0.345}$	$63.47_{0.382}$

Table 5. Investigating the features from different DINO models (ViT-B/16) optimized (training from scratch) on different datasets in a zero-shot manner. The test images are from the cleaned testing set of the Mosaics UCSD dataset.

Datasets	CoralSRT	5 points		10 points		20 points		50 points	
		mIoU	mPA	mIoU	mPA	mIoU	mPA	mIoU	mPA
ImageNet-1K [33] (1.28M)	✘	14.58	16.59	22.10	26.32	29.67	35.45	41.00	49.29
	✓	$16.00_{+1.42}$	$19.02_{+2.43}$	$23.67_{+1.57}$	$29.00_{+2.68}$	$30.80_{+1.13}$	$38.14_{+2.69}$	$41.59_{+0.59}$	$51.53_{+2.24}$
BenthicNet [51] (1.45M)	✘	14.32	16.38	21.35	25.60	28.34	34.18	39.16	47.63
	✓	$16.16_{+1.84}$	$18.82_{+2.44}$	$23.52_{+2.17}$	$28.52_{+2.92}$	$30.26_{+1.92}$	$37.05_{+2.87}$	$40.17_{+1.01}$	$49.86_{+2.23}$
Seaview [37] (1.08M)	✘	14.49	16.17	22.37	25.96	30.14	35.12	41.54	48.86
	✓	$16.63_{+2.14}$	$19.15_{+2.98}$	$24.69_{+2.32}$	$29.31_{+3.35}$	$31.98_{+1.84}$	$38.33_{+3.21}$	$42.76_{+1.22}$	$52.04_{+3.18}$
CoralWorld-0.1M (0.1M)	✘	13.99	16.04	20.54	24.93	27.06	33.17	36.90	45.93
	✓	$16.08_{+2.09}$	$18.88_{+2.84}$	$23.51_{+2.97}$	$28.51_{+3.58}$	$30.57_{+3.51}$	$37.37_{+4.2}$	$40.57_{+3.67}$	$50.27_{+4.34}$
CoralWorld-1M (1M)	✘	15.06	16.84	22.89	26.61	30.47	35.64	41.57	49.19
	✓	$17.09_{+2.03}$	$19.35_{+2.51}$	$25.05_{+2.16}$	$29.63_{+3.02}$	$32.31_{+1.84}$	$38.65_{+3.01}$	$42.82_{+1.25}$	$51.92_{+2.73}$
CoralWorld (2.654M)	✘	15.25	16.96	23.23	26.81	30.77	35.74	41.73	49.16
	✓	$17.01_{+1.76}$	$19.29_{+2.33}$	$25.08_{+1.85}$	$29.62_{+2.81}$	$32.43_{+1.66}$	$38.61_{+2.87}$	$42.66_{+0.93}$	$51.85_{+2.69}$

feature space. Our method could re-utilize the existing redundant sparse point annotations to dense semantic segmentation masks. Meanwhile, CoralSRT demonstrates a strong efficiency and flexibility for coral reef analysis, which are invaluable for the coral reef research community. **Promising sparse-to-dense conversion**. Our CoralSRT is the first model-agnostic framework to rectify the features extracted from various foundation models. The sparse-to-dense conversion also closes the performance gap with supervised semantic segmentation algorithms.

- **Without retraining or fine-tuning foundation models**. Our algorithm does not shave straightforward scaling up of the model size, dataset size, and diversity, or length of training. We try to formulate the fundamental problem for coral reef semantic segmentation from the intrinsic properties of corals and the domain requirements of the coral reef analysis community. We combined the advantages of self-supervised pre-training and supervised training for efficient feature rectification.

Modern learning-based segmentation systems require extensive data collection, time-consuming labeling and heavy computational resources. It took months or even years to curate extensive domain-specific data (*e.g.*, CoralWorld with 2.64M images), annotate data with expertise and optimize models from scratch. Such challenges make these systems **not sustainable** for new domain applications. This work proves that strong performance can be achieved without extensive curated domain-specific data, expertise involvement and powerful GPUs for optimizing/fine-tuning models. CoralSRT neither directly utilizes features from existing foundation models (FMs) nor uses SAM 2 to produce final output. Inspired by **amorphous** and **self-repeated** properties of corals, our CoralSRT uses model-generated masks as training guidance to strengthen within-segment affinity of features from any FM. With a simple lightweight $\texttt{Rec}(\cdot)$ network optimized on a single GTX3090, CoralSRT effectively transfers learned knowledge from other datasets and FMs to coral segmentation without introducing any human annotations, outperforming all existing algorithms. The demonstrated **efficiency** is the core novelty and contribution of CoralSRT.

