This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MARIS: Marine Open-Vocabulary Instance Segmentation

Bingyu Li[1,2*] Feiyu Wang[3,2] Da Zhang[4,2] Zhiyuan Zhao[2] Junyu Gao[2] Xuelong Li[2†]
[1]University of Science and Technology of China, China
[2]Institute of Artificial Intelligence (TeleAI), China Telecom, China
[3]Fudan University, China
[4]Northwestern Polytechnical University, China

Abstract

*Most existing underwater instance segmentation approaches are constrained by close-vocabulary prediction, limiting their ability to recognize novel marine categories. To support evaluation, we introduce **MARIS** (Marine Open-Vocabulary Instance Segmentation), the first large-scale fine-grained benchmark for underwater Open-Vocabulary (OV) Instance segmentation (UOVIS), featuring a limited set of seen categories and diverse unseen categories. Although OV instance segmentation has shown promise on natural images, our analysis reveals that transfer to underwater scenes suffers from severe visual degradation (e.g., color attenuation) and semantic misalignment caused by lack underwater class definitions. To address these issues, we propose a unified framework with two complementary components. The Geometric Prior Enhancement Module (**GPEM**) leverages stable part-level and structural cues to maintain object consistency under degraded visual conditions. The Semantic Alignment Injection Mechanism (**SAIM**) enriches language embeddings with domain-specific priors, mitigating semantic ambiguity and improving recognition of unseen categories. Experiments show that our framework consistently outperforms existing OV baselines both In-Domain and Cross-Domain setting on MARIS, establishing a strong foundation for future underwater perception research. The code is Here[1].*

1. Introduction

Instance segmentation in underwater imagery plays a crucial role in applications such as marine biodiversity monitoring, autonomous underwater vehicles, and environmental conservation [14, 21]. The goal of this task is to accurately localize and categorize marine objects with pixel-level instance masks. However, existing approaches heavily rely on dense pixel-wise annotations, which are extremely costly to obtain in underwater environments[19]. Furthermore, conventional models are limited by the restricted set of training categories, hindering their ability to generalize to unseen species or adapt to novel marine exploration scenarios[1, 11].

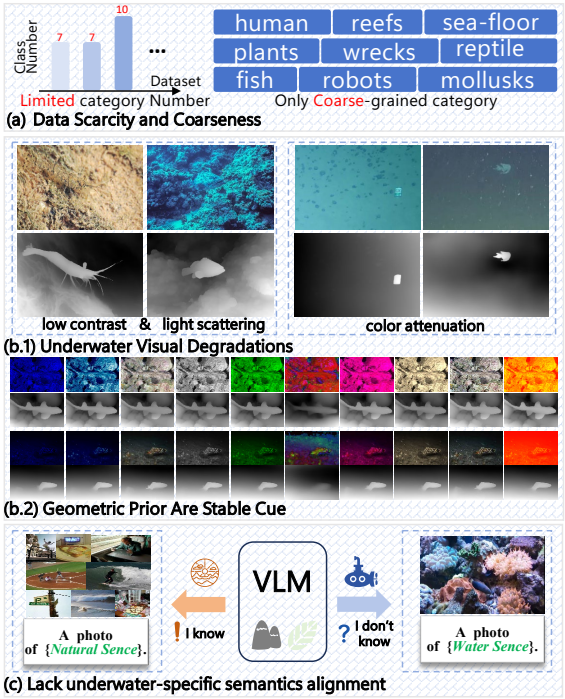

Figure 1. **The challenges of transferring OV instance segmentation to underwater scenarios** in terms of (a) datasets and (b-c) methods, which have motivated the contributions of this study.

OV learning [2, 4, 53] offers a promising solution by enabling models to recognize novel categories without exhaustive labeling or retraining. While OV segmentation models have demonstrated strong performance on terrestrial and natural images, their direct transfer to underwater imagery remains unexplored.

We analyze the OV learning paradigm in the context of

*Work done during an internship at TeleAI.
†Corresponding Author
[1]https://github.com/LiBingyu01/MARIS

underwater scenarios and identify several key challenges. The first challenge is **data scarcity and coarse-grained annotations**: OV segmentation typically relies on large-scale[32], diverse annotations[6], as illustrated in Fig. 1(a) existing underwater datasets, such as UIIS10K [21] and USIS10K [25], provide labels for only less than 20 categories. Moreover, many underwater organisms are crudely grouped into broad classes such as "fish" and "plants." For instance, `Amphora` and `Blue Parrotfish` are just categorized as "fish,". This coarse labeling severely restricts OV transfer. To overcome this limitation, we present the MARIS dataset, which introduces 158 fine-grained category labels with diverse instances, establishing the first benchmark for OV segmentation in underwater environments.

Even with sufficiently annotated data, transferring models to underwater imagery remains challenging due to the unique characteristics of underwater environments[39, 49]. Unlike terrestrial images, underwater images are captured through a medium(water) that induces significant visual degradations[2] in Fig. 1(b.1). For instance, organisms whose body colors closely resemble the surrounding environment can become visually indistinguishable, and objects may become partially or fully occluded due to lighting conditions or water turbidity. In essence, such degradations render **visual appearance cues unstable** in underwater scenes.

On the other hand, despite these visual degradations, many underwater objects retain stable geometric properties that can serve as reliable cues. As shown in Fig. 1(b.2), our preliminary visualization experiments demonstrate that although fish may lose distinctive color patterns, their body shapes and fin structures remain discernible. Likewise, coral colonies exhibit characteristic geometric growth patterns even when their surface textures are degraded. Motivated by this observation, we propose a **Geometric Prior Enhancement Module (GPEM)**, which exploits geometric priors to alleviate visual degradations in underwater imagery.

Beyond visual degradation, another distinct property of underwater imagery is **semantic ambiguity** caused by and insufficient language priors. As shown in Fig. 1(c), current VLM, trained primarily on terrestrial data, fail to capture such fine-grained marine semantics. Motivated by this, we propose a **Semantic Alignment Injection Mechanism (SAIM)**, which integrates domain-specific knowledge via prompt augmentation and embedding enrichment. By guiding the model with enriched underwater semantics, SAIM mitigates category ambiguity and improves recognition of unseen species. Together, GPEM and SAIM function complementarily, addressing the core challenges of visual degradation and semantic ambiguity in underwater imagery from distinct yet synergistic perspectives.

[2] *color attenuation*, *low contrast*, and *light scattering*

Our contributions can be summarized as follows:

- **New benchmark.** We introduce **MARIS**, the first large-scale fine-grained dataset for OV underwater instance segmentation, addressing the limitations of existing datasets with coarse-grained annotations.
- **Novel framework.** We propose two complementary modules: **GPEM**, which leverages stable geometric priors to alleviate the impact of underwater visual degradations, and **SAIM**, which integrates domain-specific semantic knowledge to resolve ambiguity in marine category recognition.
- **Comprehensive evaluation.** Extensive experiments on MARIS demonstrate that our framework achieves state-of-the-art performance on underwater instance segmentation and shows strong generalization to unseen marine categories.

