山东大学
SHANDONG UNIVERSITY

第三届先进遥感国际研讨会
暨2025年武汉遥感周

新型网络架构驱动的水下内容理解

UNDERSTANDING UNDERWATER SCENES VIA NEXT-GENERATION NETWORK DESIGNS

报告人：丛润民

山东大学控制科学与工程学院

机器智能与系统控制教育部重点实验室

—— 学无止境 气有浩然 ——

Outline

- Introduction
- Technical Methods
  - ➢ Diving Into Underwater: Segment Anything Model Guided Underwater Salient Instance Segmentation And A Large-scale Dataset (ICML’24)
  - ➢ UIS-Mamba: Exploring Mamba for Underwater Instance Segmentation via Dynamic Tree Scan and Hidden State Weaken (ACM MM’25)

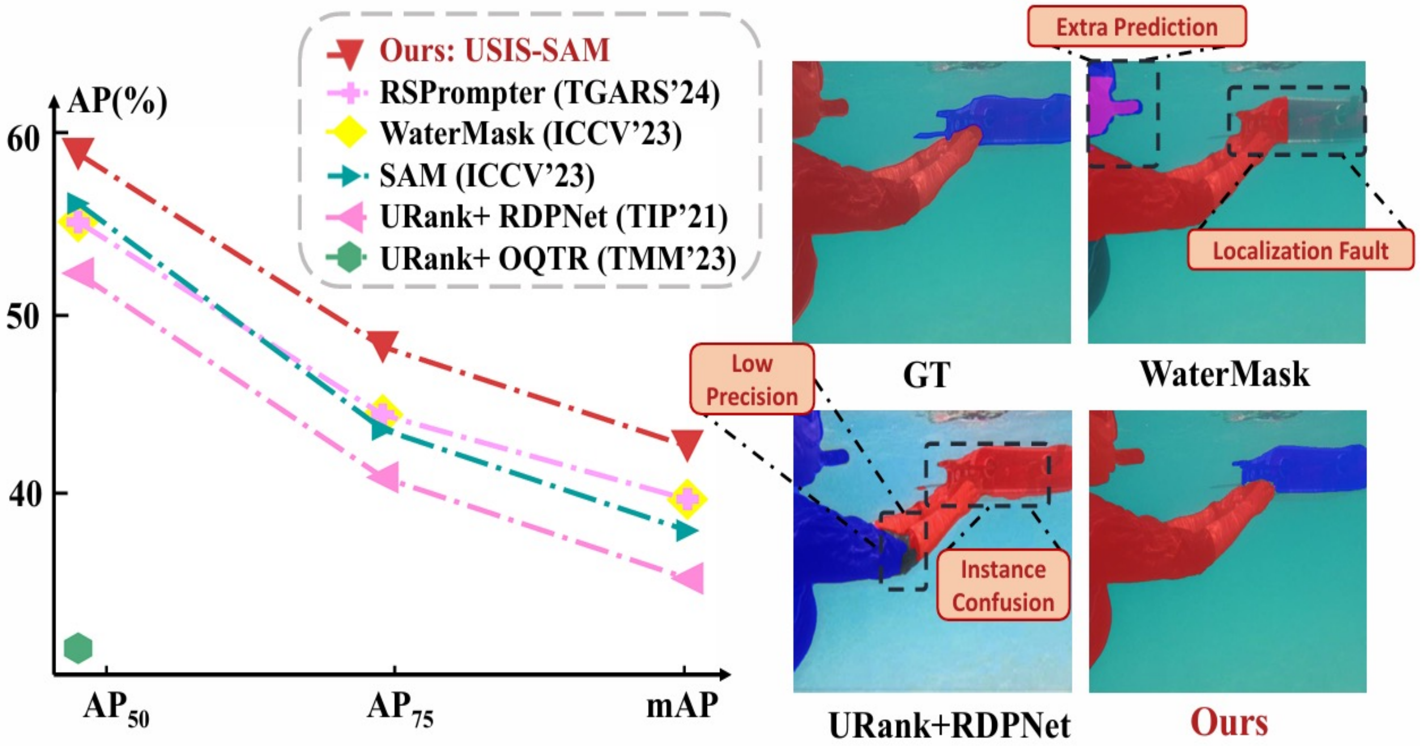
Diving into Underwater: Segment Anything Model Guided Underwater Salient Instance Segmentation and A Large-scale Dataset

Shijie Lian, Ziyi Zhang, Hua Li, Wenjie Li, Tianruo Yang, Sam Kwong, Runmin Cong

IEEE International Conference on Machine Learning, 2024

https://github.com/LiamLian0727/USIS10K

Introduction



- **Salient Instance Segmentation (SIS)**, an emerging and promising visual task, aims to segment out visually salient objects in a scene and distinguish individual salient instances, which is beneficial for vision tasks such as marine resource exploration and underwater human-computer interaction.

- However, 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**.

Motivation

- On the one hand, there is **no general underwater salient image instance segmentation dataset** to promote training and evaluation of the underwater salient instance segmentation models.
- On the other hand, even state-of-the-art SIS models trained on large-scale land-based datasets coupled with the best underwater image enhancement algorithms **cannot achieve satisfactory performance in underwater environments**.
- To alleviate this issue, we construct the **first large-scale underwater salient instance segmentation (USIS) dataset**, **USIS10K**, aiming to promote the development of salient instance segmentation for underwater tasks.
- Simultaneously, we first attempt to apply Segment Anything Model (SAM) to underwater salient instance segmentation and propose **USIS-SAM**, aiming to **improve the segmentation accuracy in complex underwater scenes**.

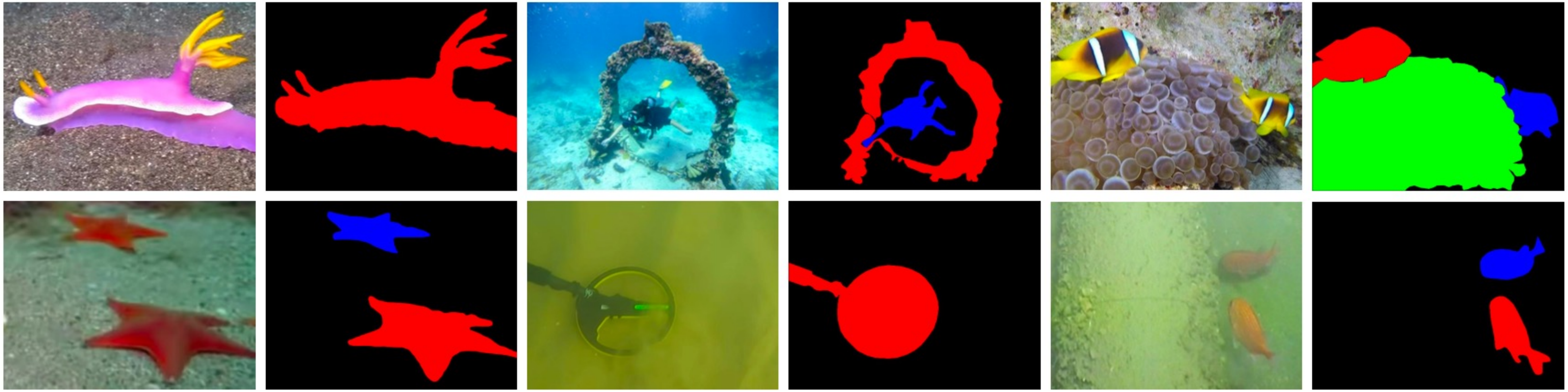
Contributions

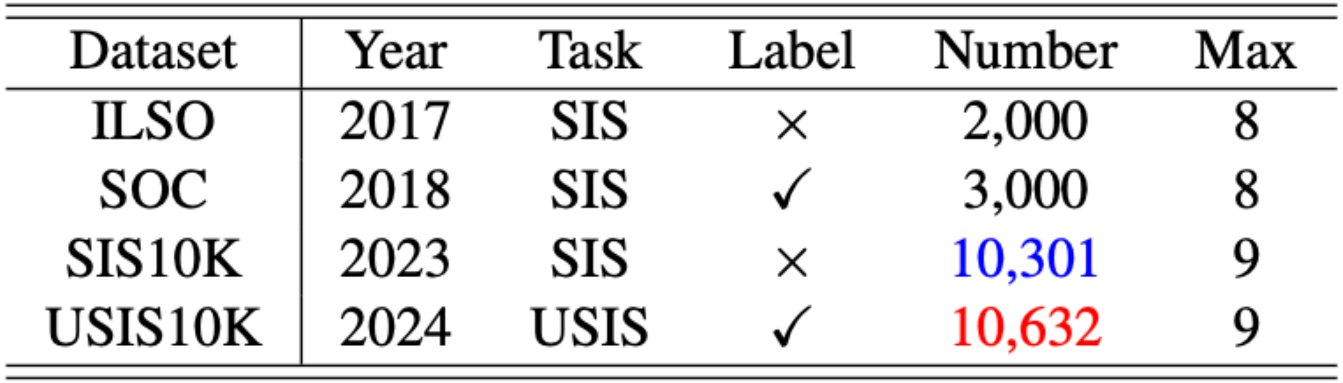
a) We construct the **first large-scale dataset, USIS10K**, for the underwater salient instance segmentation task, which contains **10,632 images and pixel-level annotations of 7 categories**. As far as we know, this is **the largest salient instance segmentation dataset**, and includes Class-Agnostic and Multi-Class labels simultaneously.

b) We propose the **first underwater salient instance segmentation model, USIS-SAM**, as far as we know. In USIS-SAM, we design **Underwater Adaptive ViT Encoder** to incorporate underwater visual prompts into network via adapters, and **Salient Feature Prompter Generator** to automatically generate salient prompters, guiding an end-to-end segmentation network.

c) Extensive public evaluation criteria and large numbers of experiments verify the effectiveness of our USIS10K dataset and USIS-SAM.

