Diving into Underwater: Segment Anything Model Guided Underwater Salient Instance Segmentation and A Large-scale Dataset

Shijie Lian [*1] **Ziyi Zhang** [*2] **Hua Li** [†1 3] **Wenjie Li** [1] **Laurence Tianruo Yang** [1 4 5] **Sam Kwong** [6]
Runmin Cong [7 8]

Abstract

With the breakthrough of large models, Segment Anything Model (SAM) and its extensions have been attempted to apply in diverse tasks of computer vision. Underwater salient instance segmentation is a foundational and vital step for various underwater vision tasks, which often suffer from low segmentation accuracy due to the complex underwater circumstances and the adaptive ability of models. Moreover, the lack of large-scale datasets with pixel-level salient instance annotations has impeded the development of machine learning techniques in this field. To address these issues, we construct the first large-scale underwater salient instance segmentation dataset (USIS10K), which contains 10,632 underwater images with pixel-level annotations in 7 categories from various underwater scenes. Then, we propose an Underwater Salient Instance Segmentation architecture based on Segment Anything Model (USIS-SAM) specifically for the underwater domain. We devise an Underwater Adaptive Visual Transformer (UA-ViT) encoder to incorporate underwater domain visual prompts into the segmentation network. We further design an out-of-the-box underwater Salient Feature Prompter Generator (SFPG) to automatically generate salient prompters instead of explicitly providing foreground points or boxes as prompts in SAM. Comprehensive experimental results show that our USIS-SAM method can achieve superior performance on USIS10K datasets compared to the state-of-the-art methods. Datasets and codes are released on Github.

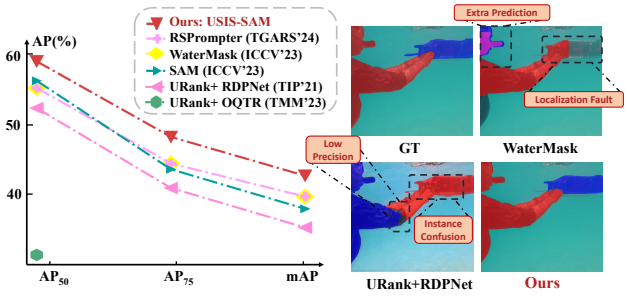


Figure 1. A simple comparison of USIS-SAM and other state-of-the-art methods trained on the USIS10K dataset. Different colors represent different salient instances. URank represents the underwater image enhancement method in UnderwaterRanker (Guo et al., 2023).

1. Introduction

In recent years, fundamental models such as Generative Pre-Trained Transformer (GPT)-4 (Achiam et al., 2023), Language Learning with Adaptive Multi-task Architecture (LLaMA) (Touvron et al., 2023), and Segment Anything Model (SAM) (Kirillov et al., 2023) have made significant progress and greatly contributed to the advancement of the society. Among them, SAM has recently performed well in many segmentation tasks due to its excellent encoder-decoder transformation framework and large training dataset SA-1B. After being fine-tuned or modified appropriately, it has strong potential for segmentation applications in medical images (Wu et al., 2023; Huang et al., 2024), remote sensing images (Chen et al., 2023a), and other fields.

To meet the growing demands for underwater exploration, the deep learning vision community is increasingly focusing on the development and application of underwater vision tasks. Salient Instance Segmentation (SIS), an emerging and promising visual task, aims to segment out visually salient objects in a scene and distinguish individual salient

[*]Equal contribution [1]Hainan University [2]The Hong Kong University of Science and Technology (Guangzhou) [3]Key Laboratory of New Generation Artificial Intelligence Technology & Its Interdisciplinary Applications (Southeast University), Ministry of Education, China [4]Huazhong University of Science and Technology [5]St. Francis Xavier University [6]LIngnan University, School of data science, Hong Kong [7]Shandong University, School of Control Science and Engineering [8]Key Laboratory of Machine Intelligence and System Control, Ministry of Education, Jinan, China. [†]Correspondence to: Hua Li <lihua@hainanu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

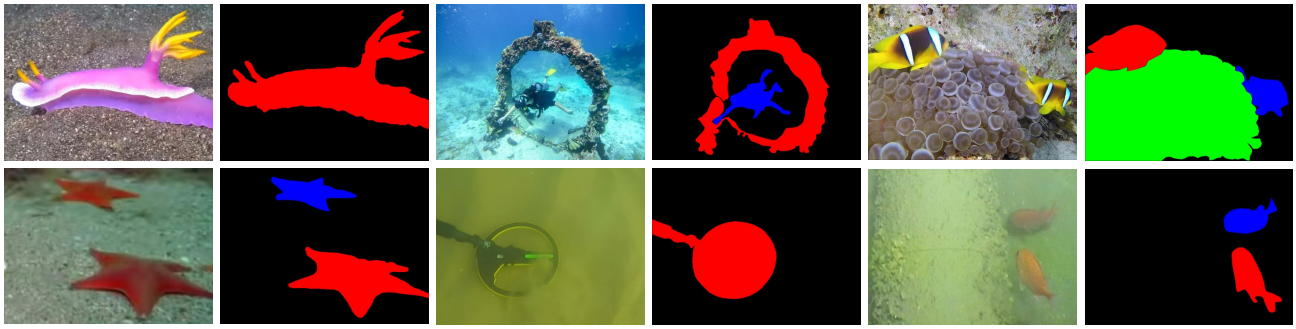


Figure 2. Examples of annotations for various salient instances in USIS10K. The image on the left is the original image and the right is the annotation mask, different colors represent different salient instances. More dataset showings can be found in Appendix A.4.

instances (Tian et al., 2022b), which is beneficial for vision tasks requiring fine-grained scene understanding. Underwater salient instance segmentation (USIS), may benefit for subsequent underwater vision tasks like marine ruins discovery, marine resources exploration, underwater human-robot interaction, and underwater image understanding. However, underwater salient instance segmentation is still a challenging task due to the absence of a large-scale underwater salient instance segmentation dataset for deep learning based methods to effectively train their models. Moreover, underwater images often suffer from unavoidable quality degradation due to wavelength attenuation and scattering. Sea snow formed by plankton in the ocean also introduces noise to the image, which impairs the imaging quality (Akkaynak et al., 2017). In addition, the specific marine species, water patterns, and large number of optical artifacts in underwater scenes make the visual content of underwater images quite different from that of land-based images (Akkaynak & Treibitz, 2018).

Directly transferring conventional SIS methods for land images to underwater scenes may struggle to achieve ideal performance attribute to the domain gap of intrinsic characteristics and extrinsic circumstances between land and underwater living. They may lose model performance due to low-quality underwater imagery and unique underwater semantic features. As a result, even state-of-the-art SIS models trained on large-scale land-based datasets are not directly applicable to underwater imagery. For example in Figure 1, the performance of state-of-the-art methods OQTR (Pei et al., 2023) and RDPNet (Wu et al., 2021) are degraded on mAP, AP_{50}, and AP_{75} when applied in underwater. Moreover, we try to combine the SIS method with enhancement pre-processing. Nevertheless, this approach does not seem to work well since it is hard to jointly train multiple tasks to achieve optimal performance. However, as shown in Figure 1 and Table 2, the performance of such approach is not even able to exceed that of WaterMask (Lian et al., 2023), an underwater instance segmentation method that is not specifically designed for saliency but trained end-to-end. Establishing a large-scale dataset for the USIS task and designing an underwater effective architecture is necessary.

To address these issues, we propose the first large-scale Underwater Image Salient Instance Segmentation dataset (USIS10K) to explore the application of the SIS field in underwater scenes. USIS10K contains 10,632 images and manually labeled pixel-level masks for salient instances in the images. While previous SIS datasets were usually not labeled with category labels for salient instances, research in the field of weakly supervised salient instance segmentation (Tian et al., 2022a) has demonstrated that category labeling helps the network to detect semantically dominant regions, offering supplementary information for SIS task training. Therefore, we labeled each salient instance with a distinct category, including fish, coral reefs, underwater plants, human divers, robots, underwater ruins, and seafloor reefs. Some samples of this dataset are shown in Figure 2. Moreover, we propose an Underwater Salient Instance Segmentation model based on SAM, called USIS-SAM. Since it has been proven that SAM has limited performance in certain tasks due to the lack of domain-specific knowledge (Chen et al., 2023b;a; Wu et al., 2023), inspired by prompt learning, we devise an Underwater Adaptive Visual Transformer (UA-ViT) encoder in the proposed USIS-SAM to adjust image features as soft prompts, aiming to facilitate the model adapt to the complex underwater scenes. In addition, SAM relies on user-supplied point or bounding box coordinates to guide network segmentation. Therefore, we also design an underwater Salient Feature Prompter Generator (SFPG) that can directly generate saliency prompts in an image as input to SAM mask decoder for end-to-end segmentation of salient instances in underwater images.

