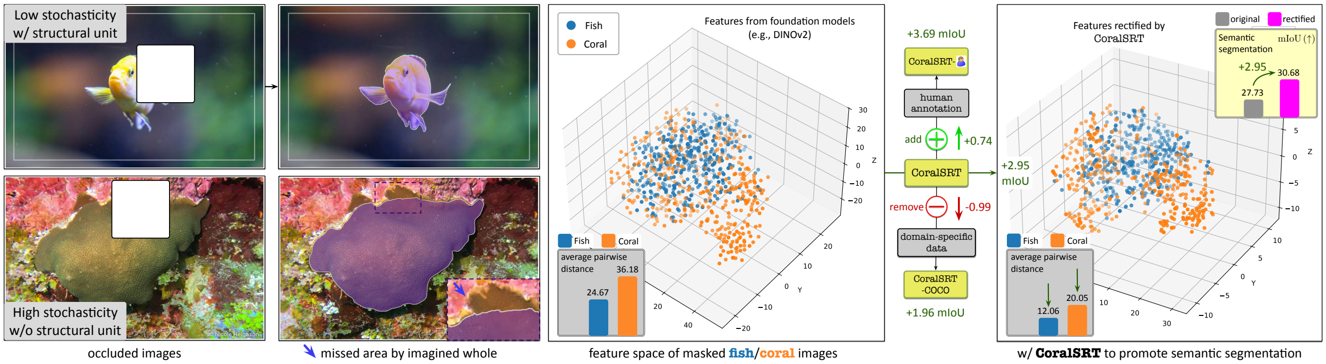
CoralSRT: Revisiting Coral Reef Semantic Segmentation by Feature Rectification via Self-supervised Guidance

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



occluded images

missed area by imagined whole

feature space of masked fish/coral images

w/ **CoralSRT** to promote semantic segmentation

Figure 1. Corals can grow in diverse shapes, textures, and regions, thus leading to high physical and appearance stochasticity. It is challenging to acquire visually consistent knowledge for segmenting corals, in contrast to segmenting objects (*e.g.*, fish). We measure the feature distribution of 400 masked fish and coral images extracted from foundation models (FMs), and found that the average pairwise distance among coral samples is higher than that of fish. We propose **CoralSRT**, an add-on self-supervised feature rectification module, to reduce the stochasticity of coral features. Our method requires no human annotations, retraining/fine-tuning FMs, or even domain-specific data. The key insight is to incorporate *self-repeated*, *asymmetric*, and *amorphous* properties of corals to strengthen within-segment affinity, leading to more efficient label propagation in feature space and producing significant semantic segmentation performance gains.

Abstract

We investigate coral reef semantic segmentation, in which multifaceted factors, like genes, environmental changes, and internal interactions, can lead to highly unpredictable growth patterns. Existing segmentation approaches in both computer vision and coral reef communities have failed to incorporate the intrinsic properties of corals, specifically their self-repeated, asymmetric, and amorphous distribution of elements, into model design. We propose ***CoralSRT****, a feature rectification module via self-supervised guidance, to reduce the stochasticity of coral features extracted by pretrained foundation models (FMs), as demonstrated in Fig. 1. Our insight is that while different corals are highly dissimilar, individual corals within the same growth exhibit strong self-affinity. Using a superset of features from FMs learned by various pretext tasks, we extract a pattern related to the intrinsic properties of each coral to strengthen within-segment affinity, aligning with centrality. We investigate features from FMs that were optimized by various pretext tasks on significantly large-scale unlabeled or labeled data, which already contain rich information for modeling both within-segment and cross-segment affinities, enabling the adaptation of FMs for coral segmentation. CoralSRT can rectify features from FMs to more efficient features for label propagation and lead to further significant semantic segmentation performance gains, all without requiring additional human supervision, retraining/finetuning FMs or even domain-specific data. These advantages help reduce human effort and the need for domain expertise in data collection and labeling. Our method is easy to implement, and also task- and model-agnostic. CoralSRT bridges the self-supervised pre-training and supervised training in the feature space, also offering insights for segmenting elements/stuffs* (e.g., *grass, plants, cells, and biofoulings).*

1. Introduction

Coral reefs [17, 24, 27, 45] are among the most diverse and valuable ecosystems on earth, creating habitats that harbor an estimated 32% of all named marine species. Performing large-scale coral reef monitoring with minimal human annotations is essentially valuable for ecological surveying. Coral reef semantic segmentation (**CRSS** for short) is widely favored and urgently required to support coverage computation [7], coral distribution [21], semantic 3D reconstruction [18, 30], and local reef research [44].

Existing CRSS works [7, 12, 35, 54, 60], that are mainly data-driven or based on Superpixels [3, 40], fail to explain the core challenges of coral segmentation, and essential differences between segmenting objects (*e.g.*, car, human, fish, *i.e.*) and corals. In this work, we first explore the characteristics of corals and how they grow, influenced by various multifaceted factors. The shapes, appearance and distribution of corals are inherently probabilistic, characterized by stochasticity and unpredictability, which can arise from a variety of sources, such as gene variance [48], environmental influences leading to dead or bleaching events [32], and biological interactions (*e.g.*, competition for space, light, and resources) causing changes of growth forms [60].

General object segmentation, while having to deal with complex object shapes, is a predictable task: we can imagine the whole fish even if it's partially occluded. In contrast, corals can grow in almost any shape, boundary, or region. There is no general knowledge that predicts how corals grow when some areas are occluded, as shown in Fig. 1. Segmenting coral reefs is more challenging than segmenting a fish because prior structural knowledge can more readily define a fish. Coral reef segmentation is not an isolated problem; similar issues arise in domains exhibiting an amorphous distribution of elements or substances [9]. These include detecting cancer cells [15], segmenting plants or grasses [39], and identifying biofoulings [29].

We revisit the CRSS task from the basic definition of segmentation, which is to group imagery into regions (also regarded as **segments**) that are homogeneous or semantic-agnostic according to some implicit criteria such as *color*, *appearance*, *shape*, *implicit semantics*, or *texture*, that can be learned and encapsulated into fixed prior knowledge.

Instead, we argue that coral segmentation is more about on-the-fly adapting to the growth pattern of a particular coral: find a basis feature to model **within-segment affinity**, grouping homogeneous pixels into the same segment. This on-the-fly adaptation also extends to reasoning **cross-segment affinity** between segments, resulting in the semantic segmentation outputs.

