MAS-SAM: Segment Any Marine Animal with Aggregated Features

Tianyu Yan[1] , **Zifu Wan**[2] , **Xinhao Deng**[1] , **Pingping Zhang**[1*] , **Yang Liu**[1] , **Huchuan Lu**[1]
[1]School of Future Technology, School of Artificial Intelligence, Dalian University of Technology
[2] Robotics Institute, Carnegie Mellon University
{2981431354,dengxh}@mail.dlut.edu.cn, zifuw@andrew.cmu.edu, {zhpp,ly,lhchuan}@dlut.edu.cn

Abstract

Recently, Segment Anything Model (SAM) shows exceptional performance in generating high-quality object masks and achieving zero-shot image segmentation. However, as a versatile vision model, SAM is primarily trained with large-scale natural light images. In underwater scenes, it exhibits substantial performance degradation due to the light scattering and absorption. Meanwhile, the simplicity of the SAM's decoder might lead to the loss of fine-grained object details. To address the above issues, we propose a novel feature learning framework named MAS-SAM for marine animal segmentation, which involves integrating effective adapters into the SAM's encoder and constructing a pyramidal decoder. More specifically, we first build a new SAM's encoder with effective adapters for underwater scenes. Then, we introduce a Hypermap Extraction Module (HEM) to generate multi-scale features for a comprehensive guidance. Finally, we propose a Progressive Prediction Decoder (PPD) to aggregate the multi-scale features and predict the final segmentation results. When grafting with the Fusion Attention Module (FAM), our method enables to extract richer marine information from global contextual cues to fine-grained local details. Extensive experiments on four public MAS datasets demonstrate that our MAS-SAM can obtain better results than other typical segmentation methods. The source code is available at https://github.com/Drchip61/MAS-SAM.

1 Introduction

Marine Animal Segmentation (MAS) is a critical and fundamental task in the field of visual intelligence and underwater robotics [Zhang *et al.*, 2024]. It aims at identifying and segmenting marine animals from underwater images or videos. Functionally, the accurate segmentation of marine animals is of great importance for various research areas, including marine biology, ecology and conservation. However, MAS presents unique challenges compared to typical terrestrial image segmentation [Zhang *et al.*, 2017a; Zhang *et al.*, 2017b; Zhang *et al.*, 2019a; Zhang *et al.*, 2021]. In fact, underwater environments are characterized by complex light scattering and absorption effects, leading to degraded image quality, reduced contrast and blurred objects. In addition, marine animals often exhibit camouflage properties, which has further complicated the segmentation task. To address these challenges, advanced perception techniques are required.

With deep Convolutional Neural Networks (CNN), many methods [Zhang *et al.*, 2018; Zhang *et al.*, 2019b; Zeng *et al.*, 2019; Zhang *et al.*, 2019c; Zhang *et al.*, 2020a; Liang *et al.*, 2022; Deng *et al.*, 2023] achieve significant improvements in the segmentation accuracy. However, the complex environment demands a more comprehensive understanding of underwater images to achieve accurate MAS. Meanwhile, CNN's local receptive fields are not well-suited to capture long-range dependencies and contextual information. Thus, it struggles to discern marine animals in underwater scenes. Recently, Segment Anything Model (SAM) [Kirillov *et al.*, 2023] is proposed and has shown great potential in general segmentation tasks. However, SAM's training scenarios primarily involve natural lighting conditions, which limits its performance in underwater environments. Besides, the simplistic decoder structure in SAM lacks the capability for generating fine-grained segmentation results.

Considering the above facts, in this work we propose a novel SAM-based feature learning framework named MAS-SAM for marine animal segmentation. More specifically, by freezing the pre-trained parameters of the SAM's encoder and introducing effective adapters, we build an Adapter-informed SAM Encoder (ASE) for extracting the unique features from marine animal images. In addition, we construct a Hypermap Extraction Module (HEM) to extract multi-scale feature maps from the new SAM's encoder. It serves as a comprehensive guidance for the subsequent mask prediction process. To improve the SAM's decoder, we introduce a Progressive Prediction Decoder (PPD) to aggregate features from the original prompt, ASE and HEM. When grafting with the Fusion Attention Module (FAM), our PPD can prioritize the importance of multi-grained feature maps and extract richer marine information from global contextual cues to fine-grained local details. Extensive experiments on four MAS datasets show that our method can consistently obtain outstanding results.

*Corresponding author

In summary, our contributions are as follows:

- We introduce a novel feature learning framework named NAS-SAM for Marine Animal Segmentation (MAS). It can custom SAM with aggregated multi-scale features for high-performance MAS.
- We propose a Hypermap Extraction Module (HEM), which generates multi-scale features based on the SAM's encoder. It serves as a comprehensive guidance for the subsequent mask prediction process.
- We propose a Progressive Prediction Decoder (PPD) to effectively improve the representation ability of SAM's decoder, capturing a wide range of marine information from global contextual cues to fine-grained local details.
- We perform extensive experiments to verify the effectiveness of the proposed modules. Our method achieves state-of-the-art performances on four MAS datasets.

2 Related Work

2.1 Marine Animal Segmentation

MAS is a specific and challenging sub-domain of image segmentation focused on marine animals. With the advance of deep learning, CNNs have become a popular choice for MAS. For example, Li *et al.* [Li *et al.*, 2021] propose an Enhanced Cascade Decoder Network (ECDNet) for accurately segmenting marine animals from complex underwater environments. Chen *et al.* [Chen *et al.*, 2022] propose to reduce the impact of water degradation diversity via a Siamese network. Recently, Cheng *et al.* [Cheng *et al.*, 2023] built a bidirectional collaborative mentoring network for leveraging structural texture and contextual clues of marine images. Fu *et al.* [Fu *et al.*, 2023] design a fusion network to learn semantic features of camouflage animals. However, the lack of global understanding ability limits the performance of these CNN-based methods. Meanwhile, vision Transformers [Dosovitskiy *et al.*, 2020] have shown promising capabilities in capturing long-range dependencies and global contextual information. They deliver much better performance in general image segmentation tasks [Zheng *et al.*, 2021; Ranftl *et al.*, 2021; Liu *et al.*, 2021b; Liu *et al.*, 2021a]. Inspired by these works, Hong *et al.* [Hong *et al.*, 2023] propose a hybrid network for underwater salient object segmentation, which jointly leverages the advantage of CNNs and Transformers. However, Transformer-based methods still require massive training data to achieve satisfactory results. Different from previous works, our method resorts to the vision fundamental model to achieve better MAS results in challenging scenes.