We have performed extensive experiments to evaluate the effectiveness of our method. First, we compared USIS-SAM with mainstream salient instance segmentation methods (Fan et al., 2019; Wu et al., 2021; Pei et al., 2023) and instance segmentation methods (Kirillov et al., 2023; Lian et al., 2023; Chen et al., 2023a) on the USIS10K to validate the effectiveness of USIS-SAM. As the simple case

shown in Figure 1 our USIS-SAM can segment more accurate masks and achieve the best quantitative performance. In addition, we also performed comparisons on the largest extant land salient instance segmentation dataset, SIS10K (Pei et al., 2023), to validate the generalization ability of our model. The main contributions are concluded as follows:

- We construct the first large-scale dataset, USIS10K, for the underwater salient instance segmentation task, which contains 10,632 images and pixel-level annotations of 7 categories. As far as we know, this is the largest salient instance segmentation dataset, and includes Class-Agnostic and Multi-Class labels simultaneously.
- We first attempt to apply SAM to underwater salient instance segmentation and propose USIS-SAM, aiming to improve the segmentation accuracy in complex underwater scenes. We design Underwater Adaptive ViT Encoder to incorporate underwater visual prompts into the network via adapters, and Salient Feature Prompter Generator to automatically generate salient prompters, guiding an end-to-end segmentation network.
- Extensive public evaluation criteria and large numbers of experiments verify the effectiveness of our USIS10K dataset and USIS-SAM.

2. Related Work

2.1. Segment Anything Model (SAM)

SAM (Kirillov et al., 2023) is a promptable model trained on a large dataset, resulting in a powerful zero-shot generalization. Currently, extension work based on SAM explores its capability boundaries, for example, image inpainting (Yu et al., 2023), medical image analysis (Ma et al., 2024; Zhang & Jiao, 2022), few-shot learning (Zhang et al., 2023; Shi et al., 2023), remote sensing segmentation (Zhang et al., 2024; Chen et al., 2023a), image editing (Xie et al., 2023) and style transfer (Liu et al., 2023). In normal scenes, SAM has excellent generalization capabilities. However, some research results show that SAM already has some limitations in special scenarios (e.g., low-contrast) and specific tasks (e.g., camouflage instance segmentation) (Chen et al., 2023b). Similarly, the complex background interference in marine environments and the widespread quality degradation of underwater images also present significant challenges to the segmentation capabilities of SAM.

2.2. Salient Instance Segmentation (SIS)

SIS is to separate salient objects while distinguishing different salient instances. Most of the existing SIS methods and datasets are designed for land scenes. ILOS dataset (Li et al., 2017) is the first SIS dataset, which contains two thousand images. S4Net (Fan et al., 2019) migrated the Mask R-CNN architecture (He et al., 2017) in instance segmentation to the SIS task, employing the ROIMasking module to exploit the inherent features of the bounding box. Similarly, RDPNet (Wu et al., 2021) augments all feature layers in the feature pyramid to perform multilevel RoIAlign for better mask prediction. The SOC dataset (Fan et al., 2018a) attempts to eliminate the data bias of the saliency dataset, which contains 3,000 images with saliency instances and 3,000 images without it. OQTR (Pei et al., 2023) introduces the transformer structure to the SIS domain, which locates salient instances by query instead of hand-designed anchors. Moreover, they provide a new dataset, SIS10K, containing 10,300 terrestrial images to train various networks.

Table 1. Comparison with existing SIS datasets. Where Label indicates whether each instance has a separate category label, Max indicates the maximum number of instances in a single image.

Dataset	Year	Task	Label	Number	Max
ILSO	2017	SIS	×	2,000	8
SOC	2018	SIS	✓	3,000	8
SIS10K	2023	SIS	×	10,300	9
USIS10K	2024	USIS	✓	10,632	9

3. USIS10K Dataset

3.1. Dataset Collection and Annotation

We have carefully selected 16,000 underwater images from many different natural underwater environments from the Internet and open source underwater datasets (Li et al., 2020; Berman et al., 2021; Lian et al., 2023; Hong et al., 2023; Islam et al., 2020). These images cover underwater vision tasks such as ocean exploration, human-robot intelligent cooperation and underwater autopilot. Then, the salient instances in these images are labeled pixel by pixel and assigned category labels. To ensure the accuracy and objectivity of the labeling as much as possible, each image will be labeled by at least 3 volunteers, and then the final labeling results will come out through voting. After removing images that did not have a consensus in the voting or were unclear the instance category, we ended up with a total of 10,632 images for the USIS10K dataset. For more in-depth information regarding the collection and annotation of the dataset, please refer to Appendix A.

3.2. Dataset Characteristics and Statistics

In this subsection, we illustrate the basic information, characteristics, and challenges of the USIS10K dataset. Some statistical information can also be seen in Figure 3.

Number and Category of Dataset. A large number of underwater images of different scenes, containing various categories, is essential to improve the generalization of the

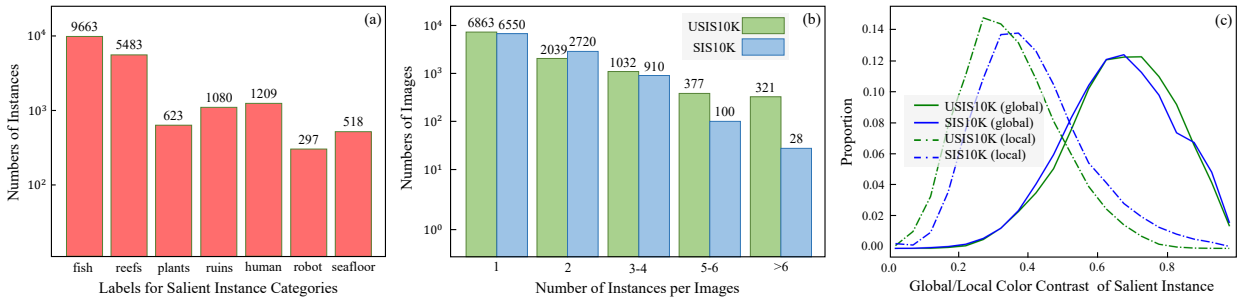
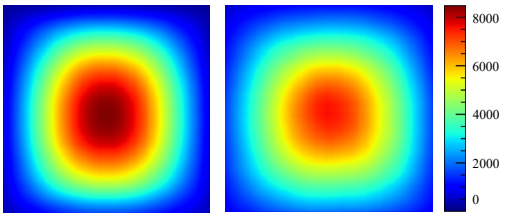


Figure 3. Essential characteristics of the USIS10K dataset. (a) The number of salient instances per category in the USIS10K dataset. (b) Distribution of the number of salient instances per image in the USIS10K dataset. (c) Comparison of USIS10K and SIS10K in global color contrast and local color contrast.



(a) SIS10K (b) USIS10K

Figure 4. A set of salient maps from our dataset and SIS10K.

network in complex marine environments and to avoid overfitting. For this purpose, we collected 10,632 images from 7 categories and divided the dataset into training, validation, and test sets in the ratio of 7:1.5:1.5. As can be seen from Table 1, USIS10K is the first salient instance segmentation dataset for underwater scenes. In addition, in order to make the USIS10K dataset applicable to more downstream tasks, all the annotations are divided into three parts: category labels, salient instance masks, and bounding boxes. This provides the possibility of using the USIS10K dataset for weakly supervised salient instance segmentation (Tian et al., 2022a), salient semantic segmentation (Fan et al., 2018b), and other downstream tasks in underwater scenes.

Number and Size of Salient Instances. In USIS10K, multiple salient instances may exist in a single image. There are 1731 images with more than 3 salient instances in our dataset, accounting for 16.3% of the total. In SIS10K (Pei et al., 2023), there are only 1038 images with more than 3 salient instances, accounting for 10.1% of the total. In the underwater salient detection dataset, USOD10K (Hong et al., 2023), there are only 722 images with more than 3 salient objects, accounting for 7.1% of the total. This not only means that our USIS10K is challenging even for state-of-the-art methods, but also allows it to be used in fields like underwater co-salient instance segmentation (Cong et al., 2022; Zhang et al., 2020; Yu et al., 2018). Furthermore, predicting instances that are too small or too large is a very common but challenging problem. In the USIS10K dataset, the average size of the salient instances is 34,336 pixels (approximately 185×185 pixels), which averaged 10.3% of the image size. There are 3053 salient instances smaller than 1% of the image area, (16.0% of the total), while there are 1733 instances larger than 30% of the image area, (9.1% of the total), which further enhances the challenge of USIS10K.

