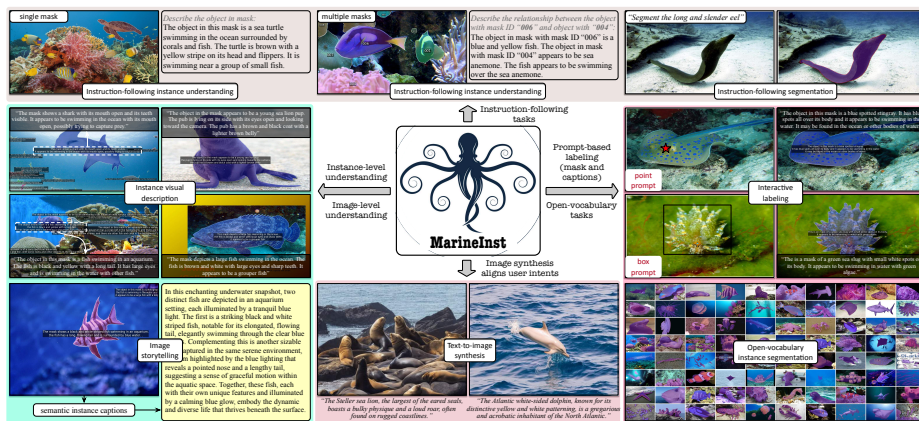


# MarineInst: A Foundation Model for Marine Image Analysis with Instance Visual Description

Ziqiang Zheng<sup>1</sup>, Yiwei Chen<sup>1</sup>, Huimin Zeng<sup>2</sup>, Tuan-Anh Vu<sup>1</sup>, Binh-Son Hua<sup>3</sup>, and Sai-Kit Yeung<sup>1</sup>

<sup>1</sup> The Hong Kong University of Science and Technology  
<sup>2</sup> Northeastern University  
<sup>3</sup> Trinity College Dublin



**Fig. 1:** We present MarineInst, a powerful and flexible marine foundation model, which could support various downstream tasks. Best viewed in color.

**Abstract.** Recent foundation models trained on a tremendous scale of data have shown great promise in a wide range of computer vision tasks and application domains. However, less attention has been paid to the marine realms, which in contrast cover the majority of our blue planet. The scarcity of labeled data is the most hindering issue, and marine photographs illustrate significantly different appearances and contents from general in-air images. Using existing foundation models for marine visual analysis does not yield satisfactory performance, due to not only the data distribution shift, but also the intrinsic limitations of the existing foundation models (*e.g.*, *lacking semantics, redundant mask generation*, or restricted to *image-level scene understanding*). In this work, we emphasize both *model* and *data* approaches for understanding marine ecosystems. We introduce **MarineInst**, a foundation model for the analysis of the marine realms with **instance visual description**, which outputs instance masks and captions for marine object instances. To

train MarineInst, we acquire **MarineInst20M**, the largest marine image dataset to date, which contains a wide spectrum of marine images with high-quality semantic instance masks constructed by a mixture of human-annotated instance masks and model-generated instance masks from our automatic procedure of *binary instance filtering*. To generate informative and detailed semantic instance captions, we use vision-language models to produce semantic richness with various granularities. Our model and dataset support a wide range of marine visual analysis tasks, from image-level scene understanding to regional mask-level instance understanding. More significantly, MarineInst exhibits strong generalization ability and flexibility to support a wide range of downstream tasks with state-of-the-art performance as demonstrated in Figure 1. Project website: <https://marineinst.hkustvgd.com>.

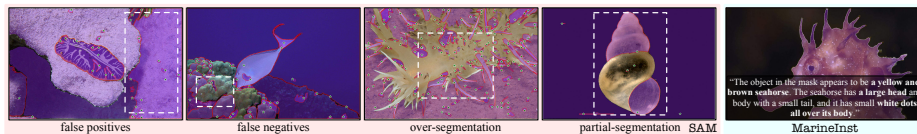
**Keywords:** Foundation model · Instance segmentation · Instance captioning · Marine visual analysis

## 1 Introduction

Marine and underwater visual analysis [7, 63, 64, 66] stands as a crucial frontier in environmental science [7], offering unparalleled insights [66] into one of the planet’s least explored but most diverse ecosystems. The oceans, covering over 70% of the Earth’s surface, are teeming with life and play a pivotal role in global climate regulation, yet remain largely unexplored and poorly understood due to their vastness and inaccessibility. Analyzing and understanding marine imagery/video has gained increasing attention in the computer vision field, such as marine object classification [71, 72], object detection [14, 15, 18], semantic segmentation [10, 23], salient object segmentation [24], underwater image restoration [5], depth estimation [56], and other researches [43, 49].

Recent foundation models [28, 31, 32, 38, 44, 70] provide a powerful and flexible solution for image analysis and understanding. Driven by a significant scale of training data [17, 29, 36, 47] and efficient deep network backbones [13, 19], foundation models demonstrate a strong generalization ability to effectively recognize unseen images and flexibility to support various downstream tasks. Two notable milestones are SAM for segmenting anything [28] and CLIP for vision-language analysis [44]. Particularly, SAM is trained on 11 million images with 1 billion masks to effectively generate precise masks for a wide spectrum of image data. CLIP bridges the image space and the textual space through contrastive learning on millions of image-text pairs from public websites. Both SAM and CLIP demonstrate strong zero-shot generalization ability. In the context of understanding marine environments, can we directly utilize these existing foundation models for marine visual analysis?

We empirically notice two challenges: one is the data distribution shift and another comes from the intrinsic challenges of marine visual data. SAM and CLIP are primarily learned on terrestrial, indoor, and outdoor images. Marine images only occupy a minority of CLIP’s training data [57]. Furthermore, it is



**Fig. 2:** Comparison between SAM and MarineInst. Best viewed in color and zoom in.

worth noting that the marine/underwater images illustrate significantly different *appearances* and *contents* from the in-air images: underwater images are usually plagued by specific conditions, *e.g.*, low visibility [40], dynamic lighting [73], light scattering, color absorption and distortion [5], and motion blur; the marine images contain more unusual and diverse contents, especially the complicated marine creatures with irregular boundaries [23], camouflaged [33], non-rigid [18], bright-colored and textured properties [4]. The untrimmed background [65] makes it difficult to consistently identify and isolate the object of interest under crowded and challenging scenes. These two challenges lead to the fact that existing foundation models cannot effectively recognize marine images. Adapting existing foundation models to handle the unique characteristics of marine imagery requires substantial modifications and domain-specific designs.

