

中图法分类号: TP391.4 文献标识码: A 文章编号: 1006-8961(XXXX)XX-0001-23

论文引用格式: Li Zhongwei, Bao Junlong, Qiao Zheng, Xu Xiangyang, Qi Yanping, Liu Xin, Yuan Dekun. Zooplankton Segmentation Based on Dual-Domain Feature Enhancement and Target-Aware Guidance[J/OL]. Journal of Image and Graphics, XXXX: 1-23. DOI: 10.11834/jig.260408.
(李忠伟, 鲍君龙, 乔征, 许向阳, 齐衍萍, 刘昕, 袁德坤. 双域特征增强与目标感知引导的浮游动物分割[J/OL]. 中国图象图形学报, XXXX: 1-23. DOI: 10.11834/jig.260408.) [DOI: 10.11834/jig.260408]

双域特征增强与目标感知引导的浮游动物分割

李忠伟¹, 鲍君龙¹, 乔征¹, 许向阳¹, 齐衍萍³, 刘昕², 袁德坤¹

1. 中国石油大学(华东), 海洋与空间信息学院, 青岛 266580; 2. 中国石油大学(华东), 青岛软件学院、计算机科学与技术学院, 青岛 266580; 3. 国家海洋局北海环境监测中心(自然资源部北海生态中心), 青岛 266033

摘要: 目的 浮游动物显微图像实例分割是实现种类识别、数量统计和形态分析的重要基础。传统人工镜检效率低且主观性强, 难以满足大规模监测需求, 自动化分析需求日益迫切。然而, 图像中存在弱边界目标、背景干扰和目标密集分布等问题, 导致现有模型分割精度受限。此外, 该领域缺乏大规模像素级实例分割数据集。方法 为解决上述问题, 构建浮游动物显微图像实例分割数据集(zooplankton microscopic image instance segmentation dataset, ZooMIIS5K), 包含 5332 张图像和 48 类浮游动物的精细像素级标注。在此基础上, 提出一种基于分割一切模型(Segment Anything Model, SAM)的浮游动物显微图像实例分割模型(zooplankton microscopic image instance segmentation model, ZooMIIS-SAM)。首先, 设计双域融合视觉Transformer(dual-domain fusion vision transformer, DDF-ViT), 缓解预训练自然图像域与显微图像域之间的特征偏移; 接着设计多尺度融合提示模块(multi-scale fusion prompt module, MSFPM), 自动生成视觉提示, 引导SAM自动理解海洋浮游动物图像; 最后, 设计目标感知模块(target-aware perception module, TAPM), 利用候选区域信息增强全局特征的实例感知能力, 提高密集场景下相邻目标的区分效果。结果 在ZooMIIS5K数据集上, ZooMIIS-SAM在所比较的卷积神经网络方法和视觉基础模型方法中取得最佳性能, 其掩码平均精度均值(mask mean average precision, mAP_{mask})达到 77.4%。消融实验验证了各模块的有效性, 可视化分析表明该方法能够更稳定地聚焦于浮游动物区域。ZooMIIS-SAM具有应用于实际浮游动物监测场景的潜力。所构建的数据集和提出的方法可为浮游动物目标计数及形态分析提供技术参考。

关键词: 浮游动物; 实例分割; Segment Anything Model; 双域特征增强; 目标感知

Zooplankton Segmentation Based on Dual-Domain Feature Enhancement and Target-Aware Guidance

Li Zhongwei¹, Bao Junlong¹, Qiao Zheng¹, Xu Xiangyang¹, Qi Yanping³, Liu Xin², Yuan Dekun¹

1. College of Oceanography and Space Information, China University of Petroleum (East China), Qingdao 266580, China; 2. Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China; 3. North China Sea Environmental Monitoring Center, State Oceanic Administration, Qingdao 266033, China

Abstract: **Objective** Zooplankton are important components of aquatic ecosystems, and changes in their abundance, species composition, and morphological characteristics can provide useful information for ecological observation and environmental assessment. Accurate analysis of zooplankton microscopic images is therefore important for species identification,

收稿日期: 2026-07-06; 修回日期: 2026-09-22

* 通信作者: 袁德坤, 男, 博士研究生, 主要研究方向为计算机视觉和多模态信息处理。E-mail: ydk_libra0903@163.com

基金项目: 山东省自然科学基金(项目编号: ZR2024MF037)

Supported by: Natural Science Foundation of Shandong Province, China (Grant No. ZR2024MF037)

© 中国图象图形学报版权所有

individual counting, and quantitative morphological measurement. At present, microscopic examination still depends heavily on trained specialists and manual operation. This process is time-consuming, subjective, and difficult to apply efficiently when large numbers of samples must be processed. Instance segmentation can simultaneously locate individual organisms and provide their pixel-level masks, making it an important technical basis for automatic zooplankton analysis. However, reliable instance segmentation of zooplankton microscopic images remains challenging. Many zooplankton species have transparent or semi-transparent bodies, resulting in weak boundaries and low contrast with the surrounding medium. Suspended particles, biological debris, uneven illumination, imaging noise, and defocus blur may produce background patterns similar to biological structures. In addition, multiple organisms may be densely distributed, overlapped, or partially occluded, which can cause ambiguous boundaries, mask merging, and incorrect instance confusion. The limited availability of large-scale datasets with pixel-level instance annotations further restricts model training and quantitative evaluation in this field. Although the Segment Anything Model (SAM) has demonstrated strong general segmentation capability, directly applying it to zooplankton microscopic images remains difficult because of the feature distribution gap between natural-image pretraining and microscopic imaging, its dependence on externally provided prompts, and its limited instance discrimination in crowded scenes. To address these problems, this study constructs a zooplankton microscopic image instance segmentation dataset and develops an automatic SAM-based framework for accurate segmentation of individual zooplankton. **Method** First, we construct a zooplankton microscopic image instance segmentation dataset (ZooMIIS5K), a large-scale zooplankton microscopic image instance segmentation dataset containing 5,332 images from 48 zooplankton categories. Each visible organism is annotated with a fine-grained pixel-level instance mask, providing a reliable benchmark for model development and evaluation. Based on ZooMIIS5K, we propose a zooplankton microscopic image instance segmentation model (ZooMIIS-SAM), an end-to-end automatic instance segmentation framework built upon SAM. The framework contains three principal components: a dual-domain fusion vision transformer (DDF-ViT), a multi-scale fusion prompt module (MSFPM), and a target-aware perception module (TAPM). DDF-ViT is designed to reduce the feature distribution gap between the natural-image pretraining domain and the zooplankton microscopic-image domain. It combines spatial-domain adaptation and frequency-domain feature enhancement to strengthen foreground responses and supplement structural and boundary information. A spatial-gated low-rank adaptation branch generates task-related feature increments with limited additional parameters, while a frequency-domain feature enhancement module improves representations of weak boundaries and fine-grained structures. Bottleneck adapters are further introduced for feature alignment and contextual compensation. DDF-ViT is deployed only in selected encoder blocks to balance domain adaptation capability and parameter cost. MSFPM is designed to replace manually specified prompts with automatically generated visual prompts. It integrates multi-level encoder features through multi-scale strip convolution, channel recalibration, and cross-level fusion. The resulting feature is used to construct a feature pyramid and generate candidate regions. Region-of-interest features are then extracted and mapped into sparse soft prompt embeddings compatible with the SAM mask decoder. TAPM is introduced before mask decoding to strengthen instance-level discrimination. It uses candidate-region features to modulate global image representations, highlighting information related to the current target and reducing interference from surrounding backgrounds and neighboring instances. The enhanced image features and automatically generated visual prompts are jointly provided to the SAM mask decoder to predict individual instance masks. **Result** Experiments on ZooMIIS5K show that ZooMIIS-SAM achieves the best performance among the compared convolutional neural network-based and visual foundation model-based methods, with a mask mean average precision (mAP_{mask}) of 77.4%. Ablation experiments further verify the effectiveness of the proposed modules. Introducing DDF-ViT produces a substantial performance improvement, confirming that spatial-frequency feature adaptation is important for reducing the domain gap and enhancing the representation of zooplankton structures. The independent comparison of MSFPM shows that its multi-scale feature enhancement improves candidate-region localization while maintaining stable mask segmentation performance. TAPM further improves instance discrimination by incorporating candidate-region information into global image features. The complete model achieves the highest segmentation accuracy when DDF-ViT, MSFPM, and TAPM are jointly employed, indicating that dual-domain feature enhancement, automatic multi-scale prompt generation, and instance-aware modulation provide complementary benefits. Component-level ablation experiments further show that the spatial gated low-rank branch,

frequency-domain feature enhancement module, and bottleneck adapters make different contributions to the final performance. The spatial branch mainly strengthens foreground responses and suppresses irrelevant background interference, whereas frequency-domain enhancement contributes to structural consistency and high-quality mask prediction. Rank analysis indicates that a moderate low-rank setting provides a favorable balance between adaptation capability and preservation of pretrained representations. Visualization results show that ZooMIS-SAM can more accurately focus on zooplankton targets, reduce irrelevant responses in complex backgrounds, and improve the separation of adjacent instances in densely distributed scenes. These results demonstrate the effectiveness of combining domain-oriented encoder adaptation, learnable multi-scale prompt generation, and target-aware instance representation. **Conclusion** ZooMIS-SAM achieves favorable performance for instance segmentation of zooplankton microscopic images on ZooMIS5K. DDF-ViT improves microscopic-domain feature representation, MSFPM generates object-related prompts automatically, and TAPM strengthens the discrimination of adjacent instances. The constructed dataset and proposed framework can provide a technical reference for zooplankton counting and morphological analysis and demonstrate potential for practical computer-assisted zooplankton monitoring. The conclusions of this study are limited to the datasets and experimental conditions evaluated, and further validation under broader imaging conditions remains necessary.