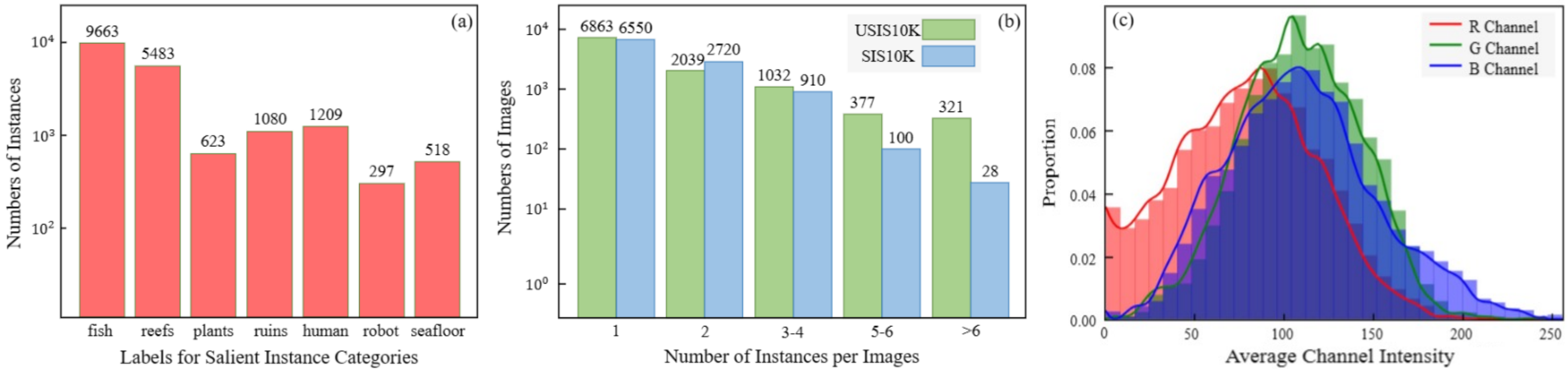
USIS10K Dataset





Dataset	Year	Task	Label	Number	Max
ILSO	2017	SIS	×	2,000	8
SOC	2018	SIS	✓	3,000	8
SIS10K	2023	SIS	×	10,301	9
USIS10K	2024	USIS	✓	10,632	9

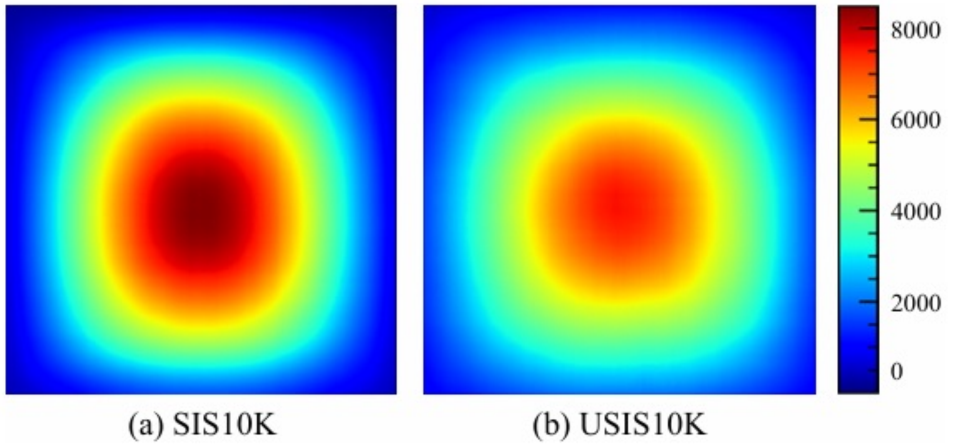
Dataset Statistic and Challenges



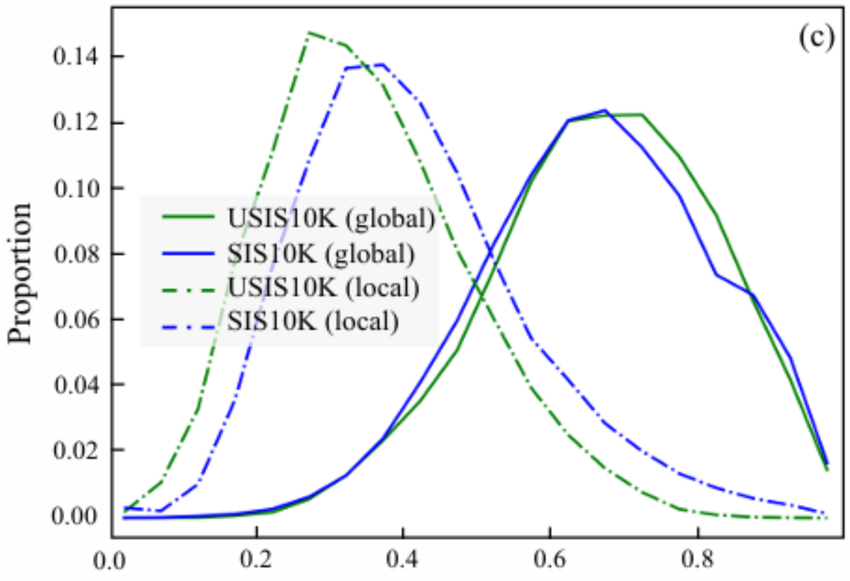
- **Challenge in the number of instance.**
  In USIS10K dataset, multiple salient instances may exist in a single image. There are 1731 images with more than 3 salient instances in our dataset, accounting for 16.3% of the total.
- **Challenges in small or large instances.**
  In USIS10K dataset, the 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).
- **Challenges in channel intensity of underwater images.**
  Optical images inevitably suffer from color attenuation due to the selective absorption of water at different wavelengths. This poses an additional challenge for the network to properly understand and handle the image color distortion caused by this attenuation

Dataset Statistic and Challenges

- **Location of Salient Objects (Less central bias).**
  In the SIS10k dataset, approximately 13.5% of the locations have fewer than 1000 instances and 32% have fewer than 2,000 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.



A set of salient maps from our dataset and SIS10K



Global/Local Color Contrast of Salient Instance

- **Color Contrast of Salient Instance.**
  Saliency is often related to the contrast between foreground and background, and it is critical to check whether salient instances are easy to detect. It can be seen that the global contrast of USIS10K is slightly higher than that of SIS10K. In addition, the local contrast of USIS10K at salient instances is lower than that of SIS10K. This poses a greater challenge in accurately segmenting the salient instance masks at the network boundary portion.

USIS-SAM

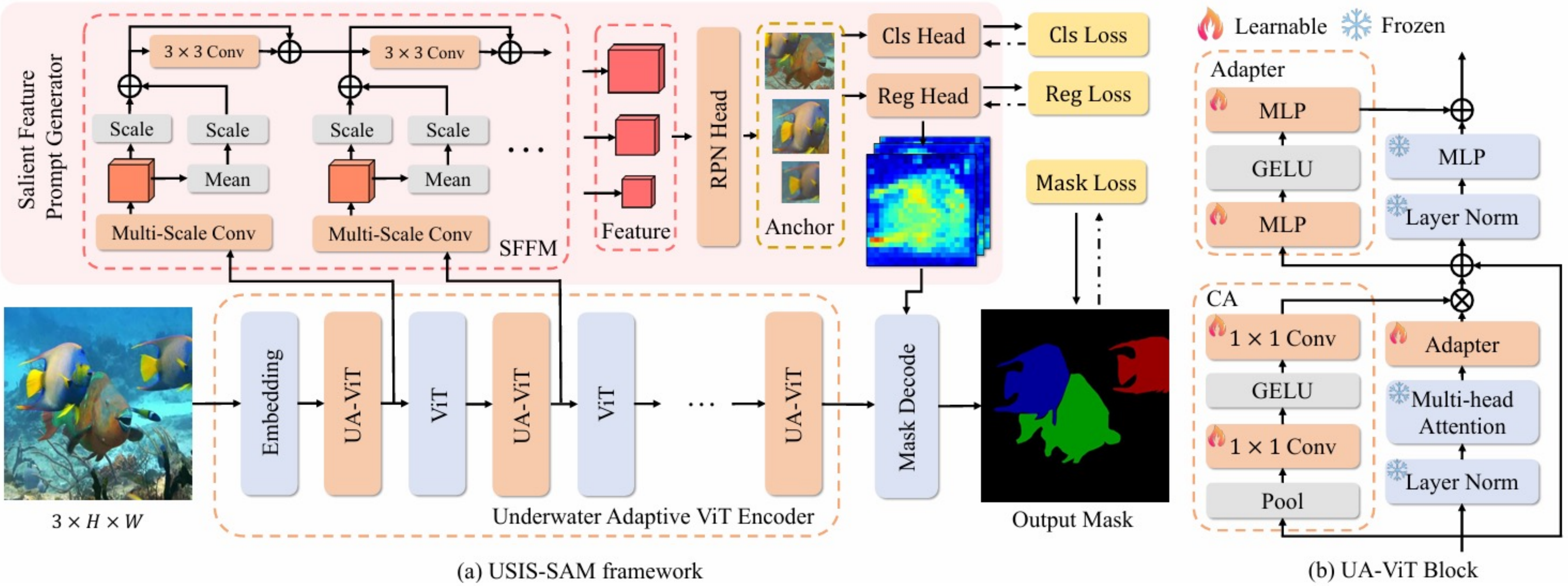


***Figure 1.* (a) USIS-SAM framework.** The USIS-SAM framework modifies the SAM by adding the Underwater Adaptive ViT Encoder and the Salient Feature Prompt Generator. **(b) The structure of UA-ViT.** In the figure, SFFM stands for Salient Feature Fusion Module, CA stands for Channel Adapter.