This motivates us to build a new marine foundation model. With substantial marine images meticulously collected, we have to determine the fundamental property/task for effective and efficient marine visual analysis. SAM is based on image segmentation, generating semantic-agnostic masks automatically or based on user prompts interactively. However, the generated masks are *inaccurate*, *without semantics*, leading to *false positives*, *false negatives* (missing unusual marine creatures), *over-segmentation*, and *partial-segmentation* (parts of objects segmented) on marine data, as shown in Figure 2. CLIP is based on a vision-language task, which computes cross-modality image-text similarity, but is limited to *image-level scene understanding*. CLIP cannot yield instance-level or region-level understanding, explicitly detecting or segmenting the interest of objects.

Taking into account such limiting factors, we selected **instance visual description** as the key task of our foundation model, simultaneously generating dense instance masks and their instance captions. Instance visual description can be viewed as the combination of instance segmentation and instance captioning built into a single objective. Unlike image-level recognition [12], object detection [14, 39, 67] with axis-aligned/oriented bounding boxes and semantic segmentation [23], instance segmentation is valuable to efficiently identify and localize diverse marine / underwater entities with complex object boundaries. Precise instance segmentation promotes comprehensive marine studies, *e.g.*, object counting [50], species identification [71, 72], biological trait detection [26], cover estimation [7], benthic composition [53], population and distribution computation [73], symbiotic relationship prediction [62] to name a few. We do **not** adopt more fine-grained segmentation (*e.g.*, segmenting the fin of fish or tentacles of octopus) as the key task since the biologically integral component is challenging

to define and generalize to a wide spectrum of marine creatures. Furthermore, existing instance segmentation [34, 35] is limited to pre-defined categories and only provides category-level semantics. Our instance visual description extends instance segmentation with open-vocabulary *instance captioning*, generating instance masks and their instance captions with semantic richness from multiple granularities. Example predictions are demonstrated in Figure 2.

Driven by such design choices, we formulate **MarineInst**, a strong foundation model to perform instance segmentation and instance captioning. To address over-segmentation and partial-segmentation issues, we propose *binary instance filtering*, a simple yet effective technique to filter out low-quality non-instance masks. Through simultaneous mask generation and binary instance filtering, our MarineInst could effectively perform precise instance mask generation. Notably, MarineInst is robust to accurately identify and delineate a wide range of marine creatures. To address the semantic-agnostic or limited semantics issue, we perform *instance captioning* to generate informative and comprehensive semantic captions for the generated instance masks by harnessing the power of vision-language models (VLMs) [31, 38, 44, 66, 70] as illustrated in Figure 3. We utilize the powerful VLMs to achieve a more plentiful granular level of understanding for instance segments within the image, going beyond mere object recognition to comprehend complex attributes and relationships depicted in visual data. MarineInst could perform instance visual descriptions of marine images with semantic richness from various semantic granularities.

Our MarineInst is driven by our constructed **MarineInst20M** dataset with semantic instance mask annotations, which is the largest marine image dataset to date. MarineInst20M, consisting of 2.4 million marine images with around 20 million instance masks, is a mixture of 1) existing public underwater/marine datasets with available various formats of annotations, 2) our collected images with manually labeled annotations, and 3) public Internet marine images with automatically generated instance masks by our MarineInst. With carefully constructed visual data with remarkable diversity, MarineInst20M could effectively alleviate the false negative issue. The formulated instance visual description not only provides a more meaningful, efficient, and valuable solution for visual analysis but also brings challenges to both computer vision and marine communities. Extensive experimental results demonstrate that our foundation model MarineInst and dataset MarineInst20M yield strong performance on vision-language tasks including salient object segmentation, underwater object detection, image/instance captioning, text-to-image synthesis, and instruction-following tasks. The main contributions of this paper are as follows:

- We propose MarineInst, a powerful and flexible marine foundation model, which could perform the instance visual description task in an automatic or interactive manner. Our instance visual description task includes instance segmentation and instance captioning.
- We propose instance segmentation with binary instance filtering to enable a strong generalization ability to unseen marine images for obtaining high-quality instance masks; we also propose to perform instance captioning for

- visual descriptions of instance masks with various granularities, yielding dense and informative mask-level semantic instance captions.
- We propose MarineInst20M, the largest documented marine image dataset to date, with remarkable visual diversity and semantic instance mask annotations.
  - We demonstrate the strong performance of MarineInst trained on MarineInst20M for various marine analysis tasks, demonstrating a wide spectrum of downstream applications in both computer vision and marine communities.

## 2 Related Work

### 2.1 Marine Visual Analysis

Marine visual analysis and understanding [8, 15, 33, 37] promote to unveil the mysteries of the oceans and harness technology to elevate marine research [33], conservation [20], and industrial endeavors [10]. Unlike in-air images, underwater images often suffer from quality degradation [5] due to scattering and absorption of light, resulting in poor contrast, blurring, and color distortion. The underwater environment is filled with moving particles, varying textures, and other organisms that can be mistaken for the target object [14]. Besides the appearance shift, marine creatures are incredibly diverse in terms of shapes, sizes, and colors [49]. Effectively handling such variability and correctly recognizing a wide range of marine creatures is a significant challenge. MarineDet [18] proposed to perform open-marine object detection and detect a wider range of marine creatures than existing close-set underwater object detection algorithms. Different from detection, instance segmentation [35] provides a feasible and effective way for object-centric instance understanding, generating precise boundaries for each instance. Through dense pixel-level semantic analysis, researchers could gain insights into social structures, predation, symbiotic relationships, and other behavioral patterns of marine creatures. Another similar line of research to our work is underwater salient object segmentation [21, 24] and salient instance segmentation [35], detecting and segmenting the salient objects from underwater visual images. However, there is no consistent and clear definition of “salient objects” and the salient objects are highly subject to humans, varying from people.

### 2.2 Foundation Model

Foundation models (*e.g.*, CLIP [44], ALIGN [25], SAM [28], and VLMs [31, 32, 38, 66, 70]) have been widely favored by the whole CV community. Optimized by millions of image-text pairs, CLIP [44] demonstrated a strong zero-shot recognition ability to unseen images. The further BLIP series [31, 32] proposed to bridge the frozen visual and language foundation models based on Q-Former. However, Both CLIP and BLIP are limited to image-level scene understanding and fail to provide fine-grained and regional instance understanding. Recent works [18, 67] proposed to utilize CLIP for open-vocabulary object detection [27, 67] and semantic segmentation [58, 59, 69], supporting to recognition of a wide range

of object categories. However, these works are mainly limited to terrestrial images [17, 36, 48, 68], only showing a limited ability to understand marine images.