We simultaneously model self-evolving features within-segment and cross-segment affinities based on recent foundation models [10, 26, 33, 38, 43] (**FMs**). Since FMs are optimized by various pretext tasks on a significant scale of training data in a self-supervised [13, 14, 50, 59] or supervised manner [26, 38], FMs offer efficient tools for domain researchers, allowing them to avoid collecting huge data and optimizing models from scratch. In a prior work, building on the promptable training, CoralSCOP [58] was devised with a parallel semantic branch as the first CRSS foundation model. In the context of coral reef analysis, can we directly utilize these existing FMs for CRSS without introducing any domain expertise or collecting huge coral reef data?

We propose **CoralSRT** (**Coral** **S**elf-supervised **R**ectification **T**raining), a *task-* and *model*-agnostic method, to adapt existing FMs for CRSS without any human annotations, re-training/finetuning FMs [47, 58] or even without coral reef images. Features from FMs already contain rich information, and FMs can produce efficient self-supervised guidance [26, 43], enabling the self-evolution of FMs. We incorporate intrinsic *self-repeated* and *amorphous* properties of corals to devise a self-supervised feature rectifier $\texttt{Rec}(\cdot)$, which was optimized by model-generated guidance, and can be strengthened by human supervision. The significant performance gains were achieved by forcing features within the semantic-agnostic segment to approach the centrality (*e.g.*, mean or median values) of the whole segment to enhance within-segment affinity.

Our CoralSRT provides the following benefits:

1) bridging self-supervised pre-training for constructing a general feature space and self-evolving guidance for an instance-specific segmentation, along with a self-adaptively constructed feature space;

2) obtaining shared commonsense knowledge for a domain with high stochasticity rather than overfitting to pre-defined specific semantics;

3) adapting existing FMs to CRSS without introducing any human annotations or finetuning FMs, demonstrating the potential of FMs for domain research without collecting data and model optimization from scratch;

4) re-utilizing readily available sparse point annotations for obtaining dense semantic segmentation outputs while possessing strong flexibility to satisfy reef research requirements. The main contributions of this work are summarized:

- We have revisited CRSS, regarding the segment as a basis to model within-segment and cross-segment affinities.
- We demonstrate that performing label propagation in the feature space enhanced by CoralSRT is more effective for sparse-to-dense conversion than promptable segmentation models due to intrinsic property of corals.
- Our method offers a novel perspective to model coral reefs and promote the semantic understanding performance of FMs with no additional labels.

We hope our research can inspire the further development of the coral reef community and other similar research fields.

2. Related Work

2.1. Coral Reef Analysis

Corals are organisms that create the building blocks of coral reef, which is a large, underwater structure made up of various coral colonies [49]. Early coral reef analysis was limited to sparse point based analysis (*e.g.*, CPCe [28]), identifying the sampled sparse points and computing coral statistics based on identified sparse points. Further works (*e.g.*, Coral-

Net [7, 12] and ReefCloud [1]) devised automatic image classification algorithms to classify fixed size image regions localized by sparse points to pre-defined semantic categories. Hence, typically less than one thousand [23] of total image pixels are actually analyzed with annotations, potentially resulting in misleading coverage statistics. Meanwhile, sparse point annotations cannot support advanced coral analysis (*e.g.*, boundary delineation [58], 3D semantic reconstruction [46], coral rugosity computation [44] and volume estimation [56]). To address these issues, coral reef semantic segmentation [25, 34, 60] aims to perform dense pixel analysis. There are mainly two lines of works: 1) **sparse-to-dense conversion** [2–4, 6, 36, 40, 42] for label propagation based on sparse point annotations and 2) **data-driven CRSS**, optimizing models with full supervision. CoralSeg [4] and Fast-MSS [36] propagated the sparse point labels based on the Superpixels [8]. For data-driven algorithms, various coral reef datasets [21, 25, 46, 58] and benchmarks [56] with predefined semantic label set have been proposed to boost the coral reef segmentation performance. CoralSCOP [58] made the first attempt to build a coral reef foundation model, with the parallel semantic branch to enable discrimination of the coral reefs from the background and further semantic classification. However, CoralSCOP is purely data-driven and it does not consider the intrinsic properties of coral reefs.

2.2. Foundation Models

Foundation models (*e.g.*, CLIP [38], DINO series [10, 19, 33], SAM series [26, 43]) have been widely adopted for visual understanding. DINO [10] has shown an emerging semantic understanding ability through knowledge distillation in a self-training pipeline. DINOv2 [33] revisited existing discriminative self-supervised methods, proposing the learning of features at both the image and patch levels, while also scaling up pre-training data. SAM [26] optimized by vast and diverse training data with full supervision, has demonstrated a strong ability to segment visual elements with precise masks in a semantic-agnostic manner. Receiving various kinds of prompts (*e.g*, point, box, and mask) from the user, SAM could yield the required mask through interactive labeling and iterative refinement. SAM 2 [43] extended SAM to the video domain by adding temporal consistency. Considering there is a huge scale of sparse point annotations available in the coral reef communities [23, 45], it is intuitive to produce dense coral reef masks by inferring SAM series or CoralSCOP with these sparse points as point prompts. However, we investigate that such promptable segmentation models are less effective for modeling coral reefs. Instead, we propose to perform label propagation in feature space of FMs to achieve CRSS.

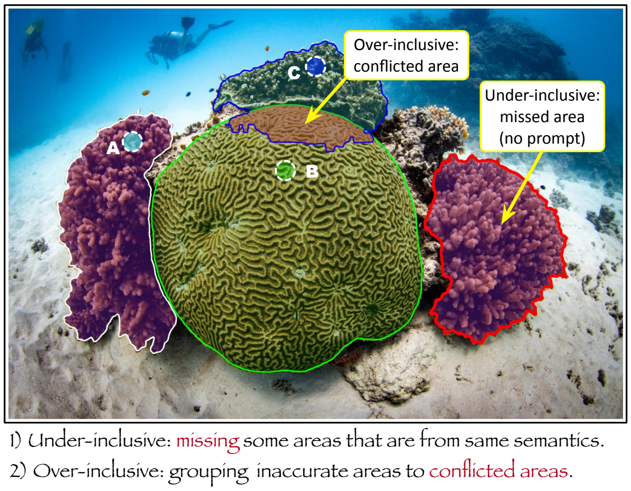

Over-inclusive: conflicted area
Under-inclusive: missed area (no prompt)
A
B
C
1) Under-inclusive: missing some areas that are from same semantics.
2) Over-inclusive: grouping inaccurate areas to conflicted areas.