2.2 SAM for Customized Tasks

Recently, SAM [Kirillov *et al.*, 2023] has drawn great attention for its powerful image understanding abilities. With appropriate prompts, it enables transfer learning across various image segmentation tasks [Zhao *et al.*, 2023; Jin *et al.*, 2023; Zhang *et al.*, 2023]. It also emerges the ability of few-shot/zero-shot segmentation, eliminating the need for task-specific retraining. However, SAM is trained on large-scale natural images. Thus, it fails to adequately capture the unique characteristics of other scenes and can hardly achieve outstanding performance on customized tasks. Meanwhile, due to the decoder's simplicity, SAM only utilizes the terminal features of vision Transformers and might lead to the loss of fine-grained object details for accurate segmentation.

To relieve the above drawbacks, some works [Zhang and Liu, 2023; Wang *et al.*, 2023; Gong *et al.*, 2023; Chen *et al.*, 2023] have integrated domain-specific cues into SAM through the utilization of simple adapters. Furthermore, Lai *et al.* [Lai *et al.*, 2023] verify that sophisticated adapters can highlight task-specific information. However, the exploration of applying SAM to MAS is rather limited. Xu *et al.* [Xu *et al.*, 2023] directly fine-tune SAM for underwater image segmentation. Zhang *et al.* [Zhang *et al.*, 2024] propose a dual-SAM structure with auto-prompts for MAS. More importantly, previous modifications on the SAM's decoder can not capture intricate details and structures of marine animals. Therefore, in this work we step further and custom SAM with aggregated multi-scale features for high-performance MAS.

3 Proposed Method

As shown in Fig. 1, our proposed method includes three main components: Adapter-informed SAM Encoder (ASE), Hypermap Extraction Module (HEM) and Progressive Prediction Decoder (PPD). These key components will be elaborated in the following sections.

3.1 Adapter-informed SAM Encoder

As stated in previous sections, although SAM has shown great potential for universal segmentation tasks, it is suboptimal to directly deploy the pre-trained SAM to MAS. To address this issue, we propose an Adapter-informed SAM Encoder (ASE), which customizes SAM for marine animal images. As shown in Fig. 1, we retain the core components of the original SAM and utilize two parameter-efficient fine-tuning mechanisms for improving the pre-trained encoder. As shown in Fig. 2, we incorporate the LoRA [Hu *et al.*, 2021] and Adapter [Houlsby *et al.*, 2019] into the Multi-Head Self-Attention (MHSA) and Feed-Forward Network (FFN) of each Transformer block, respectively. More specifically, let $\mathbf{X}_i \in \mathbb{R}^{N\times D}$ be the input of the i-th Transformer block. Here, N is the number of tokens and D denotes the embedding dimension. The MHSA layer modified by LoRA can be represented as follows:

$$\mathbf{Q}_i = W_q(\mathbf{X}_i) + W_q^{up}(W_q^{down}(\mathbf{X}_i), \tag{1}$$

$$\mathbf{K}_i = W_k(\mathbf{X}_i), \tag{2}$$

$$\mathbf{V}_i = W_v(\mathbf{X}_i) + W_v^{up}(W_v^{down}(\mathbf{X}_i), \tag{3}$$

$$\overline{\mathbf{X}}_i = MHSA\left(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i\right) + \mathbf{X}_i, \tag{4}$$

where W_q, W_k and W_v are weights of three linear projection layers to generate the original $Query$, Key and $Value$ matrices, respectively. $W_{q,v}^{down} \in \mathbb{R}^{M\times D}$ and $W_{q,v}^{up} \in \mathbb{R}^{M\times D}$ are weights of two linear projection layers to reduce and restore the feature dimension, respectively. M is the down-projection dimension. In this way, we can freeze the pre-trained weights

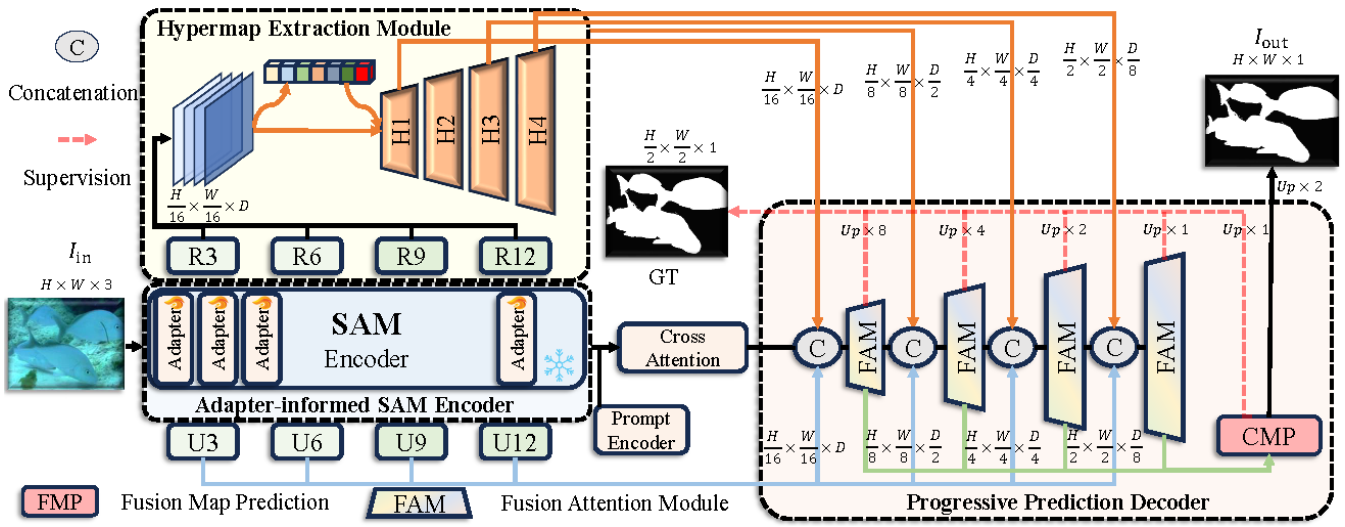

Figure 1: The overall structure of our proposed framework (MAS-SAM). It consists of three main components: Adapter-informed SAM Encoder (ASE), Hypermap Extraction Module (HEM) and Progressive Prediction Decoder (PPD).