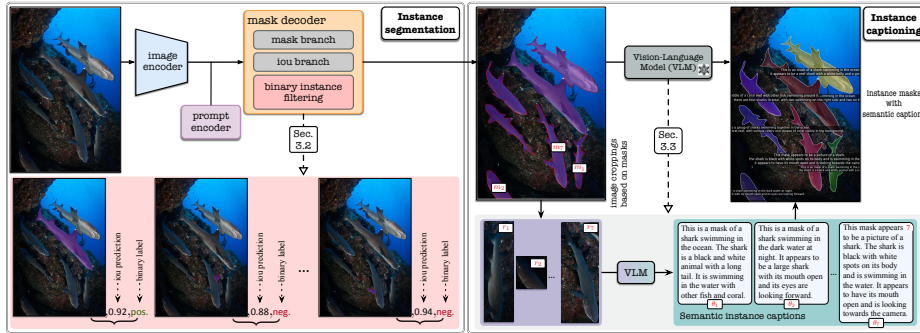
SAM [28] optimized by vast and diverse training data, has demonstrated a strong ability to segment visual elements with precise masks in a semantic-agnostic manner. Receiving various kinds of prompts (*e.g.*, point, box, and mask) from the user, SAM could yield the required mask through interactive labeling and iterative refinement. However, without any prompt, the automatically generated masks from SAM are not always meaningful or required due to the nature of lacking semantics. The redundant mask generation of SAM leads to drastic over-segmentation and SAM demonstrates a poor ability to segment camouflaged objects [52], objects with irregular boundaries [33] or remarkable pose and appearance variations [66]. The intrinsic challenges of marine images lead to significant false negative and partial-segmentation issues. Furthermore, SAM cannot provide semantic predictions for generated masks, which heavily restricts SAM from solving semantic understanding tasks. SSA [9] represents an important step towards a more sophisticated model for semantic scene understanding. However, since only assigning the semantics of textual descriptions to generated masks from SAM based on similarity computation, SSA cannot alleviate the intrinsic issues of SAM. Grounded SAM [45] combines Grounding DINO [39] and SAM to achieve text-guided segmentation while the ability to detect marine objects is limited. OVSAM [60] proposed an open-vocabulary classification head to generate the semantics for the masks labeled by the users. However, none of these existing foundation models is specially designed for marine visual analysis and thus demonstrates a poor marine analytical ability. We propose to present a powerful and flexible foundation model in the marine field.

### 3 Method

#### 3.1 Overview

Our marine foundation model has two main stages for predicting instance visual description. The first stage performs **instance segmentation** to obtain the instance masks, and the second stage performs **instance captioning** that generates instance captions on the predicted instance masks. To improve the accuracy of the instance masks, we devise a strategy for both training and inference that uses *binary instance filtering* to remove non-instance masks. We provide an overview of our MarineInst foundation model in Figure 3.

We summarize key differences between MarineInst and existing foundation models in four aspects in Table 1: 1) whether the model is performing instance understanding; 2) semantic richness; 3) semantic granularities; and 4) the inference procedure. Specifically, CLIP [44] optimized by image-level captions, cannot explicitly localize or yield fine-grained descriptions for object instances. Semantic-SAM [30] generates masks with 6 semantic granularities through a many-to-many matching design. The generated semantics of Semantic-SAM come from pre-defined object categories in existing training datasets [17, 48, 68] (mainly in-air objects). SSA [9] utilizes BLIP2 [31] for generating image captions for



**Fig. 3:** The framework overview of proposed MarineInst. There are two components in MarineInst: 1) automatic **instance segmentation** with *binary instance filtering* to remove the non-instance masks; 2) **instance captioning** to generate meaningful and comprehensive captions for generated instance masks based on frozen VLMs.

**Table 1:** Direction comparison between our MarineInst and existing foundation models. Inference: **A** - Automatic, **I** - Interactive.

Methods	Instance Understanding	Semantic richness	Semantic granularities	Inference
SAM [28]	✗	No semantics	No semantics	<b>A</b> , <b>I</b>
CLIP [44]	✗	Image-level scene understanding	Scenario visual understanding	<b>A</b>
Semantic SAM [30]	✓	Driven by pre-defined object categories	6 granularities (implicitly)	<b>A</b> , <b>I</b>
SSA [9]	✗	Lacking fine-grained information	Image-level scene understanding	<b>A</b> , <b>I</b>
OVSAM [60]	✗	Only category-level information	Pre-defined object categories (22K)	<b>I</b>
Grounded SAM [45]	✓	Provided by users	Pre-defined object categories	<b>I</b>
MarineInst	✓	Detailed and comprehensive captions	Open-ended caption generation	<b>A</b> , <b>I</b>

the masks generated from SAM. SSA then extracts nouns from generated captions and utilizes CLIP to compute the similarity between extracted nouns and cropped image regions for obtaining the final semantics. OVSAM [60] claimed to segment and recognize approximately 22 thousand classes. However, OVSAM cannot automatically generate instance masks without any prompt. Grounded SAM [45] combines Grounding DINO [39] and SAM, where Grounding DINO yields bounding box predictions based on text queries and then SAM is utilized for mask generation based on box predictions. In contrast, MarineInst automatically performs instance segmentation in an end-to-end manner.

### 3.2 Instance Segmentation

The instance segmentation component of MarineInst is built upon SAM [28]. MarineInst is continuously pre-trained on our MarineInst20M dataset to promote the ability of MarineInst to extract efficient and effective marine feature representations. To alleviate the over-segmentation and partial-segmentation issues, and also promote the accuracy of generated instance masks, we add binary instance filtering in the mask decoder as illustrated in Figure 3.

