ZMIS-SAM: Segment Anything Model Enhanced with Wavelet Transform for Zooplankton Microscopy Image Instance Segmentation

Dekun Yuan[1], Zhongwei Li[1,✉], Zheng Qiao[1], and Jie Zhang[1]

[1]College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao, 266580, China
ydk_libra0903@163.com, li.zhongwei@vip.163.com

Abstract. As primary consumers in the marine food chain, zooplankton play a crucial role in maintaining marine ecological balance. However, the Segment Anything Model (SAM) exhibits limited performance in microscopic image instance segmentation due to its lack of zooplankton-specific domain knowledge. To address these challenges, we propose a novel instance segmentation model based on SAM and wavelet transform (ZMIS-SAM), effectively tackling issues such as inaccurate classification, discontinuous segmentation of slender appendages, and incomplete boundary segmentation. Our framework incorporates three core innovations: ZM-ViT enhances SAM's capability to model zooplankton morphology and image intensity distributions through two lightweight adapters, the Neighboring Feature Aggregation Module (NFAM) improves continuous segmentation of semi-transparent slender appendages by integrating general-purpose and domain-specific features, and the Wavelet-based Multi-scale Multi-directional Feature Enhancement (WM2FE) module effectively recovers high-frequency details to refine boundary segmentation completeness. Extensive experiments demonstrate that ZMIS-SAM achieves state-of-the-art instance segmentation performance on the zooplankton dataset and exhibits strong generalization capability across multiple public cross-domain datasets. Code: ZMIS-SAM.

Keywords: Segment Anything Model (SAM) · Instance segmentation · Zooplankton recognition · Wavelet transform

1 Introduction

Zooplankton, serving as a crucial intermediate link in the marine food chain, plays a vital role in maintaining the stability of marine ecosystems [37]. Current methods for zooplankton identification can be broadly categorized into manual and intelligent approaches [33,54,55]. Traditional manual identification relies on biological experts observing and counting specimens under microscopes. While this method achieves high accuracy, it suffers from low efficiency and falls short of meeting the demands for efficient, large-scale analysis of zooplankton communities [35, 52]. With the advancement of artificial intelligence technologies,

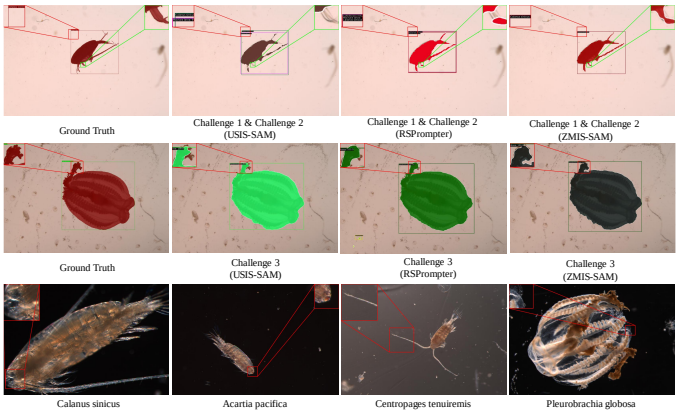

Fig. 1: Comparison of ZMIS-SAM with other state-of-the-art models on the ZMIS5K dataset. Challenge 1, Challenge 2, and Challenge 3 represent inaccurate classification, discontinuous segmentation of slender appendages, and incomplete boundary segmentation, respectively.

intelligent identification techniques based on deep learning methods, such as instance segmentation, offer a promising alternative for achieving efficient and accurate zooplankton identification.

Instance segmentation stands as one of the most challenging tasks in intelligent visual recognition, requiring not only pixel-wise dense prediction but also distinguishing between different object instances within the same category [34]. Two-stage methods, exemplified by Mask R-CNN [18], have become the mainstream paradigm in this field. These approaches explicitly generate region proposals and perform dense prediction to produce instance-level masks. While they demonstrate stronger segmentation performance on domain-specific datasets compared to single-stage methods, their effectiveness heavily depends on dataset scale, image quality, and the distribution and quantity of object instances. This characteristic makes it difficult for two-stage methods to achieve satisfactory instance segmentation performance in the special field of zooplankton. With their powerful zero-shot generalization capabilities learned from vast and diverse datasets, recent advances in large-scale pre-trained models [22,36,38] can extract transferable visual representations, providing a more efficient and flexible paradigm, for example, segmentation in specialized domains like zooplankton research.

The Segment Anything Model (SAM) [22] has emerged as a highly popular large-scale vision foundation model with strong zero-shot segmentation capabilities, allowing it to be effectively adapted to various downstream tasks such as image inpainting [10, 53] and instance segmentation [8, 24, 28]. For example, USIS-SAM [28] and RSPrompter [8] have successfully applied SAM to instance segmentation in underwater and remote sensing imagery, respectively, providing valuable references for its extension to vertical domains. However, due to

significant differences in morphological characteristics between zooplankton and typical underwater or remote sensing targets, as well as distinct imaging mechanisms in microscopy, adapting SAM to zooplankton instance segmentation poses several challenges, as illustrated in Fig. 1:

1) **Inaccurate Classification:** Zooplankton encompass a wide variety of species, and organisms from different genera or families often exhibit highly similar morphological appearances under the microscope. For instance, as shown in the third row of Fig. 1, *Calanus sinicus* and *Acartia pacifica* have similar body sizes across different magnification levels, with the key distinguishing feature being the red eye spot present in *Acartia pacifica*. However, due to the lack of fine-grained prior knowledge about zooplankton, SAM tends to misclassify *Acartia pacifica* as *Calanus sinicus* [56].
2) **Discontinuous Segmentation of Slender Appendages:** Some zooplankton possess translucent slender appendages. *Centropages tenuiremis* uses two elongated, semi-transparent antennae on its head for feeding and sensing environmental changes. However, existing SAM-based instance segmentation models fail to fully leverage features extracted by the image encoder, resulting in fragmented predictions of these fine structures.
3) **Incomplete Boundary Segmentation:** To enhance predation efficiency or avoid predators, some zooplankton have evolved transparent body structures, such as *Pleurobrachia globosa*. When adapting SAM for instance segmentation in such specialized domains, a common approach involves upsampling low-resolution feature maps to recover high-resolution details and improve model robustness. However, the upsampling process often fails to fully restore high-frequency details, making it difficult for the model to accurately separate these translucent organisms from the background, ultimately leading to incomplete boundary segmentation [14, 29].