Key words: Zooplankton; Instance segmentation; Segment Anything Model; Dual-domain feature enhancement; Target-aware perception

论文引用格式: DOI:10.11834/jig.260408

0 引言

浮游动物是海洋生态系统的重要组成部分,在海洋食物网、物质循环和生态过程中发挥着重要作用,其群落结构、丰度变化和时空分布也被广泛用于海洋生态状况与环境变化评估(Ratnarajah等,2023; Chiba等,2018)。浮游动物监测主要包括原位成像和实验室显微观察(Yuan等,2026)。原位成像适用于大范围连续观测,但在小尺度目标、复杂结构和相似物种的细粒度形态表征方面仍存在一定局限(Bergum等,2020)。实验室显微成像能够提供更丰富的形态与边界细节,更适用于物种鉴定、个体计数和形态分析(Sosa-Trejo等,2023)。然而,传统显微观察依赖专业人员进行人工鉴定和统计,在处理大规模样本时耗时费力,难以满足长期连续监测需求(Maracani等,2023)。因此,研究浮游动物显微图像的自动分析方法具有重要的实际意义(Li等,2026)。

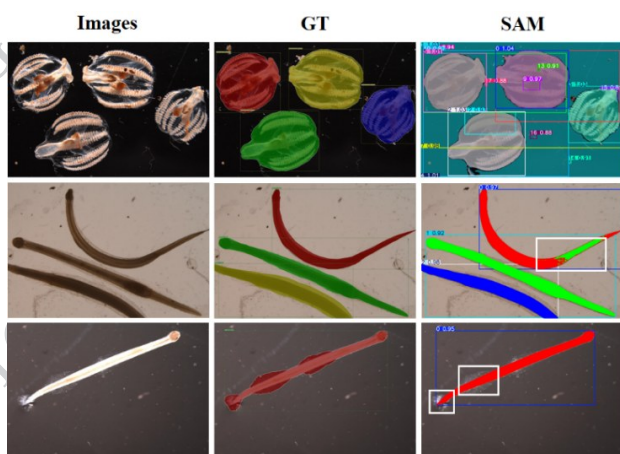
随着卷积神经网络(convolutional neural network, CNN)和视觉Transformer(vision transformer, ViT)的发展,浮游动物图像分析逐渐由人工特征方法转向深度学习方法,相关研究主要集中于图像分类和目标检测(Bureš等,2021; Yuan等,2024)。相比之下,像素级分割不仅需要识别目标类别,还需要准确提取目标轮廓,对图像分辨率、边界质量和局部

结构表达提出了更高要求。已有研究基于ZooScan图像开展了多目标实例分割(Overbeek等,2025),但总体而言,现有浮游动物图像研究仍以分类和目标检测为主,面向浮游动物的实例分割研究及公开像素级标注数据集相对较少,在一定程度上制约了浮游动物自动计数和形态分析方法的发展。

实例分割能够同时完成目标定位、类别识别和像素级轮廓提取,可为浮游动物个体计数、形态测量和群落结构分析提供更加精细的数据支持。Mask R-CNN(He等,2017)和YOLACT(Bolya等,2019)等经典方法在自然图像实例分割中取得了良好效果,但迁移至浮游动物显微图像时仍面临领域差异。近年来,分割一切模型(Segment Anything Model, SAM)(Kirillov等,2023)依托大规模预训练获得了较强的通用分割能力。为提高其在特定场景中的适用性,已有研究主要采用轻量适配器、低秩适配(low-rank adaptation, LoRA)和提示学习等方式,将SAM拓展至医学影像、遥感影像和水下图像等任务(Chen等,2023; Zhang等,2023; Chen等,2024; Lian等,2024)。此外,频域特征建模能够从频率成分角度补充空间域特征表达,为增强目标结构和边界信息提供了新的途径(Tang等,2024)。上述研究为结合空间域适配、频域特征增强和自动提示生成开展浮游动物显微图像实例分割提供了基础。

然而,自然图像预训练形成的视觉表征与浮游动物显微图像之间存在明显的域差异,直接使用

SAM 仍难以获得稳定的分割结果。如图 1 所示, SAM 在自动分割模式下容易生成大量冗余掩码,并将背景杂质或局部纹理误识别为目标;以真实边界框作为提示虽然能够减少冗余掩码,但对于弱边界和半透明目标,仍存在目标结构缺失、掩码边界不准确以及相邻实例混淆等问题。此外,交互式点提示或框提示依赖外部人工先验,难以直接满足浮游动物样本的端到端自动分析需求。



白色框突出 SAM 分割的局限;第一行为 SAM 自动分割模式的结果,第二、三行为以真实边界框作为提示的分割结果

图 1 SAM 分割结果

Figure 1 Segmentation results of SAM

在浮游动物显微图像分割任务中,仍面临以下三个关键挑战:

1)自然图像与浮游动物显微图像之间存在域差异,预训练特征难以充分表达透明或半透明目标的弱边界、低对比度和细粒度结构。

2)SAM 的自动分割模式依赖密集网格提示和后处理筛选,缺少对浮游动物目标结构与空间分布的显式建模,容易产生冗余掩码和背景误分割;交互式提示方式又依赖人工提供的点或边界框,难以实现无需人工干预的自动实例分割。

3)当多个浮游动物相互接触或局部遮挡时,全局特征和稀疏提示难以充分区分相邻目标,容易出现掩码合并、重叠和边界偏移,进而影响个体计数与形态分析。

针对上述问题,本文构建浮游动物显微图像实例分割数据集 ZooMIIS5K,并提出基于 SAM 的浮游动物显微图像实例分割框架 ZooMIIS-SAM。该框架首先通过双域融合视觉 Transformer (dual-domain

fusion vision transformer, DDF-ViT)联合适配空间域与频域特征,缓解自然图像预训练域与浮游动物显微图像域之间的特征偏移;随后利用多尺度融合提示模块 (multi-scale fusion prompt module, MSFPM)整合编码器的多层特征并自动生成视觉提示;最后通过目标感知模块 (target-aware perception module, TAPM)将候选区域信息引入全局特征,增强密集场景中的实例辨别能力。本文的主要贡献如下:

1)构建浮游动物显微图像实例分割数据集 ZooMIIS5K。该数据集包含 5,332 张图像,覆盖 48 个浮游动物类别,并提供细粒度像素级实例标注,为浮游动物显微图像实例分割研究提供了数据基础。

2)提出基于 SAM 的浮游动物显微图像实例分割框架 ZooMIIS-SAM。其中,DDF-ViT 通过空间域与频域特征的协同适配增强浮游动物主体、结构和弱边界表征;MSFPM 融合多层特征并生成自动视觉提示;TAPM 利用候选区域特征增强全局特征的实例感知能力,从而提高复杂显微场景中的自动实例分割性能。

3)在 ZooMIIS5K 上开展对比实验和消融实验,从分割精度、组件贡献、计算开销及长尾类别性能等方面验证所提方法,并在公开水下实例分割数据集 USIS10K 上进一步评价其跨场景适用性。

1 ZooMIIS-SAM

本文提出的 ZooMIIS-SAM 模型的总体结构如图 2 所示,主要包括 3 个部分:DDF-ViT、MSFPM 以及 TAPM。其中编码器中的 DDF-ViT 集成不同功能的适配器,替换掉 SAM 编码器中一些原始 ViT 结构,可以有效提取海洋浮游动物视觉特征。MSFPM 中的多尺度条状卷积用于整合编码器中的多层特征来增强多尺度特征表达,之后通过区域候选网络生成候选区域,进而自动生成视觉提示。最后通过 TAPM 获得增强的全局图像特征,将其和提示特征共同送入 SAM 解码器,从而实现整体结构端到端的实例掩码分割。

1.1 DDF-ViT 编码器

SAM 图像编码器是在大型自然图像分割数据集 SA-1B 上进行预训练的,因此具有较强的通用视觉表示能力。然而,浮游动物显微图像通常表现出目标弱边界、局部对比度低及背景干扰等特征,与

SA-1B 覆盖的自然图像场景存在明显差异, 导致原有的SAM编码器很难有效地提取浮游动物特征。

为解决此问题, 本文提出 DDF-ViT, 结构如图2(b)所示, 通过空间门控低秩适配(spatial-gated low-rank adaptation, SG-LoRA)和频域特征增强模块(frequency-domain feature enhancement module, FFEM)的协同建模, 结合空间域的目标响应适配与频域的结构边界增强, 从而缓解预训练自然图像域与显微图像域之间的特征偏移。考虑到DDF-ViT中的双域增强与特征对齐操作会引入额外参数, 本文提出的方法是从索引8对应的编码块开始, 每隔两层使用DDF-ViT替换掉原始ViT, 以控制附加参数与计算开销, 并在领域适应能力和模型复杂度之间取得平衡。