**Binary Instance Filtering.** To filter out the non-instance masks produced by MarineInst, especially the meaningless masks or the masks without instance semantics, we propose our *binary instance filtering* inside the mask decoder. We generate “negative” (non-instance) masks based on “positive” instance masks labeled by humans. We randomly sample the point prompt inside the human-annotated instance mask and infer SAM (ViT-H backbone) to generate one SAM-generated mask. If the *predicted IoU* from the IoU branch is over 0.88 (as suggested by SAM for high-quality mask generation) and the *calculated IoU* between the SAM-generated mask and the human-annotated instance mask is below a user-defined threshold (0.5 in this work), we regard this SAM-generated mask as “negative” mask. We formulate a lightweight binary filtering branch to enable MarineInst to discriminate whether the generated masks are instance masks. We conduct attention-based feature interaction between mask generation and binary instance filtering:

$$\mathcal{L}_{bin.} = -(y \log(p) + (1 - y) \log(1 - p)), p = \text{MLP}(F_I, \text{Concat}(T_M, T_P)), \quad (1)$$

where  $y$  denotes the binary ground truth.  $T_M$  and  $T_P$  are the learnable mask tokens and prompt tokens, respectively.  $F_I$  is the image embedding from the image encoder and MLP is a lightweight MLP layer. Please note that both “positive” and “negative” masks are used for optimizing our MarineInst.

### 3.3 Instance Captioning

With instance segmentation, we perform an essential step forward in conducting the comprehensive instance captioning for the generated instance masks. Different from the existing works [9, 22, 30], which utilized the pre-trained CLIP [44] for generating category proposals for the synthesized masks by computing the similarity between image regions and textual queries, we propose to generate open-ended semantic instance captions by harnessing the power of LLMs. To achieve this, we leverage the frozen VLMs to generate comprehensive and informative semantic instance captions as demonstrated in Figure 3. It is worth noting that MarineInst is flexible to various VLMs. With instance masks  $M = \{m_1, m_2, \dots, m_n\}$ , we first crop the image regions  $R = \{r_1, r_2, \dots, r_n\}$  from the whole image based on the localization information from the instance masks. Then we infer the frozen VLMs with the following prompt template: “The image is  $\langle image \rangle$ . Describe the object in this image: ”, where  $\langle image \rangle$  is the image token. The caption is generated as follows:

$$\theta^* = \arg \max_{\theta} \prod_{j=1}^N P(\theta_j | \theta_{<j}, x), \quad (2)$$

where  $x$  indicates the feature embedding of the cropped image region  $r$  after the frozen ViT image encoder and Q-former [31, 66].  $x$  is fed into a frozen LLM (*e.g.*, Vicuna [11]) for instance captioning.  $\theta^*$  is the optimal generated detailed and comprehensive captions for  $r$ . We obtain  $\Theta = \{\theta_1, \theta_2, \dots, \theta_n\}$  based on  $M$ .



Through semantic instance captioning, we transform image-level understanding to mask-level instance understanding, significantly elevating the granularities and richness of image interpretation.

### 3.4 Dataset Construction and Model Training

MarineInst20M consists of 3 main data sources: 1) existing public marine/underwater datasets (from object counting [50], retrieval [54], detection [14,37], tracking [6,61], segmentation tasks [49,73]); 2) manually collected images from public/private data and YouTube videos; and 3) public Internet images.

The existing public datasets contain various formats of annotations (point, box, and mask). For images with point/box annotations, we run inference with SAM (ViT-H) with point/box prompts to generate the instance masks. Note that we only preserve the high-quality generated masks with predicted IoU over 0.88. For images with mask annotations, we only pick up desired images with clear instance masks for training. For manually collected images, we manually label them using an internal labeling tool built upon SAM. For Internet images, we adopt crowdsourcing to scrape public Internet images with alt-texts from Flickr [1], Gettyimages [2], and Shutterstock [3].

We implement an iterative scheme to train MarineInst. We first utilize converted high-quality instance masks from existing public datasets and our manually labeled instance masks to optimize our MarineInst model. After that, we use the trained MarineInst model for automatic mask generation to enlarge the annotated masks for training. To ensure the automatically generated masks are high-quality instance masks, inspired by [16], we construct **2.65M** mask samples (**1.89M** “positive” instance masks labeled by humans or from public datasets and **0.76M** “negative” non-instance masks constructed from SAM) with binary labels for binary instance filtering. We train a binary classifier to distinguish positive and negative masks (refer to Section 3.2), and thus we can effectively alleviate over-segmentation and partial-segmentation issues by applying binary instance filtering to discard non-instance masks. It is worth noting that mask generation and binary instance filtering are conducted simultaneously. We finally generate instance masks and perform binary filtering for the public images from the public Internet. In total, we obtained 1.89M human-annotated instance masks (from both public datasets and our annotators) and **17.3M** automatically model-generated instance masks after binary instance filtering.




For instance captioning, we assign the original category label to the converted instance mask if the semantic annotations exist. For those model-generated instance masks without semantics, we run an inference with VLMs to generate the semantic captions (refer to Section 3.3). We then mix them and construct our MarineInst20M dataset, which contains around 20 million instance masks with detailed semantic instance captions. It takes approximately **4,480** human hours to label the instance masks, and **24,560** GPU hours (RTX 3090) to automatically generate instance masks with semantic captions. To cope with image credits and licenses for data redistribution, we will release MarineInst20M dataset with images in raw URL format and annotations in JSON format for research purposes.

## 4 Experiments

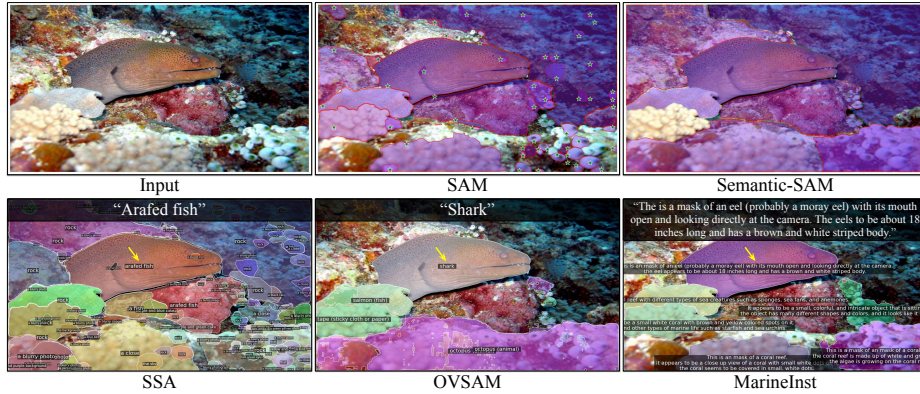
### 4.1 Implementation Details

We adopt SAM as an effective backbone for instance segmentation with binary instance filtering inside the mask decoder. MarineInst is continuously pre-trained on our MarineInst20M dataset (**2.42M** images and **19.2M** instance masks in total). We adopt the combination of point prompt (3 random points inside the mask) and box prompt as the training prompt while ignoring the mask prompt. For instance captioning, we infer frozen VLMs, such as CLIP [44], BLIP2 [31] and MarineGPT [66]. Our MarineInst is flexible for various VLMs. We adopt BLIP2 and MarineGPT as the main demonstration in this paper. We set the maximum generated tokens to 50 for generating captions. **Comparisons.** We include SAM [28], Semantic-SAM [30], SSA [9] (SAM+BLIP2 [31]+CLIP [44]) and OVSAM [60] for comparison. SAM generates masks under the automatic mode or with prompts. We set the semantic granularity of Semantic-SAM to 3 for automatically producing masks. SSA assigns semantics from BLIP2 to the generated masks from SAM. OVSAM generates masks with point or box prompts.