Figure 2. Promptable segmentation models (*e.g.*, SAM and CoralSCOP) leads to *under-inclusive* and *over-inclusive* outputs. The mask with red edge is for illustration, not model-generated.

2.3. Limitations of Existing Approaches

There remain several challenges in existing CRSS approaches. Propagating sparse points to dense masks based on Superpixels suffers from complex visual contents because Superpixels are typically generated to group pixels into visually meaningful segments based on visual features, without capturing higher-level semantics. Close-set semantic segmentation algorithms [11, 51, 54, 60] based on full supervision cannot generalize to unseen semantic categories and have a weak zero-shot ability, heavily limiting the practicality and violating the essential discovering purpose of reef research [37]. Besides, it requires retraining the models for local requirements, and it also requires collecting redundant coral reef semantic masks for retraining/fine-tuning, which involves significant human effort.

There are also barriers in using promptable-FMs to solve CRSS. We demonstrate two intrinsic limitations of promptable segmentation models (SAM series and CoralSCOP) in Fig. 2 for CRSS since corals do **not** have a visually consistent structural “unit” or “instance” to separate different masks. The semantic ambiguity and inconsistent annotations between masks lead to **under-inclusive** outputs: the generated masks are not complete or accurate in revealing the whole distribution of the corals based on given point prompts; and **over-inclusive** outputs: masks contain some unnecessary contents, *e.g.*, grouping the non-coral areas into masks and clustering different corals into the same mask, leading to the conflicted area. Instead, we perform label propagation in the feature space via feature clustering, leveraging the advantage of global visual understanding.

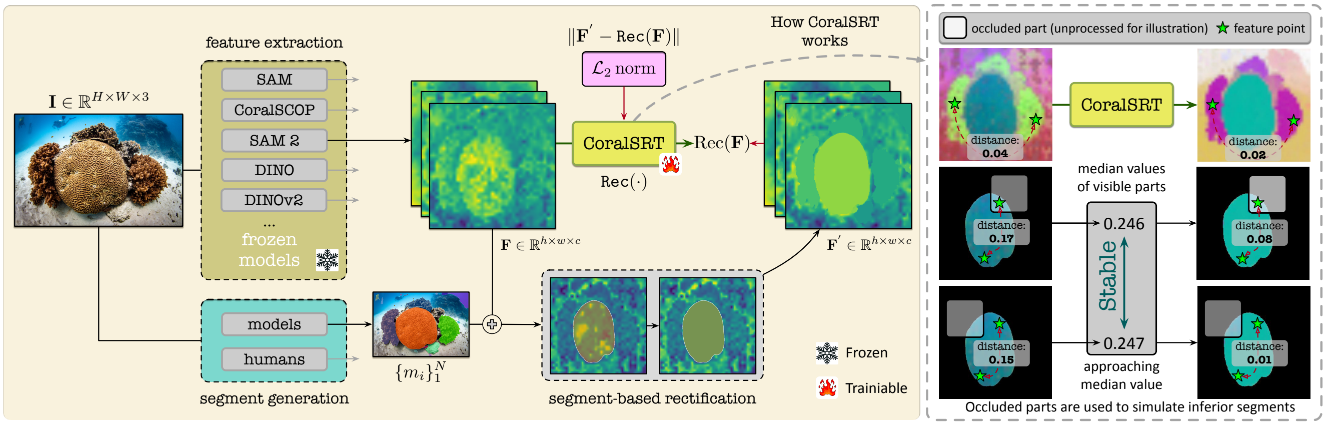

feature extraction
$\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$
SAM
CoralSCOP
SAM 2
DINO
DINOv2
...
frozen models
models
humans
segment generation
$\{m_i\}_1^N$
$\mathbf{F} \in \mathbb{R}^{h\times w\times c}$
$\|\mathbf{F}^{'} - \mathtt{Rec}(\mathbf{F})\|$
$\mathcal{L}_2$ norm
CoralSRT
Rec(·)
Rec(F)
How CoralSRT works
$\mathbf{F}^{'} \in \mathbb{R}^{h\times w\times c}$
segment-based rectification
Frozen
Trainiable
occluded part (unprocessed for illustration)
feature point
distance: 0.04
CoralSRT
distance: 0.02
median values of visible parts
distance: 0.17
0.246
Stable
0.247
distance: 0.08
distance: 0.15
distance: 0.01
approaching median value
Occluded parts are used to simulate inferior segments

Figure 3. Framework overview of proposed CoralSRT to rectify features of frozen FMs based on model-generated mask guidance or human annotations. We force features within each semantic-agnostic segment to approach its centrality to reduce the stochasticity of coral features, leading to more efficient features for label propagation in the feature space. On the right hand side, we demonstrate `Rec(·)` is learning high-dimensional features inside the segment via the centrality (*e.g.*, median value), which is stable between different inferior segments due to the intrinsic self-repeated and amorphous properties of corals.

3. Method

3.1. Segment Affinities

We define a segment as a connected region of pixels sharing the same implicit semantics. We model within-segment and cross-segment affinities to reduce intra-segment variance and enhance inter-segment differences.

Within-segment affinity. We aim to strengthen the within-segment affinity based on model-generated guidance, which was masks generated by FMs [26, 43, 58]. Considering that semantic-agnostic training [26, 43, 58] was essentially devised for grouping pixels into homogeneous regions with implicit semantics, we utilize the powerful FMs like SAM 2 [43] to automatically generate scalable dense masks/segments. Unlike existing approaches that adopt pixel-level supervision from pre-defined label sets [11, 16, 51] or semantic-agnostic training [26, 43], we propose to design a self-supervised **segment-based rectification** in the **feature space** of various FMs to strengthen within-segment affinity (details in Sec. 3.2). Our method forces the features within the same semantic-agnostic segment to approach the centrality of the whole segment, incorporating *self-repeated* and *amorphous* properties of corals. Furthermore, the mask guidance could also be from humans to embed user preferences. Our method not only leverages the compact feature space of FMs optimized with large-scale pre-training data but also makes more efficient use of the model-generated guidance in a semantic-agnostic manner, without introducing any human annotations.