2. Related Work

Underwater Segmentation Underwater scene segmentation has been supported by several datasets. Early benchmarks such as SUIM [15], MAS3K [10], and DUT-USEG [31] provided foundational data but were limited in category diversity or annotation quality. More recent efforts, including UIIS [24], UIIS10K [21], USIS10K [25], and Seaclear [7], expanded scale and scope, while USIS16K [13] further introduced large-scale pixel-level salient instance masks with multi-level labels. Nonetheless, these datasets remain constrained for OV segmentation due to coarse taxonomies and limited category coverage. Beyond data, underwater vision faces inherent challenges such as color attenuation, low contrast, and scattering. Traditional methods adapt general segmentation architectures with underwater-specific priors and enhancements [11, 24, 37, 52]. Representative models include UWSegFormer [58], UISS-Net [12], and CaveSeg [1]. Recently, Deep Learning [9, 55–57] and Vision Foundation Models (VFMs)[20, 29, 36, 44], particularly SAM-based approaches [14, 21, 25], have been adapted for underwater tasks. These developments highlight VFMs as a promising direction for robust, scalable segmentation in aquatic environments. Although underwater segmentation has progressed considerably, large-scale training for OV object segmentation remains unexplored. In this work, we take a step toward addressing this gap.

Open-Vocabulary Segmentation Open-Vocabulary Segmentation (OVS) seeks to segment image regions according to an open-world vocabulary, enabling generalization beyond pre-defined categories. Early works adapted vision-language models (VLMs)[30, 38, 40, 42, 43, 54] such as CLIP [34] to pixel-level tasks. LSeg [17] employed pixel-wise contrastive learning for zero-shot segmentation, while proposal-based approaches, including MaskFormer [3] and ZSSeg [46], generated class-agnostic masks

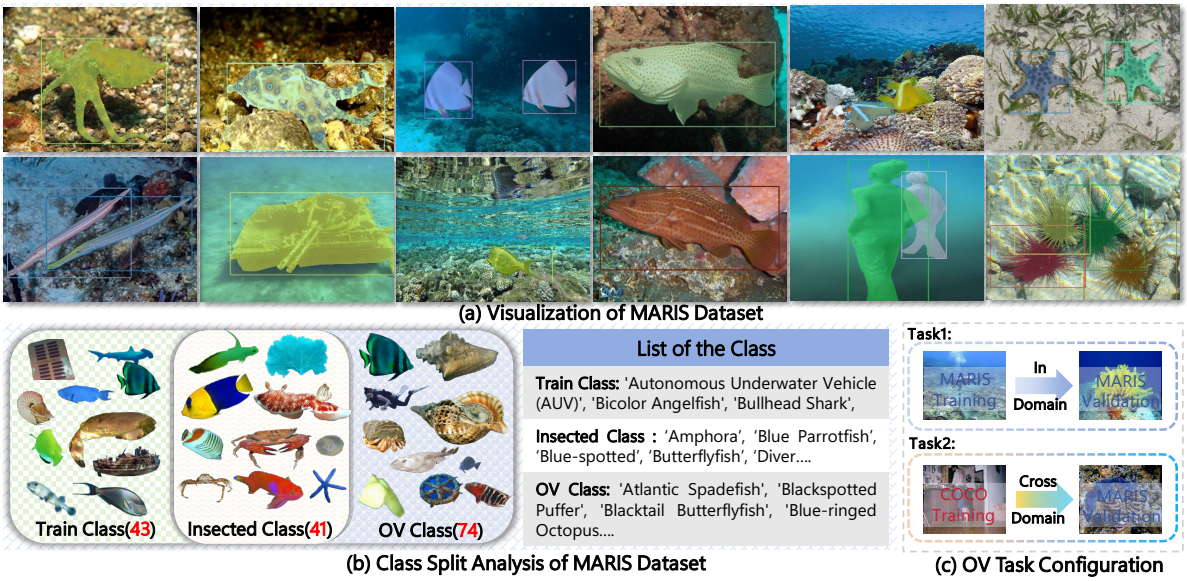

Figure 2. **Visualization and analysis of the MARIS dataset.** (a) Sample images from the MARIS dataset with object annotations. (b) Class split analysis, including Train Class, Insected Class, and OV Class. (c) Configuration of OV tasks, covering in-domain and cross-domain settings.

for subsequent classification. FreeSeg [33] unified this paradigm with a one-shot framework maintaining consistent parameters across tasks. Later methods exploited dense features and improved efficiency.MaskCLIP [8] extracted patch-level features directly from CLIP, preserving vision-language alignment. SAN [47] introduced side adapters into frozen CLIP backbones, while ODISE [45] employed diffusion-based image-text embeddings for mask generation. Other one-stage methods [16, 53], extended the single-stage paradigm by introducing a matching loss to enforce better pixel–text alignment. Recent work emphasized structural priors and cost aggregation. SCAN [28] enhanced feature quality via self-supervised learning. Other methods such as CAT-Seg and ERR-Seg [2, 4] transferred CLIP knowledge through cost aggregation without explicit mask categorization, reducing complexity [18, 41]. Other approaches, such as frequency-domain modules [48] and adaptive fusion of SAM and CLIP outputs [35], further improved generalization and adaptability. In this paper, we make the first attempt to explore the OVS task in underwater scenarios and propose a novel model paradigm to adapt OVS models to the underwater domain.

3. MARIS Benchmark

As a foundational step toward underwater OVS, we pioneer the construction of a dedicated benchmark, which incorporates precise evaluations.

3.1. Data Collection and Annotation

Our benchmark, **MARIS** (Marine Instance Segmentation), is developed to overcome the limitations of existing underwater segmentation benchmarks, which remain scarce and coarse-grained. Public datasets such as UIIS [24] and USIS10K [25] contain fewer than 20 annotated categories and group diverse organisms into broad groups such as "fish" or "plants" class. Such coarse labeling restricts OV models from generalizing to unseen or fine-grained categories. To address this gap, MARIS (Fig. 2(a)) is curated from multiple complementary sources [24, 25], including several recently released underwater datasets [13, 15, 21], which we systematically re-annotate and extend based on [13]. In total, MARIS comprises over 16K underwater images categorized into 9 super-classes and 158 fine-grained subclasses. Unlike prior benchmarks, our annotations explicitly distinguish detailed categories—for example, the "fish" super-class is refined into 76 distinct species (see Appendix for details). This ensures coverage of diverse marine organisms, artificial objects, and natural substrates. We list some of the categories in Fig. 2(b). All annotations are provided at the instance level with pixel-accurate masks, enabling detailed structural analysis. This fine-grained labeling not only enhances semantic richness but also establishes MARIS as *the first benchmark* to support rigorous evaluation of OV instance segmentation in underwater environments.