具体而言, DDF-ViT在每个ViT块中引入两个分支: SG-LoRA和FFEM, 从空域和频域增强输入特征。同时, 集成两个由多层感知机(multilayer perceptron, MLP)构成的瓶颈适配器(bottleneck adapter): 一个在融合后阶段, 另一个在前馈网络(feed-forward network, FFN)残差路径中, 实现多源特征对齐和上下文补偿。设输入特征表示为 $X \in \mathbb{R}^{H \times W \times C}$, 其中 H 和 W 分别表示特征图的空间尺寸, C 表示通道维度。DDF-ViT的整体计算过程可表示为:

$$Y = X + \text{Adapter}_1(X_{\text{att}} + \Delta X_{\text{spa}} + X_{\text{fre}}) \quad (1)$$

$$X_{\text{out}} = \text{MLP}(\text{LN}(Y)) + \text{Adapter}_2(Y) \quad (2)$$

式中, X_{att} 表示来自冻结多头注意力分支的主干特征, ΔX_{spa} 表示 SG-LoRA 产生的特征增量, LN 表示层归一化, X_{fre} 表示 FFEM 生成的增强特征。Adapter₁ 和 Adapter₂ 分别表示位于特征融合后阶段和 FFN 残差路径中的瓶颈适配器。X_{out} 表示当前模块的最终输出特征。

1) SG-LoRA: 在显微图像中, 原始 SAM 编码器的注意机制容易受到背景干扰的影响, 难以一致地聚焦于前景的浮游动物。为了解决这些问题, 本文引入 SG-LoRA 来实现对前景目标的响应。具体而言, 采用带有自适应门控机制的低秩变换生成空间特征增量, 其计算过程表示如下:

$$\Delta X_{\text{spa}} = \text{Gate}(\text{LN}(X)) \cdot \text{Up}(\text{Down}(\text{LN}(X))) \quad (3)$$

式中, Gate(·) 结合全局池化和线性映射生成一个标量系数, 用于自适应地调节低秩分支输出的权重。

Down(·) 和 Up(·) 分别表示降维投影和升维投影。

2) FFEM: 仅依赖空间分支难以同时实现噪声抑制与目标增强的解耦。为解决这一问题, 本文进一步提出频域特征增强模块, 通过自适应选择具有判别性的频率成分来增强结构与边缘表征, 从而与空间分支形成互补。具体而言, 给定输入特征 X , 首先沿空间维度逐通道执行二维实数傅里叶变换, 得到复数频谱:

$$\mathcal{F} = \text{RFFT}_2(X) \quad (4)$$

为提高数值稳定性, 对频谱中的异常幅值进行截断。随后, 将处理后频谱 $\bar{\mathcal{F}}$ 的实部和虚部沿通道维拼接:

$$Q = \text{Cat}[\text{Re}(\bar{\mathcal{F}}), \text{Im}(\bar{\mathcal{F}})] \quad (5)$$

式中, $\bar{\mathcal{F}}$ 表示经过幅值截断处理的频谱, Cat(·) 表示沿通道维拼接。

随后, 拼接后的频域特征首先进行批归一化, 随后通过两个 1×1 卷积和激活函数进行联合映射:

$$\hat{Q} = \text{Conv}_{1 \times 1} \left(\text{Clip} \left[\sigma \left(\text{Conv}_{1 \times 1} (\text{BN}(Q)) \right), 0, 1 \right] \right) \quad (6)$$

式中, BN 代表批归一化, σ 表示激活函数, Clip(·, 0, 1) 用于限制中间特征的响应范围。

最后, 将映射结果 $\hat{Q}$ 沿通道维划分为实部特征 $\hat{Q}_r$ 和虚部特征 $\hat{Q}_i$, 重新构造复数频谱, 并通过二维逆实数傅里叶变换恢复空间域特征:

$$X_{\text{fre}} = \text{IRFFT}_2 \left[\alpha \left(\hat{Q}_r + i \hat{Q}_i \right) \right] \quad (7)$$

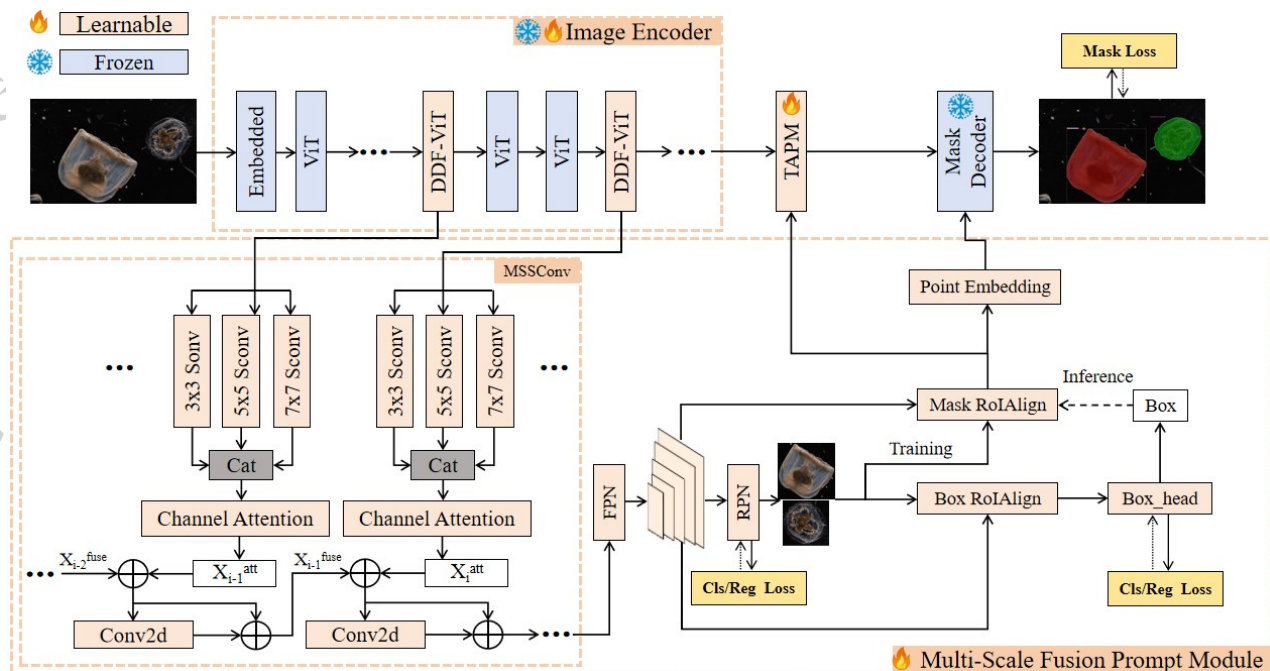
式中, α 为频域分支的缩放系数。

通过对频谱实部和虚部进行联合映射, FFEM 能够自适应调整不同频率成分的响应, 并将增强后的频域信息恢复至空间域, 从而提升模型对目标结构和边界细节的表征能力。

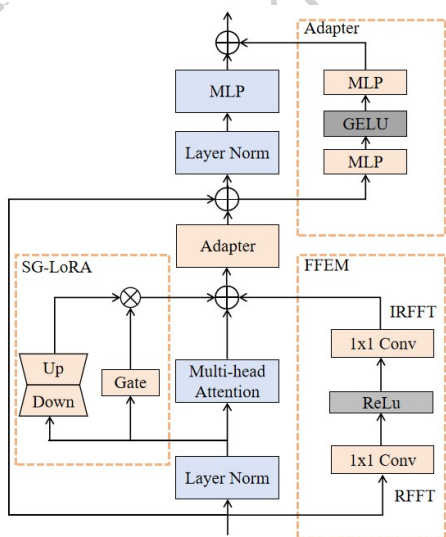
3) Bottleneck Adapter: 本文进一步引入两个瓶颈适配器。其中, Adapter₁ 位于特征融合后的阶段, 用于实现多源特征分布对齐; Adapter₂ 位于 FFN 残差路径中, 在保留原始 FFN 表征的同时增强上下文建模能力。瓶颈适配器的一般计算过程可表示为:

$$P = \text{MLP}_{\text{up}} \left(\sigma \left(\text{MLP}_{\text{down}}(F) \right) \right) \quad (8)$$

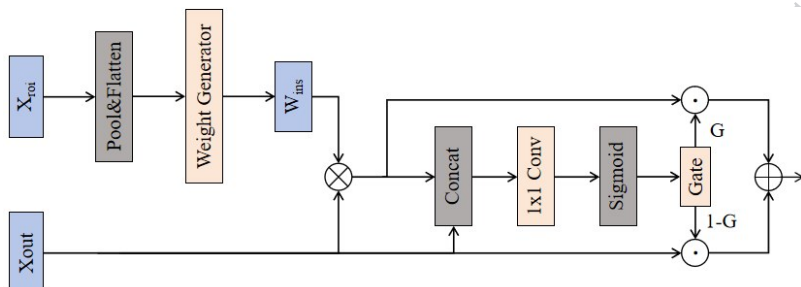
式中, F 表示适配器的输入特征。



(a) 模型整体结构图



(b) DDF-ViT模块结构图



(c) TAPM模块结构图

(a) Overall structure diagram of the ZooMIIS-SAM model; (b) DDF-ViT module structure diagram; (c) TAPM module structure diagram

图2 ZooMIIS-SAM模型整体结构图及DDF-ViT、MSFPM及TAPM模块结构图

Figure 2 Overall structure diagram of ZooMIIS-SAM model and structure diagram of DDF-ViT, MSFPM and TAPM modules

1.2 MSFPM

SAM 的自动分割模式主要基于密集网格提示和后处理筛选生成候选掩码。这种方式缺乏对浮游动物目标分布和形态结构的自适应建模,在显微图像中容易产生冗余提示、造成大量误分割掩码。而依赖用户定义的提示信息,例如点提示或边界框提

示,这又限制了其在全自动实例分割任务中的应用。为解决这一问题,本文引入 MSFPM,以自动方式生成提示特征。

具体而言,MSFPM 首先引入多尺度条状卷积(multi-scale strip convolution, MSSConv),对所选 DDF-ViT 层输出进行多尺度增强与跨层融合,得到