Cross-segment affinity. In this work, we do not explicitly model cross-segment affinity based on human supervision (*e.g.*, designing discrete one-hot labels [21, 60], combining hierarchical taxonomy [47] or text annotations [38, 57, 58]). Instead, we emphasize the importance of large-scale pre-training for modeling cross-segment affinity due to general **semantic uncertainty** and **biology-specific** features (*e.g.*, reticular pattern of corals and inefficient visual expression, discussed in Supplement). The semantic uncertainty arises from the fact that the semantic correspondences can be defined from various and complicated dimensions (*e.g.*, using general color, texture, shape, geometry, and domain-specific species, genus, and growth form metrics). The implicit semantic correspondences are also highly subject to the whole data distribution, like larger and diverse training data usually contain more comprehensive semantics. We propose to conduct label propagation [52, 55] via feature clustering in the feature space, encapsulate the model with strong flexibility to various requirements and generalization ability to unseen data. The users can also design different label sets for their local data without pre-defining fixed semantic categories, satisfying the intrinsic discovery purpose of reef research.

3.2. Self-supervised Rectification

Given image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and **frozen** feature extractor $f(\cdot)$ as illustrated in Fig. 3, where $f(\cdot)$ could be from any FMs that were optimized by various pretext tasks, we obtain features $\mathbf{F} \in \mathbb{R}^{h\times w\times c}$ at any feature size. h, w and c indicate height, width, and channel of $\mathbf{F}$. $\mathbf{I}$ is paired with a set of dense masks $\{m_i\}_1^N$ without semantics. $\{m_i\}_1^N$ are **generated by FMs** (*e.g.*, SAM series [26, 43] or domain-specific CoralSCOP [58]) or from **human annotation**. We construct target feature $\mathbf{F}^{'}$ with same shape as $\mathbf{F}$ as the self-supervised guidance:

$$\mathbf{F}^{'} = \mathbf{F} \odot m, \quad \text{for } m \in \{m_i\}_1^N, \tag{1}$$

where $\odot$ is a segment-based rectification operation (computing the *median* or *mean* value of features of the segment and assigning that value to all features within the segment) over the channels. Then we optimize our $\texttt{Rec}(\cdot)$ based on the standard squared l_2 norm:

$$\mathcal{L} = \|\mathbf{F}^{'} - \texttt{Rec}(\mathbf{F})\|. \tag{2}$$

$\texttt{Rec}(\cdot)$ is not learning to mimic segmentation masks from FMs, instead, it is learning global features inside of the mask via the centrality. We then conduct label propagation to perform sparse-to-dense conversion based on rectified features by $\texttt{Rec}(\cdot)$. Our method is easy to implement and is both *task-* and *model*-agnostic.

4. Experiments

4.1. Datasets and Comparisons

Datasets. We have curated the largest coral reef dataset to date, named **CoralWorld**, which contains **2.64** million images with no labels from across the globe. We use the CoralWorld dataset to explore the relationship between pre-training data and the constructed feature space. We adopt **CoralMask** [58] dataset for optimizing CoralSRT. We use SAM 2 [43] to generate dense masks for the CoralMask dataset, leveraging its fast inference time. Besides, it was not specifically optimized by coral reef masks, also allowing us to better dissect our method. We also used other coral reef datasets such as HKCoral [60] and Mosaics UCSD [21] for training and evaluation.

Construction of testing set. To measure the ability of various methods to generate dense semantic masks from labeled sparse points, we construct our testing set considering both *diversity* and *coverage*. Different from the existing CoralMask dataset, which only provides the binary coral reef mask annotation, the constructed testing set contains various semantic label sets according to the local requirements. We collect testing images from **10** different countries or sites. The images are from public websites/datasets or local coral reef biologists. Each subset contains at least 100 images (except *Deep Sea* set with 77 images) and the label set is site-specific. Our testing set, consisting of **1,109** images in total, is the first to include coral reef images captured from multiple countries/sites, accompanied by dense semantic mask annotations.

Evaluation metrics. We adopt mIoU and mPA as the main evaluation metrics. Different from the existing evaluation setting to compute class-level mIoU and mPA scores, we compute **image-level** mIoU and mPA scores (average IoU/PA scores of all the semantic classes within each image and we remove non-defined and background classes for evaluation) considering two factors: 1) the semantic class distribution of coral reefs is highly imbalanced, *e.g.*, some minor coral categories only appear only once; 2) it is tricky to fairly compute class-level mIoU and mPA scores for prompt-based algorithms since there will be conflicted areas between generated masks from two separated sparse points with different semantic classes. Considering these two factors, image-level mIoU and mPA are more robust, and we report the average scores of all the testing images. More details are in the supplementary file.

Comparisons. We compare our CoralSRT with existing algorithms, including the foundation models (general-purpose SAM [26], SAM 2 [43], DINO [10], DINOv2 [33], DINOv2 register [19], and coral reef foundation model CoralSCOP [58]) and specialist algorithms (CRSS and sparse-to-dense algorithms). We have also included the feature denoising method DVT [53] and feature upsampling FeatUp [22] for comparison. We adopt DeeplabV3 [11], SegFormer [51] and Mask2Former [16] for CRSS. For the sparse-to-dense algorithms, we compare Superpixel-based algorithm Fast-MSS [36] and deep learning based algorithms (PLAS [40]) and HIL [42].

Implementation details. We adopt the same architecture as DVT [53] to transform $\mathbf{F}$ into $\mathbf{F}^{'}$ except we removed the positional embedding. Unlike DVT [53], which uses a two-stage training process where the first stage, involving denoised feature production, takes around 2.5 minutes per image on a GTX 3090, making scalable training challenging with academic resources, our method only requires automatically generated dense masks from SAM 2 (around 43.6 images per minute under the same experimental conditions, making it **109** times faster than DVT for preparing target features). The core code of our method is also easy to implement with few code modifications as described in Sec. 3.2. We differentiate between models optimized with model-generated masks and human-annotated masks using CoralSRT and CoralSRT-👤 respectively. We adopt the 299,557 coral reef masks (*no semantics*) from CoralMask dataset to optimize CoralSRT-👤.

4.2. Sparse-to-dense Conversion

Comparison with Superpixel-based algorithm. We demonstrate that CoralSRT has a much stronger ability to generate accurate and precise dense masks than Fast-MSS [36] as reported in Table 1. As analyzed by [36], usually more than 300 sparse points would lead to reasonable conversion results.

Comparisons with foundation models. We compare FMs from two settings: **Prompt-based** marked with ♠ (SAM [26], SAM 2 [43] and CoralSCOP [58]) and **Feature-based** marked with ♣. In the former prompt-based setting, we use annotated sparse points as point prompts to infer the promptable segmentation models and generate dense masks. In the latter feature-based setting, we apply KNN clustering (K=1) based on features from different models and the provided labeled sparse points to generate dense masks, also specifying the feature size. We present the comparisons using the same sparse points across different settings in Table 1 and the qualitative results in Fig. 4.