3.2. Dataset Split and Experimental Settings

The MARIS dataset contains 5,712 training images and 10,439 validation images. While the initial category ratio was designed as 1:2, the presence of multiple instances per image resulted in 84 training categories and 115 validation categories, with 41 overlapping between them. Con-

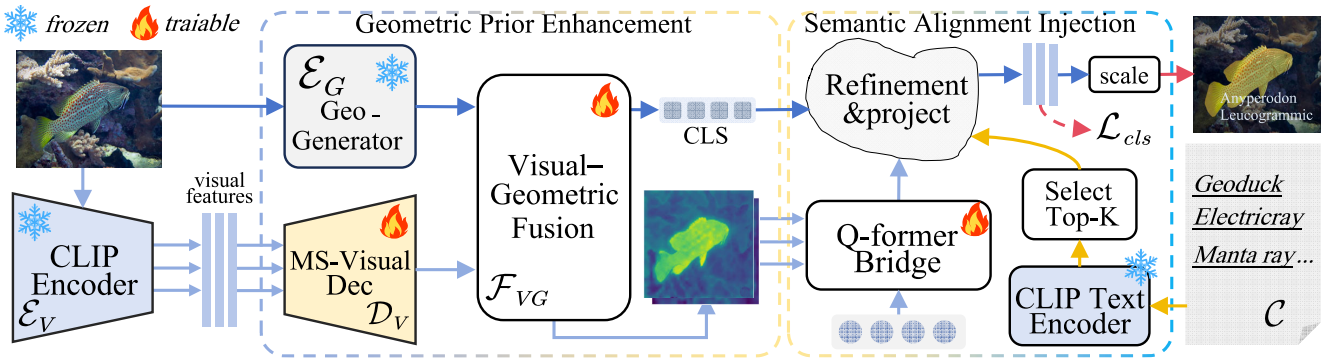

Figure 3. **Overall framework of the proposed Method.** The Geometric Prior Enhancement module strengthens structural representations via visual–geometric fusion and transformer-based query refinement. The Semantic Alignment Injection mechanism align category semantics with degraded underwater conditions.

sequently, shown in Fig. 2(b), the training set contains 43 exclusive classes, and the testing set contains 74 exclusive classes, more details are in the Appendix E-F.

3.2.1. Task Configuration

Based on this split, we define two experimental settings as illustrated in Fig. 2 (c). **In-domain.** For in-domain evaluation, models are trained on the MARIS training set and evaluated on the validation set. **Cross-domain.** To further assess cross-domain generalization, we design a more challenging setting where models are trained on COCO[27] and evaluated on the MARIS validation set. Since COCO and MARIS share no category overlap, this configuration rigorously tests the ability of models to adapt from a generic dataset to the underwater domain.

4. Method

4.1. Problem Definition

Formally, given an input image $\mathbf{I}$ and a set of textual category descriptions $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$, an OVIS model aims to produce a set of instance masks $\mathbf{M} = \{m_1, m_2, \ldots, m_k\}$ and corresponding labels $\mathbf{Y} = \{y_1, y_2, \ldots, y_k\}$, where each $y_i \in \mathcal{C}$ may represent categories that are unseen during training.

4.2. Overall Architecture

Given an input underwater image $\mathbf{I}$, the processing pipeline of MARIS can be expressed as:

$$\mathbf{F}_G = \mathcal{E}_G(\mathbf{I}), \quad \mathbf{F}_V = \mathcal{E}_V(\mathbf{I}), \tag{1}$$

where $\mathbf{F}_G$ denotes the geometric prior features extracted by the frozen Geo-Generator, and $\mathbf{F}_V$ represents the visual features from the frozen CLIP visual encoder. The multi-scale visual decoder $\mathcal{D}_V$ processes $\mathbf{F}_V$ and fuses it with $\mathbf{F}_G$ via the visual-geometric fusion module $\mathcal{F}_{VG}$:

$$\mathbf{F}_{VG} = \mathcal{F}_{VG}\big(\mathcal{D}_V(\mathbf{F}_V), \mathbf{F}_G\big), \tag{2}$$

producing the enhanced visual-geometric representation $\mathbf{F}_{VG}$ along with a global [CLS] token. The Semantic Alignment Injection Mechanism (SAIM) then refines these features with semantic embeddings $\mathbf{E}_T$ generated by the frozen CLIP text encoder:

$$(\mathbf{Y}_{\text{cls}}, \mathbf{M}) = \text{SAIM}(\mathbf{F}_{VG}, \mathbf{E}_T). \tag{3}$$

The refined feature representation $\mathbf{Y}_{\text{cls}}$ and $\mathbf{M}$ are used to jointly supervise the model through the classification loss $\mathcal{L}_{cls}$ and the mask loss $\mathcal{L}_{mask}$.

4.3. Geometric Prior Enhancement Module

The GPEM is designed for fuse multi-scale CLIP visual features with depth-derived geometric priors, producing enhanced representations that combine semantic context with structural information.

Multi-scale Visual & Geometric Generator Given hierarchical features $\{\mathbf{F}_V^{(l)}\}_{l=1}^{L}$ extracted by the frozen CLIP encoder: $\{\mathbf{F}_V^{(l)}\}_{l=1}^{L} = \mathcal{E}_V(\mathbf{I})$, we employ a multi-scale deformable attention module to refine local details and long-range dependencies. The outputs include enhanced features at each scale and an aggregated global visual representation $\mathbf{F}_m$:

$$\{\{\tilde{\mathbf{F}}_V^{(l)}\}_{l=1}^{L}, \mathbf{F}_m\} = \text{MS-DeformAttn}\Big(\{\mathbf{F}_V^{(l)}\}_{l=1}^{L}\Big). \tag{4}$$

To incorporate reliable structural cues, we use a frozen depth encoder [50, 51] to produce multi-scale geometric features $\{\mathbf{F}_G^{(l)}\}_{l=1}^{L}$ and a global depth token $\mathbf{g}_{\text{cls}}$:

$$\{\{\mathbf{F}_G^{(l)}\}_{l=1}^{L}, \mathbf{g}_{\text{cls}}\} = \mathcal{E}_G(\mathbf{I}). \tag{5}$$

Visual–Geometric Feature Fusion $\mathcal{F}_{VG}$: To integrate multi-scale visual and geometric representations, both modalities are first projected into a shared latent space:

$$\hat{\mathbf{F}}_V^{(l)} = W_V^{(l)}\tilde{\mathbf{F}}_V^{(l)}, \qquad \hat{\mathbf{F}}_G^{(l)} = W_G^{(l)}\mathbf{F}_G^{(l)}. \tag{6}$$

An adaptive weight is then computed for each scale:

$$\alpha^{(l)} = \sigma\Big(W_{\alpha}^{(l)}[\hat{\mathbf{F}}_V^{(l)} \,\|\, \hat{\mathbf{F}}_G^{(l)}]\Big), \tag{7}$$

and the fused feature is obtained as:

$$\mathbf{F}_{VG}^{(l)} = \mathrm{MLP}\Big(\hat{\mathbf{F}}_V^{(l)} + \alpha^{(l)} \odot \hat{\mathbf{F}}_G^{(l)}\Big), \tag{8}$$

where σ denotes the sigmoid function, $\|$ indicates concatenation, and $\odot$ is element-wise multiplication. This formulation allows multi-scale geometric cues to be adaptively injected, ensuring that structural depth information complements fine-grained visual features effectively.