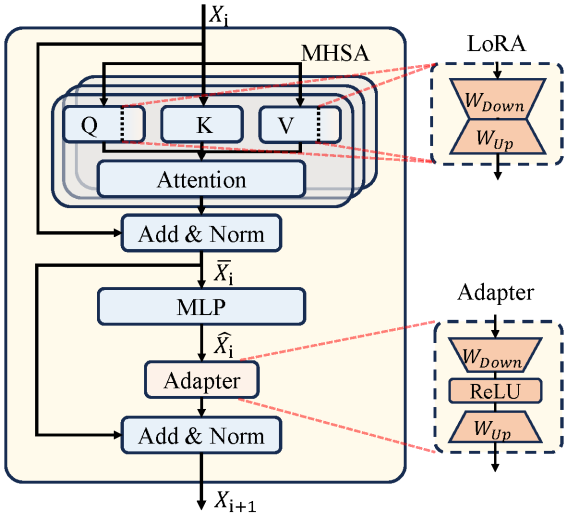

Figure 2: The enhanced Transformer block in our proposed Adapter-informed SAM Encoder.

(W_q, W_k and W_v) and utilize rank decomposition matrices to greatly reduce the number of trainable parameters.

Furthermore, we insert an Adapter into the FFN as:

$$\widehat{\mathbf{X}}_i = MLP\left(LN\left(\overline{\mathbf{X}}_i\right)\right), \tag{5}$$

$$\mathbf{X}_{i+1} = W_{adpt}^{up}\left(\sigma\left(W_{adpt}^{down}\left(\widehat{\mathbf{X}}_i\right)\right)\right) + \overline{\mathbf{X}}_i, \tag{6}$$

where LN and MLP stand for the Layer Normalization (LN) [Ba *et al.*, 2016] and Multilayer Perceptron (MLP), respectively. σ is the Rectified Linear Unit (ReLU). $W_{adpt}^{down} \in \mathbb{R}^{P\times D}$ and $W_{adpt}^{up} \in R^{P\times D}$ are weights of two linear projections to reduce and restore the feature dimension, respectively. P is the down-projection dimension. Similar to LoRA, by employing an extremely low value of the parameter P, we can achieve a parameter-efficient fine-tuning for adapting the pre-trained SAM's encoder to marine scenes.

3.2 Hypermap Extraction Module

Due to the complex underwater environment, it is of vital needs to exploit both local details and global contexts for robust and accurate MAS. As previous works point out, different Transformer layers capture different-level semantics [Van Aken *et al.*, 2019; Bin *et al.*, 2023]. Generally, shallow layers retain more local details and deep layers express more contextual information. Therefore, to enable our proposed model leverage richer marine information, we propose a Hypermap Extraction Module (HEM) to consider multi-scale feature maps from ASE. It then serves as a comprehensive guidance for the subsequent mask prediction process.

More specifically, we first feed an image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ into ASE and obtain the outputs of different Transformer layers. In this work, we select the 3-6-9-12 layers and get the multi-scale token features, i.e., $\mathbf{X}_i (i = 3, 6, 9, 12)$. Then, we reshape them to spatial feature maps $\mathbf{F}_i \in \mathbb{R}^{H/16\times W/16\times D}$. To simultaneously consider these multi-scale feature maps, we perform the following feature aggregation,

$$\mathbf{R}_i = \phi_{1\times 1}(\mathbf{F}_i), \tag{7}$$

$$\overline{\mathbf{H}}_1 = \phi_{3\times 3}[\mathbf{R}_3, \mathbf{R}_6, \mathbf{R}_9, \mathbf{R}_{12}], \tag{8}$$

where $\phi_{1\times 1}$ and $\phi_{3\times 3}$ are convolutional layers with 1×1 and 3×3 kernels, respectively. To improve the training stability, a Batch Normalization (BN) and a ReLU activation function are also introduced after the convolutional layers. $[\cdot]$ is the concatenation in channel.

Afterwards, we introduce a channel-attention layer to generate the hypermap $\mathbf{H}_j$ as follows:

$$\mathbf{H}_1 = \overline{\mathbf{H}}_1 \times \delta(\phi_{1\times 1}(GAP(\overline{\mathbf{H}}_1))) + \overline{\mathbf{H}}_1, \tag{9}$$

$$\mathbf{H}_{j+1} = \phi_{3\times 3}(\psi_{2\times 2}(\mathbf{H}_j)), j = 1, 2, 3, \tag{10}$$

where GAP is the Global Average Pooling (GAP), δ is the Sigmoid function, and $\psi_{2\times2}$ is a deconvolutional layer with a 2×2 kernel. As shown in the top-left part of Fig. 1, we can obtain multi-scale hypermaps. These hypermaps play a crucial role in improving the performance of MAS.

3.3 Progressive Prediction Decoder

Due to the significant appearance variations of marine animals, the simple decoder design in SAM struggles to achieve accurate segmentation masks. To this end, inspired by [Zhang *et al.*, 2020b; Zhang *et al.*, 2020c], we propose a Progressive Prediction Decoder (PPD) to effectively improve the prediction ability. As shown in the right part of Fig. 1, it has a pyramidal structure to progressively aggregate multi-source features from the original prompt, ASE and HEM, and obtain the final segmentation predictions.

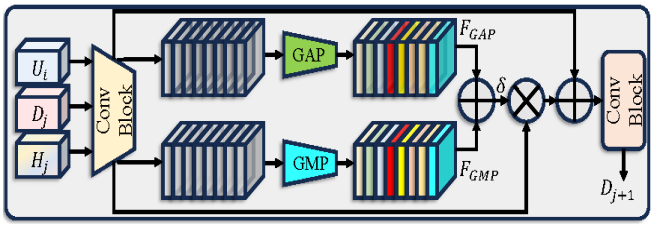

Figure 3: The structure of our Fusion Attention Module.