From the results, we have the following key observa-

Table 1. The quantitative sparse-to-dense results for various algorithms using different numbers of labeled sparse points: 5, 10, 20, 50, and 100. ♢ - Superpixed-based; ♠ - Prompt-based; ♣ - Feature-based. Best results are in bold.

Methods	Backbone	Feat. size	5 points		10 points		20 points		50 points		100 points	
			mIoU	mPA	mIoU	mPA	mIoU	mPA	mIoU	mPA	mIoU	mPA
Fast-MSS$^\diamondsuit$ [36]	–	–	2.36	22.01	4.18	24.82	7.49	26.19	13.65	27.66	17.96	28.34
SAM$^\spadesuit$ [26]	ViT-H	–	16.34	30.06	21.59	32.50	23.96	31.18	25.38	28.82	26.89	29.75
SAM 2$^\spadesuit$ [43]	Hiera-L	–	15.09	27.24	20.21	30.97	24.96	33.09	27.01	31.47	28.13	33.18
CoralSCOP$^\spadesuit$ [58]	ViT-L	–	18.14	31.78	23.56	33.67	26.16	34.98	28.78	34.67	30.18	36.78
DINO$^\clubsuit$ [10]	ViT-B/16	$32^2 \times 768$	27.51	31.54	35.32	41.37	43.22	50.69	53.65	62.15	60.61	69.86
DINOv2$^\clubsuit$ [33]	ViT-B/14	$37^2 \times 768$	27.73	30.64	35.63	40.53	43.66	49.99	53.85	61.61	60.23	68.87
DINOv2-Reg$^\clubsuit$ [19]	ViT-B/14/R4	$37^2 \times 768$	28.46	31.63	36.80	42.03	44.54	51.05	54.59	62.51	61.13	70.00
SAM$^\clubsuit$ [26]	ViT-H	$64^2 \times 1280$	24.64	29.77	31.69	39.37	39.35	48.83	50.38	61.25	58.27	69.66
CoralSRT (SAM)$^\clubsuit$	ViT-H	$64^2 \times 1280$	30.73	**35.87**	38.38	45.94	46.34	54.92	56.54	66.34	63.14	73.23
SAM 2$^\clubsuit$ [43]	Hiera-L	$64^2 \times 576$	23.59	28.47	30.47	38.17	38.26	47.18	49.95	60.25	58.39	69.21
CoralSRT (SAM 2)$^\clubsuit$	Hiera-L	$64^2 \times 576$	27.45	32.11	35.33	42.55	43.48	51.67	54.22	63.91	61.24	71.71
CoralSCOP$^\clubsuit$ [58]	ViT-L	$64^2 \times 1024$	26.18	30.71	33.56	40.46	41.10	49.71	51.38	62.25	58.97	70.45
CoralSRT (CoralSCOP)$^\clubsuit$	ViT-L	$64^2 \times 1024$	**31.93**	35.19	**40.00**	**45.50**	**48.20**	**54.97**	**58.67**	67.23	65.46	74.32
DVT (DINOv2)$^\clubsuit$ [53]	ViT-B/14	$37^2 \times 768$	30.02	32.98	38.07	43.25	45.83	52.43	55.69	63.67	61.73	70.61
CoralSRT (DINOv2)$^\clubsuit$	ViT-B/14	$37^2 \times 768$	30.68	34.53	38.95	44.86	47.35	54.15	57.46	65.57	63.40	72.07
FeatUp (DINO)$^\clubsuit$ [22]	ViT-S/16	$512^2 \times 384$	28.27	32.08	36.48	42.14	45.01	51.59	55.68	63.51	62.83	71.10
FeatUp (DINOv2)$^\clubsuit$ [22]	ViT-S/14	$518^2 \times 384$	29.40	32.45	38.01	42.97	46.42	52.44	57.11	64.51	64.32	72.31
CoralSRT$_{+\text{FeatUp}}$ (DINO)$^\clubsuit$	ViT-S/16	$224^2 \times 384$	31.24	35.07	39.68	45.18	47.47	54.26	57.14	65.35	63.18	72.17
CoralSRT$_{+\text{FeatUp}}$ (DINOv2)$^\clubsuit$	ViT-S/14	$224^2 \times 384$	31.45	35.12	39.87	45.28	48.05	54.75	58.65	**67.41**	**65.78**	**74.76**

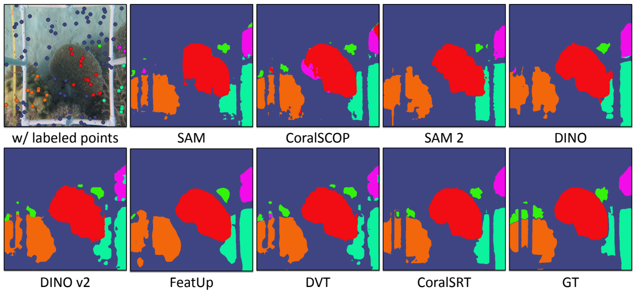


Figure 4. The sparse-to-dense conversion results of various algorithms using 100 randomly sampled labeled sparse points.

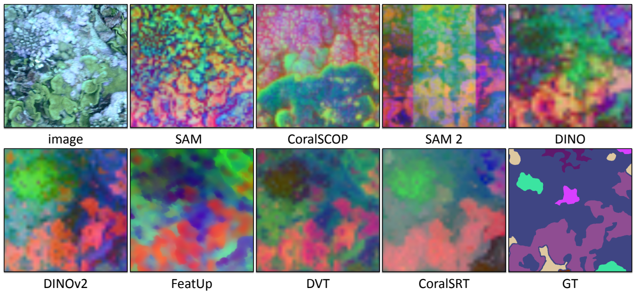


Figure 5. PCA (first 3 components) visualization of features. FeatUp, DVT and CoralSRT are using DINOv2 features.

tions: 1) Performing sparse-to-dense conversion in the feature space demonstrates a stronger ability to generate more accurate coral reef masks than promptable segmentation models due to under-inclusive and over-inclusive mask outputs. We provide more analysis in the supplementary material. 2) The features from the DINO series (DINOv2 with registers performs best among the three) possess more implicit semantic feature representations for label propagation than features from SAM series (even with a larger feature size). We attribute this to that promptable segmentation models over-emphasized local regional information under full supervision. 3) CoralSRT could effectively strengthen within-segment affinity of SAM and SAM 2 in the feature space with significant performance improvements. 4) CoralSCOP optimized by dense coral masks has more efficient features for label propagation than SAM series and demonstrated a higher upper bound when equipped with CoralSRT. 5) The high-resolution features (*e.g.*, 518 × 518) of FeatUp yield accurate dense masks and CoralSRT could beat FeatUp even with a much smaller feature size of 37 × 37 when using few sparse points (less than 50 points). 6) Combined with FeatUp for feature upsampling, CoralSRT achieves remarkable performance. To better illustrate the effectiveness of features from various FMs, we provide the PCA visualization of extracted features in Fig. 5.