Underwater Adaptive ViT Encoder

In USIS-SAM, we design the **Underwater Adaptive ViT (UA-ViT)** to **integrate underwater visual prompts into the network via adapter and channel adapter**. UA-ViT enables a more effective utilization of the SAM image encoder in underwater scenarios.

***Adapter*:**

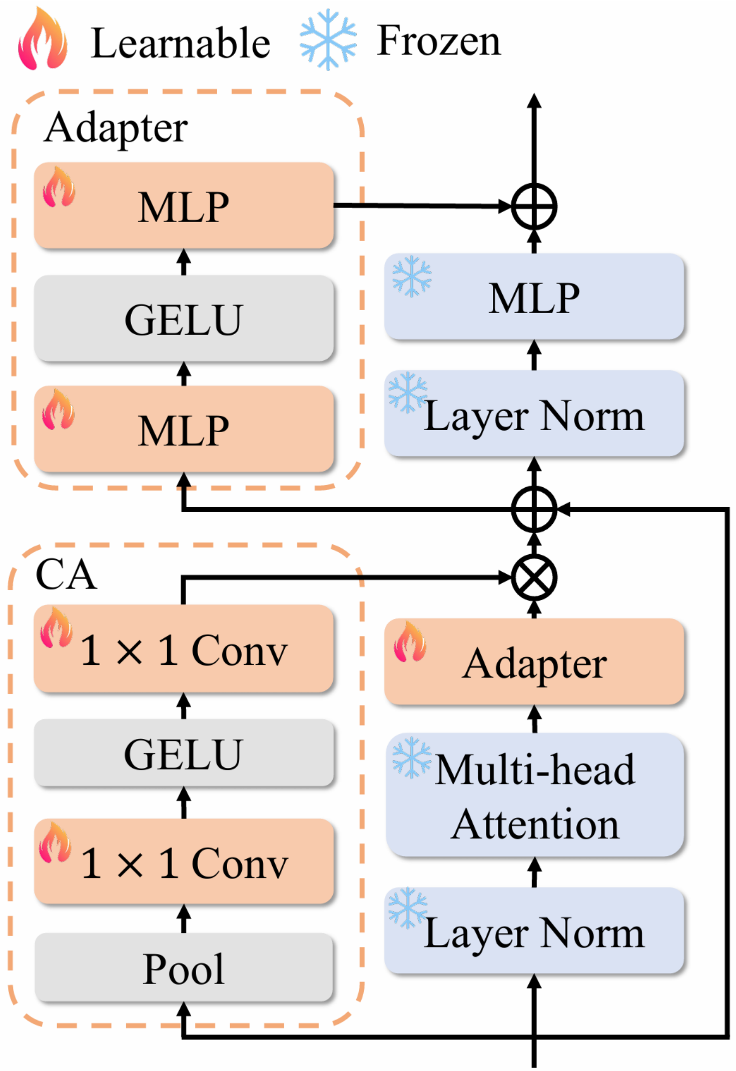
$$P = MLP_{out}\left(\sigma\left(MLP_{prompt}(F)\right)\right),$$

where F is the input feature, and P is the output prompt for each adapter layer. σ is the activation function.

***Channel Adapter*:**

$$C = F \times Conv_{up}\left(\sigma\left(Conv_{down}\left(Pool(F)\right)\right)\right),$$

where C is the output feature after channel adapter, $Conv$ is a 1×1 convolutional layer, and $Pool$ is an average pooling layer.



Salient Feature Prompt Generator

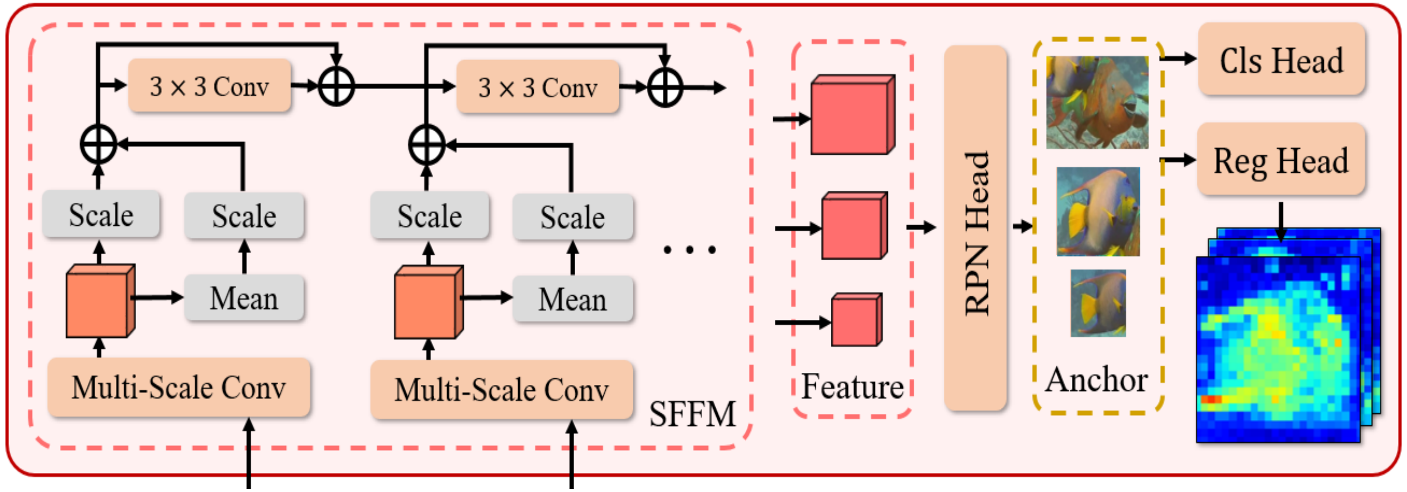


Figure 2. The structure of Salient Feature Prompt Generator (SFPG). The SFPG module efficiently filters out non-salient noise, allowing for robust feature aggregation of salient instances.

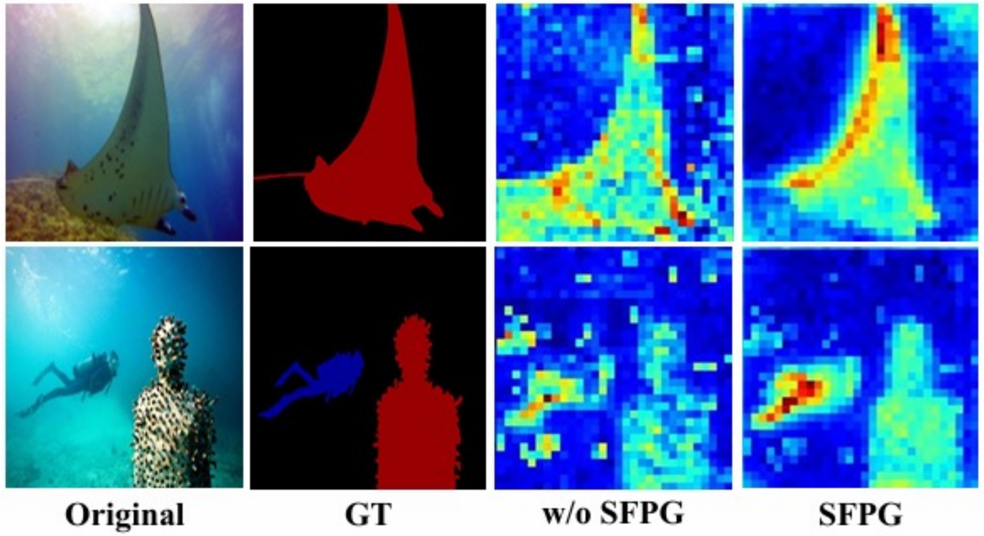
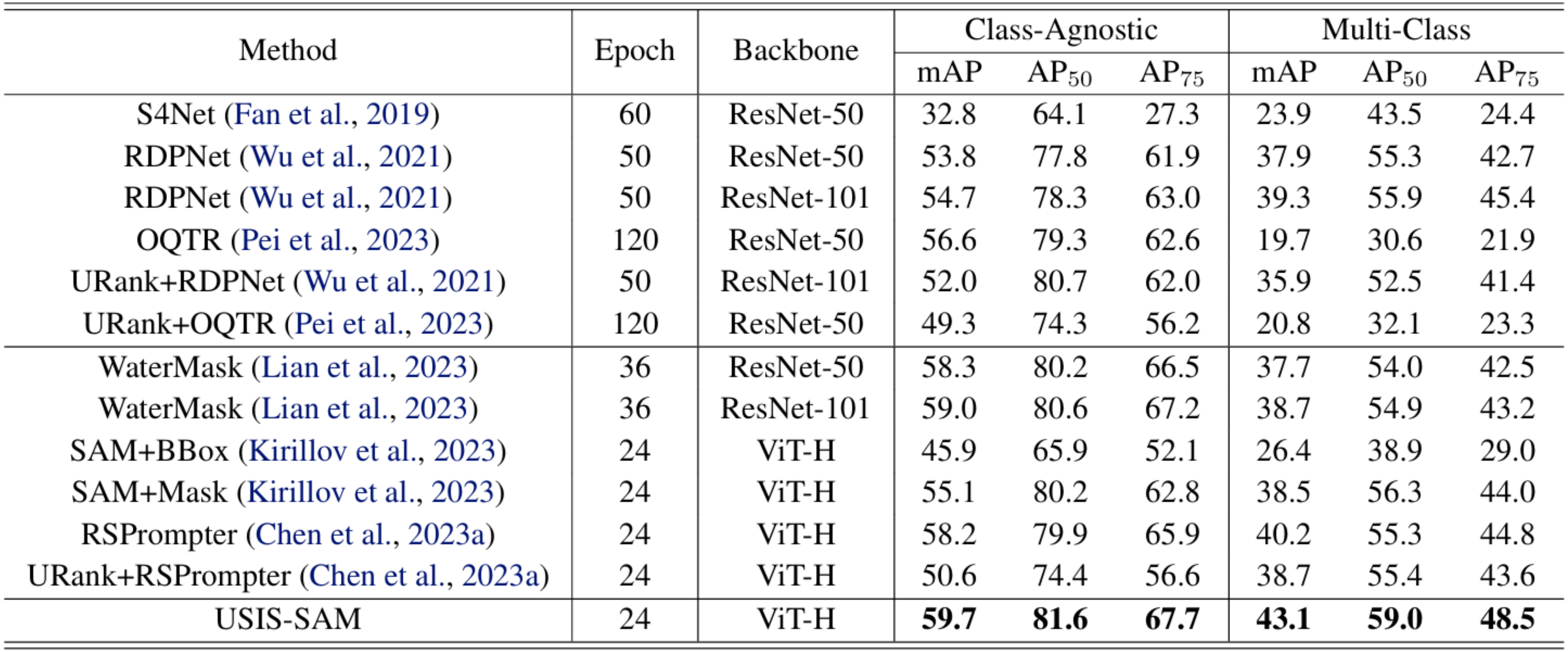