As shown in Fig. 3, we propose a Fusion Attention Module (FAM) to fully aggregate the multi-source features. More specifically, we begin by up-sampling the features from ASE and resizing the input features into the same size. Then, we fuse them as follows:

$$\mathbf{U}_i = \varphi(\mathbf{F}_i), \tag{11}$$

$$\mathbf{D}_{j+1} = FAM(\mathbf{U}_i, \mathbf{D}_j, \mathbf{H}_j)), \tag{12}$$

where $\mathbf{U}_i$ is the up-sampled feature by utilizing a bilinear interpolation φ. $\mathbf{D}_j$ is the output of the j-th pyramidal stage in the proposed PPD. For the FAM, we utilize a channel-attention to prioritize the importance of multi-source features. Residual structures are also deployed to enhance the representation ability. The procedure can be formulated as:

$$\mathbf{Y}_j = \phi_{1\times1}([\mathbf{U}_i, \mathbf{D}_j, \mathbf{H}_j]), \tag{13}$$

$$\mathbf{W}_j = \delta(\phi_{1\times1}(GAP(\mathbf{Y}_j)) + \phi_{1\times1}(GMP(\mathbf{Y}_j))), \tag{14}$$

$$\overline{\mathbf{D}}_{j+1} = W_j(\mathbf{Y}_j) + \mathbf{Y}_j, \tag{15}$$

$$\mathbf{D}_{j+1} = \phi_{1\times1}(\overline{\mathbf{D}}_{j+1}), \tag{16}$$

where GMP is the Global Max Pooling (GMP). The channel-wise weights can highlight relevant features and suppress irrelevant ones [Yan *et al.*, 2023]. Meanwhile, the attention mechanism employed by our FAM helps to capture intricate relationships between features across various scales, resulting in more coherent and informative feature representations. Thus, FAM can effectively integrate and refine the multi-source features.

Finally, to achieve the progressive prediction, we build the PPD grafted with the FAM as follows:

$$\mathbf{P}_j = \phi_{1\times1}(\varphi(FAM(\mathbf{U}_i, \mathbf{D}_j, \mathbf{H}_j))), \tag{17}$$

where $\mathbf{P}_j$ is the prediction mask of the j-th pyramidal stage. The proposed PPD facilitates the seamless aggregation of multi-source features from the original prompt, ASE and HEM, resulting in richer marine information from global contextual cues to fine-grained local details.

To further improve the prediction results, we take all the predictions at different stages and generate the final prediction as follows:

$$\mathbf{P} = \phi_{1\times1}([\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4]). \tag{18}$$

With the synergistic use of the pyramid structure and FAM, our MAS-SAM can effectively leverage the diverse information and produce highly refined and detailed segmentation masks for a wide range of marine animal shapes and sizes.

3.4 Model Training

During training, we follow previous methods [Li *et al.*, 2021; Yan *et al.*, 2022] and use deep supervision from three levels, i.e., pixel-level supervision (Binary Cross-Entropy Loss), region-level supervision (SSIM Loss) and overall-level supervision (IoU Loss). Thus, we define $\mathcal{L}^f$ or $\mathcal{L}^j$ as a combined loss with three terms:

$$\mathcal{L}^{f/j} = \mathcal{L}_{BCE} + \mathcal{L}_{SSIM} + \mathcal{L}_{IoU}, \tag{19}$$

where $\mathcal{L}^f$ and $\mathcal{L}^j$ are losses of the final prediction and the j-th stage-output, respectively. Due to the space limitation, we refer readers to [Yan *et al.*, 2022] for the exact formulas.

4 Experiments

4.1 Datasets and Evaluation Metrics

In this work, we adopt four public MAS benchmarks to evaluate the model performance. The **MAS3K** dataset [Li *et al.*, 2020] comprises 3,103 marine images, in which 193 are background images. Following the default split, we use 1,769 images for training and 1,141 images for testing. The **RMAS** dataset [Fu *et al.*, 2023] consists of 3,014 marine animal images with 2,514 images used for training and 500 images for testing. The **UFO120** dataset [Islam *et al.*, 2020] comprises 1,620 underwater images with various scenes. Following the default split, we use 1,500 images for training and 120 images for testing. The **RUWI** dataset [Drews-Jr *et al.*, 2021] is a real underwater image dataset captured under complex light conditions. The dataset consists of 700 images, splitting into 525 images for training and 175 images for testing.

Meanwhile, we adopt five metrics to evaluate the segmentation performance of different models. They are Mean Intersection over Union (mIoU), Structural Similarity Measure (S_α) [Fan *et al.*, 2017], Weighted F-measure (F_β^w) [Margolin *et al.*, 2014], Mean Enhanced-Alignment Measure (mE_ϕ) [Fan *et al.*, 2021] and mean absolute error (MAE).

Method	MAS3K					RMAS				
	mIoU	$\mathbf{S}_\alpha$	$\mathbf{F}_\beta^w$	$\mathbf{mE}_\phi$	**MAE**	**mIoU**	$\mathbf{S}_\alpha$	$\mathbf{F}_\beta^w$	$\mathbf{mE}_\phi$	**MAE**
UNet++ [Zhou *et al.*, 2018]	0.506	0.726	0.552	0.790	0.083	0.558	0.763	0.644	0.835	0.046
BASNet [Qin *et al.*, 2019]	0.677	0.826	0.724	0.862	0.046	0.707	0.847	0.771	0.907	0.032
PFANet [Zhao and Wu, 2019]	0.405	0.690	0.471	0.768	0.086	0.556	0.767	0.582	0.810	0.051
SCRN [Wu *et al.*, 2019]	0.693	0.839	0.730	0.869	0.041	0.695	0.842	0.731	0.878	0.030
U2Net [Qin *et al.*, 2020]	0.654	0.812	0.711	0.851	0.047	0.676	0.830	0.762	0.904	0.029
SINet [Fan *et al.*, 2020]	0.658	0.820	0.725	0.884	0.039	0.684	0.835	0.780	0.908	0.025
PFNet [Mei *et al.*, 2021]	0.695	0.839	0.746	0.890	0.039	0.694	0.843	0.771	0.922	0.026
RankNet [Lv *et al.*, 2021]	0.658	0.812	0.722	0.867	0.043	0.704	0.846	0.772	0.927	0.026
C2FNet [Sun *et al.*, 2021]	0.717	0.851	0.761	0.894	0.038	0.721	0.858	0.788	0.923	0.026
ECDNet [Li *et al.*, 2021]	0.711	0.850	0.766	0.901	0.036	0.664	0.823	0.689	0.854	0.036
OCENet [Liu *et al.*, 2022]	0.667	0.824	0.703	0.868	0.052	0.680	0.836	0.752	0.900	0.030
ZoomNet [Pang *et al.*, 2022]	0.736	0.862	0.780	0.898	0.032	0.728	0.855	0.795	0.915	0.022
MASNet [Fu *et al.*, 2023]	0.742	0.864	0.788	0.906	0.032	<u>0.731</u>	<u>0.862</u>	<u>0.801</u>	0.920	0.024
SETR [Zheng *et al.*, 2021]	0.715	0.855	0.789	0.917	0.030	0.654	0.818	0.747	0.933	0.028
TransUNet [Chen *et al.*, 2021]	0.739	0.861	0.805	0.919	0.029	0.688	0.832	0.776	<u>0.941</u>	0.025
H2Former [He *et al.*, 2023]	<u>0.748</u>	0.865	0.810	<u>0.925</u>	<u>0.028</u>	0.717	0.844	0.799	0.931	<u>0.023</u>
SAM [Kirillov *et al.*, 2023]	0.566	0.763	0.656	0.807	0.059	0.445	0.697	0.534	0.790	0.053
I-MedSAM[Wei *et al.*, 2023]	0.698	0.835	0.759	0.889	0.039	0.633	0.803	0.699	0.893	0.035
Med-SAM[Wu *et al.*, 2023]	0.739	0.861	<u>0.811</u>	0.922	0.031	0.678	0.832	0.778	0.920	0.027
SAM-Adapter[Chen *et al.*, 2023]	0.714	0.847	0.782	0.914	0.033	0.656	0.816	0.752	0.927	0.027
SAM-DADF [Lai *et al.*, 2023]	0.742	<u>0.866</u>	0.806	0.925	0.028	0.686	0.833	0.780	0.926	0.024
MAS-SAM	**0.788**	**0.887**	**0.840**	**0.938**	**0.025**	**0.742**	**0.865**	**0.819**	**0.948**	**0.021**