Comparison with sparse-to-dense algorithms in a zero-shot manner. We then conduct zero-shot experiments on the testing set of the Mosaics UCSD dataset [21] with the comparison of specialist algorithms: Fast-MSS [36], PLAS [40], and HIL [42] (based on DVT [53]) using different numbers of sparse points in Table 2. All algorithms utilize the same sparse points. Our CoralSRT outperforms existing algorithms in generating accurate dense semantic coral reef masks, despite not being optimized on the Mosaics UCSD dataset. Additionally, we have considered the impact of sparse point selection on segmentation results (discussed in our supplementary file).

Table 2. Quantitative zero-shot comparisons with specialist algorithms on Mosaics UCSD dataset [21].

Methods	5 points		10 points		20 points		50 points	
	mIoU	mPA	mIoU	mPA	mIoU	mPA	mIoU	mPA
Fast-MSS [36]	4.69	**24.47**	8.61	**33.67**	15.09	**41.13**	29.46	51.66
PLAS [40]	12.81	14.58	18.19	21.64	24.15	29.21	35.76	42.28
HIL [42]	16.74	18.94	24.28	28.45	31.78	37.87	42.52	50.91
FeatUp (DINO) [22]	15.78	17.81	23.62	27.60	31.62	36.92	43.22	50.80
FeatUp (DINOv2) [22]	15.93	18.02	23.65	27.66	31.72	37.34	43.73	51.40
CoralSRT (DINOv2)	**18.15**	20.98	**26.45**	30.67	**33.27**	40.01	**44.66**	**53.03**

4.3. Semantic Coral Reef Segmentation

In this section, following the setting of [4, 40], we provide **comparisons with supervised semantic segmentation algorithms**, where three widely used algorithms were conducted on HKCoral dataset [60]. We follow the official training/testing split and compare the algorithms under two settings: 1) **Pseudo dense masks**: we optimize three baseline models based on training images with generated dense masks (regarded as pseudo labels). The pseudo labels are from the sparse-to-dense conversion based on CoralSRT using different numbers of sparse points randomly sampled from the ground truth of training images. After optimization, we infer the optimized models by pseudo labels with the testing images and report the results in Table 3, where we could indirectly measure the quality of converted dense masks since high quality pseudo labels result in strong models. For better comparison, we report the results of "Oracle", which utilizes the ground truth of the training images for optimizing the baseline models. 2) **Inference only**. We consider the setting without any training/fine-tuning. We directly infer CoralSRT (denoted as CoralSRT‡) with testing images using the same numbers of sparse point annotations sampled from the ground truth of the testing images. By comparing these two settings, CoralSRT based on label propagation in the feature space of FMs is competitive or even better than supervised semantic segmentation algorithms (*e.g.*, DeeplabV3 and Mask2Former) optimized by ground truths due to larger network compactness and significant training data of FMs. The pseudo labels generated by CoralSRT exhibit high fidelity compared with ground truth masks, showing the promising potential of converting already available redundant sparse point annotations to dense masks. Finally, the competitive results of CoralSRT‡ with few point annotations also demonstrate that CoralSRT with strong flexibility can better satisfy various local reef analysis requirements without collecting dense masks and optimizing models from scratch. Above experiments are based on rectified DINOv2 features by CoralSRT.

Table 3. The CRSS performance of various algorithms on testing set of HKCoral dataset [60]. CoralSRT‡ indicates CoralSRT is inferred via a training-free manner. Best viewed in color.

Settings	DeeplabV3 [11]		SegFormer [51]		Mask2Former [16]		CoralSRT‡	
	mIoU	mPA	mIoU	mPA	mIoU	mPA	mIoU	mPA
Oracle	38.88	51.24	77.75	85.43	62.33	70.76	Uncomputable	
Pseudo dense masks from sparse-to-dense conversion via CoralSRT							*Infer only*	
10 points	26.65	44.98	51.65	67.79	19.77	25.74	53.79	62.89
20 points	32.17	55.15	61.33	75.48	34.89	43.59	62.52	71.82
50 points	33.44	49.80	68.27	79.38	53.97	63.08	70.23	79.64
100 points	35.21	50.25	70.41	79.63	60.18	69.28	75.29	83.99

Table 4. Investigating the features from different DINO models (ViT-B/16) optimized on different datasets with data scale specified (millions). Readers are suggested to compare results of 1) using different data scales; 2) with and without CoralSRT and 3) using similar data scales.

Datasets	CoralSRT	5 points		10 points		20 points		50 points	
		mIoU	mPA	mIoU	mPA	mIoU	mPA	mIoU	mPA
ImageNet-1K [20]	✘	27.51	31.54	35.32	41.37	43.22	50.69	53.65	62.15
(1.28M)	✓	30.78	35.40	38.52	45.32	45.96	54.39	55.15	65.03
BenthicNet [31]	✘	26.20	30.05	33.87	39.78	41.53	49.05	52.09	60.89
(1.45M)	✓	30.07	34.81	37.45	44.34	44.91	53.04	54.12	64.07
Seaview [23]	✘	26.71	30.09	34.71	40.06	43.08	49.82	53.96	61.76
(1.08M)	✓	30.83	35.25	38.62	45.21	46.19	54.03	55.53	65.25
CoralWorld-0.1M	✘	25.86	29.82	33.17	39.26	40.33	48.23	50.49	59.85
(0.1M)	✓	29.47	34.16	36.99	43.72	44.28	52.58	53.60	63.39
CoralWorld-1M	✘	27.37	30.93	35.39	41.02	43.66	50.64	54.12	62.18
(1M)	✓	32.03	36.19	39.60	46.06	47.10	54.65	56.28	65.83
CoralWorld	✘	27.89	31.25	36.05	41.38	44.14	50.81	54.43	62.17
(2.64M)	✓	32.28	36.36	40.15	46.13	47.42	54.83	56.41	65.51