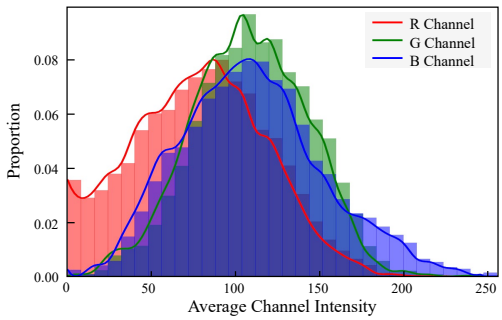
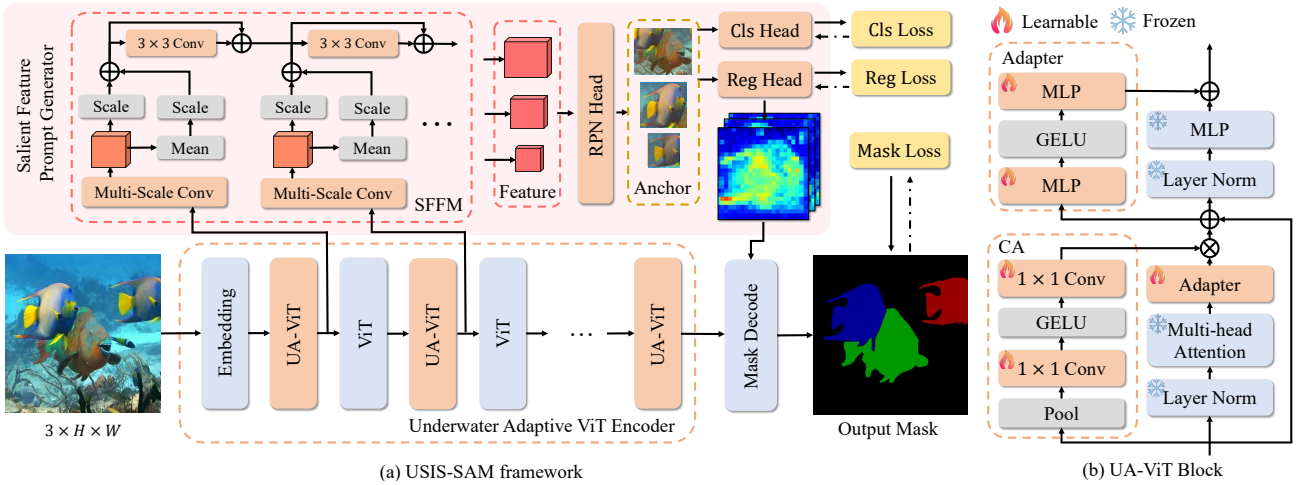


Figure 5. Average channel intensity in USIS10K with proportion.

Color Contrast of Salient Instance. Saliency is often related to the global contrast between foreground and background, and it is critical to check whether salient instances are easy to detect (Li et al., 2014). We calculated the Bhattacharyya distance (Choi & Lee, 2003) between the RGB histogram of each salient instance and the background RGB histogram in USIS10K and SIS10K dataset to measure their global contrast. It can be seen that the global contrast of USIS10K is slightly higher than SIS10K. In addition, similar to the global contrast, we also computed the local contrast in a 5×5 patch on the salient instance boundary. It can be seen that USIS10K has less local contrast than SIS10K due to characteristics like instance clustering in underwater environments (Lian et al., 2023; Hong et al., 2023). This poses a greater challenge for accurate segmentation of salient instance masks in the boundary portion of the network.

Location of Salient Objects. The central bias has been identified as one of the most significant biases in salience datasets (Fan et al., 2018a; Hong et al., 2023) due to the fact that humans have a natural tendency to perceive salient objects by focusing their attention on the center of the observed scene. Figure 4 illustrates a set of salient instance heat maps. It can be seen that while USIS10K has a smaller central deviation compared to SIS10K. Specifically, in the SIS10k dataset, approximately 13.5% of the locations have fewer than 1,000 instances and 32% have fewer than 2,000



(a) USIS-SAM framework (b) UA-ViT Block

Figure 6. (a) USIS-SAM framework. The USIS-SAM framework modifies the SAM by adding the Underwater Adaptive ViT Encoder (in Section 4.1) and the Salient Feature Prompt Generator (in Section 4.2). (b) The structure of UA-ViT. In the figure, SFFM stands for Salient Feature Fusion Module, CA stands for Channel Adapter.

instances, while in our dataset, only 2.75% of the locations have fewer than 1,000 instances and 22.5% have fewer than 2,000 instances.

Channel Intensity of Underwater Images. Optical images inevitably suffer from color attenuation due to the selective absorption of water at different wavelengths. In particular, the attenuation of the red channel is an order of magnitude greater than that of the blue and green channels (Mobley, 2001). As a result, in most cases, underwater images are bluish or greenish. To quantify this attenuation property of the USIS10K, we compute the average channel intensities and probability densities of R, G, and B for each image in Figure 5. It can be seen that R channel has the lowest intensity, but also follows the same trend as G and B channels.

4. USIS-SAM

In this section, we introduce our proposed USIS-SAM, a prompt learning method based on SAM. The framework of it is illustrated in Figure 6 (a). In USIS-SAM, we design the Underwater Adaptive ViT (UA-ViT) to integrate underwater visual prompts into the network via adapter and channel adapter. A Salient Feature Prompt Generator (SFPG) was also designed to aggregate the UA-ViT output features and generate salient prompts to guide end-to-end segmentation.

4.1. Underwater Adaptive ViT Encoder

Due to being trained on the SA-1B dataset (Kirillov et al., 2023), the largest existing natural image segmentation dataset, the SAM image encoder possesses strong capabilities in extracting image features. This makes it a valuable tool for underwater image salient segmentation. However, underwater images often suffer from color distortion, and the SAM image encoder lacks expertise in the underwater domain due to the domain bias between the SA-1B dataset and the complex marine environment. Consequently, the SAM image encoder is unable to fully extract features from underwater images, and its direct application to underwater scenes does not yield satisfactory results. To address this issue, we propose the Underwater Adaptive ViT (UA-ViT) inspired by the Parameter-Efficient Fine-Tuning (PEFT) technique (Ben Zaken et al., 2022). By incorporating a set of adapters into ViT block (as depicted in Figure 6 (b)), the UA-ViT enables a more effective utilization of the SAM image encoder in underwater scenarios.

UA-ViT learns the expertise of the underwater visual prompts via adapters and injects it into the subsequent network. Essentially, the adapter can be regarded as a soft prompt to embed the network, which takes advantage of the fact that the ViT has been trained on large-scale datasets, selectively mask or leverage the original network structure and weights by adjusting the inflow features, and enhances the model's ability to generalize to downstream tasks with only a small amount of training. Therefore, we added two adapters for each ViT block. Specifically, we place the first adapter immediately after multi-head attention layer to adjust for color distortion due to selective absorption of water at different wavelengths. The second adapter on the residual path of the frozen MLP layer, and utilizes parallel branches to aggregate the updated context by summation while maintaining the original MLP features. Where the adapter can be represented as:

$$P = MLP_{out}(\sigma(MLP_{prompt}(F))), \quad (1)$$

where MLP_{prompt} is the linear layer used to generate the underwater task prompts for each adapter, and MLP_{out} is the projection layer used to resize the adaptive features. F is the input feature, and P is the output prompt for each adapter layer. σ is the activation function.

We further propose a channel adapter (CA) module to address the frequency distribution bias due to the selective absorption of wavelengths by the water, which utilizes global attention-aware features to adaptively adjust the importance of each channel, thus significantly improving the representation capability of the network. The channel adapter can be represented as:

$$C = F * Conv_{up}(\sigma(Conv_{down}(Pool(F)))), \quad (2)$$

where C is the output feature after channel adapter, $Conv$ is a 1×1 convolutional layer, and $Pool$ is an average pooling layer. It can be seen that the channel adapter has a very similar structure to the adapter, only the MLP is replaced by the convolution. In fact, they share a similar function, and the channel adapter focuses more on the adjustment of frequency domain information to help the network adapt to the underwater environment.