融合特征。给定第 i 个 DDF-ViT 特征 $X_i \in \mathbf{R}^{C \times H \times W}$, 单尺度分支中的条带卷积可表示为:

$$\text{Sconv}_{k \times k}(X_i) = \text{Conv}_{1 \times 1} \left(\text{WConv} \left(\text{HConv} \left(\text{Conv}_{k \times k}(X_i) \right) \right) \right) \quad (9)$$

式中, $\text{Sconv}(\cdot)$ 采用标准 $k \times k$ 卷积与两个正交方向的条带卷积级联, 以增强对浮游动物细长结构和方向性信息的建模能力。

引入多尺度分支 $k \in \{3, 5, 7\}$, 以在不同感受野下提取特征。随后, 将不同尺度分支得到的特征进行拼接, 并通过 1×1 卷积进行融合, 得到聚合特征 X_i^{cat} :

$$X_i^{\text{cat}} = \text{Conv}_{1 \times 1} \left(\text{Cat}_{k \in \{3, 5, 7\}} \left(\text{Sconv}_{k \times k}(X_i) \right) \right) \quad (10)$$

为进一步筛选有效特征, 本文采用通道注意力机制对 X_i^{cat} 进行校准。具体而言, 首先分别沿空间维度对 X_i^{cat} 执行全局平均池化和全局最大池化:

$$z_i^{\text{avg}} = \text{AvgPool}(X_i^{\text{cat}}), z_i^{\text{max}} = \text{MaxPool}(X_i^{\text{cat}}) \quad (11)$$

随后, 将两类通道描述向量分别输入参数共享的两层映射网络, 并将输出相加后通过 Sigmoid 函数生成通道权重:

$$a_i = \sigma \left[W_2 \sigma(W_1 z_i^{\text{avg}}) + W_2 \sigma(W_1 z_i^{\text{max}}) \right] \quad (12)$$

式中, W_1 和 W_2 分别表示降维和升维映射矩阵。

最终, 将通道权重 a_i 沿空间维度广播, 并与 X_i^{cat} 逐通道相乘, 得到校准后的特征 X_i^{att} :

$$X_i^{\text{att}} = X_i^{\text{cat}} \odot a_i \quad (13)$$

在此基础上, 引入跨层特征融合策略, 将当前层特征 X_i^{att} 与前一层融合结果 X_{i-1}^{fuse} 进行逐元素相加, 随后通过 3×3 卷积进行残差增强, 得到本层融合特征 X_i^{fuse} :

$$Z_i = X_i^{\text{att}} + X_{i-1}^{\text{fuse}} \quad (14)$$

$$X_i^{\text{fuse}} = Z_i + \text{ReLU} \left(\text{BN} \left(\text{Conv}_{3 \times 3}(Z_i) \right) \right) \quad (15)$$

该融合结果继续传递至下一选取层, 从而逐步整合不同编码深度的语义与空间信息。对于首个选取层, 将前一层融合结果初始化为零。

在上述跨层融合基础上, 将最终融合特征进一步输入特征金字塔网络 (feature pyramid network, FPN), 以构建适用于不同尺度浮游动物实例的层级特征。具体而言, 通过两级转置卷积、一级转置卷积、恒等映射和最大池化分别产生步长为 4、8、16 和 32 的四级特征, 并通过进一步下采样获得步长为 64 的第五级特征。各级特征经 1×1 卷积完成通

道对齐, 并利用 3×3 卷积进行特征增强, 最终形成五级特征金字塔。

FPN 输出的多尺度特征同时用于区域建议网络 (region proposal network, RPN) 候选区域生成和感兴趣区域 (region of interest, ROI) 特征提取。RPN 基于多级 FPN 特征预测不同尺度的候选区域, 检测分支利用通过用于边界框预测的感兴趣区域特征对齐操作 (Box RoIAlign) 提取特征, 并将其输入检测框预测头 (Box_head); 掩码分支则通过独立的掩码感兴趣区域特征对齐操作 (Mask RoIAlign) 提取 ROI 特征。该特征一方面输入 TAPM; 另一方面输入提示映射网络, 经卷积压缩、特征展平和多层感知机映射后, 生成与 SAM 解码器提示空间一致的稀疏软提示。具体而言, 设第 i 个 ROI 特征为 F_i^{roi} , 其提示映射过程可表示为:

$$H_i = \sigma \left[\text{BN} \left(\text{Conv}_{3 \times 3}(F_i^{\text{roi}}) \right) \right] \quad (16)$$

$$P_i = \Phi \left(\text{MLP} \left[\text{Flatten}(H_i) \right] \right) \quad (17)$$

式中, $\Phi(\cdot)$ 表示特征重排与正弦非线性映射, P_i 表示生成的稀疏软提示。这些提示不对应图像中的显式二维点击坐标, 而是由提示映射网络根据 ROI 特征自适应生成的连续特征嵌入。最后, 将稀疏软提示输入解码器, 完成目标实例的掩码预测。通过上述过程, MSFPM 利用检测分支定位的实例区域自动构造稀疏软提示, 替代 SAM 原有的人工交互式提示, 实现端到端的浮游动物实例分割。

1.3 TAPM

在实例密集、边界弱、背景干扰强的显微场景中, 全局特征和稀疏提示不足以实现准确的实例识别。因此, 本文提出 TAPM, 在解码器阶段之前明确地将 ROI 特征与全局特征融合在一起, 使全局特征在进入解码器之前包含实例级先验, 从而增强解码过程中的实例辨别能力。模块结构如图 2(c) 所示。TAPM 的输入由两部分组成: 1.2 节中的 ROI 特征 X_{roi} 和编码器输出的全局特征 X_{out} 。

具体而言, 首先通过全局池化和展平操作将 $X_{\text{roi}} \in \mathbf{R}^{C \times H_r \times W_r}$ 压缩为通道描述向量。随后, 将得到的 C 维向量输入权重生成网络, 生成 C^2 个参数, 并将其重塑为实例特定的动态调制权重 $W_{\text{ins}} \in \mathbf{R}^{C \times C}$ 。其计算过程可表示为:

$$W_{\text{ins}} = \Psi \left(\text{Flatten} \left(\text{Pool}(X_{\text{roi}}) \right) \right) \quad (18)$$

式中, $\psi(\cdot)$ 表示动态权重生成函数, 由两个全连接层和 GELU 激活函数组成; $\text{Flatten}(\cdot)$ 表示展平操作; $\text{Pool}(\cdot)$ 表示池化操作。

接下来, 利用动态调制矩阵 $\mathbf{W}_{\text{ins}}$ 对全局特征 $\mathbf{X}_{\text{out}}$ 进行实例感知变换。随后, 引入门控机制, 将变换后的特征与原始全局特征进行自适应融合, 从而突出与候选实例相关的响应, 并抑制无关背景干扰。其计算过程可表示为:

$$\mathbf{G} = \sigma(\text{Conv}([\mathbf{W}_{\text{ins}} \mathbf{X}_{\text{out}}, \mathbf{X}_{\text{out}}])) \quad (19)$$

$$\mathbf{X}_{\text{aug}} = \mathbf{G} \odot (\mathbf{W}_{\text{ins}} \mathbf{X}_{\text{out}}) + (1 - \mathbf{G}) \odot \mathbf{X}_{\text{out}} \quad (20)$$

式中, $[\cdot, \cdot]$ 表示沿通道维度进行拼接, $\odot$ 表示逐元素乘法, $\mathbf{X}_{\text{aug}}$ 表示经过实例感知调制后的增强全局特征。最后, 将实例感知增强后的全局特征 $\mathbf{X}_{\text{aug}}$ 输入 SAM 解码器进行分割。

此外, 在训练阶段, Mask RoIAlign 根据正样本区域从 FPN 多尺度特征中提取 ROI 特征。推理阶段, RPN 候选区域经过 Box_head 后, Mask RoIAlign $L_{\text{rpn_cls}} = -\frac{1}{N_{\text{rpn}}} \sum_{i \in S_{\text{rpn}}} [y_i^* \log(p_i) + (1 - y_i^*) \log(1 - p_i)]$ 式中, S_{rpn} 表示参与分类损失计算的有效锚框集合, 包含采样得到的正、负样本; $N_{\text{rpn}} = |S_{\text{rpn}}|$ 表示有效锚框数量; p_i 表示第 i 个锚框被预测为前景目标的概率; $y_i^* \in \{0, 1\}$ 表示对应的真实二值标签, 其中 1 和 0 分别表示前景正样本和背景负样本。

RPN 边界框回归采用 Smooth L1 损失, 其具体形式为:

$$L_{\text{rpn_reg}} = \frac{1}{N_{\text{rpn}}} \sum_{i \in S_{\text{rpn}}} \text{SmoothL1}(t_i - t_i^*) \quad (24)$$

式中, S_{rpn}^+ 表示 RPN 阶段的正样本集合, N_{rpn}^+ 表示 RPN 阶段正样本数量; t_i 和 t_i^* 分别表示第 i 个正样本锚框的预测边界框偏移量和真实边界框偏移量。这里的边界框偏移量包含中心点坐标和宽高尺度等信息。

ROI 阶段进一步对 RPN 生成的候选区域进行分类判别、边界框精修和实例掩码预测, 其损失函数表示为:

$$L_{\text{ROI}} = L_{\text{roi_cls}} + L_{\text{roi_reg}} + L_{\text{mask}} \quad (25)$$

$$L_{\text{roi_cls}} = \frac{1}{N_{\text{roi}}} \sum_{i=1}^{N_{\text{roi}}} -\log p_{i,c_i} \quad (26)$$

$$L_{\text{roi_reg}} = \frac{1}{N_{\text{roi}}^+} \sum_{i \in S_{\text{roi}}^+} \text{SmoothL1}(d_i - d_i^*) \quad (27)$$