Geometry-based Visual & Semantic Bridge To extract effective visual representations and bridge them with semantic information, we employ a lightweight Q-Former (a N-layer transformer encoder always used in VLM [5, 23] to bridge visual and semantic features). The fused geometric-visual features $\mathbf{F}_{VG}^{(l)}$ are processed by the Q-Former to update the query embeddings $\mathbf{Q} \in \mathbb{R}^{N_Q \times C}$, and the final geometry-informed queries are obtained by aggregating outputs across all scales.

4.4. Semantic Alignment Injection Mechanism

We design the **Semantic Alignment Injection Mechanism (SAIM)** from two complementary perspectives: (1) introducing underwater-aware textual prompts and adaptive template selection, and (2) incorporating geometry-based global priors to enrich category representations.

Adaptation to Underwater Scenes Generic language prompts in VLMs often fail to capture underwater-specific semantics, where degradations such as scattering, low contrast, and color attenuation distort object appearance [22, 34]. To address this, we introduce **underwater prompts** as environment-aware priors into the text encoder. These prompts encode five complementary aspects of underwater scenes: (*i*) environmental context, (*ii*) water medium and visibility, (*iii*) illumination and perception, (*iv*) depth cues, and (*v*) scene interactions, producing refined text embeddings that are consistent with underwater visual features.

Nevertheless, upon closer examination, we found that not all templates contribute equally; some may even introduce noise under degraded conditions. For example, in low-light scenarios, certain images can be effectively matched with prompts such as `a <class> in low visibility conditions`, yet such matches tend to be diluted when averaged with other less relevant prompts. To adaptively select the most reliable templates, we compute the similarity between visual features and all textual templates for each category. We rank the templates according to the average similarity across spatial positions and select the top-N templates with the highest scores (detailed in Appendix D).

Category Discrimination We fuse the global depth token $\mathbf{g}_{\text{cls}}$ with the aggregated mask features $\mathbf{F}_m$ to obtain enhanced representations $\mathbf{F}_f$. The compact pooled feature $\mathbf{F}_c = \mathrm{Pool}(\mathbf{F}_f)$ is first combined with the adapted text embeddings $\mathbf{E}_T$ to produce the classification predictions:

$$\mathbf{Y}_{\text{cls}} = \mathbf{F}_c \odot \hat{\mathbf{E}} \in \mathbb{R}^{Q \times C}. \tag{9}$$

Meanwhile, the global depth token $\mathbf{g}_{\text{cls}}$ is fused with the aggregated mask features $\mathbf{F}_m$ to guide the query embeddings $\mathbf{Q}$ and produce the mask: $\mathbf{M} \in \mathbb{R}^{Q \times H \times W}$.

4.5. Training

During training, the model is optimized with a classification loss $\mathcal{L}_{\text{cls}}$,

$$\mathcal{L}_{\text{cls}} = \mathrm{CrossEntropy}(\mathbf{Y}_{\text{cls}}, \mathbf{Y}_{\text{gt}}). \tag{10}$$

implemented as a binary cross-entropy between the predicted and ground-truth categories, and a mask loss $\mathcal{L}_{\text{mask}}$,

$$\mathcal{L}_{\text{mask}} = \mathrm{DiceLoss}(\mathbf{M}, \mathbf{M}_{\text{gt}}) + \mathrm{BCE}(\mathbf{M}, \mathbf{M}_{\text{gt}}), \tag{11}$$

following the same formulation as MaskFormer[53] to supervise the predicted instance masks. Both losses are combined to guide the model toward accurate category recognition and precise spatial segmentation.

5. Experiments And Results

5.1. Experimental Details

All experiments are conducted on four NVIDIA RTX 4090 GPUs (24GB memory) with the batch size of 16. We evaluate two experimental settings (in- and cross-domian) to comprehensively assess the proposed approach. The reproduction of comparative methods is detailed in Appendix B.

5.2. Main Experiments

Experiments for In-Domain Task Table 1 reports results on both intersection and OV categories. MARIS consistently outperforms all competing methods under different backbones. With ConvNeXt-B, MARIS achieves 52.68 mAP on intersection classes and 39.77 mAP on OV classes, surpassing the strongest baseline by over 4 points. The improvement is further amplified with ConvNeXt-L, where MARIS reaches 61.55 mAP and 54.02 mAP on intersection and OV categories, respectively. Overall, MARIS delivers the best results across all metrics, with particularly notable gains under AP_{75}, indicating more accurate and robust mask predictions. These results demonstrate that our method effectively *enhances category discrimination and generalization*, leading to superior performance in underwater OV segmentation.

Table 1. **Comparison of in-domain open-vocabulary segmentation performance** across different methods and backbones. Our method consistently outperforms previous approaches on both ConvNext-B and ConvNext-L backbones. Rows with gray background highlight our method and its improvement over the second-best approach.