To address the aforementioned challenges, we constructed the first high-quality zooplankton microscopy image instance segmentation dataset (ZMIS5K) using imaging from a research-grade stereomicroscope. ZMIS5K comprises 5,358 pixel-level annotated images across 47 species, containing 10,228 instances in total. Furthermore, we propose a SAM-based instance segmentation model specifically designed for zooplankton microscopy images. To address the issues encountered when applying SAM to zooplankton instance segmentation, the innovations of our proposed ZMIS-SAM are as follows: **1)** For classification inaccuracy, we design two novel adapter modules, Shape Adapter (SA) and Intensity Adapter (IA), built upon the widely-used lightweight adapter tuning framework [32]. These two adapters, together with ViT [15], constitute ZM-ViT. These are designed to capture zooplankton-specific morphological characteristics and microscopy-specific channel intensity distributions, respectively. The proposed ZM-ViT integrates these adapters into selected ViT within SAM's image encoder, enabling effective domain-specific feature learning. **2)** For discontinuous segmentation of slender appendages, we design a Neighboring Feature Aggregation Module (NFAM) that effectively integrates general-purpose features from frozen ViT layers in SAM's image encoder and domain-specific features from ZM-ViT. This enhances the

model's ability to segment translucent slender appendages with improved continuity. **3)** To tackle incomplete boundary segmentation, we design a Wavelet-based Multi-scale and Multi-directional Feature Enhancement (WM2FE) module. The resulting feature enhancement module compensates for high-frequency details lost during upsampling, thereby improving segmentation accuracy along transparent or low-contrast edges. The main contributions of this work are as follows:

1) To our knowledge, ZMIS-SAM is the first large vision model for instance segmentation in zooplankton microscopy images. It shows strong performance in ZMIS5K and offers valuable insight for adapting SAM to other vertical domains.
2) We propose three key components: ZM-ViT enhances domain-specific representation learning for improved species classification, an NFAM improves segmentation continuity of translucent slender appendages by combining general and domain-specific features, and a WM2FE module recovers high-frequency details to refine boundary segmentation.
3) We introduce ZMIS5K, the species-rich, high-resolution zooplankton microscopy image dataset for instance segmentation. It includes 5,358 images across 47 species with 10,228 instances, and supports various vision tasks such as image classification, object detection, semantic segmentation, and instance segmentation. For a more detailed introduction to the ZMIS5K dataset, please refer to **Appendix A.**

2 Related work

2.1 Instance Segmentation

Instance segmentation methods are primarily categorized into single-stage and two-stage approaches. Single-stage methods [4, 6, 23, 45] simultaneously predict object category, location, and segmentation mask to achieve rapid instance segmentation, though their segmentation accuracy is lower than that of two-stage methods. Two-stage methods utilize object detection results for segmentation mask tasks, divided into bottom-up and top-down approaches. Bottom-up methods [1, 19, 25] first perform pixel-level predictions, then employ clustering methods to distinguish different instances. This approach better preserves low-level information such as details, but its poor generalization capability makes it challenging to handle complex scenes [27]. Top-down approaches [11, 12, 18, 39] first generate bounding boxes for candidate objects, then perform pixel-level segmentation within these boxes. Their strong generalization and segmentation capabilities have made them the current mainstream method for instance segmentation. Due to the vast diversity of zooplankton species requiring annotation by experienced biological experts, coupled with the necessity of specialized microscopy equipment to capture zooplankton microscopic images, no relevant instance segmentation datasets exist in the field of zooplankton research. Therefore, this paper establishes the first zooplankton microscopic image instance segmentation dataset to advance instance segmentation studies in this domain.

2.2 Segment Anything Model (SAM)

SAM [22] is a prompt-engineering-based large-scale visual segmentation model pre-trained on 11 million images and over 1.1 billion masks. Its robust zero-shot segmentation capability enables transfer to various downstream visual tasks such as image restoration [10, 53] and style transfer [30], as well as other vertical domains including medicine, remote sensing, and underwater applications [8,9,16,24,28,40,43,49,51]. For instance, FoodSAM [24] integrates SAM's robust segmentation capabilities with domain-specific semantic knowledge through semantic enhancement strategies, achieving semantic segmentation, instance segmentation, and panoramic segmentation of food images via object detectors. USIS-SAM [28] fine-tunes SAM's encoder to learn underwater domain knowledge and generates salient features using a salient feature prompt generator, enabling instance segmentation of underwater images. However, SAM's lack of domain-specific knowledge for zooplankton and the distinct distribution of zooplankton microscopic images compared to other datasets limit its performance in segmenting such images. In this study, we leverage a zooplankton microscopic image instance segmentation dataset to build the first large-scale instance segmentation model for the zooplankton domain based on SAM.

2.3 Wavelet Transform

The Wavelet Transform adaptively adjusts the analysis scale by scaling and translating a mother wavelet function, enabling multi-scale and multi-directional signal analysis. In recent years, several studies have integrated wavelet transform with deep learning to enhance model performance in image processing [2, 13, 57] and intelligent recognition tasks [3, 17, 26, 50, 58]. In the field of image segmentation, Xu *et al.* [48] addressed the problem of information loss caused by conventional downsampling by introducing a Haar wavelet-based downsampling method. As a result, it reduces feature map resolution while preserving structural and textural details, demonstrating improved performance on public semantic segmentation datasets. For challenging segmentation tasks under low-light or adverse weather conditions such as rain and fog, Wang *et al.* [46] proposed a wavelet-constrained semantic segmentation framework. Their method incorporates the dual-tree complex wavelet transform within a multi-branch architecture and employs a pixel attention mechanism to fuse information from both the wavelet and feature domains, effectively improving segmentation accuracy in complex railway scenes. Unlike the aforementioned works, which do not introduce any learnable or weighted operations, our work focuses on dynamically fusing wavelet features for instance segmentation. We propose a wavelet-based multi-scale and multi-directional feature enhancement module designed to recover fine details lost, thereby improving boundary segmentation accuracy for transparent zooplankton in complex microscopic imaging conditions.