5.2. Broader Impact

Coral reef research is essential for deepening our understanding of the marine ecosystems that are crucial to both marine life and human societies [30]. Additionally, coral reefs are among the most biodiverse ecosystems on the planet, supporting an estimated one to nine million species of marine organisms, including fish, invertebrates, algae, and microorganisms.

We then discuss the potential deployment scenarios of the proposed CoralSRT. The deployment scenarios include

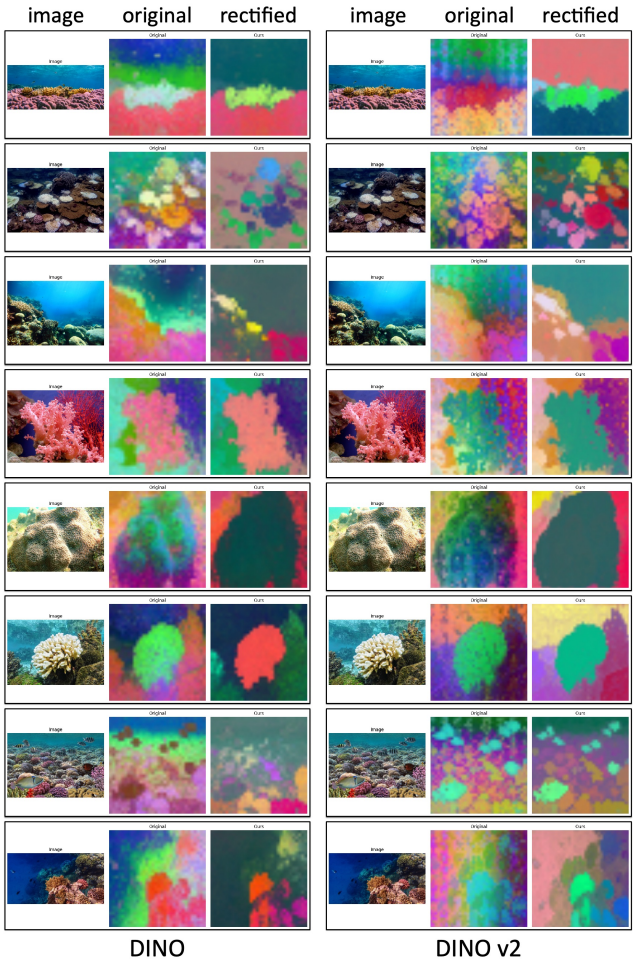

Figure 11. PCA visualization (first 3 components) of both original features and rectified features by our CoralSRT from DINO and DINOv2. CoralSRT demonstrates a strong generalization ability to unseen coral reef images.

1) sparse-to-dense conversion based on existing sparse point annotations for more accurate area cover statistics, favored by **coral biologists** to conduct coral reef surveying analysis; and 2) using these converted masks after sparse-to-dense conversion to optimize semantic segmentation algorithms while preserving user flexibility. The optimized semantic segmentation models can then automatically generate semantic predictions.

5.3. Limitations

Medium-scale testing set. First, the constructed CoralWorld testing set is still relatively small compared with the huge training data. Unlike generating the binary coral reef masks in [74], annotating the semantic coral reef masks from different semantic granularities requires significant domain expertise [34]. Currently, we cannot scale up the data annotation while considering that the coral reef images are from various sites around the world, and they require essential domain expertise for annotation.

Failure to automatically segment coral reefs without sparse points. Our method, converting the annotated sparse points to dense masks, cannot automatically generate separated coral reef masks as CoralSCOP or SAM series. CoralSRT must receive the labeled sparse points to generate the corresponding dense masks. However, we also point out that users could utilize the converted dense semantic segmentation masks from sparse point annotations to optimize semantic segmentation algorithms to alleviate such a limitation.

5.4. Future Works

Semi-supervised coral reef video segmentation. The features rectified by our CoralSRT have the potential to model the semantic correspondences between different coral reef regions and images, as supported by the PCA visualization. We leave the semi-supervised coral reef semantic segmentation with minimum human efforts or domain expertise requirements as our future work.

Multi-modal data. Some coral species can only be reliably distinguished through genetic methods (*e.g.*, DNA barcoding). Visual imagery cannot express key identifying features, hindering accurate coral species identification. However, collecting such paired RGB and DNA barcoding data requires specific domain expertise and specific devices. We leave these as our future works.