4.4. Ablation Studies

First, we aim to answer two longstanding and valuable questions: 1) *Can more efficient features be obtained based on self-supervised training by continuously increasing the scale of training data?* and 2) *Is large-scale, high-quality, domain-specific data essential?* To investigate these, we conduct experiments on ImageNet-1K [20], BenthicNet [31], Seaview [23] and our CoralWorld in Table 4, where we choose DINO [10] for experiments. We construct two subsets of our CoralWorld dataset: **CoralWorld-0.1M** and **CoralWorld-1M** with 0.1 and 1 million randomly sampled images from the whole dataset, respectively, to explore influence of data scale to optimized feature space. We use same experimental setting and train models from scratch, except for ImageNet-1K dataset, in which we employ the officially released model. We perform sparse-to-dense conversion based on features from DINO models optimized on different datasets. We summarize following insights from experimental comparisons: 1) FMs optimized by diverse and large-scale natural images (*e.g.*, ImageNet-1K) have a strong transferability to domain images: better performance than Seaview dataset with pure coral reef images. 2) Solely scaling up pre-training data could only lead to a bit of performance improvement, comparing results of CoralWorld-0.1M, CoralWorld-1M, and CoralWorld. 3) Data diversity and coverage are important for domain research, comparing results of BenthicNet, Seaview, and CoralWorld-1M with similar data scale. 4) CoralSRT achieves significant performance improvement compared with solely scaling up pre-training data or curating high-quality pre-training data. It reveals that incorporating intrinsic properties of corals into model design is more important and demonstrates the effectiveness of self-generated guidance in the feature space. 5) The features rectified by CoralSRT are also subject to the data quality and coverage

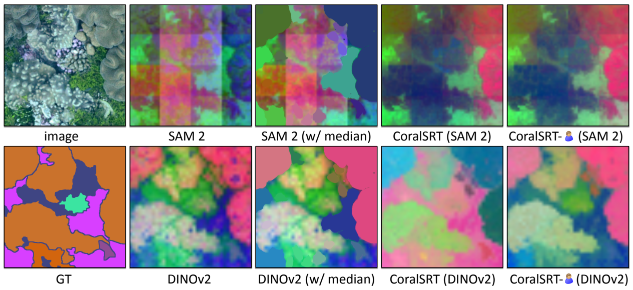


Figure 6. PCA visualization of original, rectified (median value based and without training), CoralSRT and CoralSRT-👤 features.

Table 5. **CoralSRT-👤** and **CoralSRT** denote `Rec(·)` was optimized using **human-annotated** masks and **model-generated** masks from SAM 2, respectively. CoralSRT-COCO denotes `Rec(·)` was optimized by images from the COCO-Stuff dataset.

Methods	5 points		10 points		20 points		50 points	
	mIoU	mPA	mIoU	mPA	mIoU	mPA	mIoU	mPA
SAM 2	23.59	28.47	30.47	38.17	38.26	47.18	49.95	60.25
CoralSRT-COCO$_{(\text{SAM 2})}$	26.66	30.72	34.50	40.83	42.21	49.55	53.06	62.00
CoralSRT$_{(\text{SAM 2})}$	27.45	32.11	35.33	42.55	43.48	51.67	54.22	63.91
CoralSRT-👤$_{(\text{SAM 2})}$	28.22	32.60	36.04	42.92	43.96	51.96	54.45	64.11
DINOv2	27.73	30.64	35.63	40.53	43.66	49.99	53.85	61.61
CoralSRT-COCO$_{(\text{DINOv2})}$	29.68	32.84	38.05	43.19	46.07	52.40	56.20	63.83
CoralSRT$_{(\text{DINOv2})}$	30.68	34.53	38.95	44.86	47.35	54.15	57.46	65.57
CoralSRT👤$_{(\text{DINOv2})}$	31.42	34.74	39.88	45.22	47.95	54.53	57.65	65.82

of data used for pre-training: CoralWorld-1M (w/ CoralSRT) is better than Seaview (w/ CoralSRT).

We then consider an extreme case where we have **no coral reef images**. We randomly sample 40,000 images (similar size as CoralMask for a fair comparison) from the COCO-Stuff dataset to optimize CoralSRT, named CoralSRT-COCO. Meanwhile, we also explore **gap between model-generated supervision and human annotation** for optimizing CoralSRT. We conduct these experiments based on DINOv2 and SAM 2 features. Results are presented in Table 5. By comparing the features of vanilla SAM 2 and CoralSRT$_{(\text{SAM 2})}$, we conclude that it is promising to adapt SAM 2 for CRSS task by strengthening within-segment affinity without introducing any human supervision. The performance improvement of CoralSRT$_{(\text{DINOv2})}$ over DINOv2 also reveals that we can boost the CRSS performance via rectifying DINOv2 features based on the constructed self-supervision from SAM 2. By comparing CoralSRT and CoralSRT-👤, the small performance gap indicates the low reliance of CoralSRT on human annotation, while CoralSRT could achieve promising performance gains. We also notice that CoralSRT-COCO could promote the performance without access to coral reef images. We attribute such performance improvements to shared underlying principles for grouping pixels into segments, such as *geometry*, *repeated texture*, *self-similarity*, and *biological appearance*. Our method could effectively learn such principles and achieve performance gains, even optimized by general images without any human supervision. These promising results demonstrate the values of CoralSRT to reduce human effort and the need for domain expertise in data collection and labeling.

Table 6. Dissecting effectiveness of `Rec(·)` and various rectifying operations. We directly utilize $\mathbf{F}^{'}$ for label propagation under the setting without `Rec(·)` except Vanilla setting (using $\mathbf{F}$).

Settings	`Rec(·)`	5 points		10 points		20 points		50 points	
		mIoU	mPA	mIoU	mPA	mIoU	mPA	mIoU	mPA
Vanilla	✘	27.73	30.64	35.63	40.53	43.66	49.99	53.85	61.61
Mean	✘	27.83	30.11	35.64	39.69	43.02	48.30	53.21	59.44
Median	✘	27.87	30.16	35.98	39.97	43.29	48.48	53.56	59.88
Mean	✓	30.61	34.43	38.89	44.78	47.19	54.06	57.38	65.42
Median	✓	30.68	34.53	38.95	44.86	47.35	54.15	57.46	65.57

Finally, we dissect whether `Rec(·)` of CoralSRT is necessary since we can easily obtain dense masks generated by SAM 2 for feature rectification without any training. We infer SAM 2 with the testing images to generate masks and conduct the segment-based feature rectification to obtain $\mathbf{F}^{'}$. We have also investigated the difference between using *mean* and *median* values as centrality. We conduct experiments on DINOv2 features ("*Vanilla*") and results are reported in Table 6. As reported, it is necessary to optimize `Rec(·)` since `Rec(·)` is learning shared and common knowledge inside redundant masks to reduce the stochasticity via centrality. Only conducting feature rectification with generated dense masks from SAM 2 at test time heavily depends on the fidelity of generated dense masks (potentially over-segmentation and missing coral reef regions) as illustrated in Fig. 6. Such deterministic rectification process without any training cannot help model to learn common knowledge to reduce stochasticity of features for better semantic segmentation.