### 4.2 Results on Instance Segmentation

We perform instance segmentation to evaluate the ability of various foundation models to generate precise instance masks. We adopt 500 unseen marine images with manually labeled instance masks for semantic-agnostic evaluation, where we regard all instances as the “foreground” category (semantics are ignored). We leave semantic accuracy evaluation of generated semantic captions in Section 4.4. Following evaluation metrics of the COCO dataset [36], we report the AP scores of the various algorithms under two evaluation types: *bbox* and *segm*. We compute results under three settings: “” - Automatic (no human prompts are provided); “” - Point (one random point inside the instance mask is chosen as prompt); “” - BBOX (the bounding box of the instance mask is regarded as prompt). Note that we do not compute the AP scores of evaluation type *bbox* under the “BBOX” setting since the bounding boxes are already given. For fair comparisons, all methods are evaluated with the same prompts.

The qualitative and quantitative results are illustrated in Figure 4 and Table 2, respectively. As demonstrated, SAM cannot automatically generate reasonable or accurate instance masks. Unsurprisingly, SAM obtains very low scores under the automatic setting due to over-segmentation and partial-segmentation in Table 2. With the point or box prompt, SAM could achieve reasonable performance, demonstrating its strong ability for mask generation as an interactive tool. Based on SAM, SSA generates wrong semantics based on inferior mask predictions from SAM in Figure 4. Semantic-SAM with semantic granularity 3 (“instance-level”) also fails to achieve meaningful instance segmentation and nearly cannot achieve the small- and medium-sized object instance segmentation as demonstrated in Table 2. Note that Semantic-SAM cannot generate explicit semantic captions for generated masks. Meanwhile, OVSAM demonstrates a poor



**Fig. 4:** Comparison with existing SOTA algorithms. MarineInst could effectively address the over-segmentation and partial-segmentation issues of SAM and Semantic-SAM. Meanwhile, MarineInst could generate meaningful and comprehensive semantic captions faithful to each generated instance mask, while others cannot.

**Table 2:** Instance segmentation results of various algorithms under settings: “**A**” - Automatic; “**\***” - Point; “**□**” - BBOX.

Method	AP $\uparrow$		AP <sub>s</sub> $\uparrow$		AP <sub>m</sub> $\uparrow$		AP <sub>f</sub> $\uparrow$	
	bbox	segm	bbox	segm	bbox	segm	bbox	segm
SAM <b>A</b> [28]	5.9	5.8	0.3	0.4	3.2	3.5	15.2	14.7
Semantic-SAM [30] <b>A</b>	2.5	0.0	0.0	0.0	0.0	0.0	5.3	0.0
MarineInst <b>A</b>	<b>30.8</b>	<b>32.7</b>	<b>7.6</b>	<b>8.8</b>	<b>32.1</b>	<b>35.5</b>	<b>40.2</b>	<b>40.8</b>
SAM* [28]	59.0	63.0	64.3	77.8	70.2	77.3	48.5	47.4
OVSAM* [60]	46.8	49.0	42.4	43.8	50.1	53.9	44.7	46.4
MarineInst*	<b>73.1</b>	<b>75.4</b>	<b>77.5</b>	<b>86.6</b>	<b>82.5</b>	<b>86.7</b>	<b>64.1</b>	<b>62.8</b>
SAM <b>□</b> [28]	-	93.5	-	95.3	-	95.7	-	92.2
OVSAM <b>□</b> [60]	-	79.4	-	63.1	-	81.7	-	86.9
MarineInst <b>□</b>	-	<b>95.4</b>	-	<b>96.4</b>	-	<b>97.3</b>	-	<b>93.8</b>

**Table 3:** Underwater salient object segmentation results of using different backbones on USOD10K datasets [21]. Evaluation metrics followed [21].

Method	Back.	$S_m \uparrow$	$E_e^{max} \uparrow$	$\max F \uparrow$	MAE $\downarrow$
SAM [28]	ViT-B	0.8695	0.9199	0.8445	0.0387
MarineInst	ViT-B	<b>0.8773</b>	<b>0.9276</b>	<b>0.8537</b>	<b>0.0353</b>
SAM [28]	ViT-L	0.8843	0.9279	0.8658	0.0336
MarineInst	ViT-L	<b>0.8931</b>	<b>0.9325</b>	<b>0.8713</b>	<b>0.0304</b>
SAM [28]	ViT-H	0.9034	0.9374	0.8812	0.0287
MarineInst	ViT-H	<b>0.9103</b>	<b>0.9411</b>	<b>0.8876</b>	<b>0.0256</b>

ability to generate reliable semantic category predictions (*e.g.*, misrecognizing “eel” to “shark”) for the inferred masks from point/box prompts provided by users. Furthermore, even with the point/box prompts, OVSAM still cannot achieve very precise instance segmentation since both feature extraction and mask generation abilities of OVSAM have been weakened after knowledge distillation. MarineInst demonstrated much stronger instance segmentation ability under all three settings. Under the most challenging automatic setting, MarineInst achieves **30.8/32.7** AP scores while SAM only has 5.9/5.8. MarineInst effectively alleviates the over-segmentation and partial-segmentation issues. It is also worth noting that MarineInst could generate instance masks with corresponding informative semantic descriptions, which enable the extraction of nuanced information, including appearance, pose, color, and other attributes in Figure 4.



**Fig. 5:** (a) Image storytelling based on MarineInst. (b) Marine text-to-image synthesis based on stable diffusion model [46] (“*stable-diffusion-v1-5*”).