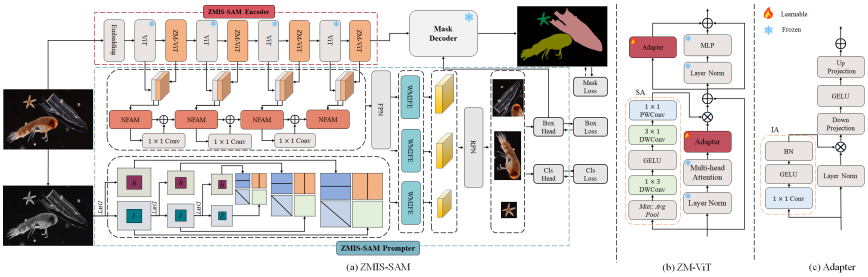

Fig. 2: Overview of ZMIS-SAM architecture. ZMIS-SAM includes ZMIS-SAM Encoder, ZMIS-SAM Prompter, and Mask Decoder.

3 Methodology

In this section, we will introduce our proposed ZMIS-SAM model, a SAM-based framework that is designed specifically for instance segmentation of zooplankton microscopic images. The subsections are as follows: 3.1 Overall Framework of ZMIS-SAM, 3.2 Zooplankton Micrograph Adaptive Vision Transformer (ZM-ViT), 3.3 Neighboring Feature Aggregation Module (NFAM), and 3.4 Wavelet-based Multi-scale and Multi-directional Feature Enhancement (WM2FE). For more prior knowledge, the loss function and evaluation metrics of the model, please refer to **Appendix B.**

3.1 Overall Framework of ZMIS-SAM

The overall architecture of ZMIS-SAM is illustrated in Fig. 2, which consists of three key components: ZM-ViT, NFAM, and WM2FE. Specifically, ZM-ViT incorporates two lightweight adapters, the Shape Adapter (SA) and the Intensity Adapter (IA). The SA refines the representation of zooplankton morphological structures, while the IA adjusts the channel-wise feature intensity distribution. Together, they guide SAM to capture domain-specific characteristics better and extract more discriminative zooplankton features. The NFAM aggregates the zooplankton-specific features extracted by ZM-ViT with the general visual features from adjacent frozen ViT layers, followed by a Feature Pyramid Network [29] to obtain multi-scale representations. Furthermore, the WM2FE module performs wavelet decomposition on the multi-scale features to enhance both low-frequency contextual cues and high-frequency directional details, effectively compensating for the edge information lost during the encoding process. Finally, the features from the encoder and the embeddings from the prompt encoder are fed into the frozen mask decoder of SAM to produce fine-grained zooplankton instance segmentation results.

3.2 ZM-ViT: Zooplankton Micrograph Adaptive ViT

The SAM's image encoder can extract rich and general-purpose visual representations that are effective across many vertical domains. However, zooplank-

ton in marine environments often exhibit slender appendages, transparent or semi-transparent bodies, and significant illumination variations caused by the microscopic imaging environment. Moreover, there exists a notable domain gap between the natural images used for SAM pretraining and the microscopic zooplankton images. These factors make the original SAM image encoder less effective in capturing the fine-grained morphological characteristics of zooplankton, leading to suboptimal segmentation performance. To address this issue, we propose ZM-ViT, which integrates two lightweight adapters into the ViT blocks of SAM, as shown in Fig. 2(b). This design enables ZM-ViT to effectively leverage the pretrained SAM image encoder while adapting to the zooplankton domain, allowing for more discriminative feature extraction and improved instance segmentation accuracy.

The proposed ZM-ViT can effectively learn discriminative features of zooplankton in microscopic imagery through Parameter-Efficient Fine-Tuning [44]. Inspired by USIS-SAM, we integrate two lightweight adapters into the ViT architecture, namely the Intensity Adapter (IA) and the Shape Adapter (SA). Unlike the adapter design in USIS-SAM, we introduce a parallel branch that combines Conv–GELU–BN operations with the Layer Normalization (LN) residual pathway for the features fed into the adapter. This design forms the IA, which dynamically adjusts the channel-wise feature intensity distribution. The placement of IA within the ViT block remains consistent with that of the adapter in USIS-SAM: one IA module is inserted after the multi-head self-attention layer to modulate the channel intensity distribution, and another is integrated in parallel with the frozen MLP residual branch to aggregate and update contextual information more effectively. The IA can be formulated as follows:

$$F_{IA} = \mathcal{U}_p(\mathcal{D}_p(BN(LN(x) * \mathcal{C}(x)))) + F_{in} \tag{1}$$

where x and $F_{IA} \in \mathbb{R}^{N \times C}$ denote the input and output features, respectively. $\mathcal{C}(\cdot)$ denotes the convolutional branch consisting of Conv–GELU–BN, $\mathcal{U}_p(\cdot)$ and $\mathcal{D}_p(\cdot)$ denote the up projection and down projection, respectively. $BN(\cdot)$ and $LN(\cdot)$ denote the batch normalization and layer normalization, respectively.

Since zooplankton have diverse morphological structures, we further design a Shape Adapter (SA) based on a striped depth-wise separable convolution. The SA can effectively enhance the model's ability to represent the shape structural characteristics of zooplankton, while the depthwise separable convolution greatly reduces computational complexity. The SA can be expressed as follows:

$$F_{SA} = \sum_{max;avg} \mathcal{C}_1((Pool_i(x))) \tag{2}$$

where x and $F_{SA} \in \mathbb{R}^{N \times C}$ denote the input and output features, respectively. $\mathcal{C}_1(\cdot)$ denotes the PWConv-DWConv-GELU-DWConv operation. $Pool(\cdot)$ denotes average pooling or maximum pooling.

To enable the SAM image encoder to extract optimal feature representations, we use ViT-H as the backbone of the SAM image encoder. Furthermore, to ensure a fair comparison with the baseline experiments, ZM-ViT replaces one frozen ViT layer every two layers, starting from the eighth layer.

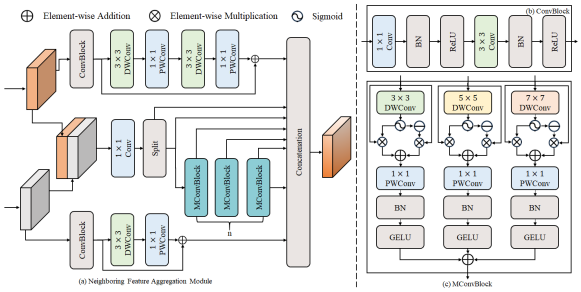

Fig. 3: Overview of Neighboring Feature Aggregation Module (NFAM) architecture.