Figure 3. Visualize features generated by the Salient Feature Prompt Generator.

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. Therefore, **we design the Salient Feature Prompt Generator to directly predict prompts embedding of salient instances, enabling end-to-end performing the USIS task**

Experiments



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**

***Table 1*. Quantitative comparisons with state-of-the-arts on the USIS10K datasets.** Urank stands for an underwater image enhancement method in UnderwaterRanker (AAAI 2023 oral), SAM+BBox uses inference results from Faster RCNN as prompts for prediction, SAM+Mask stands for Mask RCNN networks use SAM as backbone. The RSPrompter in the table is the RSPrompter-anchor framework.

Experiments

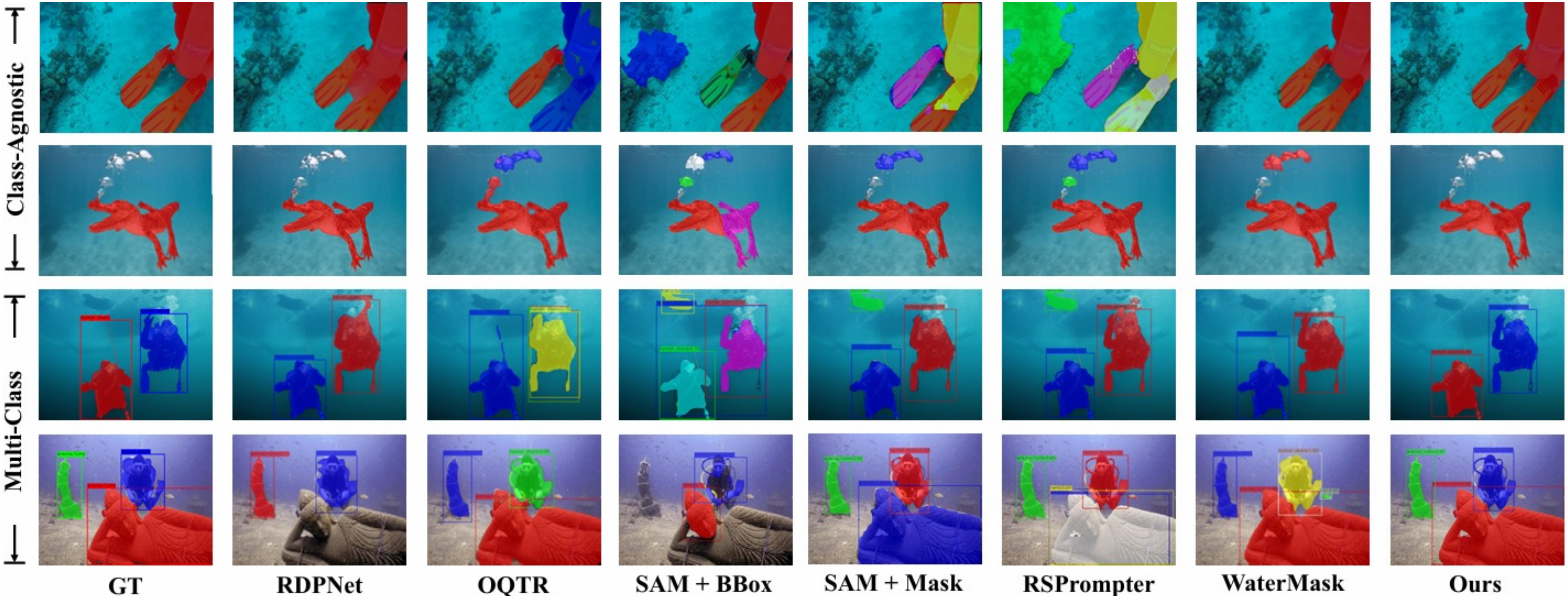
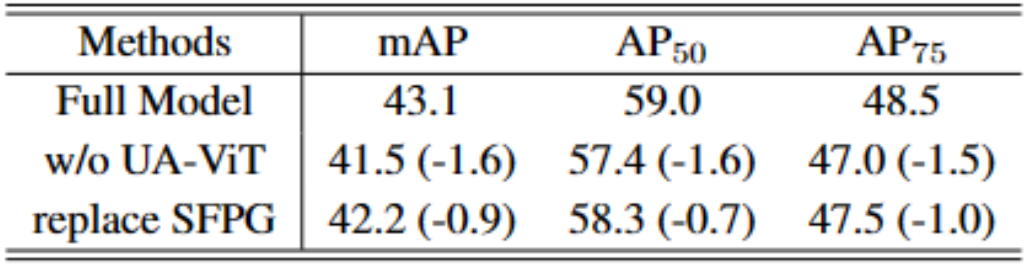


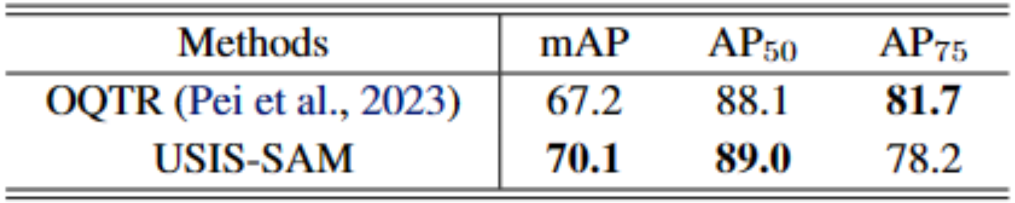
***Figure 4.* Qualitative comparison on the USIS10K dataset.** Each salient instance is represented by a unique color, and the segmented mask is superimposed on the image.

Ablation Study



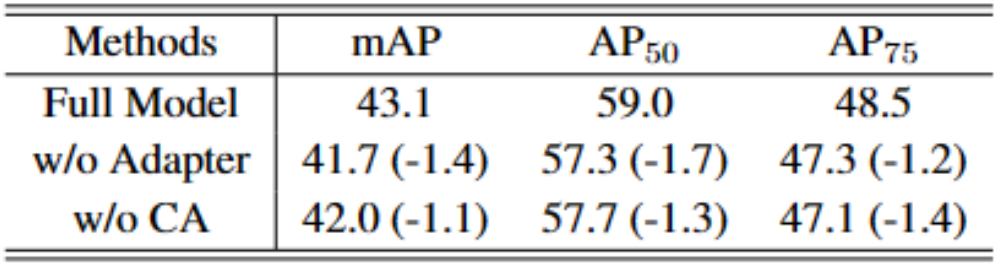
Methods	mAP	AP_{50}	AP_{75}
Full Model	43.1	59.0	48.5
w/o UA-ViT	41.5 (-1.6)	57.4 (-1.6)	47.0 (-1.5)
replace SFPG	42.2 (-0.9)	58.3 (-0.7)	47.5 (-1.0)

Table 2. Effectiveness of each component in USIS-SAM, replace SFPG means to use Multi-scale Feature Enhancer Module in RSPrompter (TGARS’24) instead of SFPG.



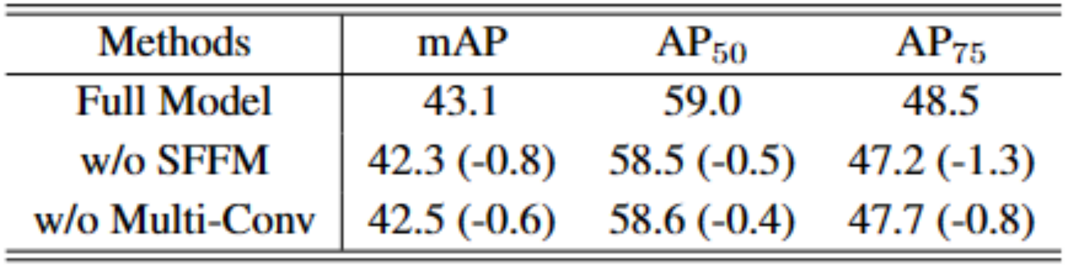
Methods	mAP	AP_{50}	AP_{75}
OQTR (Pei et al., 2023)	67.2	88.1	**81.7**
USIS-SAM	**70.1**	**89.0**	78.2

Table 3. Generalization Ability of USIS-SAM. Quantitative comparisons with state-of-the-art methods on SIS10K indicate that USIS-SAM did not overfit our dataset.