### 4.3 Downstream Tasks

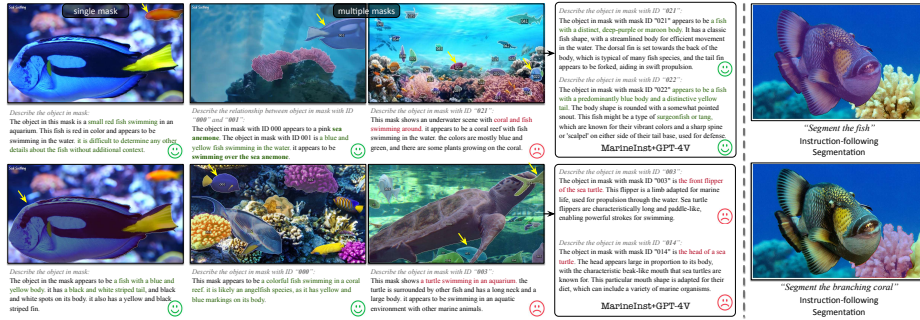
We demonstrate the robustness of our foundation model with following downstream tasks: (1) salient object segmentation; (2) semantic instance captioning and image storytelling; (3) text-to-image synthesis; and (4) instruction-following tasks. We also demonstrate underwater object detection in the supplementary.

**Salient object segmentation.** We utilize MarineInst for underwater salient object segmentation on the USOD10K dataset [21]. We regard the image size as the box prompt to generate the salient object mask. SAM is included for comparison. The heavy image encoders of both SAM and MarineInst are frozen and we only optimize the prompt encoder and mask decoder. We demonstrate that MarineInst is more effective than SAM on underwater visual analysis in Table 3, indicating a stronger feature extraction ability.

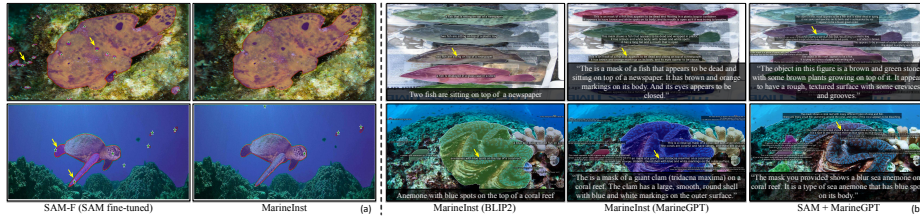
**Semantic instance captioning and storytelling.** MarineInst could *automatically* generate the instance mask with semantic captions, making it an effective and comprehensive image analysis and labeling tool, significantly reducing the need for manual annotations. We could perform comprehensive and informative **image storytelling** by asking ChatGPT [41] to generate image-level captions based on generated instance-level semantic captions, as demonstrated in Figure 5(a). Compared to mask-aligned descriptions [55], our captions are automatically generated.

**Text-to-image synthesis.** The constructed MarineInst20M dataset has a significant value for promoting marine text-to-image synthesis. The semantic captions for the close-up images that contain rich information (including the pose, color, texture, boundary, and appearance of objects), will guide a stronger model to generate photo-realistic and reasonable image outputs in Figure 5(b).

**Instruction-following tasks.** We provide instruction-following instance understanding results in Figure 6 under two settings: *single mask* and *multiple masks* (assigned with mask IDs). The latter setting enables multiple instance understanding simultaneously. After the instruction-following tuning, MarineInst demonstrates a powerful ability to understand the instances visualized in mask due to the help of visual prompts [51]. However, MarineInst could only work ef-



**Fig. 6:** The results of instruction-following instance understanding and segmentation. Texts in green are correct responses and red are wrong responses.



**Fig. 7:** (a) Effects of binary instance filtering. (b) Effects of different VLMs on generating instance captions for the generated masks.

fectively under scenarios that are not crowded. When there are multiple instances small and crowded, MarineInst still fails to localize the object by mask IDs. We attribute this failure to the poor OCR ability of frozen ViT backbones of the existing VLMs, which are mainly optimized by image-level captions. We finally combine MarineInst (generating instance masks) with GPT-4V [42]. GPT-4V demonstrates a stronger ability to localize and understand user-selected instance masks. However, we also noticed that GPT-4V would fail with self-occlusions between different instances. Our observations and dataset provide valuable insights and contributions to marine instance understanding. We have also provided the instruction-following segmentation results in Figure 6.

#### 4.4 Ablation Studies and Further Analysis

##### Effects of binary instance filtering and model-generated annotations.

We aim to evaluate the effects of the proposed binary instance filtering and the model-generated annotations. We directly fine-tune SAM (SAM-F) on our MarineInst20M as a baseline. Please note that we also utilize the non-instance masks for optimizing MarineInst with binary instance filtering. We report the automatic instance segmentation results of 500 unseen testing images in Table 4

**Table 4:** Effects of binary instance filtering and model-generated instance masks. **SAM-F**: SAM fine-tuned on MarineInst20M.

Method	binary instance filtering	human-annotated instance masks	model-generated instance masks	AP $\uparrow$	
				bbox	segm
SAM <sup>Ⓞ</sup> [28]	✗	✗	✗	5.9	5.8
SAM-F <sup>Ⓞ</sup>	✗	✓	✗	23.0	24.8
MarineInst <sup>Ⓞ</sup>	✓	✓	✗	28.2	30.1
SAM-F <sup>Ⓞ</sup>	✗	✓	✓	24.0	25.8
MarineInst <sup>Ⓞ</sup>	✓	✓	✓	<b>30.8</b>	<b>32.7</b>

**Table 5:** User studies. The average satisfactory scores of 1,000 mask-caption pairs (higher is better).

Method	Satisfactory Score
SSA [9]	1.14
OVSAM [60]	1.56
MarineInst (BLIP2 [31])	2.31
MarineInst (MarineGPT [66])	<b>3.52</b>

(more results left in supplementary). With the binary instance filtering, MarineInst demonstrates larger performance improvements over SAM-F, effectively alleviating the over-segmentation and partial-segmentation illustrated in Figure 7(a) as well. We also observe that the model-generated annotations could further improve instance segmentation performance for both SAM-F and MarineInst.

**Comparing different VLMs.** We use our foundation model with different VLMs to evaluate the ability to generate instance captions. We select BLIP2 as a generic model and MarineGPT as a marine-specific model for captioning. We found that our foundation model could support both captioning models, but using a specific VLM like MarineGPT yields long and comprehensive captions with superior performance than BLIP2, as shown in Figure 7(b).