根据最终保留的检测框提取 ROI 特征并输入 TAPM。RPN、检测分支与包含 TAPM 的掩码分支在同一网络中联合优化。掩码分支的梯度可通过 TAPM 和 ROI 特征提取路径回传至 FPN 及其上游可训练模块, 但由于候选框筛选、匹配和采样等过程不构成连续可微路径, 该梯度不会沿候选框生成路径直接监督 RPN Head, RPN 主要由自身的分类损失和边界框回归损失进行优化。

1.4 损失函数

ZooMIIS-SAM 采用两阶段实例分割结构, 其训练目标由 RPN 阶段损失与 ROI 阶段损失共同组成, 可表示为:

$$L_{\text{total}} = \lambda_1 L_{\text{RPN}} + \lambda_2 L_{\text{ROI}} \quad (21)$$

式中, $\lambda_1 = \lambda_2 = 1$, RPN 阶段用于生成候选区域, 其损失由前景/背景分类损失和候选框回归损失组成:

$$L_{\text{RPN}} = L_{\text{rpn_cls}} + L_{\text{rpn_reg}} \quad (22)$$

RPN 分类采用二分类交叉熵损失, 其具体形式为:

$$(23)$$

$$L_{\text{mask}} = \frac{1}{N_{\text{roi}}^+} \sum_{i \in S_{\text{roi}}^+} \text{BCE}(\mathbf{M}_i, \hat{\mathbf{M}}_i) \quad (28)$$

式中, $L_{\text{roi_cls}}$ 为候选区域分类损失, 用于区分 48 类浮游动物前景类别及背景类别; $L_{\text{roi_reg}}$ 为 ROI 边界框回归损失; L_{mask} 为实例掩码分割损失。 N_{roi} 和 N_{roi}^+ 分别表示 ROI 阶段参与训练的候选区域数量和正样本候选区域数量; S_{roi}^+ 表示 ROI 阶段正样本集合; c_i 表示第 i 个 ROI 的真实类别, p_{i,c_i} 表示该 ROI 被预测为真实类别 c_i 的概率; d_i 和 d_i^* 分别表示第 i 个正样本 ROI 的预测边界框偏移量和真实边界框偏移量; $\mathbf{M}_i$ 和 $\hat{\mathbf{M}}_i$ 分别表示第 i 个正样本 ROI 对应的真实实例掩码和预测实例掩码; $\text{BCE}(\cdot)$ 表示在掩码像素上计算的二值交叉熵损失。回归损失和掩码损失仅在正样本候选区域上计算, 背景候选区域仅参与分类损失。

2 实验

为了验证所提出的 ZooMIIS-SAM 在浮游动物显微图像分割中的有效性, 我们在自建的 ZooMIIS5K 数据集上进行了对比实验和消融研究。此外, 在公开的水下数据集 USIS10K 上进一步进行了实验, 验

证了所提模型的跨场景泛化能力。

水域。自然资源部北海局北海生态中心为样本采集和物种鉴定提供了技术支持。研究团队使用

2.1 数据集

样本采集自黄海和渤海黄河三角洲附近的沿海

表 1 ZooMIIS5K 中 48 个类别的实例数量分布
Table 1 Instance distribution of the 48 categories in ZooMIIS5K

1个

Unit: instances					
类别	实例数量	类别	实例数量	类别	实例数量
Calanus sinicus	1184	Themisto gracilipes	205	Sugiura chengshanense	62
Sagitta crassa	1154	Clytia folleata	205	Centropages tenuiremis	59
Eirene menoni	683	Oikopleura longicauda	194	Polychaeta larva later	56
Macrura larva	597	Tornaria larva	180	Remains	52
Pleurobrachia globosa	519	Trochophora larva	163	Megalopa larva	49
Muggiaea atlantica	497	Turritopsis nutricula	149	Paracalanus parvus	47
Penilia avirostris	419	Bougainvillia muscus	139	Acetes larva	42
Fritillaria formica	394	Fish eggs	137	Oithona plumifera	32
Doliioletta gegenbauri	304	Proboscidaactyla flavicirrata	131	Centropages abdominalis	27
Evadne tergestina	283	Jellyfish larva	116	Calanopia thompsoni	24
Ophioplutues larva later	271	Alima larva	107	Centropages dorsispinatus	24
Ophioplutues larva early	252	Porcellana zoea larva	101	Pontellopsis tenuicauda	21
Brachyura zoea larva	248	Obelia dichotoma	97	Clytia hemisphaerica	18
Aequorea conica	240	Fish larva	78	Actinotrocha larva	16
Acartia pacifica	232	Varitentaculata yantaiensis	69	Ectopleura bimanatus	10
Labidocera rotunda	228	Polychaeta larva early	65	Euphausia pacifica	4

徕卡 M205A 立体显微镜对放置在培养皿中的浮

游动物样本进行成像。成像参数,包括照度和焦距,根据生物体的形态特征及其焦平面位置进行动态调整。在明光场和暗光场条件下获取图像。然后使用 Labelme 工具进行像素级实例注释。该数据集包含 5,332 张图像、10,184 个像素级标注实例,涵盖 48 个浮游动物类别,平均每类包含约 212 个实例。各类别的具体实例数量如表 1 所示。数据集按照 8:2 的比例划分为训练集和验证集,训练集用于模型参数优化,验证集不参与梯度更新,用于实验配置比较、训练过程监测,并以验证集上的掩码分割性能为依据选择最佳检查点。考虑到 ZooMIIS5K 存在明显的类别长尾分布,部分低频类别的样本数量较少,进一步划分独立测试集可能造成少数类别在各子集中

的样本不足,因此本文未单独设置测试集。验证集同时用于模型选择与性能评价,所报告结果可能存在一定的乐观偏差。所有对比方法均采用相同的数据划分、模型选择规则和评价指标,以保证横向比较的一致性。本数据集可同时支持实例分割、目标检测和物种分类等任务。

在标注质量控制方面,标注人员经过两周的统一培训后,依据统一的物种鉴定与像素级轮廓标注规范完成初始标注。对于透明、弱边界及相互遮挡的目标,以显微图像中可辨识的实际轮廓作为标注依据。初始标注完成后,由具备浮游动物鉴定经验的专业人员对物种类别、实例完整性和边界准确性进行复核,对于存在类别或轮廓争议的样本,经讨论确认后统一修正,从而保证数据集标注标准的一致性。为进一步量化标注者间的一致性,本文从数据

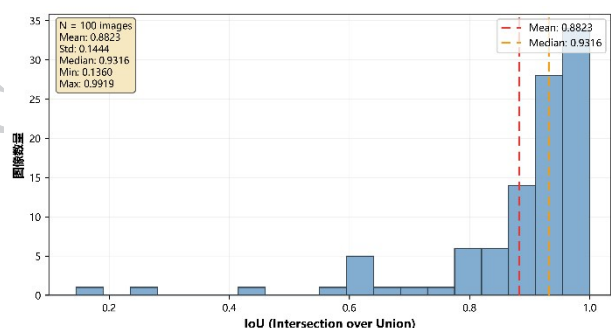


图3 标注者间掩码一致性(IoU)分布

Figure 3 Distribution of inter-annotator mask agreement measured by IoU

集中随机抽取100幅图像,由3名经过统一培训的标注者在互不参考彼此结果的情况下独立完成实例掩码标注。对于每幅图像,首先将同一标注者标注的全部实例掩码合并为二值前景掩码,然后计算3名标注者之间3种两两组合的交并比(intersection over union, IoU),并将其平均值作为该图像的标注一致性得分。由此得到100个图像级一致性得分,其分布如图3所示。100幅图像的一致性得分均值为0.8823,中位数为0.9316,标准差为0.1444,表明不同标注者对大部分图像中目标前景区域及总体边界的标注具有较高一致性。

2.2 实验设置

所有实验均采用通用目标数据集(Microsoft Common Objects in Context, COCO)的实例分割评价协议。其中, mAP_{mask} 表示在IoU阈值为0.50~0.95、步长为0.05时计算的掩码平均精度均值; AP_{mask}^{50} 和 AP_{mask}^{75} 分别表示IoU阈值为0.50和0.75时的掩码平均精度。ZooMIIS-SAM及对比模型均基于PyTorch和MMDetection(Chen等,2019)框架实现。在训练和推理过程中,输入图像统一调整为512×512,并采用随机水平翻转、随机缩放和随机裁剪等数据增强策略。由于结构限制,YOLACT保留其默认输入分辨率550×550,以保证其原型掩码分支与主干网络之间的兼容性。所有模型均在NVIDIA 3090 GPU(24 GB)上训练,单张批次大小设置为2,训练轮数为300。优化器采用AdamW,初始学习率设置为 1×10^{-4} ,采用多步学习率衰减策略调整学习率。对于ZooMIIS-SAM,训练过程中冻结SAM预训练权重。

2.3 对比实验

提出的ZooMIIS-SAM是将基于SAM基础模型

的提示学习应用于海洋浮游动物场景的尝试,重点探索其在该领域的潜力。为了验证其性能,我们在ZooMIIS5K数据集上比较了ZooMIIS-SAM与基于CNN的架构和基于基础视觉模型的方法。