In order to minimize the number of parameters, we replace only some of the ViT blocks in the SAM encoder with UA-ViT blocks. Specifically, when we use ViT-H (Dosovitskiy et al., 2021) as the backbone of SAM encoder, we replace one layer every two layers starting from the eighth layer.

4.2. Salient Feature Prompt Generator (SFPG)

The USIS task needs the model to automatically recognize and segment each salient object in underwater images. However, SAM requires the user to explicitly provide foreground points, boxes, or texts as prompts to guide the model segmentation. Simply adding the object detection network to identify the location of significant objects and feeding the bounding boxes as prompts into the SAM's prompt encoder would significantly increase the complexity of the model and severely limit the model optimization. Therefore, we design the Salient Feature Prompt Generator to directly predict prompts embedding of salient instances.

The SFPG first designs a Salient Feature Fusion Module (SFFM) to fuse the features output from each UA-ViT block and align them with the prompts required by the SAM decoder. We first aggregate multi-scale features through multi-scale convolutions, suppress irrelevant features, and reasonably capture significant information in images. The multi-scale features of the ith-layer UA-ViT F_i^m can be obtained by:

$$F_i^m = \sum_{s\in3,5,7} Conv_{s*s}(CA(F_i)), \quad (3)$$

where F_i is the features from the ith-layer UA-ViT, $Conv_{s*s}$ are the multi-scale convolutions (3×3, 5×5, and 7×7), and CA means channel adapter. We then balance the multi-scale features using the average residuals to dampen the noise in the features, which can be formulated as:

$$F_i^a = \lambda F_i^m + (1-\lambda)Avg(F_i^m), \quad (4)$$

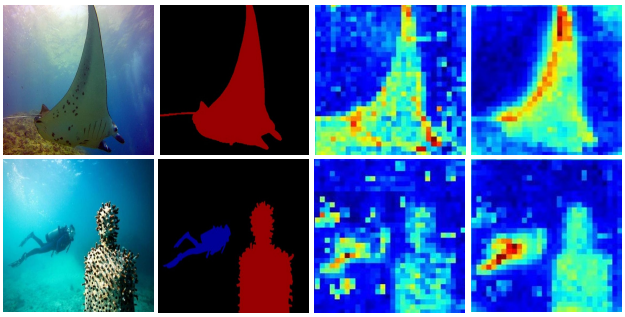


Figure 7. Visualize features generated by the SFPG. The SFPG can aggregate features on salient instances.

where Avg is the global average pooling, λ is a hyperparameter that controls the noise suppression effect, which in this paper is 0.8. In addition, the ith F_i^a will also be fused with the previous feature, which can be denoted as:

$$F_i^{a*} = F_i^a + Conv_{3*3}(F_{i-1}^{a*}). \quad (5)$$

Finally, since the saliency of the instance varies with the image scale, we use multi-scale deconvolution to $2\times$ and $4\times$ upsampling of the final output F_n^{a*} of the SFFM, which is used in the subsequent RPN header (Ren et al., 2015).

As a visual illustration in Figure 7, we can clearly observe a significant change in the distribution of salient features with application SFPG module. This is due to the ability provided by the SFPG module to aggregate different receptive field features, where large receptive field features help the model to locate salient regions while filtering out noise, and small receptive field features contribute to supplementing the image details with fine-grained semantic granularity.

Loss Function. The loss function of the USIS-SAM framework is similar to the Mask RCNN (He et al., 2017) and consists of several components i.e., localization loss, classification loss, regression loss and segmentation loss. Therefore, the cumulative loss can be expressed as:

$$\mathcal{L} = \frac{\sum_i^N \mathcal{L}_{rpn}^i}{N} + \frac{\sum_i^N \mathbb{1}(i)\cdot(\mathcal{L}_{cls}^i + \mathcal{L}_{reg}^i + \mathcal{L}_{seg}^i)}{\sum_i^N \mathbb{1}(i)}, \quad (6)$$

where N is the total number of anchor boxes predicted by the RPN head and $\mathbb{1}(\cdot)$ is the indicator function, which outputs 1 when the anchor box is salient and otherwise outputs 0. $\mathcal{L}_{rpn}^i$ denotes the region proposal loss, with the same configuration as Mask RCNN (He et al., 2017). $\mathcal{L}_{cls}^i$ denotes the cross-entropy loss, computed between the predicted categories and the ground truth categories. $\mathcal{L}_{reg}^i$ is the Smooth L1 loss (Girshick, 2015), computed based on the bounding box coordinate offset between the prediction and the ground truth. $\mathcal{L}_{seg}^i$ is the binary cross-entropy loss between the USIS-SAM output salient mask and the ground truth mask.

Table 2. Quantitative comparisons with state-of-the-arts on the USIS10K datasets. URank stands for the underwater image enhancement method in UnderwaterRanker (Guo et al., 2023). SAM+BBox uses inference results from Faster RCNN (Ren et al., 2015) as prompts for prediction, SAM+Mask stands for Mask RCNN networks (He et al., 2017) use SAM as the backbone. The RSPrompter in the table is the RSPrompter-anchor framework.

Method	Epoch	Backbone	Class-Agnostic			Multi-Class		
			mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
S4Net (Fan et al., 2019)	60	ResNet-50	32.8	64.1	27.3	23.9	43.5	24.4
RDPNet (Wu et al., 2021)	50	ResNet-50	53.8	77.8	61.9	37.9	55.3	42.7
RDPNet (Wu et al., 2021)	50	ResNet-101	54.7	78.3	63.0	39.3	55.9	45.4
OQTR (Pei et al., 2023)	120	ResNet-50	56.6	79.3	62.6	19.7	30.6	21.9
URank+RDPNet (Wu et al., 2021)	50	ResNet-101	52.0	80.7	62.0	35.9	52.5	41.4
URank+OQTR (Pei et al., 2023)	120	ResNet-50	49.3	74.3	56.2	20.8	32.1	23.3
WaterMask (Lian et al., 2023)	36	ResNet-50	58.3	80.2	66.5	37.7	54.0	42.5
WaterMask (Lian et al., 2023)	36	ResNet-101	59.0	80.6	67.2	38.7	54.9	43.2
SAM+BBox (Kirillov et al., 2023)	24	ViT-H	45.9	65.9	52.1	26.4	38.9	29.0
SAM+Mask (Kirillov et al., 2023)	24	ViT-H	55.1	80.2	62.8	38.5	56.3	44.0
RSPrompter (Chen et al., 2023a)	24	ViT-H	58.2	79.9	65.9	40.2	55.3	44.8
URank+RSPrompter (Chen et al., 2023a)	24	ViT-H	50.6	74.4	56.6	38.7	55.4	43.6
USIS-SAM	24	ViT-H	**59.7**	**81.6**	**67.7**	**43.1**	**59.0**	**48.5**

5. Experiments

We evaluated our USIS-SAM compared with other state-of-the-art methods on USIS10K. First, since previous SIS datasets are usually class-agnostic, the experiments only require the network to segment each salient instance in the image without predicting their classes. We followed previous conventions in the SIS field and performed class-agnostic salient instance segmentation. Then, we also require the network to perform multi-class salient instance segmentation, i.e., predicting instance labels while segmenting salient instances, in order to evaluate the overall capability of the network. We follow OQTR (Pei et al., 2023) and adopt the standard mask AP metrics (Lin et al., 2014) as evaluation Criteria, including mAP, AP_{50} and AP_{75}.

5.1. Implementation Details

In this paper, all backbones and hyperparameters of the methods and comparison algorithms are the same as original paper, except for our newly designed parts. We implemented USIS-SAM with PyTorch and MMDetection (Chen et al., 2019), and trained it over 24 epochs on 6 NVIDIA 3090 GPUs, using the AdamW optimizer with a starting learning rate of 1e-4 and weight decay of 1e-3. We also used the MindSpore Lite tool[1] to implement our network. Unless otherwise stated, the hyperparameters of all ablation experiments were consistent with the comparison experiments with multi-class salient instance segmentation. For data enhancement, both our and the comparison algorithms use random horizontal flipping, random scaling, and random cropping. During training, the original parameters of SAM module were frozen for our USIS-SAM and other compared methods.