Table 1: Performance comparison on MAS3K and RMAS. The best and second results are in bold and underlined, respectively.

4.2 Implementation Details

Our model is implemented with the PyTorch toolbox and one RTX 3090 GPU. For our model, the SAM's encoder and prompt encoder are initialized from the pre-trained SAM-B [Kirillov *et al.*, 2023], while other proposed modules are randomly initialized. During the training process, we freeze the SAM's encoder and only fine-tune the remaining modules. The widely-used AdamW optimizer [Loshchilov and Hutter, 2017] is adopted to update the model parameters. The initial learning rate and weight decay are set to 0.001 and 0.1, respectively. We reduce the learning rate by a factor of 10 at every 20 epochs. The total number of training epochs is set to 50. The mini-batch size is set to 8. The input images are uniformly resized to $512 \times 512 \times 3$.

4.3 Comparison with State-of-the-arts

In this section, we compare our method with other approaches on four public MAS datasets. Specifically, we follow previous works, *i.e.*, MASNet [Fu *et al.*, 2023] and ECDNet [Li *et al.*, 2021], to include compared methods/tools. MASNet and ECDNet are typical MAS methods. Other compared methods are modified for underwater scene segmentation. Especially, some methods (*e.g.*, SINet [Fan *et al.*, 2020], C2FNet [Sun *et al.*, 2021]) focus on camouflage object detection for underwater images. Thus, it is very reasonable to compare with them. I-MedSAM [Wei *et al.*, 2023], MedSAM [Wu *et al.*, 2023], SAM-Adapter [Chen *et al.*, 2023] and SAM-DADF [Lai *et al.*, 2023] are similar to our method in addressing the adaptability issue of SAM for downstream tasks.

Quantitative Comparisons. Tab. 1 and Tab. 2 illustrate the quantitative results of compared methods. They clearly demonstrate that our method outperforms other methods on all five metrics across four datasets. These results show the superiority of our method in terms of overall completeness, structural accuracy, and pixel-wise precision.

Compared with CNN-based methods, our method delivers significant advantages in multi-scale and long-range modeling capabilities. On the large-scale MAS3K dataset, our method achieves the best mIoU, S_α , F_β^w and mE_ϕ values. It demonstrates an improvement of approximately 3-4% across various metrics. Compared with Transformer-based methods, our method shows a 2-3% gain across multiple metrics on the MAS3K dataset. Moreover, significant improvements are also observed on other three datasets. Compared with SAM-based methods, our method efficiently injects specific marine scene information into SAM through two concise adapter modules. In short, our design notably enhances the model's segmentation ability and achieves a considerable performance gain over other approaches.

Qualitative Comparisons. To demonstrate the superiority of our method more intuitively, we present visual comparisons of different methods in Fig. 4. The visualizations provide a clearer advantage of our method compared with the previous approaches. There are challenging images with high background clutters and abundant details. However, our method still generates better segmentation masks.

4.4 Ablation Studies

On the MAS3K dataset, we conduct experiments to verify the effectiveness of each module in our model. More results are in the supplementary material. The model (A) is directly deploying the pre-trained SAM. Visual results are in Fig. 5.

Method	UFO120					RUWI				
	mIoU	S_α	F_β^w	mE_ϕ	MAE	mIoU	S_α	F_β^w	mE_ϕ	MAE
UNet++ [Zhou *et al.*, 2018]	0.412	0.459	0.433	0.451	0.409	0.586	0.714	0.678	0.790	0.145
BASNet [Qin *et al.*, 2019]	0.710	0.809	0.793	0.865	0.097	0.841	0.871	0.895	0.922	0.056
PFANet [Zhao and Wu, 2019]	0.677	0.752	0.723	0.815	0.129	0.773	0.765	0.811	0.867	0.096
SCRN [Wu *et al.*, 2019]	0.678	0.783	0.760	0.839	0.106	0.830	0.847	0.883	0.925	0.059
U2Net [Qin *et al.*, 2020]	0.680	0.792	0.709	0.811	0.134	0.841	0.873	0.861	0.786	0.074
SINet [Fan *et al.*, 2020]	0.767	0.837	0.834	0.890	0.079	0.785	0.789	0.825	0.872	0.096
PFNet [Mei *et al.*, 2021]	0.570	0.708	0.550	0.683	0.216	0.864	0.883	0.870	0.790	0.062
RankNet [Lv *et al.*, 2021]	0.739	0.823	0.772	0.828	0.101	0.865	0.886	0.889	0.759	0.056
C2FNet [Sun *et al.*, 2021]	0.747	0.826	0.806	0.878	0.083	0.840	0.830	0.883	0.924	0.060
ECDNet [Li *et al.*, 2021]	0.693	0.783	0.768	0.848	0.103	0.829	0.812	0.871	0.917	0.064
OCENet [Liu *et al.*, 2022]	0.605	0.725	0.668	0.773	0.161	0.763	0.791	0.798	0.863	0.115
ZoomNet [Pang *et al.*, 2022]	0.616	0.702	0.670	0.815	0.174	0.739	0.753	0.771	0.817	0.137
MASNet [Fu *et al.*, 2023]	0.754	0.827	0.820	0.879	0.083	0.865	0.880	0.913	0.944	0.047
SETR [Zheng *et al.*, 2021]	0.711	0.811	0.796	0.871	0.089	0.832	0.864	0.895	0.924	0.055
TransUNet [Chen *et al.*, 2021]	0.752	0.825	0.827	0.888	0.079	0.854	0.872	0.910	0.940	0.048
H2Former [He *et al.*, 2023]	0.780	0.844	0.845	0.901	0.070	0.871	0.884	0.919	0.945	0.045
SAM [Kirillov *et al.*, 2023]	0.681	0.768	0.745	0.827	0.121	0.849	0.855	0.907	0.929	0.057
I-MedSAM [Wei *et al.*, 2023]	0.730	0.818	0.788	0.865	0.084	0.844	0.849	0.897	0.923	0.050
Med-SAM [Wu *et al.*, 2023]	0.774	0.842	0.839	0.899	0.072	0.877	0.885	0.921	0.942	0.045
SAM-Adapter [Chen *et al.*, 2023]	0.757	0.829	0.834	0.884	0.081	0.867	0.878	0.913	0.946	0.046
SAM-DADF [Lai *et al.*, 2023]	0.768	0.841	0.836	0.893	0.073	0.881	0.889	0.925	0.940	0.044
MAS-SAM	**0.807**	**0.861**	**0.864**	**0.914**	**0.063**	**0.902**	**0.894**	**0.941**	**0.961**	**0.035**