如表2所示,WaterMask优于基准Mask R-CNN,这表明了针对特定成像场景的任务特定架构设计的重要性。此外,USIS-SAM通过更强的骨干网络进一步提高了分割性能,表明高质量的视觉特征提取在复杂环境中起着至关重要的作用。最后,当高质量的视觉特征提取器与目标感知模块协同工作时,我们的方法达到了最佳性能,证明了全局-局部协同微调在处理复杂显微场景中的有效性和合理性。不同方法的定性比较如图4所示。

2.4 消融实验

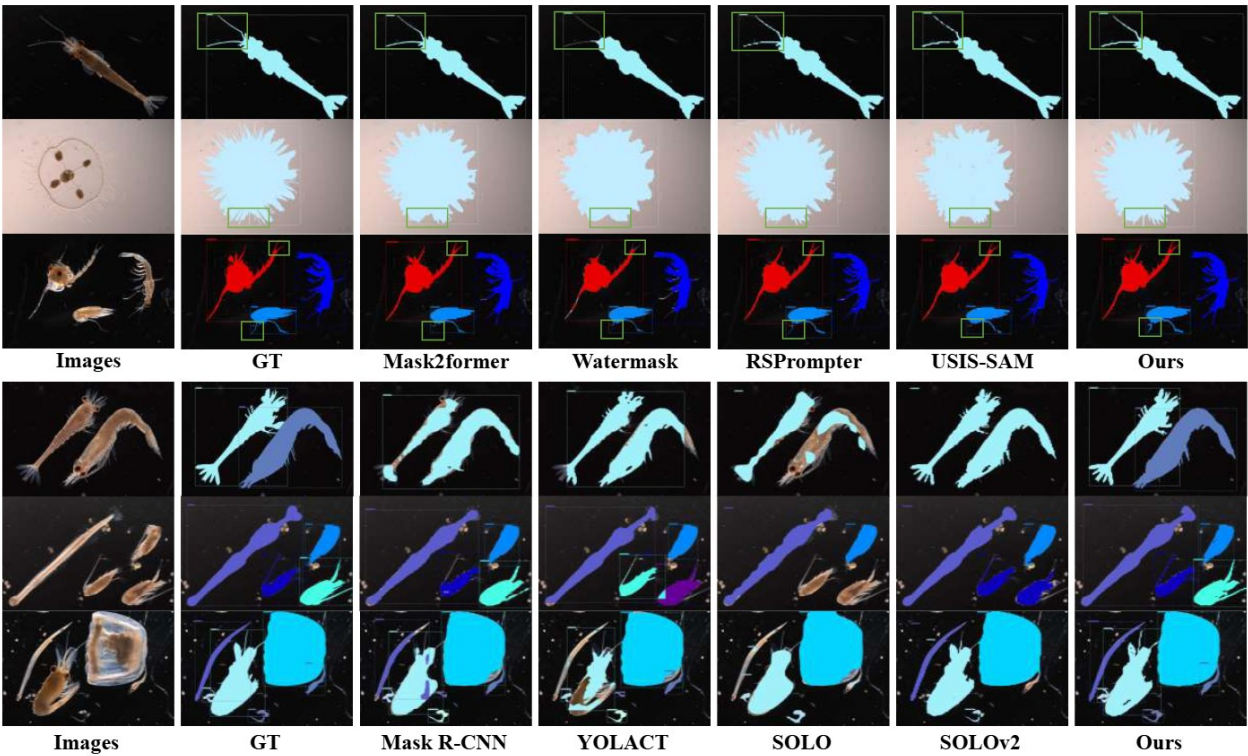
2.4.1 不同结构的有效性

1)整体组件消融分析。为分析模型中每个组成部分的单独贡献,本文进行了消融研究,并增加了检测框指标 mAP_{box} 、 AP_{box}^{50} 和 AP_{box}^{75} 。其中, mAP_{box} 表示在IoU阈值为0.50~0.95、步长为0.05时计算的检测框平均精度均值, AP_{box}^{50} 和 AP_{box}^{75} 分别表示IoU阈值为0.50和0.75时的检测框平均精度。首先,为评价MSSConv的作用,本文构建了采用1×1卷积进行通道映射并通过平均操作完成跨层融合的退化基线,同时保留后续FPN、RPN、RoIAlign及掩膜解码流程,以保证两组实验均具备完整的自动实例分割能力。如表3所示,与普通1×1卷积相比,引入MSSConv后, mAP_{box} 和 AP_{box}^{75} 分别提高2.2和1.6个百分点。与此同时, AP_{box}^{50} 基本不变。这表明MSSConv带来的增益主要体现在更严格交并比条件下的目标定位,而在较宽松的定位阈值下提升并不明显。其可能原因是,冻结编码器输出的多层特征已包含目标位置、尺度和区域范围等通用信息,MSSConv通过不同尺度和方向的空间卷积进一步组织这些特征,增强了对目标空间范围及结构方向的表征,从而改善了RPN候选区域生成与边界框回归。相比之下,MSSConv对掩膜指标的提升相对有限。虽然MSSConv比普通1×1卷积具有更强的局部与多尺度建模能力,但其输入仍来自冻结的SAM编码器,对浮游动物显微图像中弱边界和细粒度形态等领域特有信息的表达可能受到一定限制;仅依靠后续的多尺度特征处理,难以充分弥补预训练域与目标域之间的表征差异。

表2 不同模型在 ZooMIIS5K 数据集上的性能对比
Table 2 Performance comparison of different models on the ZooMIIS5K dataset

Unit: %					/%	
模型	期刊/会议	Backbone	Params	mAP_{mask}	AP_{mask}^{50}	AP_{mask}^{75}
Mask R-CNN(He 等, 2017)	ICCV	ResNet-101	63M	68.9	89.9	71.8
YOLACT(Bolya 等, 2019)	ICCV	ResNet-101	49M	70.7	90.3	75.8
SOLO(Wang 等, 2020)	ECCV	ResNet-101	79M	57.8	87.2	58.2
SOLOv2(Wang 等, 2020)	NeurIPS	ResNet-101	65M	68.4	88.4	73.3
Mask2Former(Cheng 等, 2022)	CVPR	ResNet-101	63M	71.9	86.4	77.9
WaterMask(Lian 等, 2023)	ICCV	ResNet-101	67M	71.4	91.0	76.0
RSPrompter(Chen 等, 2024)	TGRS	ViT-H	632M	71.2	91.2	78.2
USIS-SAM(Lian 等, 2024)	ICML	ViT-H	701M	72.5	93.8	78.4
ZooMIIS-SAM		ViT-H	891M	77.4	96.3	85.1

注:加粗数值表示该列最优



为更直观展示本文方法在弱边界目标上的改善效果,图中对部分对比模型的典型分割差异区域进行了红框标注

图4 不同模型在 ZooMIIS5K 数据集上的实例分割可视化结果

Figure 4 Visualization results of instance segmentation by different models on the ZooMIIS5K dataset

在 Baseline (MSSConv) 的基础上加入 DDF-ViT 后, mAP_{mask} 由 68.6% 提高至 76.5%, mAP_{box} 由 77.9% 提高至 86.6%, 分别提升 7.9 和 8.7 个百分点; AP_{mask}^{75} 和 AP_{box}^{75} 也分别提高 9.2 和 4.6 个百分点。该结果说明, 通过 DDF-ViT 对预训练编码器进行领域适配, 能

够增强模型对显微图像领域特征的提取能力, 缓解自然图像预训练特征与浮游动物显微图像之间的分布差异。

单独加入 TAPM 后, mAP_{mask} 由 68.6% 提高至 71.3%, mAP_{box} 由 77.9% 提高至 80.3%, 分别提升

表3 MSSConv、DDF-ViT、TAPM模块对模型分割性能影响比较
Table 3 Effect of MSSConv , DDF-ViT and TAPM on segmentation performance

模型设置	Unit: %					
	mAP _{mask}	AP ⁵⁰ _{mask}	AP ⁷⁵ _{mask}	mAP _{box}	AP ⁵⁰ _{box}	AP ⁷⁵ _{box}
Baseline (1 x 1 Conv + Average Fusion)	68.2	90.5	74.0	75.7	93.2	87.6
Baseline (MSSConv)	68.6	90.6	74.5	77.9	93.0	89.2
Baseline (MSSConv) + DDF-ViT	76.5	95.0	83.7	86.6	95.5	93.8
Baseline (MSSConv) + TAPM	71.3	92.5	77.4	80.3	93.4	90.7
Baseline (MSSConv) + DDF-ViT + TAPM	77.4	96.3	85.1	89.1	96.5	95.4

注:加粗数值表示该列最优值

2.7和2.4个百分点,表明TAPM能够增强实例感知全局特征。在已采用DDF-ViT的基础上进一步加入TAPM后,mAP_{mask}由76.5%提高至77.4%,mAP_{box}由86.6%提高至89.1%;其中,AP⁵⁰_{mask}和AP⁷⁵_{mask}分别提高1.3和1.4个百分点,mAP_{box}提高2.5个百分点。这进一步说明,TAPM能够在领域适配特征的基础上增强目标实例感知与提示表达,有助于改善实例定位与掩膜分割性能。

2)强基线下的组件有效性分析。为验证本文模块在强基线上的有效性,本文以USIS-SAM为基础,分别引入TAPM和DDF-ViT,并在统一输入尺寸下统计分割性能和浮点运算量(floating-point operations,FLOPs),结果如表4所示。