References

[1] 100 island challenge. https://100islandchallenge.org/. 5, 6

[2] Youtube. https://www.youtube.com/. 5, 6

[3] Tensorflow - help protect the great barrier reef. https://www.kaggle.com/competitions/tensorflow-great-barrier-reef. 5

[4] inaturalist. https://www.inaturalist.org/. 5, 6

[5] Healthy and bleached corals image classification. https://www.kaggle.com/datasets/vencerlanz09/healthy-and-bleached-corals-image-classification, . 5, 6

[6] Corals classification. https://www.kaggle.com/datasets/aneeshdighe/corals-classification, .

[7] Bhd corals. https://www.kaggle.com/datasets/sonainjamil/bhd-corals, .

[8] Coral reef bleaching. https://universe.roboflow.com/sabrina-eiiwb/coral-reef-bleaching, .

[9] Coral reef bleach detection. https://universe.roboflow.com/coralreef/coral-reef-bleach-detection, .

[10] Coral reef monitoring computer vision project. https://universe.roboflow.com/vesit-w2gzf/coral-reef-monitoring-byok1, .

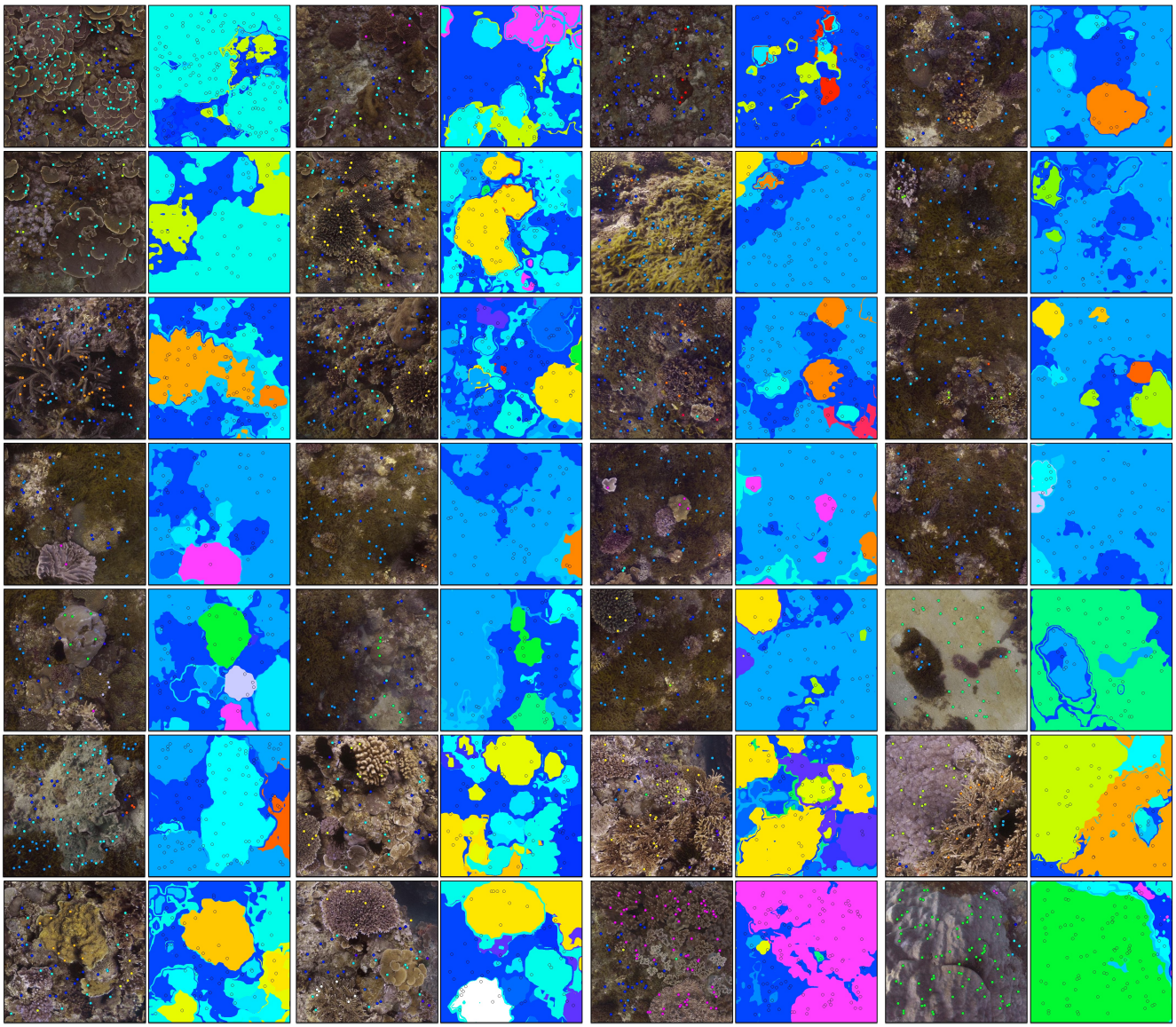

Figure 12. The zero-shot sparse-to-dense conversion results on the Seaview [37] dataset. Each image was paired with 100 sparse point annotations. CoralSRT provided a feasible and reasonable way to re-utilize the already available sparse points to dense masks without introducing any human supervision. The generated dense masks are valuable for the 3D semantic reconstruction of the reef ecosystem and a reliable way for area cover computation without human annotation.