3.3 NFAM: Neighboring Feature Aggregation Module

To achieve automatic instance segmentation, existing approaches usually treat the representations extracted from certain ViT layers of the SAM image encoder as pseudo-prompts. However, when transferring SAM to domain-specific instance segmentation tasks, current methods either rely solely on the frozen ViT features with general representations or only exploit the domain-specific representations obtained from the adapter. Both strategies fail to fully leverage the complementary information between general and domain-specific representations for generating effective prompt features.

We propose an NFAM that comprehensively aggregates both the general representations extracted from the frozen ViT layers and the domain-specific features captured by the fine-tuned ZM-ViT, thereby producing more effective prompt representations and enhancing the instance segmentation performance of SAM. As illustrated in Fig. 3, NFAM processes the domain-specific features $\{F_{sdf}\}_{i=j}^{L} \in \mathbb{R}^{C_i \times H_i \times W_i}$, the general features $\{F_{gf}\}_{i=j-1}^{L-1} \in \mathbb{R}^{C_i \times H_i \times W_i}$, and their concatenated representations $F_{cf} \in \mathbb{R}^{2C_i \times H_i \times W_i}$ in parallel. The domain-specific and general features are respectively refined using a ConvBlock (Fig. 3 (b)) and a depthwise separable convolution to enhance local feature extraction, while residual connections preserve fine-grained spatial information.

$$\tilde{F_{sdf}} = ConvBlock(F_{sdf}) + \mathcal{C}_2(ConvBlock(F_{sdf})) \tag{3}$$

$$\tilde{F_{gf}} = ConvBlock(F_{gf}) + \mathcal{C}_3(ConvBlock(F_{gf})) \tag{4}$$

$$F_{cf} = Concat(F_{sdf}, F_{gf}) \tag{5}$$

where $\tilde{F_{sdf}}$ and $\tilde{F_{gf}}$ domain-specific and general features after local enhancement, respectively. $\mathcal{C}_2(\cdot)$ and $\mathcal{C}_3(\cdot)$ denote the PWConv-DWConv-PWConv-DWConv and the PWConv-DWConv operation, respectively. In parallel, the concatenated features are passed through a 1×1 convolution to balance the channel dimensions. The resulting feature map is then divided into n parts, each processed

by an independent multi-scale convolutional block (Fig. 3 (c)) with different receptive fields to capture spatial details at multiple scales.

$$[F^{(1)}, ..., F^{(n)}] = Split_n(Conv_{1\times1}(F_{cf})) \tag{6}$$

$$M^{(j)} = MConvBlock(F^{(j)}), j = 1, ..., n \tag{7}$$

where $Split_n(\cdot)$ denotes dividing the input representation into n parts. $F^{(j)}$ denotes the j-th characteristic obtained from the division. $M^{(j)}$ denotes the j-th part after multi-scale convolution processing.

In the MConvBlock, depth-wise convolutions with different receptive fields are first used to extract multi-scale features from the input features separately. Then, a gating mechanism is employed to dynamically adjust the weights of features at different scales. Pointwise convolution, Batch Normalization, and GELU activation follow each receptive field is utilized to output the enhanced representation.

$$\psi_k = \sigma(DW_{k\times k}(x)) \tag{8}$$

$$x_k^{gate} = (\psi_k * x + (1 - \psi_k) * x)) \tag{9}$$

$$B_k(x) = GELU(BN(PW_{1\times1}(x_k^{gate})) \tag{10}$$

$$MConvBlock(x) = \sum_{k=3}^{K} B_k(x), k = 3, 5, 7 \tag{11}$$

where ψ_k denotes the gating attention weights under different receptive fields. $\sigma(\cdot)$ and $GELU(\cdot)$ denote the activation function of the sigmoid and GELU, respectively. x_k^{gate} denotes the gating characteristic. $BN(\cdot)$ denotes the Batch Normalization. $B_k(x)$ refers output feature of the receptive field k branch. Finally, the outputs from all branches are concatenated to form the aggregated representation F_a.

$$F_a = Concat(M^{(1)}, M^{(2)}, ..., M^{(n)}, \tilde{F_{sdf}}, \tilde{F_{gf}}) \tag{12}$$

3.4 WM2FE: Wavelet-based Multi-scale and Multi-directional Feature Enhancement

After obtaining the aggregated features, existing SAM-based instance segmentation methods usually employ a simple FPN network to extract multi-scale features and generate candidate prompts for enhancing model robustness. However, in the FPN, upsampling low-resolution features to high-resolution ones via deconvolution may lead to the loss of fine-grained details. As a multi-scale and multi-directional image analysis tool, the wavelet transform can provide high-frequency details in vertical, horizontal, and diagonal directions. Therefore, we

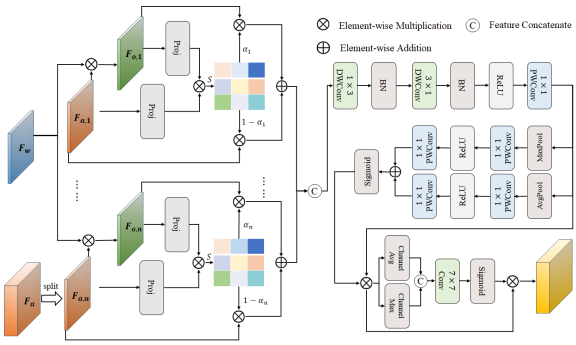

Fig. 4: Overview of Wavelet-based Multi-scale and Multi-directional Feature Enhancement (WM2FE) architecture.

design a Wavelet-based Multi-scale Feature Enhancement (WM2FE) module to compensate for the missing details during upsampling and improve the instance segmentation performance for zooplankton.

As illustrated in Fig. 4, WM2FE enhances multi-scale feature representation through dynamic fusion of aggregated features and wavelet-weighted features. Specifically, the aggregated feature F_a is divided into n parts $\{F_{a,1}, F_{a,2}, ...F_{a,n}\}$ along the wavelet feature F_w dimension, and each part is modulated by the corresponding wavelet feature to strengthen the representation of transparent boundaries of zooplankton. Then, linear learnable parameters are applied to dynamically fuse the aggregated features and wavelet-enhanced features. Finally, the n parts are concatenated along the channel dimension, and a directional depthwise separable convolution is employed to enhance directional perception.

$$F_{o,i} = F_{a,i} \odot F_w \tag{13}$$

$$F_{m,i} = \alpha_i F_{out,i} + (1 - \alpha_i) F_{a,i}, i = 1, 2, ..., n \tag{14}$$

$$F_m = \mathcal{C}_4(Concat(F_{m,1}, F_{m,2}, ..., F_{m,n})) \tag{15}$$

where $\mathcal{C}_4(\cdot)$ denotes the PWConv-ReLu-DWConv-BN-DWConv operation. $\odot$ denotes element-wise multiplication, α_i denotes the linear learnable parameter of the i-th part. $F_w = Contact[LL; LH; HL; HH]$ denotes the wavelet features.