Method	Publication	Backbone	Intersection Class			Open-Vocabulary Class			Overall Class		
			mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
OVSeg[26]	CVPR'23	ViT-B	37.52	48.51	43.26	27.21	33.65	30.38	30.95	39.02	35.47
ODISE[45]	CVPR'23	ViT-B	41.89	50.74	46.83	30.26	35.68	32.54	34.71	41.56	38.12
SAN[47]	CVPR'23	ViT-B	43.26	52.18	48.05	31.57	37.09	34.02	36.05	43.06	39.26
FCCLIP[53]	NeurIPS'23	ConvNext-B	47.78	57.22	52.44	34.53	39.84	37.15	39.26	46.03	42.60
MAFT+[16]	ECCV'24	ConvNext-B	48.15	58.26	54.57	35.72	40.67	38.88	40.08	47.16	43.33
EOVSeg[32]	AAAI'25	ConvNext-B	37.98	48.95	41.55	27.48	33.89	29.56	31.22	39.26	33.83
Our Method	—	ConvNext-B	**52.68**	**61.56**	**57.33**	**39.77**	**45.78**	**42.68**	**44.37**	**51.41**	**47.90**
Ours vs 2nd	—	—	↑4.53	↑3.30	↑2.76	↑4.05	↑5.11	↑3.80	↑4.29	↑4.25	↑4.57
OVSeg[26]	CVPR'23	ViT-B	48.96	57.92	53.64	44.63	51.89	48.25	46.41	54.23	50.36
ODISE[45]	CVPR'23	ViT-B	49.32	58.75	54.26	45.18	52.64	48.93	46.95	55.02	51.07
SAN[47]	CVPR'23	ViT-B	50.17	59.63	55.08	46.05	53.47	49.76	47.78	55.86	51.92
FCCLIP[53]	NeurIPS'23	ConvNext-L	54.29	63.33	58.37	50.99	58.66	54.57	52.17	60.33	55.92
MAFT+[16]	ECCV'24	ConvNext-L	55.32	64.24	59.42	51.54	59.44	55.74	53.41	61.36	58.88
EOVSeg[32]	AAAI'25	ConvNext-L	51.72	63.16	55.57	48.32	57.26	51.53	49.53	59.36	53.04
Our Method	—	ConvNext-L	**61.55**	**71.02**	**66.04**	**54.02**	**61.54**	**57.44**	**56.71**	**64.92**	**60.51**
Ours vs 2nd	—	—	↑6.23	↑6.78	↑6.62	↑2.48	↑2.10	↑1.70	↑3.30	↑3.56	↑1.63

Table 2. **Cross-domain open-vocabulary segmentation results.** All models are trained on COCO and evaluated on the MARIS validation set. Rows with gray background highlight our method and its improvement over the second-best approach.

Method	Publication	Backbone	Overall Class		
			mAP	AP_{50}	AP_{75}
OVSeg[26]	CVPR'23	ViT-B	18.95	24.30	19.82
ODISE[45]	CVPR'23	ViT-B	18.51	23.86	19.40
SAN[47]	CVPR'23	ViT-B	19.18	24.63	20.05
FCCLIP[53]	NeurIPS'23	ConvNeXt-B	29.79	36.12	33.50
MAFT+[16]	ECCV'24	ConvNeXt-B	30.05	36.57	34.11
EOVSeg[32]	AAAI'25	ConvNeXt-B	18.90	25.91	21.19
Our Method	—	ConvNeXt-B	**32.62**	**39.60**	**36.65**
Ours vs 2nd	—	—	↑2.57	↑3.03	↑2.54
OVSeg[26]	CVPR'23	ViT-B	30.65	40.78	37.90
ODISE[45]	CVPR'23	ViT-B	32.82	41.95	37.01
SAN[47]	CVPR'23	ViT-B	34.05	42.20	38.26
FCCLIP[53]	NeurIPS'23	ConvNeXt-L	39.46	46.39	43.62
MAFT+[16]	ECCV'24	ConvNeXt-L	40.27	47.89	45.72
EOVSeg[32]	AAAI'25	ConvNeXt-L	35.90	45.33	40.11
Our Method	—	ConvNeXt-L	**46.18**	**54.34**	**51.11**
Ours vs 2nd	—	—	↑5.91	↑6.45	↑5.39

Experiments for Cross-Domain Task Table 2 reports the results of cross-domain OVS, where models are trained on COCO and evaluated on the MARIS validation set. As expected, transferring models across domains leads to a clear performance drop, reflecting the large domain gap between terrestrial and underwater imagery. Methods such as MAFT+ and FCCLIP demonstrate relatively strong generalization, achieving around 30% mAP with ConvNeXt-B backbones. However, EOVSeg struggles significantly, indicating that techniques relying heavily on domain-specific cues may fail in cross-domain scenarios. In contrast, our proposed MARIS framework achieves the best performance across both ConvNeXt-B and ConvNeXt-L backbones, surpassing previous methods by a consistent margin. In particular, MARIS improves the overall mAP from 30.05 to 32.62 with ConvNeXt-B and from 40.27 to 46.18 with ConvNeXt-L, highlighting its *effectiveness in handling the severe visual degradations and semantic discrepancies of underwater environments*.

5.3. Ablation Experiments

Ablation Study of GPEM and SAIM Table 3 reports the impact of GPEM and SAIM on segmentation performance. The baseline without either module achieves the lowest scores. Incorporating improves Intersection Class metrics, while SAIM mainly benefits intersection Class AP_{50} and Overall Class mAP. Notably, the integration of GPEM or SAIM particularly strengthens the model's ability to generalize to OV classes. Combining both modules leads to the best results, with intersection Class mAP of 61.55% and OV Class mAP of 54.02%, demonstrating their complementary effects for enhancing both intersection and OV segmentation.

Effectiveness of Underwater Prompts and Template Selection Table 4 evaluates different underwater prompt strategies. Adding underwater prompts (UW) already improves all metrics compared to using no prompts. Further, template selection consistently boosts performance. Notably, mixed selection strategy not only enhances general segmentation accuracy but also strengthens OV class performance, demonstrating its effectiveness for handling diverse

Table 3. **Ablation study on the effectiveness of GPEM and SAIM components.** All experiments use large backbones for both $\mathcal{E}_G$ and $\mathcal{E}_V$. Rows with gray background indicate the combination of both components, achieving the best performance.

GPEM	SAIM	Intersection Class			Open-Vocabulary Class			Overall Class		
		mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
✗	✗	54.29	63.33	58.37	50.99	58.66	54.57	52.17	60.33	55.92
✓	✗	60.05	68.62	64.61	52.19	58.63	56.05	54.99	62.19	59.10
✗	✓	60.88	70.07	64.84	52.16	58.89	55.59	55.27	62.88	58.89
✓	✓	**61.55**	**71.02**	**66.04**	**54.02**	**61.54**	**57.44**	**56.71**	**64.92**	**60.51**

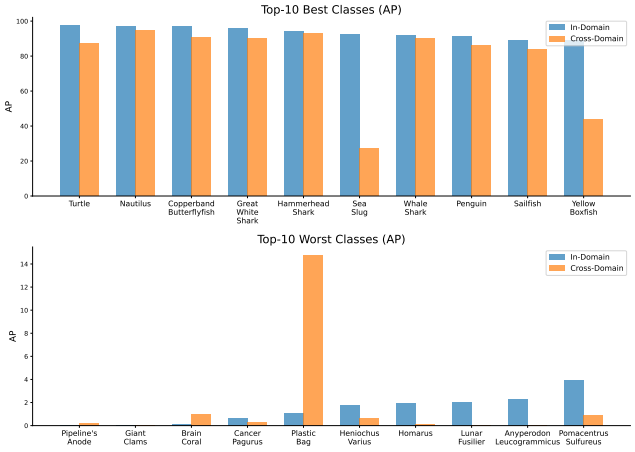

Figure 4. **Top-10 Best and Worst Classes:** Comparison of in-domain and cross-domain AP, illustrating performance drops and gains with geometric-enhanced fusion.

underwater scenes.