[1] https://www.mindspore.cn/

5.2. Experimental Results

In this subsection, we will test class-agnostic salient instance segmentation and multi-class salient instance segmentation for USIS-SAM and other state-of-the-art methods. The relevant results of the quantitative analysis and qualitative comparison can be found in Table 2 and Figure 8. More experiments can be found in Appendix B.

Class-Agnostic Salient Instance Segmentation. Class-agnostic salient instance segmentation can be essentially understood as foreground instance segmentation focused on salient parts in images. Thus, the model needs a robust capability to localize salient regions, and the segmented salient instance masks must strictly align with the ground truth to achieve good experimental results in this task. It can be seen that USIS-SAM outperforms the best-performing SIS method, OQTR, by 3.1, 2.3, and 5.1 AP in mAP, AP_{50}, and AP_{75}. This indicates that the AP advantage of USIS-SAM is due to the ability to segment higher precision masks in the underwater environment. In contrast, USIS-SAM has an advantage of 0.7, 1.0, and 0.5 AP in mAP, AP_{50}, and AP_{75} compared to the underwater instance segmentation method WaterMask. At this point, the AP advantage of USIS-SAM is based on the accurate identification of salient areas. Regarding the RSPrompter, also built based on SAM, USIS-SAM outperforms it by 1.5, 1.7, and 1.8 AP on mAP, AP_{50}, and AP_{75}. This improvement is attributed to better learning of the underwater visual prompt with the assistance

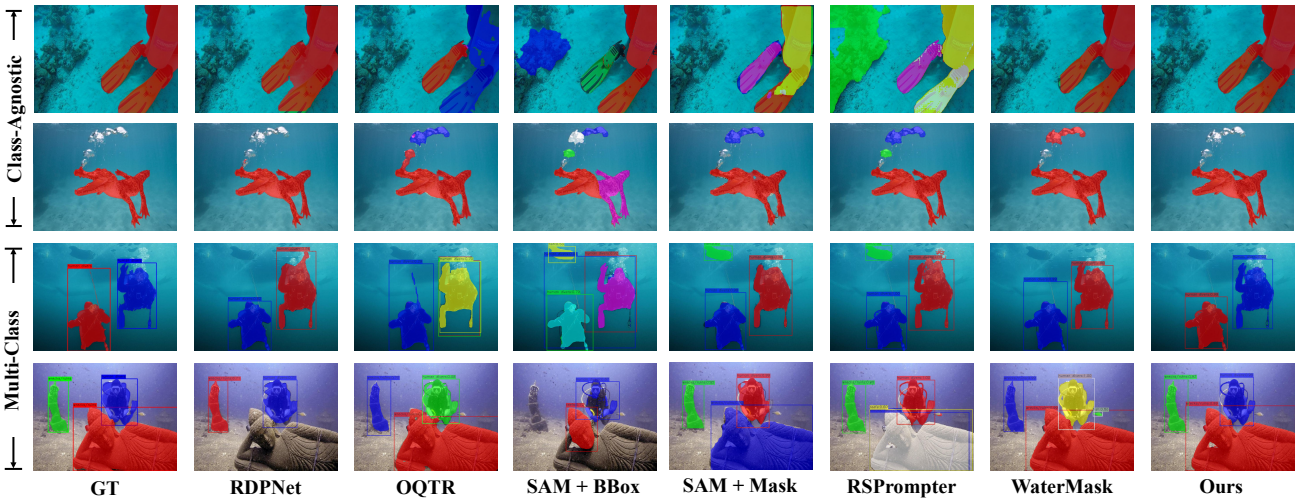


Figure 8. Qualitative comparison on the USIS10K dataset. Each salient instance is represented by a unique color, and the segmented mask is superimposed on the image. More quantitative comparisons can be found in Figure 10 in the Appendix.

Table 3. Effectiveness of each component in USIS-SAM, replacing SFPG means to use Multi-scale Feature Enhancer Module (Chen et al., 2023a) instead of SFPG.

Methods	mAP	AP_{50}	AP_{75}
Full Model	43.1	59.0	48.5
w/o UA-ViT	41.5 (-1.6)	57.4 (-1.6)	47.0 (-1.5)
replace SFPG	42.2 (-0.9)	58.3 (-0.7)	47.5 (-1.0)

of the Underwater Adaptive ViT Encoder in USIS-SAM, enhancing the network's accuracy in segmentation.

Multi-Class Salient Instance Segmentation. Multi-class salient instance segmentation can be considered as a fusion of tasks from Class-Agnostic Salient Instance Segmentation and Salient Instance Class Prediction. From Table 2 in the Multi-class case, USIS-SAM continues to gain an AP advantage due to better mask accuracy. For OQTR, S4Net, the single-stage network coupling classification and segmentation, USIS-SAM lead them by 22.4 AP and 19.5 AP. Furthermore, USIS-SAM also gains a 4.1, 3.3, and 4.0 AP lead over two-stage network RDPNet for mAP, AP_{50}, and AP_{75}. Moreover, USIS-SAM achieved a 4.4 and 5.1 AP lead in mAP over WaterMask and RSPrompter, which were originally designed for multi-class instance segmentation tasks. USIS-SAM leads them by 5.1 and 6.2 AP on AP_{50}, respectively. This is benefit from the extra labels that help the network localize semantically dominant regions.

5.3. Ablation Studies

In this subsection, we perform ablation experiments on Underwater Adaptive ViT Encoder and Salient Feature Prompt Generator, and the relevant data can be found in Table 3. We also performed comparisons on the largest land salient instance segmentation dataset, SIS10K (Pei et al., 2023), to validate the generalization ability of our model. More detailed ablation experiments are in Appendix B.2.

Effectiveness of Underwater Adaptive ViT Encoder. We validate the performance of the Underwater Adaptive ViT Encoder by reverting the UA-ViT Block added in USIS-SAM back to the original ViT Block. In other words, at this stage, USIS-SAM will utilize the pre-trained ViT-H as the image encoder. It can be observed that the model achieves a 1.6 AP improvement after incorporating the UA-ViT Block. This demonstrates that the UA-ViT Block introduces underwater visual information to the Underwater Adaptive ViT Encoder through adapters, enabling the network to effectively handle complex marine scenes such as marine snow, light scattering, optical artifacts, etc. Consequently, it enhances the network's capability to extract features and facilitates network inference.

Effectiveness of Salient Feature Prompt Generator. We evaluate the effect of the Salient Feature Prompt Generator by replacing it. After replacing, the network will utilize the Multi-scale Feature Enhancer (Chen et al., 2023a) to aggregate multi-layer features. After the replacement, the network's mAP, AP_{50}, and AP_{75} are reduced by 0.9, 0.7, and 1.0 AP, respectively. What makes this possible is that SFPG do not simply fuse backbone features, it also enables the network to filter attention from non-significant regions and direct segmentation gravity to salient locations.

Generalization Ability of USIS-SAM. Since there is no other underwater instance segmentation dataset and it is also a thought-provoking idea whether USIS-SAM can be directly applied to land imagery, we retrained USIS-SAM on saliency instance segmentation dataset SIS10K to validate the generalization ability of our model. In Table 4, we can see that, contributing to the excellent segmentation ability of

Table 4. Quantitative comparisons with state-of-the-arts methods on the SIS10K (Pei et al., 2023). The hyperparameters and other settings for the methods in the table are the same as in Table 2.

Methods	mAP	AP_{50}	AP_{75}
OQTR (Pei et al., 2023)	67.2	88.1	**81.7**
USIS-SAM	**70.1**	**89.0**	78.2

SAM and the saliency region localization ability of SFPG, USIS-SAM still achieves good results compared to OQTR, and gains on AP50, but is slightly weaker on AP75. This may be due to domain knowledge learned in UA-ViT is the same as SAM encoder learned on SA-1B dataset, and the guiding effect of UA-ViT on the network is weakened. This shows that USIS-SAM did not overfit our dataset.

6. Discussion

USIS-SAM proposed in this work is devised as the baseline of USIS10K dataset for underwater scenes, its prompt learning architecture may be feasible when transferred to other domains. For example, dark light and hazing/raining scenes share the similar atmospheric scattering model (Akkaynak & Treibitz, 2019; Zhao et al., 2021) with underwater scenes. Hence, UA-ViT's adapter architecture may be conducive to them automatically map color distortion features to corresponding enhancement features. Meanwhile, the Salient Feature Prompt Generator merges multi-layer and multi-scale features to provide salient prompts, can serve as an out-of-the-box module to facilitate other SIS methods to further explore and utilize SAM.