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)

Table 4. 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 SFFM	42.3 (-0.8)	58.5 (-0.5)	47.2 (-1.3)
w/o Multi-Conv	42.5 (-0.6)	58.6 (-0.4)	47.7 (-0.8)

Table 5. 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-scale convolution module.

Conclusion and Future Work

- We have constructed the **first general underwater salient image instance segmentation dataset** with pixel-level annotations, which enables us to **comprehensively explore the underwater salient instance segmentation task**.
- we first attempt to apply Segment Anything (SAM) model to underwater salient instance segmentation and propose **USIS-SAM**, aiming to improve the segmentation accuracy in complex underwater scenes. Extensive experiments have validated the **effectiveness** and **generalizability** of USIS-SAM.
- In future work, we plan to extend the USIS datasets to **broader and more challenging underwater images and underwater videos**.

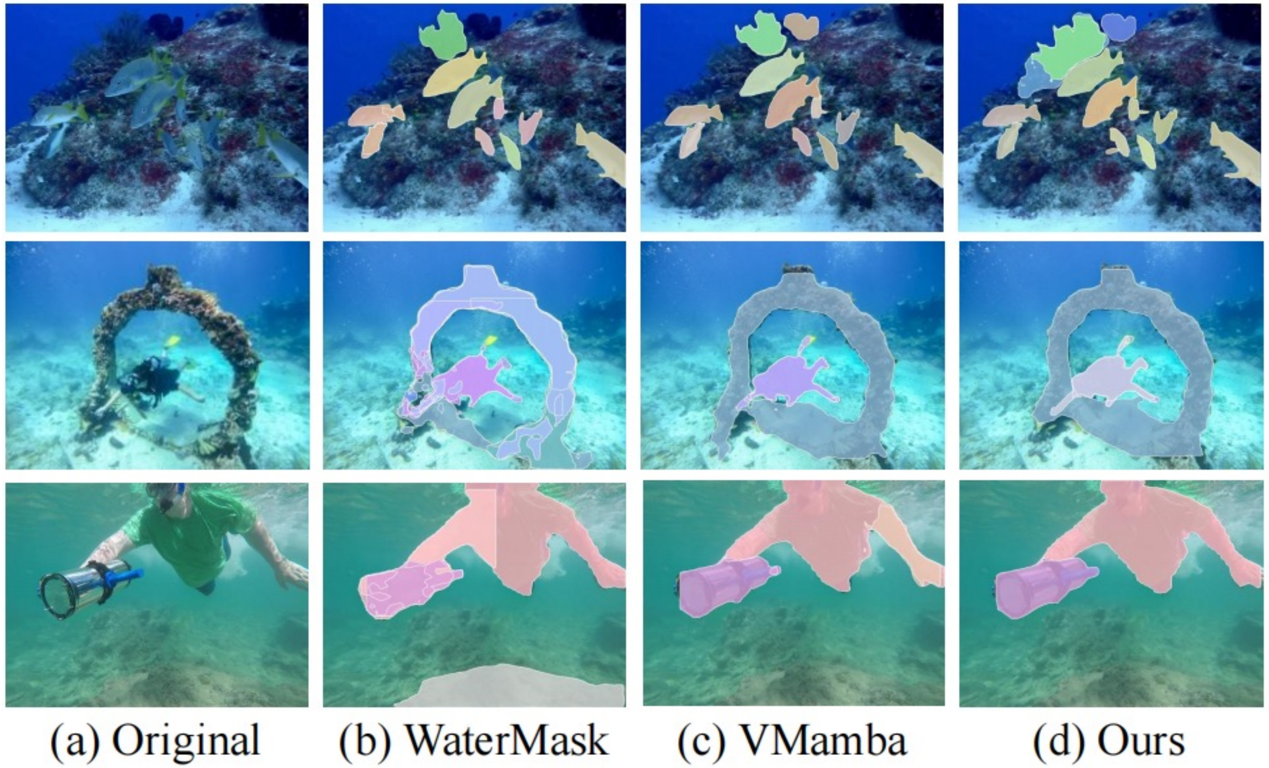
ACM multimedia
Dublin, Ireland 27-31.10.2025

UIS-Mamba: Exploring Mamba for Underwater Instance Segmentation via Dynamic Tree Scan and Hidden State Weaken

Runmin Cong, Zongji Yu, Hao Fang, Haoyan Sun, Sam Kwong

*ACM International Conference on Multimedia, **2025***

Introduction



- To address the growing demand for underwater exploration, we have increasingly focused on developing **underwater visual tasks**, particularly the UIS tasks.
- **Mamba** is highly suitable for processing image segmentation tasks with long sequence features

- **Underwater Instance Segmentation (UIS)** and **Underwater Salient Instance Segmentation (USIS)** are emerging tasks: UIS aims to segment all instance objects in underwater scenes with category distinction, while USIS specifically targets visually salient objects by differentiating prominent instances.

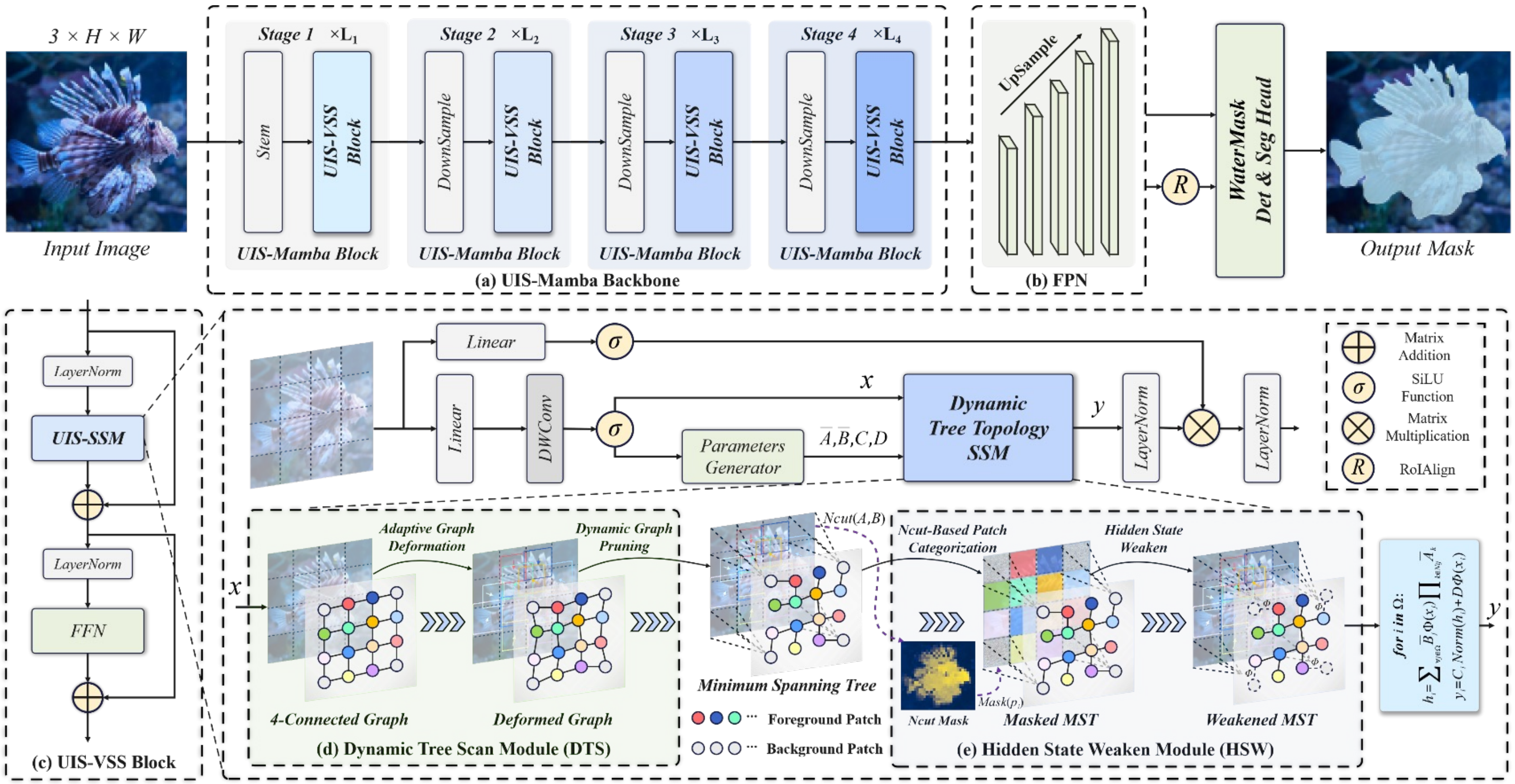
Contributions

a) We reveal the *limitations* of existing *vision Mamba* in underwater scenes and propose first Mamba-based underwater instance segmentation model **UIS-Mamba**, which improves the vision Mamba's ability to *understand the features of underwater instance objects* and provides new insights for Mamba's migration to underwater tasks.

b) We design **Dynamic Tree Scan (DTS)** module and **Hidden State Weaken (HSW)** module to introduce vision Mamba into underwater scenes. The **DTS** module allows the patches to *dynamically offset and scale*, and the **HSW** module weakens the interference of the complex underwater background on the *hidden state update*.