Table 4. **Ablation study on prompt strategies.** All experiments use base $\mathcal{E}_G$ and $\mathcal{E}_V$ models. The gray row highlights our final selection strategy. Bold values indicate the best results per column.

Method	Intersection Class			Open-Vocabulary Class			Overall Class		
	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
Template	51.92	60.74	56.31	37.92	42.82	40.60	42.91	49.21	46.20
UWTemplate	**53.99**	**62.92**	58.10	38.29	43.88	40.97	43.89	50.67	47.08
Selection	53.80	62.35	**59.04**	**39.40**	**44.99**	**42.35**	**44.54**	**51.17**	**48.30**

Ablation Experiments of $\mathcal{E}_G$ size Table 5 shows that larger $\mathcal{E}_G$ (vitl) with Convnext-L yields the best in-domain results, while vitb consistently outperforms in cross-domain settings. This indicates that vitl benefits from higher capacity under matched distributions, but vitb strikes a better balance between capacity and generalization, reducing overfitting to in-domain patterns.

Ablation Experiments of Different feature fusion method Table 6 presents the ablation study on the proposed GPEM and SAIM. Without either component, the baseline achieves 52.17% mAP overall. Introducing GPEM brings a clear improvement, raising the overall mAP to 54.99%, which demonstrates its effectiveness in injecting global prompts to reduce domain discrepancies.

Table 5. **Ablation study on the Different $\mathcal{E}_G$ and $\mathcal{E}_V$ size.**

$\mathcal{E}_G$	$\mathcal{E}_V$	in-Domain			Cross-Domain		
		mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
vits	ConvNext-B	42.36	48.83	45.64	30.82	37.62	34.93
vitb	ConvNext-B	**44.54**	51.17	**48.30**	**32.62**	**39.60**	**36.65**
vitl	ConvNext-B	44.37	**51.41**	47.90	32.07	38.55	35.73
vits	Convnext-L	54.22	62.27	57.81	45.75	54.10	50.40
vitb	Convnext-L	55.22	63.37	59.32	**46.18**	**54.34**	**51.11**
vitl	Convnext-L	**56.71**	**64.92**	**60.51**	43.70	51.18	47.98

Table 6. **Performance and efficiency comparison of different fusion methods.** We report overall-class metrics along with GFLOPS and model size. Rows with gray background indicate our proposed fusion method.

Method	mAP	AP_{50}	AP_{75}	GFLOPS	Params (M)
MLP	43.87	50.73	47.36	364G	21.72
add	43.52	50.54	46.81	362G	20.94
alphafusion	**44.54**	**51.17**	**48.30**	365G	22.51

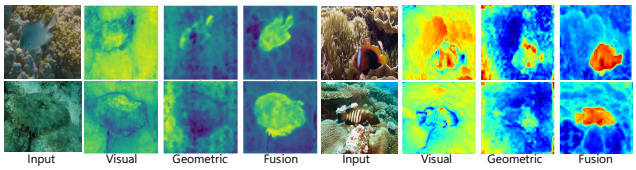

Figure 5. **Qualitative Results** of visual information, geometric information, and their geometric-enhanced fusion, demonstrating clear improvements (viridis on the left and jet on the right).

5.4. Per-Class Performance Analysis

The Fig. 4 highlights the top-10 and bottom-10 classes in terms of AP (More in Appendix J). Overall, high-frequency and visually distinctive categories (e.g., *Shark*, *Turtle*, *Dolphin*) achieve consistently high AP across settings, indicating strong generalization. In contrast, rare or visually ambiguous categories (e.g., *Sponges*, *Anemonefish variants*, *Small invertebrates*) exhibit large performance gaps, reflecting the challenges of recognition in underwater scenes.

5.5. Cross-Domain and In-Domain Analysis

Overall Performance Degradation in Cross-Domain In general, cross-domain performance is lower than in-domain, confirming the effectiveness of domain-specific knowledge. This suggests that incorporating more marine knowledge could further improve cross-domain generalization. On the other hand, it also indicates that our model, trained on natural scenes, can achieve effective cross-domain recognition.

Per-Class Failure Case Analysis We observed several failure cases where AP approaches zero, mostly corresponding to highly specialized species. Small fish such as *Lunar Fusilier* and *Pomacentrus Leucogrammicus* are not well captured by existing VLMs, likely due to insufficient

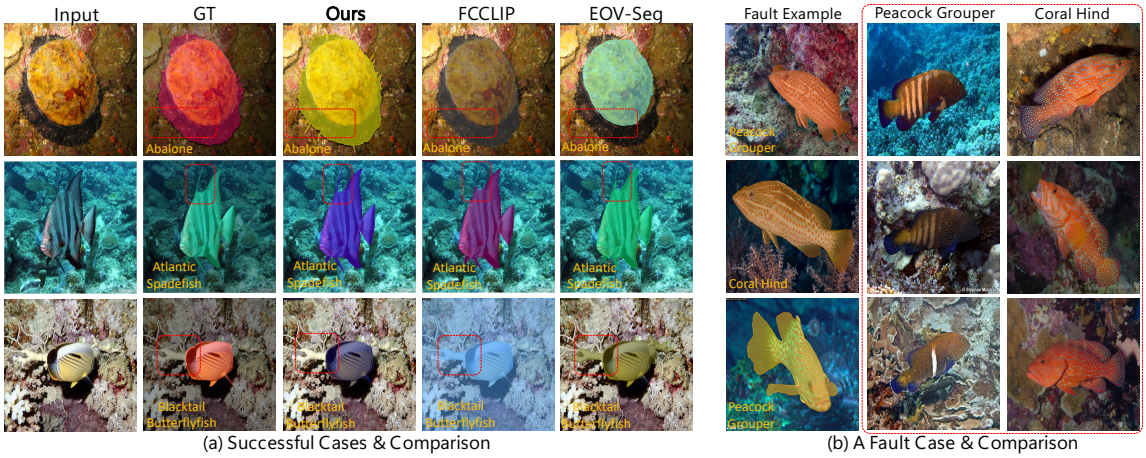

Figure 6. **(a)** Successful cases and comparisons of our method with other approaches. **(b)** A fault case where the model misclassifies `Anyperodon Leucogrammicus` as `Peacock Grouper` or `Coral Hind`. Visually, these species share similarities, which likely leads to confusion in the model's prediction.

semantic encoding. These cases highlight the challenges in cross-domain generalization caused by missing semantic alignment.

Cross-Domain Outperforming In-Domain Interestingly, *Plastic Bag* achieves higher AP in cross-domain evaluation, likely because this object also appears in natural scenes (e.g., COCO dataset). This demonstrates that our model can effectively recognize objects in a new domain if they have been seen during training.