[11] Hard coral and soft coral computer vision project. https://universe.roboflow.com/dc3group14/test2-mt4jr, . 5, 6

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

[13] Shutterstock. https://www.shutterstock.com/. 5, 6

[14] Encyclopedia of life. http://eol.org, 2018. 5, 6

[15] Derya Akkaynak and Tali Treibitz. Sea-thru: A method for removing water from underwater images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1682–1691, 2019. 7

[16] Inigo Alonso, Ana Cambra, Adolfo Munoz, Tali Treibitz, and Ana C Murillo. Coral-segmentation: Training dense labeling models with sparse ground truth. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2874–2882, 2017. 1

[17] 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. 1

[18] 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. 1, 5

[19] Oscar Beijbom, Peter J Edmunds, Chris Roelfsema, Jennifer

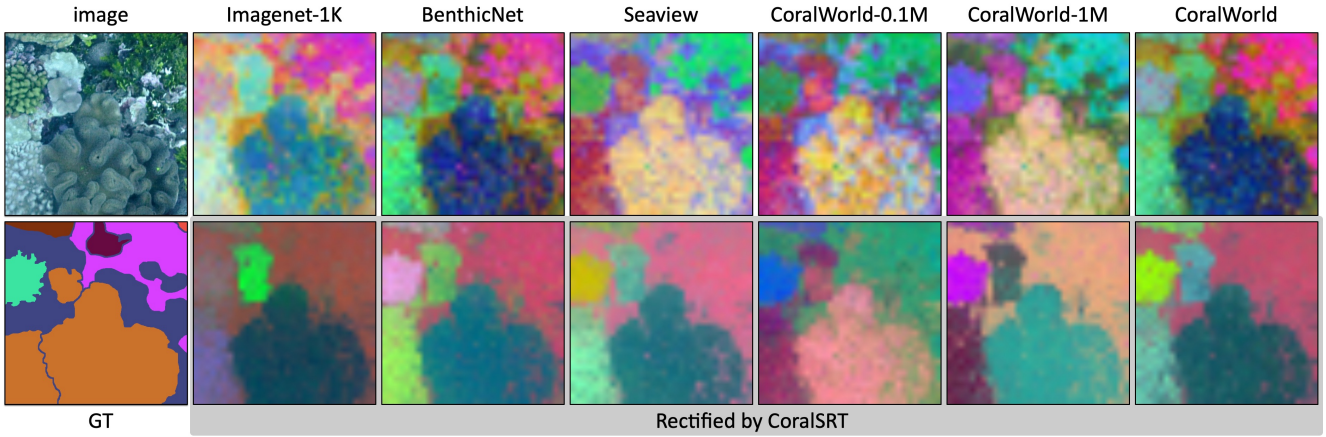

Figure 13. We investigate the DINO features (ViT/B16) on Mosaics UCSD dataset. The PCA visualization (first 3 components) of both original features and rectified features by our CoralSRT from DINO features pre-trained on different datasets. The original image and the corresponding semantic ground truth are also provided for better comparison.