Table 2: Performance comparison on UFO120 and RUWI. The best and second results are in bold and underlined, respectively.

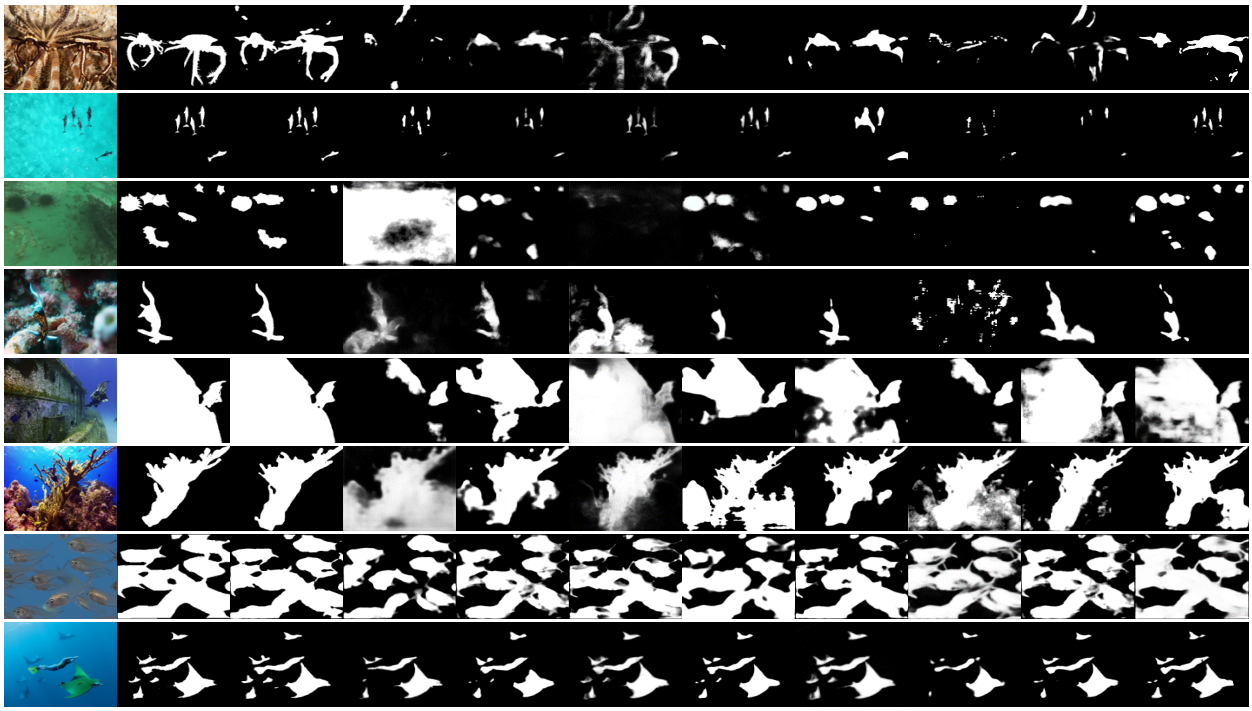

Figure 4: Visual comparison of predicted segmentation masks with different methods. The images in the 1-2nd rows are from the MAS3K, 3-4th are from RMAS, 5-6th are from UFO120 and the 7-8th are from RUWI. Best view by zooming in.

Effects of Adapters in ASE. In the 1-3 rows of Tab. 3, we show the impact of adopters in the ASE. By incorporating two effective adapters, our method has a significant improvement in terms of all evaluation metrics. These results indicate that it

	Method						MAS3K				
	LoRA	**Adapter**	**FPN**	**HEM**	**FAM**	**ML**	**mIoU**	$\mathbf{S}_\alpha$	$\mathbf{F}_\beta^w$	$\mathbf{mE}_\phi$	**MAE**
(A)	×	×	×	×	×	×	0.566	0.763	0.656	0.807	0.059
(B)	✓	×	×	×	×	×	0.742	0.867	0.806	0.919	0.029
(C)	✓	✓	×	×	×	×	0.755	0.868	0.813	0.924	0.028
(D)	✓	✓	✓	×	×	×	0.763	0.871	0.819	0.922	0.028
(E)	✓	✓	✓	✓	×	×	0.771	0.875	0.827	0.926	0.028
(F)	✓	✓	✓	✓	✓	×	0.781	0.881	0.834	0.933	0.025
(G)	✓	✓	✓	✓	✓	✓	0.788	0.887	0.840	0.938	0.025

Table 3: Performance comparisons of using different modules on MAS3K dataset.

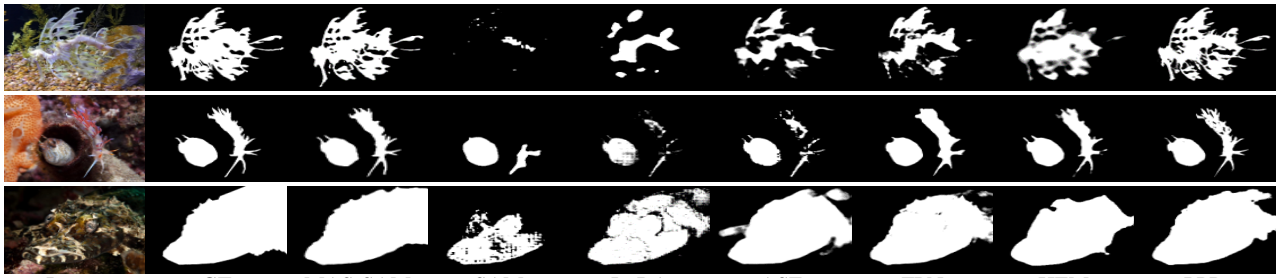

Figure 5: Visual comparison of predicted segmentation masks with different modules. Best view by zooming in.

is very necessary and powerful to inject marine environment information into the fundamental model.