5.6. Analysis of Inference Efficiency and Model Complexity

As shown in Table 7, our method consistently achieves higher in-domain mAP across different backbones. Despite the performance gains, it maintains lower GFLOPS and significantly fewer trainable parameters compared to previous approaches.

Table 7. **Comparison of different methods on overall-class mAP (%) using various backbones.** In-domain (id) performance is reported. Rows with gray background indicate our proposed method.

Method	Backbone	mAP (id)	FLOPS	Trainable Params.	FPS
MAFT+	ConvNext-B	40.08	210G	108.66M	12.20
OVSeg	-	39.26	1.84T	408.55M	-
Our Method (vits)	ConvNext-B	42.36	259G	22.12M	10.53
Our Method (vitb)	ConvNext-B	44.54	365G	22.51M	9.90
Our Method (vitl)	ConvNext-B	44.37	721G	22.77M	7.52
MAFT+	ConvNext-L	53.41	368G	223.22M	9.52
OVSeg	-	39.26	1.84T	408.55M	-
Our Method (vits)	ConvNext-L	54.22	416G	22.33M	8.85
Our Method (vitb)	ConvNext-L	55.22	522G	22.82M	8.20
Our Method (vitl)	ConvNext-L	56.71	878G	23.09M	6.49

5.7. Qualitative Results.

Qualitative Performance on Visual-Geometric Fusion. The qualitative comparisons in Fig. 5 demonstrate that integrating visual and geometric information consistently outperforms using either modality alone.

Qualitative Performance on Segmentation Maps. In the successful cases (Fig. 6(a)), we compare our method with other state-of-the-art approaches, namely FCCLIP and EOV-Seg. For diverse underwater organisms like Abalone, Atlantic Spadefish, and Blacktail Butterflyfish, our method demonstrates superior segmentation performance. More qualitative results are in the Appendix I.

Fault Cases Analysis & Comparison. As shown in the failure case (Fig. 6(b)), our model misclassifies `Anyperodon Leucogrammicus` as `Peacock Grouper` or `Coral Hind`, mainly due to their grouper-like morphology with colorful, patterned bodies. This highlights the need for future models to better disentangle visual similarity from semantic distinctiveness.

6. Conclusion

We introduced MARIS, the first large-scale fine-grained benchmark for open-vocabulary underwater instance segmentation, addressing the limitations of existing datasets with coarse-grained labels. Our framework integrates **GPEM** to leverage stable geometric cues and **SAIM** to enrich language priors, improving segmentation under challenging underwater conditions. Overall, MARIS and the proposed framework provide a robust benchmark and methodology for open-vocabulary segmentation in challenging underwater scenarios.

Limitation: While MARIS covers diverse categories, extreme environments and rare species remain underrepresented, which may limit generalization. Future work will focus on expanding the dataset and enhancing model robustness in such scenarios.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China under Grants 62306241 and U62576284.

References

[1] Adnan Abdullah, Titon Barua, Reagan Tibbetts, Zijie Chen, Md Jahidul Islam, and Ioannis Rekleitis. Caveseg: Deep semantic segmentation and scene parsing for autonomous underwater cave exploration. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3781–3788. IEEE, 2024. 1, 2

[2] Lin Chen, Qi Yang, Kun Ding, Zhihao Li, Gang Shen, Fei Li, Qiyuan Cao, and Shiming Xiang. Efficient redundancy reduction for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2501.17642*, 2025. 1, 3

[3] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021. 2

[4] Seunghyun Cho, Hyunjung Shin, Seunghoon Hong, Anurag Arnab, Paul H. Seo, and Seon Joo Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024. 1, 3

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023. 5

[6] Zhengyuan Ding, Jingdong Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation maskclip. *arXiv preprint arXiv:2208.01343*, 2022. 2

[7] Antun DJuravs, Ben J Wolf, Athina Ilioudi, Ivana Palunko, and Bart De Schutter. A dataset for detection and segmentation of underwater marine debris in shallow waters. *Scientific data*, 11(1):921, 2024. 2

[8] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10995–11005, 2023. 3

[9] Songcheng Du, Yang Zou, Zixu Wang, Xingyuan Li, Ying Li, Changjing Shang, and Qiang Shen. Unsupervised hyperspectral image super-resolution via self-supervised modality decoupling. *International Journal of Computer Vision*, 2026. 2

[10] Zhenqi Fu, Ruizhe Chen, Yue Huang, En Cheng, Xinghao Ding, and Kai-Kuang Ma. Masnet: A robust deep marine animal segmentation network. *IEEE Journal of Oceanic Engineering*, 49(3):1104–1115, 2023. 2

[11] Huilin Ge and Jiali Ouyang. Underwater image segmentation via the progressive network of dual iterative complement enhancement. *Expert Systems with Applications*, 266:126049, 2025. 1, 2

[12] ZhiQian He, LiJie Cao, JiaLu Luo, XiaoQing Xu, JiaYi Tang, JianHao Xu, GengYan Xu, and ZiWen Chen. Uissnet: Underwater image semantic segmentation network for improving boundary segmentation accuracy of underwater images. *Aquaculture International*, 32(5):5625–5638, 2024. 2

[13] Lin Hong, Xin Wang, Yihao Li, and Xia Wang. Usis16k: High-quality dataset for underwater salient instance segmentation. *arXiv preprint arXiv:2506.19472*, 2025. 2, 3

[14] Yang Hong, Xiaowei Zhou, Ruzhuang Hua, Qingxuan Lv, and Junyu Dong. Watersam: Adapting sam for underwater object segmentation. *Journal of Marine Science and Engineering*, 12(9):1616, 2024. 1, 2

[15] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 1769–1776. IEEE, 2020. 2, 3

[16] Siyu Jiao, Hongguang Zhu, Jiannan Huang, Yao Zhao, Yunchao Wei, and Shi Humphrey. Collaborative vision-text representation optimizing for open-vocabulary segmentation. In *European Conference on Computer Vision*, 2024. 3, 6

[17] Bowen Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 2

[18] Bingyu Li, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Fgaseg: Fine-grained pixel-text alignment for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2501.00877*, 2025. 3

[19] Bingyu Li, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. U3m: Unbiased multiscale modal fusion model for multimodal semantic segmentation. *Pattern Recognition*, page 111801, 2025. 1

[20] Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhuo Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, et al. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*, 2026. 2

[21] Hua Li, Shijie Lian, Zhiyuan Li, Runmin Cong, and Sam Kwong. Uwsam: Segment anything model guided underwater instance segmentation and a large-scale benchmark dataset. *arXiv preprint arXiv:2505.15581*, 2025. 1, 2, 3

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 5

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 5

[24] Shijie Lian, Hua Li, Runmin Cong, Suqi Li, Wei Zhang, and Sam Kwong. Watermask: Instance segmentation for underwater imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1305–1315, 2023. 2, 3