7. Conclusions

In this paper, we construct the first large-scale salient instance segmentation dataset for underwater images, USIS10K, with pixel-level annotations, which enables us to explore the application of salient instance segmentation techniques in underwater environments. In addition, we also propose USIS-SAM for underwater salient instance segmentation based on SAM. Extensive experiments demonstrate the effectiveness of the USIS10K dataset and validate the performance of USIS-SAM.

Acknowledgements

This work was supported in part by the National Key R&D Program of China 2022ZD0118300; in part by the National Natural Science Foundation of China under Grant 62201179; in part by the Innovation Platform for "New Star of South China Sea" of Hainan Province under Grant No. NHXXRCXM202306; in part by the specific research fund of The Innovation Platform for Academicians of Hainan Province under Grant No.YSPTZX202410; in part by the Research Start-up Fund of Hainan University (No. KYQD(ZR)-22015); in part by the Taishan Scholar Project of Shandong Province under Grant tsqn202306079, in part by Xiaomi Young Talents Program, in part by the Hong Kong GRF-RGC General Research Fund under Grant 11209819 and Grant 11203820.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Akkaynak, D. and Treibitz, T. A revised underwater image formation model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

Akkaynak, D. and Treibitz, T. Sea-thru: A method for removing water from underwater images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1682–1691, 2019.

Akkaynak, D., Treibitz, T., Shlesinger, T., Loya, Y., Tamir, R., and Iluz, D. What is the space of attenuation coefficients in underwater computer vision? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

Ben Zaken, E., Goldberg, Y., and Ravfogel, S. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1–9. Association for Computational Linguistics, 2022.

Berman, D., Levy, D., Avidan, S., and Treibitz, T. Underwater single image color restoration using haze-lines and a new quantitative dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2822–2837, 2021. doi: 10.1109/TPAMI.2020.2977624.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C. C., and Lin, D. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., and Shi, Z. Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *arXiv preprint arXiv:2306.16269*, 2023a.

Chen, T., Zhu, L., Deng, C., Cao, R., Wang, Y., Zhang, S., Li, Z., Sun, L., Zang, Y., and Mao, P. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3367–3375, 2023b.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1290–1299, June 2022.

Choi, E. and Lee, C. Feature extraction based on the bhattacharyya distance. *Pattern Recognition*, 36(8):1703–1709, 2003. ISSN 0031-3203. doi: https://doi.org/10.1016/S0031-3203(03)00035-9.

Cong, R., Yang, N., Li, C., Fu, H., Zhao, Y., Huang, Q., and Kwong, S. Global-and-local collaborative learning for co-salient object detection. *IEEE Transactions on Cybernetics*, 53(3):1920–1931, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

ETI BioInformatics. Marine Species Identification Portal. http://species-identification.org/, 2009. Accessed: 10-9-2023.

Fan, D.-P., Cheng, M.-M., Liu, J.-J., Gao, S.-H., Hou, Q., and Borji, A. Salient objects in clutter: Bringing salient object detection to the foreground. In *Proceedings of the European Conference on Computer Vision*, September 2018a.

Fan, R., Hou, Q., Cheng, M.-M., Yu, G., Martin, R. R., and Hu, S.-M. Associating inter-image salient instances for weakly supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pp. 367–383, 2018b.

Fan, R., Cheng, M.-M., Hou, Q., Mu, T.-J., Wang, J., and Hu, S.-M. S4net: Single stage salient-instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.

Fang, Y., Yang, S., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., and Liu, W. Instances as queries. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6910–6919, October 2021.

Girshick, R. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

Guo, C., Wu, R., Jin, X., Han, L., Zhang, W., Chai, Z., and Li, C. Underwater ranker: Learn which is better and how to be better. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):702–709, Jun. 2023.

He, K., Gkioxari, G., Dollar, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Oct 2017.

Hong, L., Wang, X., Zhang, G., and Zhao, M. Usod10k: A new benchmark dataset for underwater salient object detection. *IEEE Transactions on Image Processing*, pp. 1–1, 2023. doi: 10.1109/TIP.2023.3266163.

Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al. Segment anything model for medical images? *Medical Image Analysis*, 92:103061, 2024.

Islam, M. J., Luo, P., and Sattar, J. Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception. In *Proceedings of the Robotics: Science and Systems*, 2020.

Kirillov, A., Wu, Y., He, K., and Girshick, R. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9799–9808, 2020.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollar, P., and Girshick, R. Segment anything. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4015–4026, October 2023.

Li, C., Guo, C., Ren, W., Cong, R., Hou, J., Kwong, S., and Tao, D. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 29:4376–4389, 2020. doi: 10.1109/TIP.2019.2955241.

Li, G., Xie, Y., Lin, L., and Yu, Y. Instance-level salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

Li, Y., Hou, X., Koch, C., Rehg, J. M., and Yuille, A. L. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 280–287, 2014.

Lian, S., Li, H., Cong, R., Li, S., Zhang, W., and Kwong, S. Watermask: Instance segmentation for underwater imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1305–1315, October 2023.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Liu, S., Ye, J., and Wang, X. Any-to-any style transfer: Making picasso and da vinci collaborate. *arXiv e-prints*, pp. arXiv–2304, 2023.

Ma, J., He, Y., Li, F., Han, L., You, C., and Wang, B. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Mobley, C. D. Radiative transfer in the ocean. *Encyclopedia of ocean sciences*, 4:2321–2330, 2001.

Oceana Inc. The Ocean Animal Encyclopedia. https://oceana.org/marine-life/, 2001. Accessed: 10-10-2023.

Pei, J., Cheng, T., Tang, H., and Chen, C. Transformer-based efficient salient instance segmentation networks with orientative query. *IEEE Transactions on Multimedia*, 25:1964–1978, 2023. doi: 10.1109/TMM.2022.3141891.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.

Shi, P., Qiu, J., Abaxi, S. M. D., Wei, H., Lo, F. P.-W., and Yuan, W. Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. *Diagnostics*, 13(11): 1947, 2023.

Tian, X., Xu, K., Yang, X., Yin, B., and Lau, R. W. Learning to detect instance-level salient objects using complementary image labels. *International Journal of Computer Vision*, 130(3):729–746, 2022a.

Tian, X., Xu, K., Yang, X., Yin, B., and Lau, R. W. H. Learning to detect Instance-Level salient objects using complementary image labels. *International Journal of Computer Vision*, 130(3):729–746, March 2022b.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y., and Arbel, T. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv e-prints*, pp. arXiv–2304, 2023.

Wu, Y.-H., Liu, Y., Zhang, L., Gao, W., and Cheng, M.-M. Regularized densely-connected pyramid network for salient instance segmentation. *IEEE Transactions on Image Processing*, 30:3897–3907, 2021. doi: 10.1109/TIP.2021.3065822.

Xie, D., Wang, R., Ma, J., Chen, C., Lu, H., Yang, D., Shi, F., and Lin, X. Edit everything: A text-guided generative system for images editing. *arXiv preprint arXiv:2304.14006*, 2023.

Yu, H., Zheng, K., Fang, J., Guo, H., Feng, W., and Wang, S. Co-saliency detection within a single image. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., and Chen, Z. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.

Zhang, Q., Cong, R., Hou, J., Li, C., and Zhao, Y. Coadnet: Collaborative aggregation-and-distribution networks for co-salient object detection. *Advances in Neural Information Processing Systems*, 33:6959–6970, 2020.

Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Dong, H., Qiao, Y., Gao, P., and Li, H. Personalize segment anything model with one shot. In *The Twelfth International Conference on Learning Representations*, 2023.

Zhang, X., Liu, Y., Lin, Y., Liao, Q., and Li, Y. Uv-sam: Adapting segment anything model for urban village identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 22520–22528, 2024.

Zhang, Y. and Jiao, R. How segment anything model (sam) boost medical image segmentation: A survey. *Available at SSRN 4495221*, 2022.

Zhao, S., Zhang, L., Shen, Y., and Zhou, Y. Refinednet: A weakly supervised refinement framework for single image dehazing. *IEEE Transactions on Image Processing*, 30: 3391–3404, 2021.

A. More Details and Samples about USIS10K

A.1. Dataset Collection.