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, 2, 5, 6

[20] 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):23166, 2016. 1

[21] 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. 1

[22] Michael Bewley, Ariell Friedman, Renata Ferrari, Nicole Hill, Renae Hovey, Neville Barrett, Ezequiel M Marzinelli, Oscar Pizarro, Will Figueira, Lisa Meyer, et al. Australian sea-floor survey data, with images and expert annotations. *Scientific data*, 2(1):1–13, 2015. 5, 6

[23] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 2, 3

[24] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 7, 10

[25] 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. 10

[26] 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 (ICCV)*, pages 3693–3702, 2021. 1

[27] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[28] 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, 10

[29] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023. 3

[30] 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. 13

[31] Daniel D Conley and Erin NR Hollander. A non-destructive method to create a time series of surface area for coral using 3d photogrammetry. *Frontiers in Marine Science*, page 974, 2021. 1

[32] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 2

[33] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 13

[34] 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, 6, 8, 9, 12, 13, 14

[35] Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T Freeman. Featup: A model-agnostic framework for features at any resolution. *ICLR*, 2024. 10, 12, 13

[36] Manuel González-Rivero, Oscar Beijbom, Alberto Rodriguez-Ramirez, Tadzio Holtrop, Yeray González-Marrero, Anjani Ganase, Chris Roelfsema, Stuart Phinn, and Ove Hoegh-Guldberg. Scaling up ecological measurements of coral reefs using semi-automated field image collection and analysis. *Remote Sensing*, 8(1):30, 2016. 1

[37] Manuel González-Rivero, Alberto Rodriguez-Ramirez, Oscar Beijbom, Peter Dalton, Emma V Kennedy, Benjamin P Neal, Julie Vercelloni, Pim Bongaerts, Anjani Ganase, Dominic EP Bryant, et al. Seaview survey photo-quadrat and image classification dataset. 2019. 1, 9, 11, 13, 15

[38] 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

[39] Hongyong Han, Wei Wang, Gaowei Zhang, Mingjie Li, and Yi Wang. Cross-domain coral image classification using dual-stream hierarchical neural networks. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, pages 945–952, 2024. 1

[40] Brian M Hopkinson, Andrew C King, Daniel P Owen, Matthew Johnson-Roberson, Matthew H Long, and Suchendra M Bhandarkar. Automated classification of three-dimensional reconstructions of coral reefs using convolutional neural networks. *PloS one*, 15(3):e0230671, 2020. 1

[41] Terry P Hughes, James T Kerry, and Tristan Simpson. Large-scale bleaching of corals on the great barrier reef. *Ecology*, 99(2), 2018. 1

[42] Varun Jampani, Deqing Sun, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Superpixel sampling networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 352–368, 2018. 2

[43] Andrew King, Suchendra M Bhandarkar, and Brian M Hopkinson. Deep learning for semantic segmentation of coral reef images using multi-view information. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–10, 2019. 1, 2

[44] 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, 7

[45] Nancy Knowlton, Emily Corcoran, Thomas Felis, Jasper de Goeij, Andréa Grottoli, Simon Harding, Joan Kleypas, Anderson Mayfield, Margaret Miller, David Obura, et al. Rebuilding coral reefs: a decadal grand challenge. 2021. 1

[46] 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. 1

[47] Daniel Langenkämper, Martin Zurowietz, Timm Schoening, and Tim W Nattkemper. Biigle 2.0-browsing and annotating large marine image collections. *Frontiers in Marine Science*, 4:83, 2017. 1

[48] 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. 1

[49] 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. 2, 10

[50] Jiajun Liu, Brano Kusy, Ross Marchant, Brendan Do, Torsten Merz, Joey Crosswell, Andy Steven, Nic Heaney, Karl von Richter, Lachlan Tychsen-Smith, et al. The csiro crown-of-thorn starfish detection dataset. *arXiv preprint arXiv:2111.14311*, 2021. 6

[51] Scott C Lowe, Benjamin Misiuk, Isaac Xu, Shakhboz Abdulazizov, Amit R Baroi, Alex C Bastos, Merlin Best, Vicki Ferrini, Ariell Friedman, Deborah Hart, et al. Benthicnet: A global compilation of seafloor images for deep learning applications. *arXiv preprint arXiv:2405.05241*, 2024. 9, 13

[52] 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. 1

[53] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[54] G Pavoni, M Corsini, M Callieri, M Palma, and R Scopigno. Semantic segmentation of benthic communities from ortho-mosaic maps. In *Underwater 3D Recording and Modelling*, pages 151–158. Copernicus GmbH, 2019. 1

[55] Jordan Pierce, Mark J Butler IV, Yuri Rzhanov, Kim Lowell, and Jennifer A Dijkstra. Classifying 3-d models of coral reefs using structure-from-motion and multi-view semantic segmentation. *Frontiers in Marine Science*, page 1623, 2021. 1