[25] Shijie Lian, Ziyi Zhang, Hua Li, Wenjie Li, Laurence Tianruo Yang, Sam Kwong, and Runmin Cong. Diving into

underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset. *arXiv preprint arXiv:2406.06039*, 2024. 2, 3

[26] Feng Liang, Baitao Wu, Xinyu Dai, Kuan Li, Yue Zhao, Han Zhang, Peng Zhang, Peter Vajda, and Daniel Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 6

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[28] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3500, 2024. 3

[29] Yuanzhe Liu, Jingyuan Zhu, Yuchen Mo, Gen Li, Xu Cao, Jin Jin, Yifan Shen, Zhengyuan Li, Tianjiao Yu, Wenzhen Yuan, et al. Palm: Progress-aware policy learning via affordance reasoning for long-horizon robotic manipulation. *arXiv preprint arXiv:2601.07060*, 2026. 2

[30] Weijian Ma, Shizhao Sun, Tianyu Yu, Ruiyu Wang, Tat-Seng Chua, and Jiang Bian. Thinking with blueprints: Assisting vision-language models in spatial reasoning via structured object representation, 2026. 2

[31] Zhiwei Ma, Haojie Li, Zhihui Wang, Dan Yu, Tianyi Wang, Yingshuang Gu, Xin Fan, and Zhongxuan Luo. An underwater image semantic segmentation method focusing on boundaries and a real underwater scene semantic segmentation dataset. *arXiv preprint arXiv:2108.11727*, 2021. 2

[32] Hongwei Niu, Jie Hu, Jianghang Lin, Guannan Jiang, and Shengchuan Zhang. Eov-seg: Efficient open-vocabulary panoptic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6254–6262, 2025. 2, 6

[33] Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, et al. Freeseg: Unified, universal and open-vocabulary image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19446–19455, 2023. 3

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 5

[35] Xudong Shan, Di Wu, Guorong Zhu, Yong Shao, Nong Sang, and Changxin Gao. Open - vocabulary semantic segmentation with image embedding balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28412–28421, 2024. 3

[36] Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Matthew Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in vlms. *arXiv preprint arXiv:2506.21656*, 2025. 2

[37] Pengfei Shi, Shen Shao, Yueyue Liu, Xinnan Fan, and Yuanxue Xin. Crackinst: a real-time instance segmentation method for underwater dam cracks. *IEEE Transactions on Instrumentation and Measurement*, 2024. 2

[38] Xiu Su, Qinghua Mao, Zhongze Wu, Xi Lin, Shan You, Yue Liao, and Chang Xu. Large language models driven neural architecture search for universal and lightweight disease diagnosis on histopathology slide images. *npj Digital Medicine*, 8(1):682, 2025. 2

[39] Quang Trung Truong, Wong Yuk Kwan, Duc Thanh Nguyen, Binh-Son Hua, and Sai-Kit Yeung. Autv: Creating underwater video datasets with pixel-wise annotations. *arXiv preprint arXiv:2503.12828*, 2025. 2

[40] Zhongze Wu, Hongyan Xu, Yitian Long, Shan You, Xiu Su, Jun Long, Yueyi Luo, and Chang Xu. Detecting any instruction-to-answer interaction relationship: Universal instruction-to-answer navigator for med-vqa. In *Forty-first International Conference on Machine Learning*, 2024. 2

[41] Bo Xie, Jie Cao, Jing Xie, Fahad Shahbaz Khan, and Youtao Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3426–3436, 2024. 3

[42] Yuechen Xie, Jie Song, Huiqiong Wang, and Mingli Song. Training data provenance verification: Did your model use synthetic data from my generative model for training? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23817–23827, 2025. 2

[43] Yuechen Xie, Xiaoyan Zhang, Yicheng Shan, Hao Zhu, Rui Tang, Rong Wei, Mingli Song, Yuanyu Wan, and Jie Song. Spatialqa: A benchmark for evaluating spatial logical reasoning in vision-language models. *arXiv preprint arXiv:2602.20901*, 2026. 2

[44] Feng Xu, Guangyao Zhai, Xin Kong, Tingzhong Fu, Daniel FN Gordon, Xueli An, and Benjamin Busam. Stare-vla: Progressive stage-aware reinforcement for fine-tuning vision-language-action models. *arXiv preprint arXiv:2512.05107*, 2025. 2

[45] Jingyi Xu, Shu Liu, Arash Vahdat, Woojin Byeon, Xinyong Wang, and Stefano De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 3, 6

[46] Ming Xu, Zhen Zhang, Feng Wei, Yixuan Lin, Yukun Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 2

[47] Ming Xu, Zhen Zhang, Feng Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 3, 6

[48] Wenhan Xu, Chen Wang, Xin Feng, Runze Xu, Lei Huang, Zhen Zhang, Lei Guo, and Shuaicheng Xu. Generalization

boosted adapter for open - vocabulary segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 3

[49] Xizhe Xue, Yang Zhou, Dawei Yan, Ying Li, Haokui Zhang, and Rong Xiao. Uvlm: Benchmarking video language model for underwater world understanding. *arXiv preprint arXiv:2507.02373*, 2025. 2

[50] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024. 4

[51] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 4

[52] Dewei Yi, Hasan Bayarov Ahmedov, Shouyong Jiang, Yiren Li, Sean Joseph Flinn, and Paul G Fernandes. Coordinate-aware mask r-cnn with group normalization: A underwater marine animal instance segmentation framework. *Neurocomputing*, 583:127488, 2024. 2

[53] Qingyi Yu, Jiahao He, Xin Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023. 1, 3, 5, 6

[54] Pingrui Zhang, Yifei Su, Pengyuan Wu, Dong An, Li Zhang, Zhigang Wang, Dong Wang, Yan Ding, Bin Zhao, and Xuelong Li. Cross from left to right brain: Adaptive text dreamer for vision-and-language navigation. *arXiv preprint arXiv:2505.20897*, 2025. 2

[55] Ruilin Zhang, Haiyang Zheng, and Hongpeng Wang. Cnmbi: Determining the number of clusters using center pairwise matching and boundary filtering. In *International Conference on Advanced Data Mining and Applications*, pages 262–277. Springer, 2023. 2

[56] Ruilin Zhang, Haiyang Zheng, and Hongpeng Wang. Tdec: Deep embedded image clustering with transformer and distribution information. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 280–288, 2023.

[57] Haiyang Zheng, Ruilin Zhang, and Hongpeng Wang. Deep image clustering based on curriculum learning and density information. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 330–338, 2024. 2

[58] Xin Zuo, Jiaran Jiang, Jifeng Shen, and Wankou Yang. Improving underwater semantic segmentation with underwater image quality attention and muti-scale aggregation attention. *Pattern Analysis and Applications*, 28(2):1–12, 2025. 2