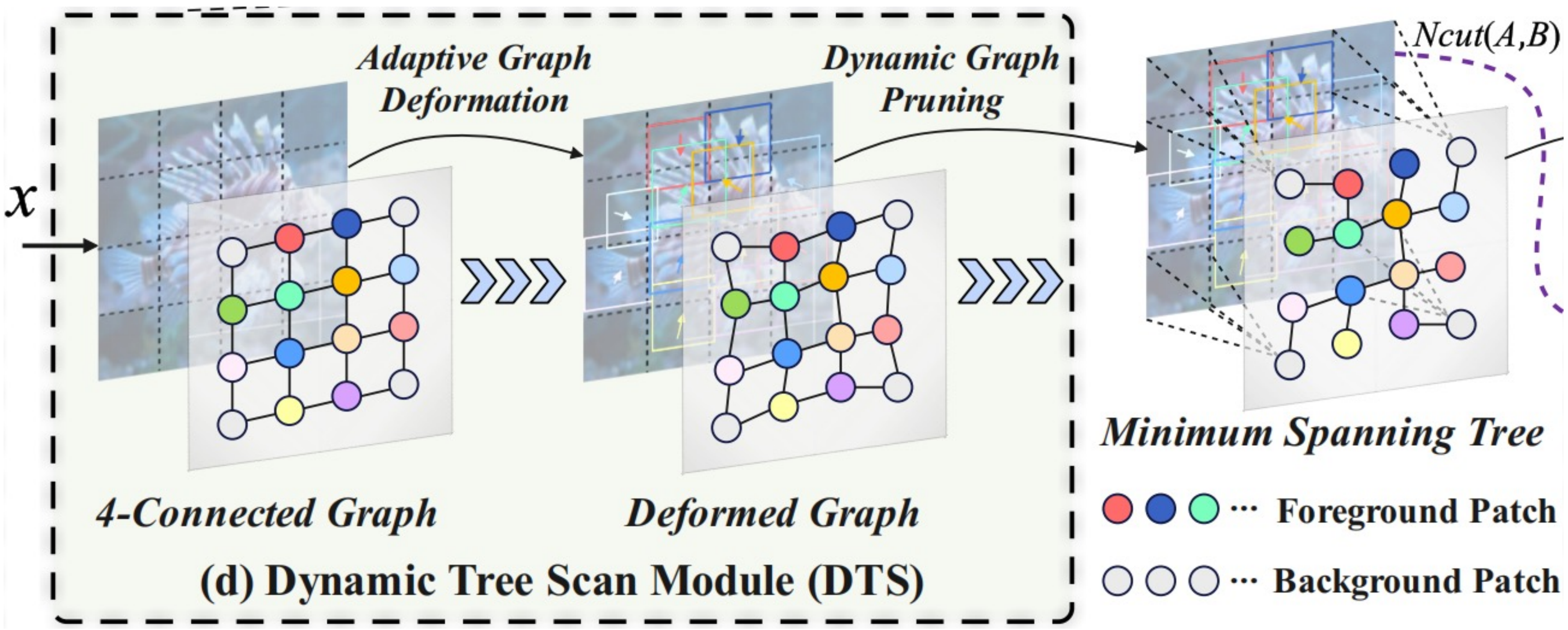
c) Experiments on the **UIIS** and **USIS10K** regarding instance segmentation and salient instance segmentation show that our proposed **UIS-Mamba** achieves *state-of-the-art segmentation performance*.

UIS-Mamba



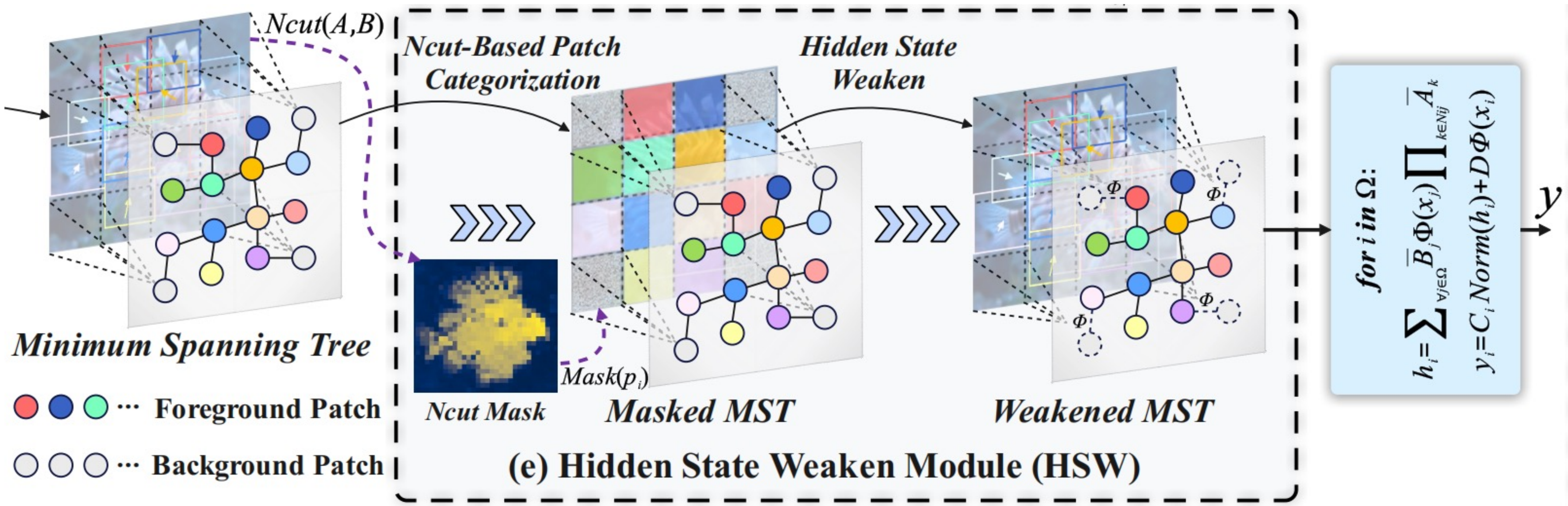
- **Underwater image degradation：** $I(x) = J(x) \cdot e^{-\beta d(x)} + B_\infty (1 - e^{-\beta d(x)})$

Dynamic Tree Scan



- **Adaptive Graph Deformation** dynamically adjusts the feeling field to preserve topology.
- **Dynamic Graph Pruning** eliminates redundant feature connections.

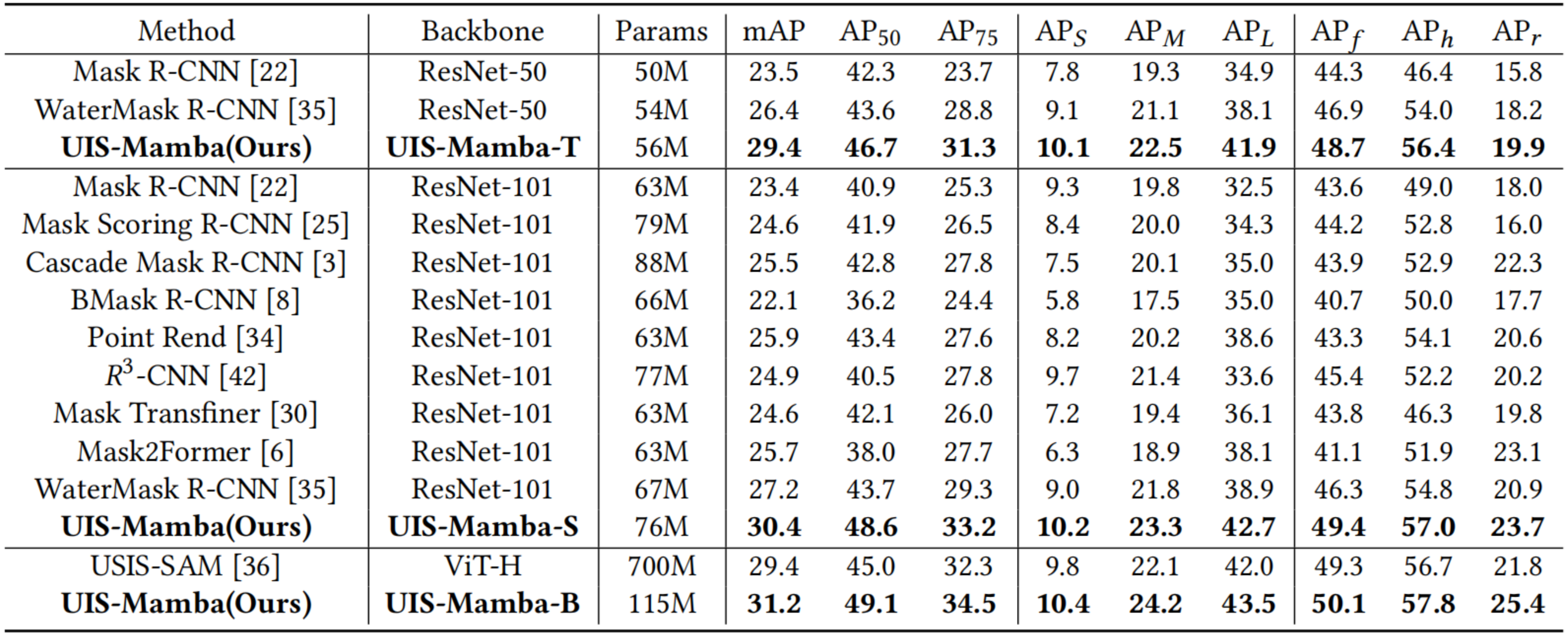
Hidden State Weaken


Ncut(A,B)
Ncut-Based Patch Categorization
Hidden State Weaken
Minimum Spanning Tree
●●● ··· Foreground Patch
○○○ ··· Background Patch
Ncut Mask
Mask(p_i)
Masked MST
Φ
Weakened MST
(e) Hidden State Weaken Module (HSW)
for i in Ω:
$h_i=\sum_{\forall j\in\Omega}\bar{B}_j\Phi(x_j)\prod_{k\in N_{ij}}\bar{A}_k$
$y_i=C_i Norm(h_i)+D\Phi(x_i)$
y

- ➢ **Ncut-Based Patch Categorization** perform the category determination of patches through a foreground-background separation Ncut algorithm.
- ➢ **Hidden State Weaken** suppresses the influence of background patches on hidden state updates

UIIS Experiment

Table 1: Results on UIIS with our UIS-Mamba.