[56] Jordan P Pierce, Yuri Rzhanov, Kim Lowell, and Jennifer A Dijkstra. Reducing annotation times: Semantic segmentation of coral reef survey images. In *Oceans*, pages 1–9. IEEE, 2020. 1, 13

[57] 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 *ICML*, 2021. 3

[58] 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*, 7(3):8291–8298, 2022. 1, 2, 13

[59] Scarlett Raine, Ross Marchant, Brano Kusy, Frederic Maire, Niko Sunderhauf, and Tobias Fischer. Human-in-the-loop

segmentation of multi-species coral imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2723–2732, 2024. 1, 2, 10, 13

[60] Ahmad Rafiuddin Rashid and Arjun Chennu. A trillion coral reef colors: Deeply annotated underwater hyperspectral images for automated classification and habitat mapping. *Data*, 5(1):19, 2020. 5, 6, 9

[61] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv:2408.00714*, 2024. 2, 3

[62] Tiny Remmers, Alana Grech, Chris Roelfsema, Sophie Gordon, Marine Lechene, and Renata Ferrari. Close-range underwater photogrammetry for coral reef ecology: a systematic literature review. *Coral Reefs*, 43(1):35–52, 2024. 1

[63] Hugh Runyan, Vid Petrovic, Clinton B Edwards, Nicole Pedersen, Esmeralda Alcantar, Falko Kuester, and Stuart A Sandin. Automated 2d, 2.5 d, and 3d segmentation of coral reef pointclouds and orthoprojections. *Frontiers in Robotics and AI*, 9:884317, 2022. 1, 2

[64] Jonathan Sauder, Guilhem Banc-Prandi, Anders Meibom, and Devis Tuia. Scalable semantic 3d mapping of coral reefs with deep learning. *Methods in Ecology and Evolution*, 15(5): 916–934, 2024. 1

[65] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19412–19424, 2024. 3

[66] Carden C Wallace. *Staghorn corals of the world: a revision of the coral genus Acropora (Scleractinia; Astrocoeniina; Acroporidae) worldwide, with emphasis on morphology, phylogeny and biogeography*. CSIRO publishing, 1999. 4

[67] 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, 10

[68] Yaofeng Xie, Lingwei Kong, Kai Chen, Ziqiang Zheng, Xiao Yu, Zhibin Yu, and Bing Zheng. Uveb: A large-scale benchmark and baseline towards real-world underwater video enhancement. In *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 7

[69] Fengting Yang, Qian Sun, Hailin Jin, and Zihan Zhou. Superpixel segmentation with fully convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13964–13973, 2020. 2

[70] Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas J. Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Dvt: Denoising vision transformers. 2024. 2, 9, 10, 12

[71] Hanqi Zhang, Ming Li, Jiageng Zhong, and Jiangying Qin. Cnet: A novel seabed coral reef image segmentation approach based on deep learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 767–775, 2024. 2

[72] Ziqiang Zheng, Xie Yaofeng, Liang Haixin, Yu Zhibin, and Sai-Kit Yeung. Coralvos: Dataset and benchmark for coral video segmentation. *arXiv:2310.01946*, 2023. 2

[73] Ziqiang Zheng, Yiwei Chen, Huimin Zeng, Tuan-Anh Vu, Binh-Son Hua, and Sai-Kit Yeung. Marineinst: A foundation model for marine image analysis with instance visual description. *ECCV*, 2024. 2, 10

[74] Ziqiang Zheng, Haixin Liang, Binh-Son Hua, Yue Him Wong, Put ANG Jr, Apple Pui Yi CHUI, and Sai-Kit Yeung. CoralSCOP: Segment any COral image on this planet. In *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 5, 6, 9, 12, 14

[75] 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. 1

[76] Jiageng Zhong, Ming Li, Hanqi Zhang, and Jiangying Qin. Fine-grained 3d modeling and semantic mapping of coral reefs using photogrammetric computer vision and machine learning. *Sensors*, 23(15):6753, 2023. 1

[77] Zheng Ziqiang, Liang Haixin, Hei Wut Fong, Him Wong Yue, Pui Yi CHUI Apple, and Yeung Sai-Kit. Hkcoral: Benchmark for dense coral growth form segmentation in the wild. In *IEEE Journal of Oceanic Engineering (JOE)*, 2024. 1, 2, 8, 9