Effects of Key Modules. In the 3-6 rows of Tab. 3, we examine the impact of other key modules. By constructing a PPD, the model can capture more local details, resulting in a 0.8% increase in mIoU. Subsequently, HEM enables the full utilization of multi-level features extracted by the ASE, serving as informative guidance for the subsequent prediction structure. This enhancement leads to a noticeable improvement in performance, better than typical FPN structures. Lastly, we leverage the fused information through the FAM. Compared with the original SAM decoder, our method achieves great improvements in all metrics.

Effects of Different Losses. In the 6-7 rows of Tab. 3, we demonstrate the significance of utilizing multi-level supervision during the training process. By employing three levels of supervision, our method shows a notable improvement.

4.5 More Discussions

In this work, our aim is to adapt SAM to MAS. However, there are three key issues. 1) Domain differences. SAM is pre-trained on natural light scenes, while MAS always suffers from the refraction and absorption of light. Thus, directly employing SAM is hard to achieve good performance for MAS. 2) Camouflage behaviours. Marine animals usually exhibit camouflage behaviours, and large variations in shapes and sizes. The original SAM is not good at modeling these appearances. 3) Missing details. With cross-attentions and only two deconvolutions, the original SAM's decoder leads to a substantial loss of object details.

To address above issues, we present three contributions as follows: 1) We inject marine domain knowledge into the SAM backbone through efficient adapters, thus making SAM more adaptable to marine scene tasks. 2) We employ a HEM to fully leverage multi-scale information. This helps our model to accurately locate marine animals against complex appearances. 3) We propose a PPD structure to progressively integrate feature maps from different levels. With three-grained supervision losses, it can capture the intricate details of marine life extensively. The above facts make our work valuable for underwater intelligence.

5 Conclusion

In this work, we propose a novel feature learning framework named MAS-SAM for MAS. Specifically, we first propose an Adapter-informed SAM Encoder (ASE) for underwater scenes. Then, we introduce a Hypermap Extraction Module (HEM) to generate multi-scale features for a comprehensive guidance. Finally, we propose a Progressive Prediction Decoder (PPD) to aggregate the multi-scale features and predict the final segmentation results. When grafting with the Fusion Attention Module (FAM), our approach extracts richer marine information from global contextual cues to fine-grained local details. Extensive experiments on four MAS datasets show the effectiveness of our MAS-SAM.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62101092), the Fundamental Research Funds for the Central Universities (No. DUT23YG232) and the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No. VRLAB2022C02).

References

[Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[Bin *et al.*, 2023] Yi Bin, Haoxuan Li, Yahui Xu, Xing Xu, Yang Yang, and Heng Tao Shen. Unifying two-stream encoders with transformers for cross-modal retrieval. *arXiv preprint arXiv:2308.04343*, 2023.

[Chen *et al.*, 2021] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[Chen *et al.*, 2022] Ruizhe Chen, Zhenqi Fu, Yue Huang, En Cheng, and Xinghao Ding. A robust object segmentation network for underwater scenes. In *ICASSP*, pages 2629–2633, 2022.

[Chen *et al.*, 2023] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Shangzhan Zhang, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam fails to segment anything?– sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. *arXiv preprint arXiv:2304.09148*, 2023.

[Cheng *et al.*, 2023] Jinguang Cheng, Zongwei Wu, Shuo Wang, Cédric Demonceaux, and Qiuping Jiang. Bidirectional collaborative mentoring network for marine organism detection and beyond. *TCSVT*, 2023.

[Deng *et al.*, 2023] Xinhao Deng, Pingping Zhang, Wei Liu, and Huchuan Lu. Recurrent multi-scale transformer for high-resolution salient object detection. In *ACM MM*, pages 7413–7423, 2023.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Drews-Jr *et al.*, 2021] Paulo Drews-Jr, Isadora de Souza, Igor P Maurell, Eglen V Protas, and Silvia S C. Botelho. Underwater image segmentation in the wild using deep learning. *Journal of the Brazilian Computer Society*, 27:1–14, 2021.

[Fan *et al.*, 2017] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4548–4557, 2017.

[Fan *et al.*, 2020] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, pages 2777–2787, 2020.

[Fan *et al.*, 2021] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *Scientia Sinica Informationis*, 6(6), 2021.

[Fu *et al.*, 2023] 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*, 2023.

[Gong *et al.*, 2023] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. *arXiv preprint arXiv:2306.13465*, 2023.

[He *et al.*, 2023] Along He, Kai Wang, Tao Li, Chengkun Du, Shuang Xia, and Huazhu Fu. H2former: An efficient hierarchical hybrid transformer for medical image segmentation. *TMI*, 2023.

[Hong *et al.*, 2023] Lin Hong, Xin Wang, Gan Zhang, and Ming Zhao. Usod10k: a new benchmark dataset for underwater salient object detection. *TIP*, 2023.

[Houlsby *et al.*, 2019] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, pages 2790–2799, 2019.

[Hu *et al.*, 2021] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2021.

[Islam *et al.*, 2020] Md Jahidul Islam, Peigen Luo, and Junaed Sattar. Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception. *arXiv preprint arXiv:2002.01155*, 2020.

[Jin *et al.*, 2023] Zheyan Jin, Shiqi Chen, Yueting Chen, Zhihai Xu, and Huajun Feng. Let segment anything help image dehaze. *arXiv preprint arXiv:2306.15870*, 2023.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023.

[Lai *et al.*, 2023] Yingxin Lai, Zhiming Luo, and Zitong Yu. Detect any deepfakes: Segment anything meets face forgery detection and localization. In *CCBR*, pages 180–190, 2023.

[Li *et al.*, 2020] Lin Li, Eric Rigall, Junyu Dong, and Geng Chen. Mas3k: An open dataset for marine animal segmentation. In *International Symposium on Benchmarking, Measuring and Optimization*, pages 194–212. Springer, 2020.