**Effects of instance masks.** We experiment with how instance masks affect caption quality. We combine SAM with MarineGPT to generate masks with semantics in Figure 7(b). When being provided with unsatisfactory masks from SAM, MarineGPT cannot generate accurate and meaningful captions. Therefore, our foundation model is valuable in that generating more accurate instance masks, effectively alleviates error accumulation for caption generation.

**Accuracy of semantic captions.** We perform user studies to evaluate the accuracy of the generated semantics produced by SSA, OVSAM, MarineInst (BLIP2), and MarineInst (MarineGPT). We randomly picked up 1,000 mask-caption pairs generated by each algorithm from the whole pool. For subject fidelity, we asked 3 students from the marine biology field to answer 1,000 scoring questions, totaling 12,000 answers. The students are asked to answer the question: “Please give your satisfactory score (from 1 to 5) based on the correctness, helpfulness, and information richness of generated captions for the instance mask”. The quantitative comparisons are provided in Table 5.

## 5 Conclusion

In this work, we have proposed the marine foundation model called MarineInst to perform marine image analysis, which could generate instance masks with semantics. Besides, the constructed MarineInst20M dataset could significantly promote the performance of various downstream tasks. Our work paves the way for future exploration of marine image analysis.

**Acknowledgment.** We thank the valuable discussions and suggestions from Yang Wu and Jianbo Shi. We also thank all the volunteer students who helped do the annotations and evaluations. The work was partially supported by the Innovation and Technology Support Programme of the Innovation and Technology Fund (Ref: ITS/200/20FP) and the Marine Conservation Enhancement Fund (MCEF20107 and MCEF23EG01) and an internal grant from HKUST (R9429). Binh-Son Hua is supported by the Science Foundation Ireland under the SFI Frontiers for the Future Programme (22/FFP-P/11522).

## References

1. Flickr. <https://www.flickr.com/>
2. Getty images. <https://www.gettyimages.com/>
3. Shutterstock. <https://www.shutterstock.com/>
4. Encyclopedia of life. <http://eol.org> (2018)
5. Akkaynak, D., Treibitz, T.: Sea-thru: A method for removing water from underwater images. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 1682–1691 (2019)
6. Alawode, B., Guo, Y., Ummar, M., Werghi, N., Dias, J., Mian, A., Javed, S.: Utb180: A high-quality benchmark for underwater tracking. In: Asian Conference on Computer Vision (ACCV). pp. 3326–3342 (2022)
7. Beijbom, O., Edmunds, P.J., Roelfsema, C., Smith, J., Kline, D.I., Neal, B.P., Dunlap, M.J., Moriarty, V., Fan, T.Y., Tan, C.J., et al.: Towards automated annotation of benthic survey images: Variability of human experts and operational modes of automation. *PloS one* **10**(7), e0130312 (2015)
8. Bovcon, B., Muhovič, J., Perš, J., Kristan, M.: The mastr1325 dataset for training deep usv obstacle detection models. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3431–3438. IEEE (2019)
9. Chen, J., Yang, Z., Zhang, L.: Semantic segment anything. <https://github.com/fudan-zvg/Semantic-Segment-Anything> (2023)
10. Cheng, Y., Zhu, J., Jiang, M., Fu, J., Pang, C., Wang, P., Sankaran, K., Onabola, O., Liu, Y., Liu, D., Bengio, Y.: Flow: A dataset and benchmark for floating waste detection in inland waters. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10953–10962 (October 2021)
11. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. <https://vicuna.lmsys.org> (2023)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255. Ieee (2009)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Fan, B., Chen, W., Cong, Y., Tian, J.: Dual refinement underwater object detection network. In: European Conference on Computer Vision (ECCV). pp. 275–291. Springer (2020)

15. Fulton, M., Hong, J., Islam, M.J., Sattar, J.: Robotic detection of marine litter using deep visual detection models. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 5752–5758. IEEE (2019)
16. Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C.C.T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., et al.: Textbooks are all you need. arXiv preprint arXiv:2306.11644 (2023)
17. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5356–5364 (2019)
18. Haixin, L., Ziqiang, Z., Zeyu, M., Yeung, S.K.: Marinedet: Towards open-marine object detection. arXiv preprint arXiv:2310.01931 (2023)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
20. Hong, J., Fulton, M., Sattar, J.: Trashcan: A semantically-segmented dataset towards visual detection of marine debris. arXiv preprint arXiv:2007.08097 (2020)
21. Hong, L., Wang, X., Zhang, G., Zhao, M.: Usod10k: a new benchmark dataset for underwater salient object detection. IEEE Transactions on Image Processing (TIP) (2023)
22. Huynh, D., Kuen, J., Lin, Z., Gu, J., Elhamifar, E.: Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7020–7031 (2022)
23. Islam, M.J., Edge, C., Xiao, Y., Luo, P., Mehtaz, M., Morse, C., Enan, S.S., Sattar, J.: Semantic segmentation of underwater imagery: Dataset and benchmark. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1769–1776. IEEE (2020)
24. Islam, M.J., Wang, R., Sattar, J.: Svam: saliency-guided visual attention modeling by autonomous underwater robots. Robotics: Science and Systems (2022)
25. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning (ICML). pp. 4904–4916. PMLR (2021)
26. Khan, F.F., Li, X., Temple, A.J., Elhoseiny, M.: Fishnet: A large-scale dataset and benchmark for fish recognition, detection, and functional trait prediction. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20496–20506 (2023)
27. Kim, D., Angelova, A., Kuo, W.: Region-aware pretraining for open-vocabulary object detection with vision transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11144–11154 (2023)
28. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
29. Li, C., Liu, H., Li, L., Zhang, P., Aneja, J., Yang, J., Jin, P., Hu, H., Liu, Z., Lee, Y.J., et al.: Elevater: A benchmark and toolkit for evaluating language-augmented visual models. Advances in Neural Information Processing Systems (Neurips) **35**, 9287–9301 (2022)
30. Li, F., Zhang, H., Sun, P., Zou, X., Liu, S., Yang, J., Li, C., Zhang, L., Gao, J.: Semantic-sam: Segment and recognize anything at any granularity. arXiv preprint arXiv:2307.04767 (2023)



31. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International Conference on Machine Learning (ICML)* (2023)
32. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: *International Conference on Machine Learning (ICML)*. pp. 12888–12900. PMLR (2022)
33. Li, L., Dong, B., Rigall, E., Zhou, T., Dong, J., Chen, G.: Marine animal segmentation. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)* **32**(4), 2303–2314 (2021)
34. Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y.: Fully convolutional instance-aware semantic segmentation. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 2359–2367 (2017)
35. Lian, S., Li, H., Cong, R., Li, S., Zhang, W., Kwong, S.: Watermask: Instance segmentation for underwater imagery. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 1305–1315 (2023)
36. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: *European Conference on Computer Vision (ECCV)*. pp. 740–755. Springer (2014)
37. Lin, X., Sanket, N.J., Karapetyan, N., Aloimonos, Y.: Oysternet: Enhanced oyster detection using simulation. In: *IEEE International Conference on Robotics and Automation (ICRA)*. pp. 5170–5176. IEEE (2023)
38. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *Neural Information Processing Systems (Neurips)* (2023)
39. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023)
40. Marques, T.P., Albu, A.B.: L2uwe: A framework for the efficient enhancement of low-light underwater images using local contrast and multi-scale fusion. In: *IEEE/CVF conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. pp. 538–539 (2020)
41. OpenAI: Introducing chatgpt (2022), <https://openai.com/blog/chatgpt>
42. OpenAI: Gpt-4 technical report (2023)
43. Palnitkar, A., Kapu, R., Lin, X., Liu, C., Karapetyan, N., Aloimonos, Y.: Chat-sim: Underwater simulation with natural language prompting. *arXiv preprint arXiv:2308.04029* (2023)
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning (ICML)*. pp. 8748–8763. PMLR (2021)
45. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded sam: Assembling open-world models for diverse visual tasks (2024)
46. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 10684–10695 (June 2022)
47. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* (2021)
48. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: *IEEE/CVF international conference on computer vision (CVPR)*. pp. 8430–8439 (2019)

49. Shi, Z., Guan, C., Li, Q., Liang, J., Cao, L., Zheng, H., Gu, Z., Zheng, B.: Detecting marine organisms via joint attention-relation learning for marine video surveillance. *IEEE Journal of Oceanic Engineering* **47**(4), 959–974 (2022)
50. Sun, G., An, Z., Liu, Y., Liu, C., Sakaridis, C., Fan, D.P., Van Gool, L.: Indiscernible object counting in underwater scenes. In: *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)* (2023)
51. Sun, Z., Fang, Y., Wu, T., Zhang, P., Zang, Y., Kong, S., Xiong, Y., Lin, D., Wang, J.: Alpha-clip: A clip model focusing on wherever you want. *arXiv preprint arXiv:2312.03818* (2023)
52. Tang, L., Xiao, H., Li, B.: Can sam segment anything? when sam meets camouflaged object detection. *arXiv preprint arXiv:2304.04709* (2023)
53. Tebbett, S.B., Connolly, S.R., Bellwood, D.R.: Benthic composition changes on coral reefs at global scales. *Nature ecology & evolution* **7**(1), 71–81 (2023)
54. Truong, Q.T., Vu, T.A., Ha, T.S., Lokoč, J., Wong, Y.H., Joneja, A., Yeung, S.K.: Marine video kit: a new marine video dataset for content-based analysis and retrieval. In: *International Conference on Multimedia Modeling (MMM)*. pp. 539–550. Springer (2023)
55. Urbanek, J., Bordes, F., Astolfi, P., Williamson, M., Sharma, V., Romero-Soriano, A.: A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. *arXiv preprint arXiv:2312.08578* (2023)
56. Varghese, N., Kumar, A., Rajagopalan, A.: Self-supervised monocular underwater depth recovery, image restoration, and a real-sea video dataset. In: *IEEE/CVF International Conference on Computer Vision (CVPR)*. pp. 12248–12258 (2023)
57. Xu, H., Xie, S., Tan, X.E., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying clip data. *arXiv preprint arXiv:2309.16671* (2023)
58. Xu, X., Xiong, T., Ding, Z., Tu, Z.: Masqclip for open-vocabulary universal image segmentation. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 887–898 (2023)
59. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *arXiv preprint arXiv:2308.02487* (2023)
60. Yuan, H., Li, X., Zhou, C., Li, Y., Chen, K., Loy, C.C.: Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. *arXiv preprint* (2024)
61. Zhang, X., Zeng, H., Liu, X., Yu, Z., Zheng, H., Zheng, B.: In situ holothurian noncontact counting system: A general framework for holothurian counting. *IEEE Access* **8**, 210041–210053 (2020)
62. Zheng, Z., Chen, Y., Zhang, J., Vu, T.A., Zeng, H., Tim, Y.H.W., Yeung, S.K.: Exploring boundary of gpt-4v on marine analysis: A preliminary case study. *arXiv preprint arXiv:2401.02147* (2024)
63. Zheng, Z., Ha, T.S., Chen, Y., Liang, H., Chui, A.P.Y., Wong, Y.H., Yeung, S.K.: Marine video cloud: A cloud-based video analytics platform for collaborative marine research. In: *OCEANS*. pp. 1–6. IEEE (2023)
64. Zheng, Z., Liang, H., Hua, B.S., Wong, Y.H., Ang, P., Chui, A.P.Y., Yeung, S.K.: Coralscop: Segment any coral image on this planet. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 28170–28180 (2024)
65. Zheng, Z., Xin, Z., Yu, Z., Yeung, S.K.: Real-time gan-based image enhancement for robust underwater monocular slam. *Frontiers in Marine Science* (2023)
66. Zheng, Z., Zhang, J., Vu, T.A., Diao, S., Tim, Y.H.W., Yeung, S.K.: Marinegpt: Unlocking secrets of ocean to the public. *arXiv preprint arXiv:2310.13596* (2023)

67. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16793–16803 (2022)
68. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR). pp. 633–641 (2017)
69. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision (ECCV). pp. 696–712. Springer (2022)
70. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
71. Zhuang, P., Wang, Y., Qiao, Y.: Wildfish: A large benchmark for fish recognition in the wild. In: ACM international conference on Multimedia (ACM MM). pp. 1301–1309 (2018)
72. Zhuang, P., Wang, Y., Qiao, Y.: Wildfish++: A comprehensive fish benchmark for multimedia research. *IEEE Transactions on Multimedia (TMM)* **23**, 3603–3617 (2020)
73. Ziqiang, Z., Yaofeng, X., Haixin, L., Zhibin, Y., Yeung, S.K.: Coralvos: Dataset and benchmark for coral video segmentation. arXiv preprint arXiv:2310.01946 (2023)