To further focus on informative regions, inspired by [47], we introduce channel and spatial attention mechanisms in WM2FE. The channel attention module learns channel-wise weights by feeding both average-pooled and max-pooled features into a shared MLP followed by a sigmoid activation function.

$$M_c = \sigma(\sum_{avg;max} \mathcal{C}_5(F_m)) \tag{16}$$

$$F^{'} = M_c \otimes F_m \tag{17}$$

where $F^{'}$ denotes the refined features. M_c denotes the channel attention map and σ denotes the sigmoid function. $\mathcal{C}_5(\cdot)$ denotes the PWConv-ReLu-DWConv-BN-DWConv-BN-DWConv operation.

Subsequently, the spatial attention mechanism aggregates average-pooled and max-pooled features along the channel dimension, followed by a 7×7 convolution and sigmoid activation to generate the spatial attention map:

$$M_s = \sigma(f^{7\times7}([\mathcal{A}(F^{'}); \mathcal{M}(F^{'})])) \tag{18}$$

$$F_{fe} = M_s \otimes F^{'} \tag{19}$$

4 Experiments

In this section, we evaluate the performance of ZMIS-SAM on the ZMIS5K dataset and compare it with SOTA instance segmentation models, and further conduct ablation experiments. In addition, we conducted generalization experiments on public datasets from other domains, as detailed in **Appendix D.**

4.1 Datasets

The ZMIS5K dataset contains a total of 5,358 images, including 47 species of zooplankton, and was divided into 4,262 images for the training set and 1,096 images for the test set. For a more detailed introduction to the ZMIS5K dataset, please refer to **Appendix A.**

4.2 Implementation details

We build our ZMIS-SAM model based on the MMDetection framework [7], and most of the comparison models in this paper are also trained and evaluated within the same framework. ZMIS-SAM and other SAM-based comparison models are trained on 6 NVIDIA RTX 3090 GPUs, with a batch size of 2 per GPU, for 400 epochs. All input images are resized to 512 × 512 pixels, and standard data augmentation techniques, including random flipping, random scaling, and random cropping, are applied to enhance the model's generalization capability. We use the AdamW optimizer and adopt automatic mixed precision to accelerate training and reduce GPU memory consumption. The initial learning rate is set to 2.0e-4. In the first stage, a linear warm-up is performed over the first 50 iterations, followed by a cosine decay schedule to gradually reduce the learning rate in the second stage.

4.3 Experiments Result

Quantitative Results. We compare our ZMIS-SAM with fourteen state-of-the-art instance segmentation models on the ZMIS5K test set. To explore the optimal segmentation performance of each model, traditional instance segmentation

Table 1: A comparison with SOTA methods on ZMIS5K is provided. Results highlighted in red indicate the best performance, while those in blue denote the second-best.

Method	Backbone	Params	mAP	AP_{50}	AP_{75}
Mask R-CNN *(ICCV'17)* [18]	ResNet-101	63M	66.4	90.9	69.0
Cascade Mask R-CNN *(CVPR'18)* [5]	ResNet-101	88M	66.0	89.5	68.3
Mask Scoring R-CNN *(CVPR'19)* [20]	ResNet-101	79M	67.0	89.8	69.1
CondInst *(ECCV'20)* [41]	ResNet-101	63M	68.3	91.7	70.9
R^3-CNN *(CAIP'21)* [39]	ResNet-101	77M	66.6	90.0	70.0
ConvNeXt *(CVPR'22)* [31]	ConvNeXt-T	63M	68.4	92.4	70.3
Mask2Former *(CVPR'22)* [11]	ResNet-101	63M	70.7	87.4	76.1
WaterMask *(ICCV'23)* [27]	ResNet-101	67M	60.7	82.2	64.4
ViT-UWA *(TIP'26)* [21]	ViT-B	129M	62.3	90.7	61.7
SAM+Bbox *(ICCV'23)* [22]	ViT-H	641M	60.2	82.9	65.2
SAM+Mask *(ICCV'23)* [22]	ViT-H	641M	70.1	**94.0**	75.5
RSPrompter *(TGRS'24)* [8]	ViT-H	632M	**71.8**	92.2	**76.9**
USIS-SAM *(ICML'24)* [28]	ViT-H	700M	70.4	91.7	75.8
UCIS-SAM *(TIP'26)* [42]	ViT-H	723M	66.9	80.4	72.1
ZMIS-SAM	ViT-H	758M	**73.6**	**94.6**	**80.7**

methods are implemented with ResNet-101 as the backbone, while SAM-based methods adopt ViT-H, the largest variant of the SAM backbone, to ensure fair comparison. As shown in Tab. 1, our proposed ZMIS-SAM achieves the best overall performance across all instance segmentation metrics. Traditional methods such as Mask R-CNN, CondInst, and ConvNeXt exhibit limited accuracy, with mAP values ranging from 66.0% to 68.4%. Among traditional models, Mask2Former achieves the highest 70.7%, 87.4%, and 76.1% AP in mAP, AP_{50} and AP_{75}, however, ZMIS-SAM surpasses it by 2.9% in mAP, 7.2% in AP_{50}, and 4.6% in AP_{75}, demonstrating the superior capability of our model in capturing fine-grained object boundaries. Compared with traditional models, SAM-based instance segmentation approaches achieve better performance owing to their strong feature extraction and prompt-based representation learning. Nevertheless, our ZMIS-SAM further improves segmentation accuracy with only a small increase in model parameters, outperforming the best SAM-based model RSPrompter by 1.8% in mAP, 2.4% in AP_{50}, and 3.8% in AP_{75}.