[Li *et al.*, 2021] Lin Li, Bo Dong, Eric Rigall, Tao Zhou, Junyu Dong, and Geng Chen. Marine animal segmentation. *TCSVT*, 32(4):2303–2314, 2021.

[Liang *et al.*, 2022] Zijian Liang, Pengjie Wang, Ke Xu, Pingping Zhang, and Rynson WH Lau. Weakly-supervised salient object detection on light fields. *TIP*, 31:6295–6305, 2022.

[Liu *et al.*, 2021a] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *ICCV*, pages 4722–4732, 2021.

[Liu *et al.*, 2021b] Zhengyi Liu, Yuan Wang, Zhengzheng Tu, Yun Xiao, and Bin Tang. Tritransnet: Rgb-d salient object detection with a triplet transformer embedding network. In *ACM MM*, pages 4481–4490, 2021.

[Liu *et al.*, 2022] Jiawei Liu, Jing Zhang, and Nick Barnes. Modeling aleatoric uncertainty for camouflaged object detection. In *WACV*, pages 1445–1454, 2022.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Lv *et al.*, 2021] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *CVPR*, pages 11591–11601, 2021.

[Margolin *et al.*, 2014] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, pages 248–255, 2014.

[Mei *et al.*, 2021] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *CVPR*, pages 8772–8781, 2021.

[Pang *et al.*, 2022] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *CVPR*, pages 2160–2170, 2022.

[Qin *et al.*, 2019] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, pages 7479–7489, 2019.

[Qin *et al.*, 2020] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *PR*, 106:107404, 2020.

[Ranftl *et al.*, 2021] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pages 12179–12188, 2021.

[Sun *et al.*, 2021] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for camouflaged object detection. *arXiv preprint arXiv:2105.12555*, 2021.

[Van Aken *et al.*, 2019] Betty Van Aken, Benjamin Winter, Alexander Löser, and Felix A Gers. How does bert answer questions? a layer-wise analysis of transformer representations. In *ACM CIKM*, pages 1823–1832, 2019.

[Wang *et al.*, 2023] Lin Wang, Xiufen Ye, Liqiang Zhu, Weijie Wu, Jianguo Zhang, Huiming Xing, and Chao Hu. When sam meets sonar images. *arXiv preprint arXiv:2306.14109*, 2023.

[Wei *et al.*, 2023] Xiaobao Wei, Jiajun Cao, Yizhu Jin, Ming Lu, Guangyu Wang, and Shanghang Zhang. I-medsam: Implicit medical image segmentation with segment anything. *arXiv preprint arXiv:2311.17081*, 2023.

[Wu *et al.*, 2019] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *ICCV*, pages 7264–7273, 2019.

[Wu *et al.*, 2023] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[Xu *et al.*, 2023] Muduo Xu, Jianhao Su, and Yutao Liu. Aquasam: Underwater image foreground segmentation. *arXiv preprint arXiv:2308.04218*, 2023.

[Yan *et al.*, 2022] Tianyu Yan, Zifu Wan, and Pingping Zhang. Fully transformer network for change detection of remote sensing images. In *ACCV*, pages 1691–1708, 2022.

[Yan *et al.*, 2023] Tianyu Yan, Zifu Wan, Pingping Zhang, Gong Cheng, and Huchuan Lu. Transy-net: Learning fully transformer networks for change detection of remote sensing images. *TGRS*, 2023.

[Zeng *et al.*, 2019] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *ICCV*, pages 7234–7243, 2019.

[Zhang and Liu, 2023] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

[Zhang *et al.*, 2017a] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, pages 202–211, 2017.

[Zhang *et al.*, 2017b] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, pages 212–221, 2017.

[Zhang *et al.*, 2018] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *CVPR*, pages 1741–1750, 2018.

[Zhang *et al.*, 2019a] Lu Zhang, Zhe Lin, Jianming Zhang, Huchuan Lu, and You He. Fast video object segmentation via dynamic targeting network. In *ICCV*, pages 5582–5591, 2019.

[Zhang *et al.*, 2019b] Lu Zhang, Jianming Zhang, Zhe Lin, Huchuan Lu, and You He. Capsal: Leveraging captioning to boost semantics for salient object detection. In *CVPR*, pages 6024–6033, 2019.

[Zhang *et al.*, 2019c] Pingping Zhang, Wei Liu, Huchuan Lu, and Chunhua Shen. Salient object detection with lossless feature reflection and weighted structural loss. *TIP*, 28(6):3048–3060, 2019.

[Zhang *et al.*, 2020a] Lu Zhang, Jianming Zhang, Zhe Lin, Radomír Měch, Huchuan Lu, and You He. Unsupervised video object segmentation with joint hotspot tracking. In *ECCV*, pages 490–506, 2020.

[Zhang *et al.*, 2020b] Pingping Zhang, Wei Liu, Yinjie Lei, and Huchuan Lu. Semantic scene labeling via deep nested level set. *TITS*, 22(11):6853–6865, 2020.

[Zhang *et al.*, 2020c] Pingping Zhang, Wei Liu, Yinjie Lei, Hongyu Wang, and Huchuan Lu. Rapnet: Residual atrous pyramid network for importance-aware street scene parsing. *TIP*, 29:5010–5021, 2020.

[Zhang *et al.*, 2021] Pingping Zhang, Wei Liu, Yi Zeng, Yinjie Lei, and Huchuan Lu. Looking for the detail and context devils: High-resolution salient object detection. *TIP*, 30:3204–3216, 2021.

[Zhang *et al.*, 2023] Lian Zhang, Zhengliang Liu, Lu Zhang, Zihao Wu, Xiaowei Yu, Jason Holmes, Hongying Feng, Haixing Dai, Xiang Li, Quanzheng Li, et al. Segment anything model (sam) for radiation oncology. *arXiv preprint arXiv:2306.11730*, 2023.

[Zhang *et al.*, 2024] Pingping Zhang, Tianyu Yan, Yang Liu, and Huchuan Lu. Fantastic animals and where to find them: Segment any marine animal with dual sam. *CVPR*, 2024.

[Zhao and Wu, 2019] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *CVPR*, pages 3085–3094, 2019.

[Zhao *et al.*, 2023] Qihan Zhao, Xiaofeng Zhang, Hao Tang, Chaochen Gu, and Shanying Zhu. Enlighten-anything: When segment anything model meets low-light image enhancement. *arXiv preprint arXiv:2306.10286*, 2023.

[Zheng *et al.*, 2021] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, pages 6881–6890, 2021.

[Zhou *et al.*, 2018] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *MICCAI*, pages 3–11, 2018.