Method	Backbone	Params	mAP	AP$_{50}$	AP$_{75}$	AP$_{S}$	AP$_{M}$	AP$_{L}$	AP$_{f}$	AP$_{h}$	AP$_{r}$
Mask R-CNN [22]	ResNet-50	50M	23.5	42.3	23.7	7.8	19.3	34.9	44.3	46.4	15.8
WaterMask R-CNN [35]	ResNet-50	54M	26.4	43.6	28.8	9.1	21.1	38.1	46.9	54.0	18.2
UIS-Mamba(Ours)	**UIS-Mamba-T**	56M	**29.4**	**46.7**	**31.3**	**10.1**	**22.5**	**41.9**	**48.7**	**56.4**	**19.9**
Mask R-CNN [22]	ResNet-101	63M	23.4	40.9	25.3	9.3	19.8	32.5	43.6	49.0	18.0
Mask Scoring R-CNN [25]	ResNet-101	79M	24.6	41.9	26.5	8.4	20.0	34.3	44.2	52.8	16.0
Cascade Mask R-CNN [3]	ResNet-101	88M	25.5	42.8	27.8	7.5	20.1	35.0	43.9	52.9	22.3
BMask R-CNN [8]	ResNet-101	66M	22.1	36.2	24.4	5.8	17.5	35.0	40.7	50.0	17.7
Point Rend [34]	ResNet-101	63M	25.9	43.4	27.6	8.2	20.2	38.6	43.3	54.1	20.6
R^3-CNN [42]	ResNet-101	77M	24.9	40.5	27.8	9.7	21.4	33.6	45.4	52.2	20.2
Mask Transfiner [30]	ResNet-101	63M	24.6	42.1	26.0	7.2	19.4	36.1	43.8	46.3	19.8
Mask2Former [6]	ResNet-101	63M	25.7	38.0	27.7	6.3	18.9	38.1	41.1	51.9	23.1
WaterMask R-CNN [35]	ResNet-101	67M	27.2	43.7	29.3	9.0	21.8	38.9	46.3	54.8	20.9
UIS-Mamba(Ours)	**UIS-Mamba-S**	76M	**30.4**	**48.6**	**33.2**	**10.2**	**23.3**	**42.7**	**49.4**	**57.0**	**23.7**
USIS-SAM [36]	ViT-H	700M	29.4	45.0	32.3	9.8	22.1	42.0	49.3	56.7	21.8
UIS-Mamba(Ours)	**UIS-Mamba-B**	115M	**31.2**	**49.1**	**34.5**	**10.4**	**24.2**	**43.5**	**50.1**	**57.8**	**25.4**

mAP: Mean Average Precision

AP50: Average Precision at 50% IOU

AP75: Average Precision at 75% IOU

For UIS-Mamba-T backbone, our method achieves 29.4, 46.7, and 31.3 AP in mAP, AP50, and AP75, which is 3.0, 3.1, and 2.5 higher than SOTA method Water Mask R-CNN with ResNet-50 backbone

UIIS Experiment

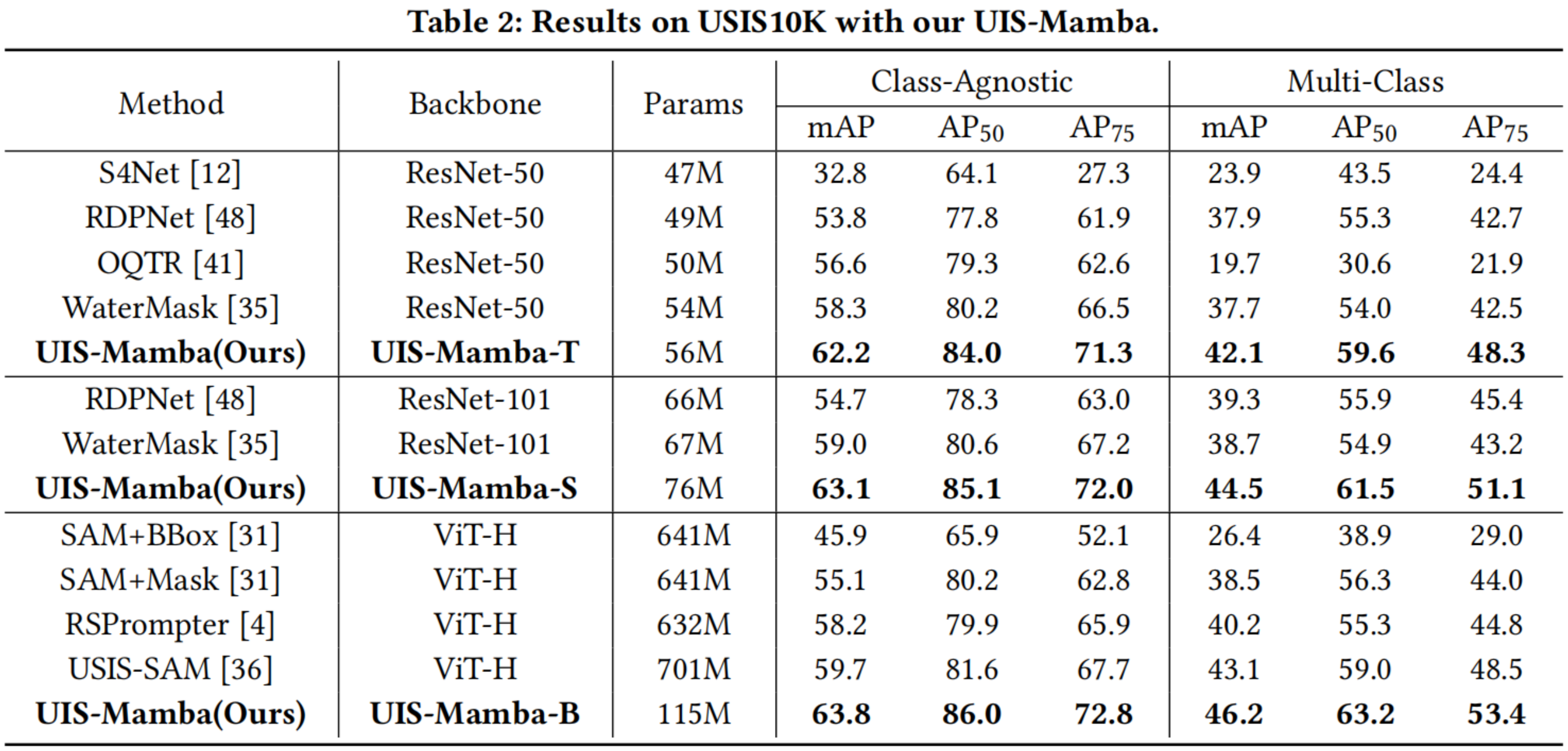


Table 2: Results on USIS10K with our UIS-Mamba.

Method	Backbone	Params	Class-Agnostic			Multi-Class		
			mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
S4Net [12]	ResNet-50	47M	32.8	64.1	27.3	23.9	43.5	24.4
RDPNet [48]	ResNet-50	49M	53.8	77.8	61.9	37.9	55.3	42.7
OQTR [41]	ResNet-50	50M	56.6	79.3	62.6	19.7	30.6	21.9
WaterMask [35]	ResNet-50	54M	58.3	80.2	66.5	37.7	54.0	42.5
UIS-Mamba(Ours)	**UIS-Mamba-T**	56M	**62.2**	**84.0**	**71.3**	**42.1**	**59.6**	**48.3**
RDPNet [48]	ResNet-101	66M	54.7	78.3	63.0	39.3	55.9	45.4
WaterMask [35]	ResNet-101	67M	59.0	80.6	67.2	38.7	54.9	43.2
UIS-Mamba(Ours)	**UIS-Mamba-S**	76M	**63.1**	**85.1**	**72.0**	**44.5**	**61.5**	**51.1**
SAM+BBox [31]	ViT-H	641M	45.9	65.9	52.1	26.4	38.9	29.0
SAM+Mask [31]	ViT-H	641M	55.1	80.2	62.8	38.5	56.3	44.0
RSPrompter [4]	ViT-H	632M	58.2	79.9	65.9	40.2	55.3	44.8
USIS-SAM [36]	ViT-H	701M	59.7	81.6	67.7	43.1	59.0	48.5
UIS-Mamba(Ours)	**UIS-Mamba-B**	115M	**63.8**	**86.0**	**72.8**	**46.2**	**63.2**	**53.4**

Class-agnosticsalient instance segmentation can be essentially understood as foreground instance segmentation exclusively focusing on salient regions in images.