5. Discussions and Conclusion

Broader impact. One significant contribution of CoralSRT to reef community is its flexibility and scalability to convert redundant sparse point annotations to dense masks without introducing any annotation or retraining/finetuning. Our solution and insight can also be valuable for segmenting stuffs (such as cells, seagrass [39, 41], algae [5], and plants). **Limitations**. Our method, converting the annotated sparse points to dense masks, cannot automatically generate separated coral reef masks as CoralSCOP or SAM series.

Conclusion. In this work, we revisited existing CRSS and promptable segmentation algorithms and found that existing algorithms did not incorporate intrinsic properties of corals into model design. We proposed a simple formulation for CRSS, with the segment as a basis to model both within-segment and cross-segment affinities. We propose CoralSRT to strengthen within-segment affinity and leverage FMs to model cross-segment affinity to preserve a strong flexibility. Our algorithm does not shave straightforward scaling up of dataset size or introducing additional human annotations, while achieving significant performance gains over existing approaches. These findings suggest a promising path towards segmenting "stuffs" and conducting self-evolving of FMs for better semantic understanding performance.

6. Acknowledgment

We thank the valuable discussions and suggestions from Manuel Aranda, David Suggett, Guoxin Cui and Eveline van der Steeg. The work was partially supported by the Marine Conservation Enhancement Fund (MCEF20107 and MCEF22112), an internal grant from HKUST (R9429), HKUST Marine Robotics and Blue Economy Technology Grant. The authors would like to express their sincere gratitude to the "Sustainable Smart Campus as a Living Lab" (SSC) program at HKUST for its invaluable support. This program and staff not only provided essential funding and coordination, but also integrates sustainability into campus operations, which served as the living demonstration of the principles underpinning this research.

References

[1] Reefcloud. https://reefcloud.ai/. 3

[2] Iñigo Alonso and Ana C Murillo. Semantic segmentation from sparse labeling using multi-level superpixels. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5785–5792. IEEE, 2018. 3

[3] 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. 2

[4] 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. 3, 7

[5] Pooja Baweja and Dinabandhu Sahoo. Classification of algae. *The algae world*, pages 31–55, 2015. 8

[6] 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. 3

[7] 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

[8] 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

[9] 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

[10] 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, 3, 5, 6, 7

[11] 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. 3, 4, 5, 7

[12] 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. 2, 3

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607, 2020. 2

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

[15] Xiaowei Chen, Xiaobo Zhou, and Stephen TC Wong. Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy. *IEEE transactions on biomedical engineering*, 53(4):762–766, 2006. 2

[16] 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. 4, 5, 7

[17] 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. 1

[18] 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. 2

[19] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 3, 5, 6

[20] 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. 7

[21] 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, 4, 5, 6, 7

[22] 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. 5, 6, 7

[23] Manuel González-Rivero, Alberto Rodriguez-Ramirez, Oscar Beijbom, Peter Dalton, Emma V Kennedy, Benjamin P Neal, Julie Vercelloni, Pim Bongaerts, Anjani Ganase, Do-

minic EP Bryant, et al. Seaview survey photo-quadrat and image classification dataset. 2019. 3, 7

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

[25] 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. 3

[26] 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, 5, 6

[27] 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

[28] 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

[29] Lutz MK Krause, Emily Manderfeld, Patricia Gnutt, Louisa Vogler, Ann Wassick, Kailey Richard, Marco Rudolph, Kelli Z Hunsucker, Geoffrey W Swain, Bodo Rosenhahn, et al. Semantic segmentation for fully automated macrofouling analysis on coatings after field exposure. *Biofouling*, 39 (1):64–79, 2023. 2

[30] 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

[31] 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. 7

[32] 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

[33] 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, 3, 5, 6

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

[35] 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. 2

[36] 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. 3, 5, 6, 7

[37] Laetitia Plaisance, M Julian Caley, Russell E Brainard, and Nancy Knowlton. The diversity of coral reefs: what are we missing? *PloS one*, 6(10):e25026, 2011. 3

[38] 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. 2, 3, 4

[39] Scarlett Raine, Ross Marchant, Peyman Moghadam, Frederic Maire, Brett Kettle, and Brano Kusy. Multi-species seagrass detection and classification from underwater images. In *Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2020. 2, 8

[40] 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. 2, 3, 5, 6, 7

[41] Scarlett Raine, Ross Marchant, Brano Kusy, Frederic Maire, and Tobias Fischer. Image labels are all you need for coarse seagrass segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5943–5952, 2024. 8

[42] 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. 3, 5, 6, 7

[43] 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, 4, 5, 6

[44] Tiny Remmers, Nader Boutros, Mathew Wyatt, Sophie Gordon, Maren Toor, Chris Roelfsema, Katharina Fabricius, Alana Grech, Marine Lechene, and Renata Ferrari. Rapidbenthos: Automated segmentation and multi-view classification of coral reef communities from photogrammetric reconstruction. *Methods in Ecology and Evolution*, 2024. 2, 3

[45] 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. 1, 3

[46] 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. 3

[47] 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. 2, 4

[48] 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. 2

[49] Clive CR Wilkinson et al. *Status of coral reefs of the world: 2004*. Australian Institute of Marine Science (AIMS), 2004. 2

[50] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 2

[51] 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. 3, 4, 5, 7

[52] Huzheng Yang, James Gee, and Jianbo Shi. Alignedcut: Visual concepts discovery on brain-guided universal feature space. *arXiv preprint arXiv:2406.18344*, 2024. 4

[53] 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. 5, 6

[54] 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, 3

[55] Zhongheng Zhang. Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, 4(11): 218, 2016. 4

[56] Ziqiang Zheng, Xie Yaofeng, Liang Haixin, Yu Zhibin, and Sai-Kit Yeung. Coralvos: Dataset and benchmark for coral video segmentation. *arXiv:2310.01946*, 2023. 3

[57] 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. 4

[58] 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, 3, 4, 5, 6

[59] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2

[60] 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. 2, 3, 4, 5, 7