We collected about 30,000 images from the Internet and open source underwater datasets from various different domains, such as underwater image enhancement (Li et al., 2020; Berman et al., 2021), underwater instance segmentation (Lian et al., 2023), and underwater salience detection (Hong et al., 2023; Islam et al., 2020). These images come from many different natural underwater environments, covering underwater vision tasks such as ocean exploration, human-computer intelligence cooperation, and underwater autopilot. After that, we allowed two volunteers to filter the alternative images for duplicates, corruptions, and images without salient objects. Finally, we carefully annotated the remaining 16,000 images.

A.2. Dataset Annotation.

We assembled 17 volunteers to annotate the dataset. Before annotation, these volunteers were trained in the following three parts: 1) the categories of various common organisms in underwater scenes. 2) the process of previous dataset construction in the field of SIS and the basic methods for image annotation. 3) the general knowledge of visible salience. We use sparse annotation polygons to annotate each instance in the dataset. Annotated data will be stored in the mainstream COCO-style format (Lin et al., 2014) for ease of use by most common frameworks and models. Each image will be annotated by at least 3 volunteers. For those images collected from the Underwater Instance Segmentation dataset, 3 volunteers will vote on the saliency of each instance, and all instances with more than 2 votes will be retained. For images collected from the underwater salient detection dataset, two volunteers will re-label the salient object bounding box at the instance level, and a third volunteer will select the best annotation and refine it. For the other images, three volunteers will annotate salient objects in the image at the instance level, and a fourth volunteer will select, refine, and merge their annotations. To ensure the precision of the labeled categories, we follow the guidance of the literature (Oceana Inc., 2001; ETI BioInformatics, 2009) to categorize potentially confusing objects. In addition, we also filtered out the images for which we could not reach a consensus on instance category or salience. Finally, we are left with 10,632 images that constitute the USIS10K dataset

A.3. Category labels contained in USIS10K

Table 5 lists all the category labels contained in the USIS10K dataset along with their detailed category definitions. As can be seen, the USIS10K dataset contains object categories such as fish, coral reefs, aquatic plants, and wrecks/ruins. These categories constitute the primary research components for marine ruins discovery, marine resources exploration, and marine ecological preservation. Moreover, the dataset contains pixel annotations for human divers, robots, and seafloor, serving as crucial research targets for human-robot-object collaboration.

Table 5. The category labels in USIS10K dataset with their definitions.

Category	Descriptions
Fish	Underwater vertebrates, e.g., fish, turtles
Reefs	Underwater invertebrates and coral reefs
Aquatic plants	Aquatic plants and flora
Wrecks/ruins	Wrecks, ruins and damaged artifacts
Human divers	Human divers and their equipment
Robots	Underwater robots like AUV, ROV
Sea-floor	Rocks and reefs on the seafloor

A.4. More Samples about USIS10K

More visual samples from the USIS10K dataset are presented in Figure 9. It is evident that the images in the USIS10K dataset contain a wide variety of underwater scenes with many different saturations, chromatic aberration, and quality degradations. This diversity is very effective for training salient instance segmentation networks suitable for general underwater scenes.

Furthermore, the USIS10K dataset also includes a large number of challenging images meticulously crafted to comprehensively evaluate the model's performance. As in Figure 9, the surface reflection in the first row's, first, and second columns has the potential to disturb the model's saliency localization. The intricate boundaries of aquatic plants in the third column of the third row and reefs in the second column of the fourth row present a more significant challenge to the segmentation

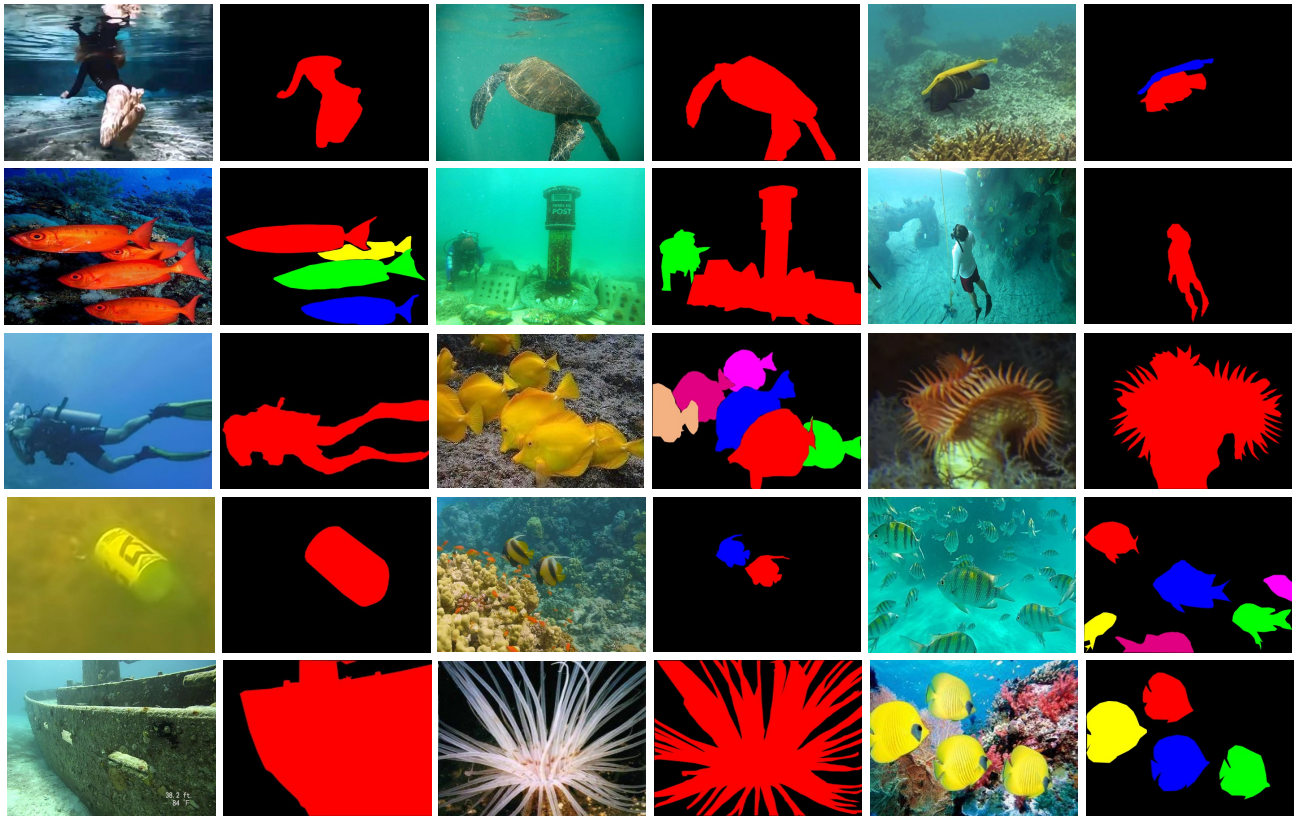


Figure 9. **More visual samples of annotated images with pixel-level salient instance segmentation in USIS10K.** The image on the left is the original image and the right is the annotation mask, different colors represent different salient instances.

accuracy of the network. Additionally, the overlapping fish in the second column of the third row necessitates that the network accurately distinguishes each salient instance.

B. More Qualitative Comparison and ablation Experiments

B.1. More Qualitative Comparison

We also show more visual comparisons on USIS10K dataset in Figure 10 to demonstrate the effectiveness of our model. It can be seen that compared to salient instance segmentation networks such as RDPNet (Wu et al., 2021) and OQTR (Pei et al., 2023), benefiting from our Underwater Adaptive ViT Encoder, USIS-SAM can better adapt to underwater environments and segment the accurate instance boundaries (as shown in Figure 10, rows 1 and 3). And compared with instance segmentation methods such as WaterMask (Lian et al., 2023), RSPrompter (Chen et al., 2023a), Salient Feature Prompt Generator brings USIS-SAM a more powerful salient region localization capability, which can better identify the instance's saliency (as shown in Figure 10, row 6).

B.2. More Ablation Experiments

In this subsection, we perform more detailed ablation experiments on each component in the Underwater Adaptive ViT Encoder and Salient Feature Prompt Generator.

Number of parameters for USIS-SAM. In order to minimize the training consumption, we do not tune all the parameters in the original SAM architecture, but freeze the pre-trained SAM weights and insert some simple but effective modules at specific locations in the SAN to modify the SAM for underwater salience instance segmentation. We show the number of parameters for each component in USIS-SAM in Table 6. It can be seen that the trainable parameters of USIS-SAM constitute

only 9.3% of the total. This means that instead of retraining or fine-tuning SAM, we can quickly apply USIS-SAM to underwater tasks by simply freezing SAM and specifically training our additional introduced components.