单独引入TAPM后,模型性能未得到明显提升,说明TAPM的作用受到输入ROI特征质量的影响。引入DDF-ViT后,FLOPs增加9.3%,mAP_{mask}和AP⁷⁵_{mask}分别提高3.3和3.4个百分点,表明空间域与频域联合适配能够增强浮游动物的结构与边界表征。在此基础上加入TAPM,仅增加约1.0%的FLOPs,AP⁷⁵_{mask}提高2.0个百分点。严格重叠阈值下的增益更加明显,说明DDF-ViT增强的目标结构与边界信息为TAPM提供了更具判别性的ROI特征,使其能够更加准确地调制候选实例对应的全局特征。完整ZooMIS-SAM的FLOPs较USIS-SAM增加15.0%,而mAP_{mask}和AP⁷⁵_{mask}分别提高4.9和6.7个百分点。上述结果表明,模型性能变化与计算量增幅并不呈简单对应关系,其性能增益主要与DDF-ViT的双域特征适配及其与TAPM的协同作用有关。

3)DDF-ViT子模块消融分析。为了进一步分析DDF-ViT中各子模块的独立贡献和协同效应,也对

表4 强基线下不同组件的分割性能与计算开销对比
Table 4 Comparison of segmentation performance and computational cost of different components based on the strong baseline

模型	FLOPs	mAP _{mask}	AP ⁵⁰ _{mask}	AP ⁷⁵ _{mask}
USIS-SAM	945.3	72.5	93.8	78.4
USIS-SAM+TAPM	955.4	72.0	92.2	77.8
USIS-SAM+DDF-ViT	1033.0	75.8	94.7	81.8
USIS-SAM+ TAPM + DDF-ViT	1043.2	76.6	95.6	83.8
ZooMIS-SAM	1087.3	77.4	96.3	85.1

注:加粗数值表示该列最优值;分割性能指标的单位为%;FLOPs在输入尺寸为512×512、批次大小为1的条件下统计,单位为G

其进行消融实验。如表5所示,不同组件对DDF-ViT的性能贡献存在差异。仅使用Adapter时,模型mAP_{mask}为70.6%,说明单纯依赖轻量适配结构难以充分建模浮游动物显微图像中的复杂目标特征。相比之下,单独引入SG-LoRA后,模型mAP_{mask}提升至75.3%,表明空间低秩适配分支能够有效增强前景目标响应。单独使用FFEM时,模型mAP_{mask}为71.5%,整体提升相对有限,这是因为频域分支主要用于增强边界与结构细节,而边界区域在整个实例掩码中占比较小,因此其对整体mAP_{mask}的直接提升相对有限。然而,从AP⁷⁵_{mask}指标可以看出,FFEM具有重要作用。与仅使用Adapter相比,单独引入FFEM后AP⁷⁵_{mask}从76.1%提升至78.3%,说明频域增强有助于改善高质量掩码预测中的边界完整性和结构一致性。

当Adapter、FFEM和SG-LoRA三者共同使用时,模型取得最佳性能。这表明DDF-ViT的性能提

升主要来自空间域前景响应增强,同时频域结构补充和瓶颈适配器的特征细化能够进一步提升特征表达的完整性。总体来看,三类组件并非独立发挥作用,而是在空间目标增强、频域细节补充和多源特征融合之间形成互补,从而提升模型在浮游动物显微图像实例分割任务中的表现。

表5 DDF-ViT 内部子模块的有效性分析

Table 5 Effectiveness analysis of internal submodules in DDF-ViT

Unit: %					
Adapter	FFEM	SG-LoRA	mAP _{mask}	AP _{mask} ⁵⁰	AP _{mask} ⁷⁵
-	-	√	75.3	94.6	81.3
-	√	-	71.5	93.9	78.3
√	-	-	70.6	91.8	76.1
√	√	-	75.1	94.3	81.2
√	√	√	76.5	95.0	83.7

注:加粗数值表示该列最优值;“√”表示使用该模块;“-”表示未使用该模块

2.4.2 不同DDF-ViT插入策略的影响

为研究DDF-ViT的插入位置、间隔和数量对模型性能的影响,本文在保持其他训练配置一致的前提下设置五组对比实验,结果如表6所示。其中,[s : 32 : k]表示从ViT-H的索引s开始,以k为间隔设置DDF-ViT。以下分别从实验结果和编码器结构两个方面进行分析。

1)实验结果分析。首先,对比覆盖范围相同而插入间隔不同的[0 : 32 : 2]和[0 : 32 : 4]可以发现,当DDF-ViT数量由8个增加至16个时,AP_{mask}⁷⁵提高1.0个百分点,说明过于稀疏的插入方式难以在编码过程中形成持续的双域特征交互。然而,性能并未随DDF-ViT数量单调提升。包含16个DDF-ViT的[0 : 32 : 2]配置,其mAP_{mask}仍比仅包含12个DDF-ViT的[8 : 32 : 2]低1.6个百分点,说明除模块数量外,插入位置也是影响性能的重要因素。

进一步比较不同起始位置可以发现,[8 : 32 : 2]取得了最优结果,mAP_{mask}达到77.4%。当起始位置提前至索引0时,mAP_{mask}下降至75.8%;当起始位置后移至索引16和24时,mAP_{mask}分别下降至76.0%和75.4%。此外,[0 : 32 : 4]与[16 : 32 : 2]

均包含8个DDF-ViT,但后者的mAP_{mask}和AP_{mask}⁵⁰分别提高0.7和1.0个百分点,进一步说明在模块数量相同的情况下,不同插入位置仍会产生不同的分割效果。

表6 不同DDF-ViT占比对ZooMHS-SAM性能的影响

Table 6 Effectiveness of different DDF-ViT proportions in the performance of ZooMHS-SAM

Unit: %			
特征层范围	mAP _{mask}	AP _{mask} ⁵⁰	AP _{mask} ⁷⁵
[0 : 32 : 2]	75.8	94.5	83.2
[0 : 32 : 4]	75.3	93.8	82.2
[8 : 32 : 2]	77.4	96.3	85.1
[16 : 32 : 2]	76.0	94.8	82.8
[24 : 32 : 2]	75.4	94.0	82.2

注:加粗数值表示该列最优值

2)编码器结构分析。SAM的ViT-H包含32个Transformer块,并在索引7、15、23和31处设置全局注意力层,其余层主要采用窗口注意力。索引8位于第一次全局注意力之后,使DDF-ViT能够在初步全局信息交互的基础上参与后续特征编码。在第一次全局注意力完成之前,编码器主要形成边缘、纹理、灰度变化和局部结构等基础视觉特征。虽然这些信息是识别浮游动物轮廓与细长附肢的重要基础,但对于透明或半透明浮游动物而言,目标主体、弱边界与显微背景可能产生相近的局部响应。在缺少充分上下文的情况下,模型难以可靠判断这些响应来源于前景结构、真实边界还是背景干扰。因此,过早引入DDF-ViT可能同时增强目标相关信息与相似的背景响应。经过索引7处的第一次全局注意力后,局部特征通过跨窗口交互获得了初步的上下文信息,为区分前景主体、目标边界与背景干扰提供了更充分的依据。此时,SG-LoRA与FFEM以并行方式分别开展空间域和频域适配。其中,SG-LoRA在保留预训练权重的基础上,利用初步形成的上下文信息进行低秩增量调整,强化浮游动物前景主体响应并抑制无关背景干扰,从而提高模型对目标区域的空间辨别能力;FFEM则对中间特征进行频域建模,自适应筛选与目标形态、结构细节和弱边界相关的有效频率成分,以补充空间分支对透明边界与细

粒度结构的表达。两个并行分支分别侧重前景主体的空间表征和结构边界的频率表征,并通过双域特征融合形成互补,从而增强模型对浮游动物整体形态与复杂边界的建模能力。

此外,[8 : 32 : 2]采用间隔式设置,使双域适配后的特征能够在相邻原始 ViT 块中继续进行上下文传播与特征整合,再进入下一 DDF-ViT 完成后续调整,从而形成领域适配与预训练特征编码交替进行的过程。当起始位置后移至索引 16 或 24 时,DDF-ViT 参与后续编码的深度和特征传递范围相应缩短。因此,[8 : 32 : 2]从第一次全局信息交互后开始进行双域增强,并持续覆盖后续编码阶段,既保留了初始局部特征的提取过程,又为双域特征的交互与传递提供了较为充分的编码深度。

2.4.3 典型困难场景下的模块有效性分析

为进一步验证模型各组成部分在典型困难场景中的具体作用,本文分别针对弱边界和密集实例场景开展定量分析。其中,通过构建低局部对比度弱边界子集,分析 DDF-ViT 对模糊边界的增强能力;同时,根据图像中的真实实例数量划分不同实例密度

子集,分析 TAPM 对实例辨别和掩码预测的影响。

为分析 DDF-ViT 对弱边界目标的增强作用,本文按照目标边界邻域的局部对比度对图像进行排序,并选取局部对比度最低的 25% 样本构成弱边界子集。在完整验证集和弱边界子集上,分别采用不同边界带宽计算 Boundary IoU,结果如表 7 所示。其中,D 表示图像对角线长度,边界带宽分别设置为 0.15%D、0.30%D、0.50%D 和 0.60%D。在完整验证集上,引入 DDF-ViT 后,模型在四种边界带宽下的 Boundary IoU (Cheng 等, 2021) 分别由 19.9%、36.6%、48.3% 和 53.5% 提高至 24.1%、44.1%、57.0% 和 62.4%,对应提升 4.2、7.5、8.7 和 8.9 个百分点。在弱边界子集上,DDF-ViT 带来的提升更加明显,Boundary IoU 分别提高 7.6、13.5、16.5 和 17.3 个百分点。与完整验证集相比,DDF-ViT 在弱边界子集上的增益更大,表明双域特征增强能够利用频域结构信息与空间域语义信息的互补性,改善低对比度目标的边界响应,从而提升模型对弱边界浮游动物的轮廓刻画能力。