Qualitative Results. We further conduct qualitative comparisons between ZMIS-SAM and several top-performing instance segmentation models on the ZMIS5K test set. To investigate the limitations of SAM when transferred to vertical domains, we visualize the segmentation results and analyze them in detail. As shown in Fig. 5, the proposed ZMIS-SAM not only distinguishes different zooplankton species accurately but also captures distinctive morphological characteristics such as slender appendages, while performing fine-grained boundary segmentation. Specifically, in Fig. 5 (a), our method successfully differentiates between *Jellyfish larvae* and *Sugiura chengshanense*, while avoiding false segmentation of unannotated organisms. In contrast, Mask2Former incorrectly segments unidentified organisms (e.g., *Fish eggs*) as known species. In Fig. 5 (b), ZMIS-

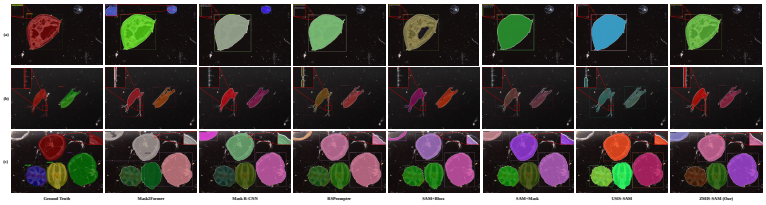

Fig. 5: Quantitative comparison between the ZMIS-SAM model and other state-of-the-art models on the ZMIS5K dataset. (a), (b), and (c) denote inaccurate classification, discontinuous segmentation of slender appendages, and incomplete boundary segmentation, respectively. Higher-resolution visualizations can be found in the appendix.

SAM effectively captures the relationship between the main body and slender appendages of *Calanus sinicus*, achieving complete and coherent segmentation. Although RSPrompter and USIS-SAM can also segment the appendages, their results are fragmented and discontinuous, unlike our method's continuous representation. In Fig. 5 (c), ZMIS-SAM demonstrates superior boundary refinement ability. For example, Mask R-CNN fails to delineate the boundaries of *Pleurobrachia globosa*, while our approach achieves precise and consistent edge segmentation. These qualitative comparisons further confirm that ZMIS-SAM excels in recognizing fine morphological details, maintaining boundary integrity, and suppressing false detections, addressing key challenges in SAM's transfer to specialized domains such as microscopic zooplankton segmentation.

4.4 Ablation experiments on ZMIS-SAM

In this section, we conduct a series of ablation studies on the ZMIS5K dataset to evaluate the design of ZMIS-SAM systematically. The experiments are designed to investigate: 1) the effectiveness of individual modules and 2) the influence of different wavelet basis functions on the final segmentation results. More ablation experiments and parametric analysis are provided in **Appendix C.**

Table 2: Albation study of our contributions.

ZM-ViT	NFAM	WM2FE	mAP	AP_{50}	AP_{75}
			70.4	91.7	75.8
✓			71.7	92.0	77.7
✓	✓		73.0	94.4	78.7
✓		✓	72.7	94.1	79.4
✓	✓	✓	**73.6**	**94.6**	**80.7**

Table 3: Albation study of wavelet function for the WM2FE.

φ	mAP	AP_{50}	AP_{75}
rbio1.1	72.7	93.8	79.2
bior1.1	72.0	92.6	78.4
haar	**73.6**	**94.6**	**80.7**

Effectiveness of Individual Modules. We perform an incremental ablation study to systematically evaluate the individual contribution and combined effect of each proposed module for zooplankton instance segmentation in microscopy images. The complete results are presented in Tab. 2, where we sequentially evaluate the following configurations: 1) whether to use ZM-ViT, 2) whether to incorporate the NFAM, and 3) whether to employ the WM2FE module. The results demonstrate that each module consistently improves instance

segmentation performance. Using USIS-SAM with a ViT-H backbone as the baseline, we observe that replacing its original ViT with our ZM-ViT brings a 1.3% mAP gain with only a marginal increase in parameters. Building upon ZM-ViT, the further addition of NFAM and WM2FE improves mAP by 1.3% and 1.0%, respectively. When both NFAM and WM2FE are integrated, ZMIS-SAM achieves mAP, AP_{50}, and AP_{75} scores of 73.6%, 94.6%, and 80.7%, respectively, establishing a new state-of-the-art in zooplankton instance segmentation with only a modest parameter overhead.

Optimal Wavelet Function for the WM2FE. We evaluate the impact of different wavelet basis functions on the instance segmentation performance of ZMIS-SAM, with experimental results summarized in Tab. 3. Different wavelets exhibit distinct frequency domain response characteristics when extracting image features, which influences the model's ability to capture fine object details. Experimental results indicate that using the Haar wavelet in the WM2FE module yields the best performance across mAP, AP_{50}, and AP_{75}, achieving scores of 73.6%, 94.6%, and 80.7%, respectively, surpassing both rbio1.1 and bior1.1 wavelets. We attribute this improvement to the high sensitivity of the Haar wavelet to edge structures, which enables more effective capture of zooplankton contour and boundary information, thereby enhancing the instance segmentation capability of ZMIS-SAM.

5 Limitations

Although the ZMIS-SAM model can perform instance segmentation on microscopic images of zooplankton, differences in task definition prevent the framework from separately segmenting the heads and antennae of zooplankton. In future work, we will introduce taxonomic and morphological text information to construct a multimodal framework for the zooplankton domain to achieve fine-grained segmentation of zooplankton body structures.

6 Conclusion

In this study, we introduce the first high-quality zooplankton microscopy image segmentation dataset (ZMIS5K), spanning 47 species with 5,358 images and 10,228 instances. More importantly, we propose ZMIS-SAM, the first SAM-based instance segmentation model specifically designed for zooplankton microscopy images. In ZMIS-SAM, we innovatively propose three modules: ZM-ViT, NFAM, and WM2FE, which effectively address the issues of inaccurate classification, discontinuous segmentation of slender appendages, and incomplete boundary segmentation encountered when transferring SAM to the zooplankton domain. Extensive experiments demonstrate that ZMIS-SAM achieves state-of-the-art performance on the ZMIS5K dataset and exhibits strong generalization capability on cross-domain benchmarks.

Acknowledgements

This work is supported in part by the Natural Science Foundation of Shandong Province (Grant No. ZR2024MF037). The North China Sea Ecological Center of the Ministry of Natural Resources strongly supports this work. The center provided samples for zooplankton photography and the microscopic imaging equipment required for this study. In addition, teachers, including Xiaoli Song and Yanping Qi, offered detailed guidance on zooplankton biology.