Multi-class salient instance segmentation can be conceptually considered as a fusion of tasks from class-independent salient instance segmentation and salient instance class prediction

Ablation Study

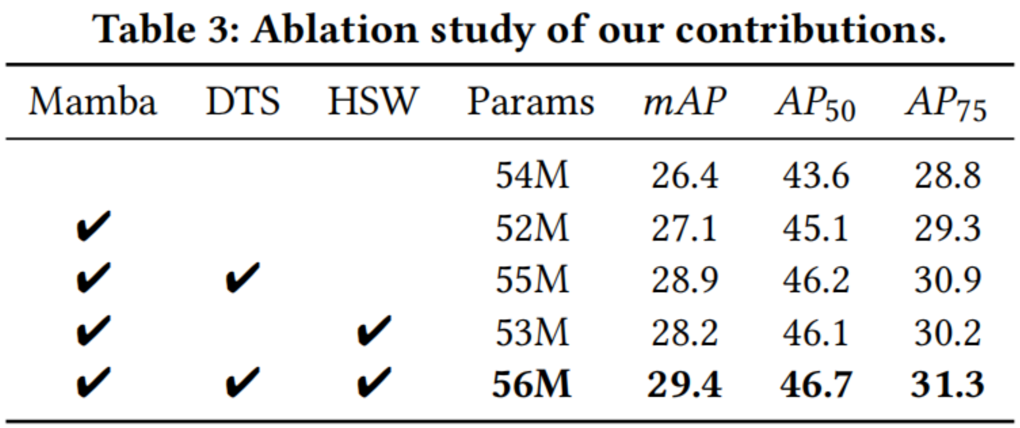


Table 3: Ablation study of our contributions.

Mamba	DTS	HSW	Params	*mAP*	AP_{50}	AP_{75}
			54M	26.4	43.6	28.8
✔			52M	27.1	45.1	29.3
✔	✔		55M	28.9	46.2	30.9
✔		✔	53M	28.2	46.1	30.2
✔	✔	✔	**56M**	**29.4**	**46.7**	**31.3**

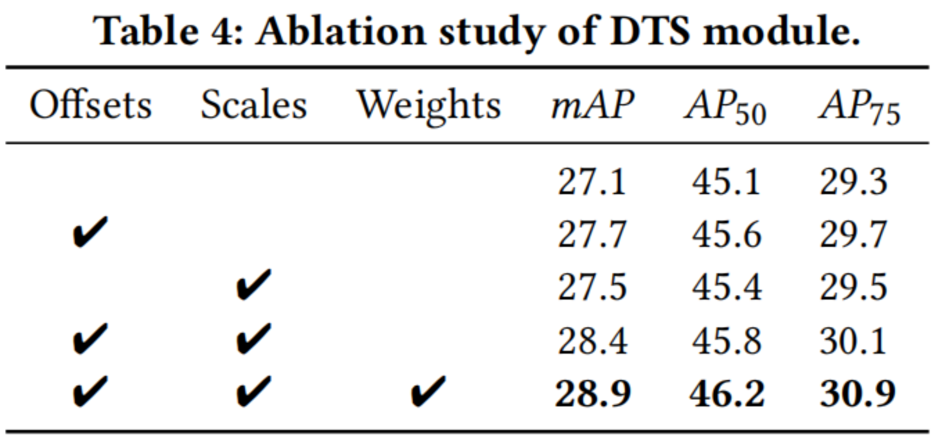


Table 4: Ablation study of DTS module.

Offsets	Scales	Weights	*mAP*	AP_{50}	AP_{75}
			27.1	45.1	29.3
✔			27.7	45.6	29.7
	✔		27.5	45.4	29.5
✔	✔		28.4	45.8	30.1
✔	✔	✔	**28.9**	**46.2**	**30.9**

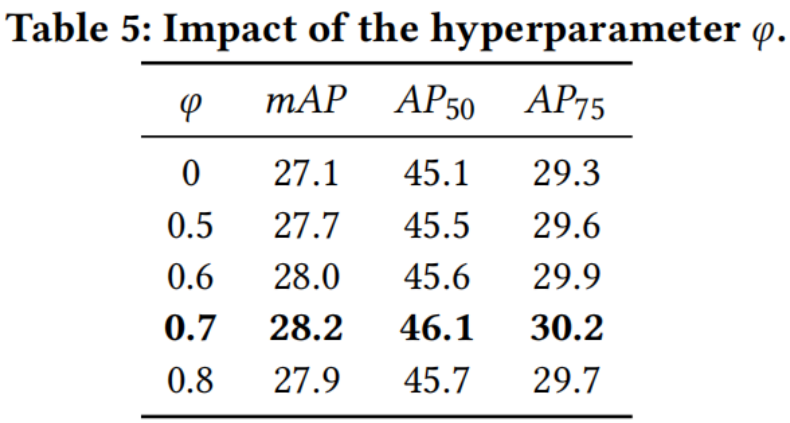


Table 5: Impact of the hyperparameter φ.

φ	*mAP*	AP_{50}	AP_{75}
0	27.1	45.1	29.3
0.5	27.7	45.5	29.6
0.6	28.0	45.6	29.9
0.7	**28.2**	**46.1**	**30.2**
0.8	27.9	45.7	29.7

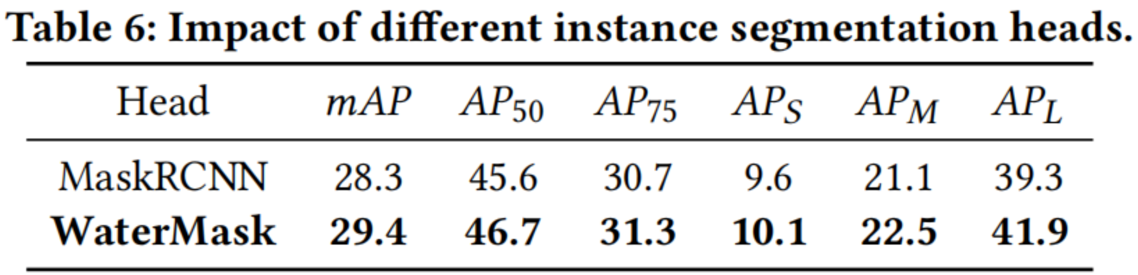


Table 6: Impact of different instance segmentation heads.

Head	*mAP*	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MaskRCNN	28.3	45.6	30.7	9.6	21.1	39.3
WaterMask	**29.4**	**46.7**	**31.3**	**10.1**	**22.5**	**41.9**

Qualitative Comparison

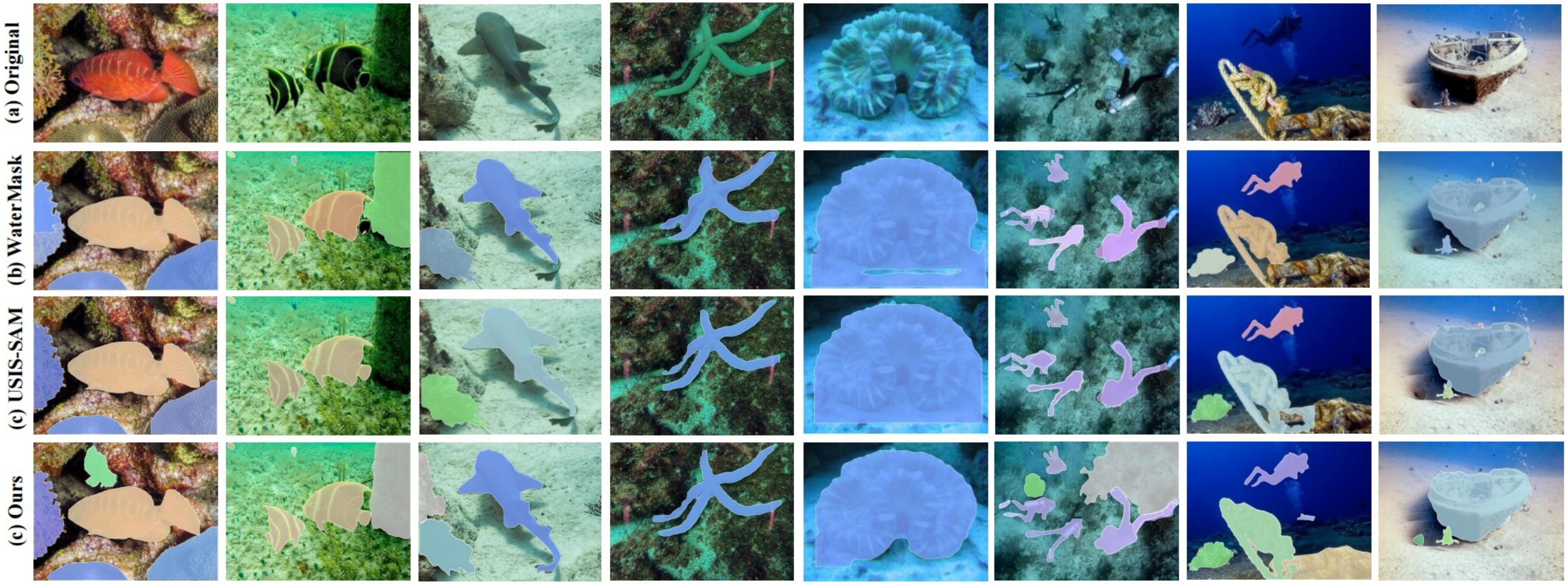


Figure 3: Qualitative comparison on the UIIS and USIS10K dataset. Each class of instance in the same image is represented by a unique color, and the segmented mask is superimposed on the image.

Conclusion

- This paper propose the first Mamba-based underwater instance segmentation model, UIS-Mamba, to port Mamba to underwater tasks through two improved modules to address the problems of segmentation confusion and semantic ambiguity in challenging underwater scenes.
- The DTS module allows the graph structure to be dynamically shifted and scaled to guide the minimum spanning tree and provides dynamic local sense fields. The HSW module suppresses interference from complex backgrounds and focuses the information flow of state propagation.
- UIS-Mamba achieves state-of-the-art performance on both the UIIS and the USIS10K datasets, while keeping the number of parameters and computational complexity low.

向海而兴
向海图强

THANKS FOR WATCHING

李华 副教授

于宗吉 本科生

Sam Kwong 教授（香港工程科学院院士）