Table 6. Number of parameters for each component of USIS-SAM, where * stands for trainable parameters.

Modules	Segment Anything Model	UA-ViT*	SPFG*
Params	641.0 M	30.5 M	29.5 M

Underwater Adaptive ViT Encoder. We show the results of the following ablation experiments about the Underwater Adaptive ViT Encoder in Table 7. First, we assess the effect of the Adapter by removing it from the UA-ViT. The results indicate that the absence of the Adapter prevents the network from learning underwater-specific visual prompts due to the freezing of the image encoder, resulting in performance degradation of 1.4, 1.7, and 1.2 AP on the mAP, AP_{50}, and AP_{75}, respectively. Additionally, we also removed the Channel Adapter to demonstrate its effectiveness. It can be seen that after removing it, the network performance decreases by 1.1, 1.3 AP, and 1.4 AP on mAP, AP_{50}, and AP_{75}. This decline is attributed to the network's inability to accurately identify the importance of each channel at this time.

Table 7. Effectiveness of each component in Underwater Adaptive ViT Encoder, w/o Adapter and w/o CA denote the removal of Adapters and Channel Adapter.

Methods	mAP	AP_{50}	AP_{75}
Full Model	43.1	59.0	48.5
w/o Adapter	41.7 (-1.4)	57.3 (-1.7)	47.3 (-1.2)
w/o CA	42.0 (-1.1)	57.7 (-1.3)	47.1 (-1.4)

Salient Feature Prompt Generator. Table 8 shows the results of ablation experiments concerning the Salient Feature Prompt Generator. First, we evaluate the effectiveness of the Salient Feature Fusion Module by removing it. The results demonstrate that upon removing it, the model is unable to effectively aggregate the multilayer features of ViT, which affects its ability to accurately localize the salient regions in the image, resulting in a decrease in the AP values of mAP, AP_{50} and AP_{75} by 0.8, 0.5 and 1.3, respectively. In addition, we explored the effect of replacing multi-scale convolution with simple 3×3 convolution after upsampling in the Salient Feature Prompt Generator. After replacement, the segmentation accuracy of the model decreased due to the reduced receptive field, decreasing by 0.6, 0.4, and 0.8 AP in mAP, AP_{50}, and AP_{75}, respectively.

Table 8. Effectiveness of each component in Salient Feature Prompt Generator, w/o SFFM and w/o Multi-Conv denote the removal of the salient feature fusion module and multi-up sampling module.

Methods	mAP	AP_{50}	AP_{75}
Full Model	43.1	59.0	48.5
w/o SFFM	42.3 (-0.8)	58.5 (-0.5)	47.2 (-1.3)
w/o Multi-Up	42.5 (-0.6)	58.6 (-0.4)	47.7 (-0.8)

Generalization Ability in similar tasks. We also conducted experiments on the underwater instance segmentation task, which is similar to the underwater salient instance segmentation task, to validate the generalization ability of our model. To do so, we retrain our model on the underwater instance segmentation dataset UIIS (Lian et al., 2023), and compare it to other state-of-the-arts underwater instance segmentation algorithms in Table 9. It can be seen that compared to WaterMask (Lian et al., 2023), which is designed for underwater instance segmentation tasks, USIS-SAM obtains a lead of 2.2 AP, 1.3 AP and 3.0 AP on mAP, AP50 and AP75, respectively. This is attributed to the underwater vision knowledge that UA-ViT introduces to the network and the strong segmentation capability of SAM, which helps USIS-SAM to achieve favorable capabilities in similar tasks.

Table 9. Quantitative comparisons with state-of-the-arts methods on underwater instance segmentation dataset UIIS (Lian et al., 2023). The hyperparameters and other settings for USIS-SAM in the table are the same as in Table 2.

Methods	Epoch	mAP	AP_{50}	AP_{75}
Mask RCNN (He et al., 2017)	36	23.4	40.9	25.3
Point Rend (Kirillov et al., 2020)	36	25.9	43.4	27.6
QueryInst (Fang et al., 2021)	36	26.0	42.8	27.3
Mask2Former (Cheng et al., 2022)	36	25.7	38.0	27.7
RDPNet (Wu et al., 2021)	50	20.6	38.7	19.4
WaterMask (Lian et al., 2023)	36	27.2	43.7	29.3
USIS-SAM	24	**29.4**	**45.0**	**32.3**

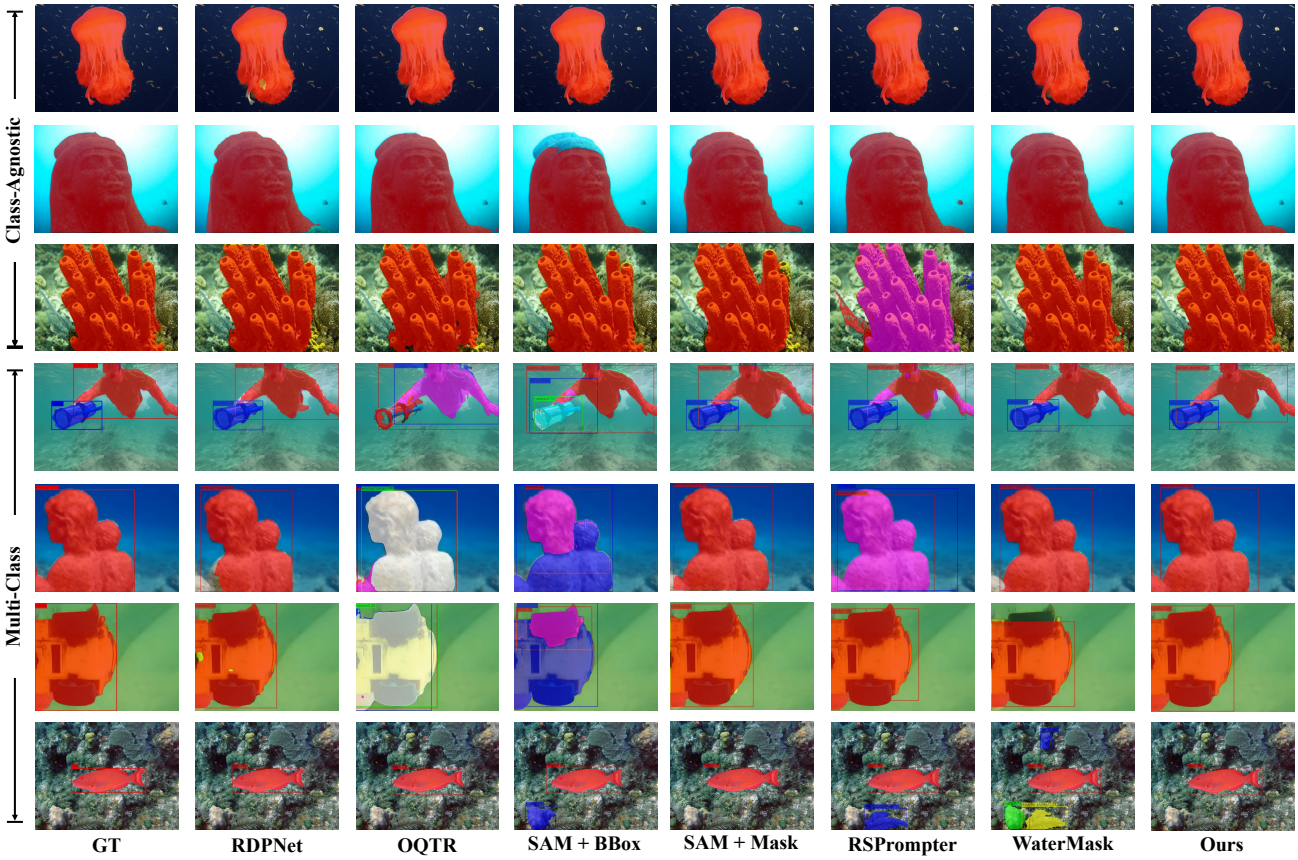


Figure 10. More qualitative comparison on USIS10K dataset. Each salient instance is represented by a unique color, and the segmented mask is superimposed on the image. SAM+BBox uses inference results from Faster RCNN(Ren et al., 2015) as prompts for prediction, SAM+Mask stands for Mask RCNN networks (He et al., 2017) use SAM as the backbone. The RSPrompter in the table is the RSPrompter-anchor framework.