References

1. Arnab, A., Torr, P.H.S.: Bottom-up instance segmentation using deep higher-order crfs. arXiv (2016), http://arxiv.org/abs/1609.02583 4
2. Balster, E., Zheng, Y., Ewing, R.: Combined spatial and temporal domain wavelet shrinkage algorithm for video denoising. IEEE Transactions on Circuits and Systems for Video Technology **16**(2), 220–230 (2006). https://doi.org/10.1109/TCSVT.2005.857816 5
3. Bi, Q., You, S., Gevers, T.: Learning generalized segmentation for foggy-scenes by bi-directional wavelet guidance. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 801–809 (2024). https://doi.org/10.1609/aaai.v38i2.27838 5
4. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: Yolact: Real-time instance segmentation. In: 2019 IEEE International Conference on Computer Vision (ICCV) (2019). https://doi.org/10.1109/ICCV.2019.00925 4
5. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6154–6162 (2018). https://doi.org/10.1109/CVPR.2018.00644 12
6. Chen, H., Sun, K., Tian, Z., Shen, C., Huang, Y., Yan, Y.: BlendMask: Top-down meets bottom-up for instance segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8570–8578 (2020). https://doi.org/10.1109/CVPR42600.2020.00860 4
7. 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., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv (2019), https://arxiv.org/abs/1906.07155 11
8. Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., Shi, Z.: Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–17 (2024). https://doi.org/10.1109/TGRS.2024.3356074 2, 5, 12
9. Chen, T., Zhu, L., Ding, C., Cao, R., Wang, Y., Zhang, S., Li, Z., Sun, L., Zang, Y., Mao, P.: Sam-adapter: Adapting segment anything in underperformed scenes. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 3359–3367 (2023). https://doi.org/10.1109/ICCVW60793.2023.00361 5
10. Chen, W., Vong, Y.J., Kuo, S.Y., Ma, S., Wang, J.: Robustsam: Segment anything robustly on degraded images. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4081–4091 (2024). https://doi.org/10.1109/CVPR52733.2024.00391 2, 5

11. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1280–1289 (2022). https://doi.org/10.1109/CVPR52688.2022.00135 4, 12
12. Cheng, B., Schwing, A.G., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. In: Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS) (2021) 4
13. Ding, S., Wang, Q., Guo, L., Li, X., Ding, L., Wu, X.: Wavelet and adaptive coordinate attention guided fine-grained residual network for image denoising. vol. 34, pp. 6156–6166 (2024). https://doi.org/10.1109/TCSVT.2023.3348804 5
14. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(2), 295–307 (2016). https://doi.org/10.1109/TPAMI.2015.2439281 3
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16 x 16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations (ICLR) (2021) 3
16. Fang, H., Zhang, T., Zhou, X., Zhang, X.: Learning better video query with sam for video instance segmentation. IEEE Transactions on Circuits and Systems for Video Technology **35**(4), 2963–2974 (2025). https://doi.org/10.1109/TCSVT.2024.3361076 5
17. Finder, S.E., Amoyal, R., Treister, E., Freifeld, O.: Wavelet convolutions for large receptive fields. In: Computer Vision – ECCV 2024. pp. 363–380. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-72949-2_21 5
18. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.322 2, 4, 12
19. Hu, J., Lu, Y., Zhang, S., Cao, L.: Istr: Mask-embedding-based instance segmentation transformer. IEEE Transactions on Image Processing **33**, 2895–2907 (2024). https://doi.org/10.1109/TIP.2024.3385980 4
20. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring r-cnn. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6402–6411 (2019). https://doi.org/10.1109/CVPR.2019.00657 12
21. Jia, Y., Lin, Q., Li, H., Li, Y., Kwong, S., Cong, R.: Vit-uwa: Vision transformer underwater-adapter for dense predictions beneath the water surface. IEEE Transactions on Image Processing **35**, 4012–4026 (2026). https://doi.org/10.1109/TIP.2026.3682126 12
22. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3992–4003 (2023). https://doi.org/10.1109/ICCV51070.2023.00371 2, 5, 12
23. Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9796–9805 (2020). https://doi.org/10.1109/CVPR42600.2020.00982 4
24. Lan, X., Lyu, J., Jiang, H., Dong, K., Niu, Z., Zhang, Y., Xue, J.: Foodsam: Any food segmentation. IEEE Transactions on Multimedia **27**, 2795–2808 (2025). https://doi.org/10.1109/TMM.2023.3330047 2, 5
25. Lazarow, J., Xu, W., Tu, Z.: Instance segmentation with mask-supervised polygonal boundary transformers. In: 2022 IEEE/CVF Conference on Computer Vision

and Pattern Recognition (CVPR). pp. 4382–4391 (2022). https://doi.org/10.1109/CVPR52688.2022.00434 4
26. Li, Q., Shen, L.: Wavesnet: Wavelet integrated deep networks for image segmentation. In: Pattern Recognition and Computer Vision. pp. 325–337. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-18916-6_27 5
27. Lian, S., Li, H., Cong, R., Li, S., Zhang, W., Kwong, S.: Watermask: Instance segmentation for underwater imagery. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1305–1315 (2023). https://doi.org/10.1109/ICCV51070.2023.00126 4, 12
28. Lian, S., Zhang, Z., Li, H., Li, W., Yang, L.T., Kwong, S., Cong, R.: Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset. In: Proceedings of the 41st International Conference on Machine Learning (ICML). pp. 29545–29559 (2024) 2, 5, 12
29. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 936–944 (2017). https://doi.org/10.1109/CVPR.2017.106 3, 6
30. Liu, S., Ye, J., Wang, X.: Any-to-any style transfer: Making picasso and da vinci collaborate. arXiv (2023), https://arxiv.org/abs/2304.09728 5
31. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11966–11976 (2022). https://doi.org/10.1109/CVPR52688.2022.01167 12
32. Lu, Y., Liu, J., Zhang, Y., Liu, Y., Tian, X.: Prompt distribution learning. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5196–5205 (2022). https://doi.org/10.1109/CVPR52688.2022.00514 3
33. Masoudi, M., Giering, S.L., Eftekhari, N., Massot-Campos, M., Irisson, J.O., Thornton, B.: Optimizing plankton image classification with metadata-enhanced representation learning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **17**, 17117–17133 (2024). https://doi.org/10.1109/JSTARS.2024.3424498 1
34. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(7), 3523–3542 (2022). https://doi.org/10.1109/TPAMI.2021.3059968 2
35. Pu, Y., Feng, Z., Wang, Z., Yang, Z., Li, J.: Anomaly detection for in situ marine plankton images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 3661–3671 (2021). https://doi.org/10.1109/ICCVW54120.2021.00409 1
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). vol. 139, pp. 8748–8763 (2021) 2
37. Ratnarajah, Lavenia, A.A., Rana, A., Angus, B., Sonia, B., Nicholas J, B., Kim S, C., Gabrielle, C., Astrid, E., Jason D, G., Maria, I., Nurul Huda Ahmad, J., David, L., Fabien, M., Erik, O., Clare, P., Sophie, R., Anthony J, S., Katrin, S., Lars, S., Kerrie M, Y., Guang, Y., Lidia: Monitoring and modelling marine zooplankton in a changing climate. Nature Communications **14** (2023). https://doi.org/10.1038/s41467-023-36241-5 1

38. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. arXiv (2024), https://arxiv.org/abs/2408.00714 2
39. Rossi, L., Karimi, A., Prati, Andrea", e.N., Panayides, A., Theocharides, T., Lanitis, A., Pattichis, C., Vento, M.: Recursively refined r-cnn: Instance segmentation with self-roi rebalancing. In: Computer Analysis of Images and Patterns (CAIP). pp. 476–486 (2021). https://doi.org/10.1007/978-3-030-89128-2_46 4, 12
40. Shan, Z., Liu, Y., Zhou, L., Yan, C., Wang, H., Xie, X.: Ros-sam: High-quality interactive segmentation for remote sensing moving object. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3625–3635 (2025). https://doi.org/10.1109/CVPR52734.2025.00343 5
41. Tian, Z., Shen, C., Chen, H.: Conditional convolutions for instance segmentation. In: Computer Vision – ECCV 2020. pp. 282–298 (2020). https://doi.org/10.1007/978-3-030-58452-8_17 12
42. Wang, C., Li, H., Li, C., Liu, H., Tang, X., Kwong, S.: Expose camouflage in the water: Underwater camouflaged instance segmentation and dataset. IEEE Transactions on Image Processing **35**, 3283–3298 (2026). https://doi.org/10.1109/TIP.2026.3675502 12
43. Wang, D., Zhang, J., Du, B., Xu, M., Liu, L., Tao, D., Zhang, L.: Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS). vol. 36, pp. 8815–8827 (2023) 5
44. Wang, L., Chen, S., Jiang, L., Pan, S., Cai, R., Yang, S., Yang, F.: Parameter-efficient fine-tuning in large models: A survey of methodologies. arXiv (2024), https://arxiv.org/abs/2410.19878 7
45. Wang, X., Kong, T., Shen, C., Jiang, Y., Li, L.: SOLO: Segmenting objects by locations. In: Computer Vision – ECCV 2020. pp. 649–665 (2020). https://doi.org/10.1007/978-3-030-58523-5_38 4
46. Wang, Z., Liao, Z., Wang, P., Chen, P., Luo, W.: Wavecrnet: Wavelet transform-guided learning for semantic segmentation in adverse railway scenes. IEEE Transactions on Intelligent Transportation Systems **26**(6), 8794–8809 (2025). https://doi.org/10.1109/TITS.2025.3543925 5
47. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Computer Vision – ECCV 2018. pp. 3–19. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1 10
48. Xu, G., Liao, W., Zhang, X., Li, C., He, X., Wu, X.: Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. Pattern Recognition **143**, 109819 (2023). https://doi.org/10.1016/j.patcog.2023.109819 5
49. Xu, Y., Tang, J., Men, A., Chen, Q.: Eviprompt: A training-free evidential prompt generation method for adapting segment anything model in medical images. IEEE Transactions on Image Processing **33**, 6204–6215 (2024). https://doi.org/10.1109/TIP.2024.3482175 5
50. Yang, Y., Jiao, L., Liu, X., Liu, F., Yang, S., Li, L., Chen, P., Li, X., Huang, Z.: Dual wavelet attention networks for image classification. IEEE Transactions on Circuits and Systems for Video Technology **33**(4), 1899–1910 (2023). https://doi.org/10.1109/TCSVT.2022.3218735 5
51. Ye, M., Zhang, J., Liu, J., Liu, C., Yin, B., Liu, C., Du, B., Tao, D.: Hi-sam: Marrying segment anything model for hierarchical text segmentation. IEEE Trans-

actions on Pattern Analysis and Machine Intelligence **47**(3), 1431–1447 (2025). https://doi.org/10.1109/TPAMI.2024.3495831 5
52. Ye, X., Wang, H., Yao, J.: A lightweight deep learning network for the precise detection and classification of variable plankton. Engineering Applications of Artificial Intelligence **163**, 112719 (2026). https://doi.org/10.1016/j.engappai.2025.112719 1
53. Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., Chen, Z.: Inpaint anything: Segment anything meets image inpainting. arXiv (2023), https://arxiv.org/abs/2304.06790 2, 5
54. Yu, Y., Lv, Q., Li, Y., Wei, Z., Dong, J.: Phytracker: An online tracker for phytoplankton. IEEE Transactions on Circuits and Systems for Video Technology **35**(3), 2932–2944 (2025). https://doi.org/10.1109/TCSVT.2024.3498349 1
55. Yuan, D., Qi, Y., Zhang, J., Li, Z.: Planktontnet: Rethinking plankton classification from a global view with swin-transformer. In: OCEANS 2025 Brest. pp. 1–5 (2025). https://doi.org/10.1109/OCEANS58557.2025.11104797 1
56. Zhang, C., Liu, L., Cui, Y., Huang, G., Lin, W., Yang, Y., Hu, Y.: A comprehensive survey on segment anything model for vision and beyond. arXiv (2023), https://arxiv.org/abs/2305.08196 3
57. Zhang, W., Zhou, L., Zhuang, P., Li, G., Pan, X., Zhao, W., Li, C.: Underwater image enhancement via weighted wavelet visual perception fusion. IEEE Transactions on Circuits and Systems for Video Technology **34**(4), 2469–2483 (2024). https://doi.org/10.1109/TCSVT.2023.3299314 5
58. Zou, W., Gao, H., Yang, W., Liu, T.: Wave-mamba: Wavelet state space model for ultra-high-definition low-light image enhancement. In: Proceedings of the 32nd ACM International Conference on Multimedia. p. 1534–1543 (2024). https://doi.org/10.1145/3664647.3681580 5