表 7 DDF-ViT 在完整验证集和弱边界子集上的边界分割性能
Table 7 Boundary segmentation performance of DDF-ViT on the full validation set and weak-boundary subset

		Unit: %			
模型设置	Evaluation set	Boundary IoU			
		0.15%D	0.30%D	0.50%D	0.60%D
Baseline (MSSConv)	Full validation set	19.9	36.6	48.3	53.5
Baseline (MSSConv) + DDF-ViT	Full validation set	24.1	44.1	57.0	62.4
Baseline (MSSConv)	Weak-boundary subset	13.1	25.2	35.6	40.7
Baseline (MSSConv) + DDF-ViT	Weak-boundary subset	20.7	38.7	52.1	58.0

注:在相同数据集下,加粗数值表示该列最优值;D 表示图像对角线长度

为分析 TAPM 在不同实例密度场景中的作用,本文根据每幅图像包含的真实实例数量,将验证集划分为包含 1-2 个实例的低密度子集和包含不少于 3 个实例的较高密度子集,并分别对 Baseline 和加入 TAPM 后的模型进行评估,结果如表 8 所示。

在 1-2 个实例的子集上,引入 TAPM 后,mAP_{mask}、AP⁵⁰_{mask} 和 AP⁷⁵_{mask} 分别由 72.0%、93.5% 和 81.1% 提高至 74.9%、94.5% 和 83.7%,对应提升 2.9、1.0 和 2.6 个百分点。这说明即使在实例数量较少的场景

中,TAPM 仍能利用 ROI 特征对全局图像嵌入进行目标相关调制,从而强化当前实例在全局特征中的响应。

在不少于 3 个实例的子集上,Baseline 的 mAP_{mask} 和 AP⁷⁵_{mask} 分别为 68.1% 和 71.8%,低于低密度子集,说明实例数量增加会加大相邻目标区分和精细边界预测的难度。加入 TAPM 后,mAP_{mask}、AP⁵⁰_{mask} 和 AP⁷⁵_{mask} 分别提高至 70.1%、92.9% 和 74.3%,对应提升 2.0、0.4 和 2.5 个百分点。其中,AP⁷⁵_{mask} 获得较为明显的

提升,表明 TAPM 提供的实例级先验有助于模型在多实例场景中区分当前目标与邻近实例,并改善较严格重叠标准下的掩码预测质量。综合两类子集的结果可知,TAPM 不仅适用于密集实例场景,也能在低密度场景中发挥目标特征强化作用。

表 8 TAPM 在不同实例密度场景下的分割性能
Table 8 Segmentation performance of TAPM under different instance densities

/%

Unit: %

模型设置	1-2个实例			大于等于3个实例		
	mAP _{mask}	AP ⁵⁰ _{mask}	AP ⁷⁵ _{mask}	mAP _{mask}	AP ⁵⁰ _{mask}	AP ⁷⁵ _{mask}
Baseline (MSSConv)	72.0	93.5	81.1	68.1	92.5	71.8
Baseline (MSSConv) + TAPM	74.9	94.5	83.7	70.1	92.9	74.3

注:加粗数值表示该列最优值

2. 4. 4 ZooMII5K 数据集的长尾分布分析

ZooMII5K 的类别分布呈现明显的长尾特征。为规范评价类别不均衡对实例分割性能的影响,本文参照长尾实例分割研究中的大词汇量实例分割 (Large Vocabulary Instance Segmentation, LVIS) 和长尾 COCO (COCO Long-Tailed, COCO-LT) 频次分组方式。LVIS 根据类别在训练集中出现的图像数量划

分 Rare、Common 和 Frequent 组 (Gupta 等, 2019); COCO-LT 根据训练实例数量划分 G1~G4 组,并采用 AP_{mask} 评价各组的实例分割性能 (Wang 等, 2020)。考虑到 ZooMII5K 仅包含 48 个类别,本文进一步根据训练集频次分布,将训练实例数不少于 150、40~149 和少于 40 的类别分别划分为头部类、中部类和尾部类。不同频次分组方式下的结果如表 9 所示。

表 9 不同长尾频次分组方式下 ZooMII5K 的 AP_{mask}
Table 9 AP_{mask} of ZooMII5K under different long-tailed frequency-grouping protocols

评估协议	分组	划分依据	类别数	AP _{mask}
LVIS	Rare	1 ~ 10 张训练图像	2	93.8
	Common	11 ~ 100 张训练图像	24	74.1
	Frequent	>100 张训练图像	22	79.5
COCO-LT	G1	1 ~ 19 个训练实例	6	79.1
	G2	20 ~ 399 个训练实例	37	76.1
	G3	400 ~ 7999 个训练实例	5	85.1
	G4	≥8000 个训练实例	0	—
自定义	尾部类	<40 个训练实例	12	71.6
	中部类	40 ~ 149 个训练实例	18	80.1
	头部类	≥150 个训练实例	18	78.6

注:AP_{mask} 表示对应频次组内所有类别掩码 AP 的算术平均值;“—”表示该频次组不包含符合条件的类别;

由表 9 可知,在 LVIS 和 COCO-LT 频次分组方式下,低频类别组的 AP_{mask} 相对较高。其主要原因是按照上述协议划分后,LVIS 的 Rare 组和 COCO-LT 的 G1 组分别仅包含 2 个和 6 个类别,组内平均结果容易受到少数类别分割难度的影响。例如,LVIS 的 Rare 组包含双手外肋水母 (Ectopleura bimanatus),

该类目标主体形态较稳定、外部轮廓相对连续,且目标边缘与暗色背景之间具有较明显的亮度差异,因此即使训练样本较少,也能取得较高的分割精度,并在一定程度上抬高本组的平均 AP_{mask}。而在与 ZooMII5K 训练集分布相适应的分组下,头部类、中部类和尾部类分别包含 18、18 和 12 个类别,其 AP_{mask}

分别为78.6%、80.1%和71.6%，尾部类别的整体分割性能相对较低，这表明类别频次可能会对尾部类别的整体表现产生一定影响。

图5给出了三个代表性类别。多毛类幼体早期(Polychaeta larva early)属于中部类,但其大量细长且向外辐射的刚毛容易发生漏分割或边界不完整;太平洋纺锤水蚤(Acartia pacifica)属于头部类,但其透明身体、细长触角及多处附肢增加了边界提取难度,二者的分割性能均相对较低。相反,尾部类双手外肋水母(Ectopleura bimanatus)虽然训练样本较少,但其主体形态较稳定、轮廓相对连续,且目标边缘与暗色背景之间具有较明显的亮度差异,因而取得了较高的分割性能。进一步对训练实例数量与单类 AP_{mask} 进行 Spearman 秩相关分析(Spearman, 1904),得到 $\rho = 0.100$ 、 $p = 0.499$,表明二者未呈现显著的单调相关关系。该结果说明类别频次并非决定分割性能的唯一因素,透明度、边界连续性、细长附肢及形态复杂度等视觉特性也可能对类别级性能产生重要影响。

本文通过随机翻转、尺度变换、随机裁剪及SAM预训练特征迁移缓解有限样本造成的过拟合。需要指出的是,部分尾部类别的实例较少,也与其在实际采样过程中的出现频率较低有关,这在一定程度上反映了真实海洋生态监测中的物种分布特征。在实际监测中,低频物种同样具有重要的识别与记录价值,因此模型不仅需要准确处理常见物种,还应具备识别少样本、低频物种的能力。后续将通过持续采集稀有类别样本、类别均衡采样及针对性数据增强等方式,进一步提高模型对尾部类别的分割性能与评价稳定性。

图5 浮游动物显微图像
Figure 5 Microscopic images of zooplankton



2.4.5 SG-LoRA 秩分析

为确定 SG-LoRA 中秩设置,本文在仅保留 SG-LoRA 模块的条件下,对不同秩值 $r \in \{1, 2, 4, 8, 16,$

32}进行了对比实验,结果如表10所示。

表10 SG-LoRA 秩值对模型分割性能的影响
Table 10 Effect of SG-LoRA rank on model segmentation performance

Unit: %			
秩(r)	mAP_{mask}	AP_{mask}^{50}	AP_{mask}^{75}
1	74.3	93.5	79.1
2	74.3	93.6	79.4
4	74.7	94.3	79.1
8	75.3	94.6	81.3
16	74.6	93.7	78.2
32	74.9	93.8	81.4

注:加粗数值表示该列最优值

实验结果表明,过低的秩值对权重更新空间限制较强,难以充分适应浮游动物目标在形态和边界特征方面的变化。随着秩值增大,模型在 $r=8$ 时取得最佳综合性能,相较于 $r=4$, AP_{mask}^{75} 提高2.2个百分点,说明适当提高秩值能够为 SG-LoRA 提供较充分的领域特征调整空间,从而改善模型的预测质量。当秩值进一步增大至16和32时,模型性能未继续提升,说明 SG-LoRA 的适配效果并不随秩值增加而单调增强。秩值过低会限制领域特征调整能力,而过高的秩值则可能引入冗余的更新自由度,影响预训练知识与任务特定适配之间的平衡。

2.4.6 RPN 候选框召回率对TAPM性能的影响

推理阶段,Mask RoIAlign根据检测分支输出的最终检测框提取ROI特征并输入TAPM,因此,能够进入TAPM的有效实例受到上游RPN候选区域覆盖能力的影响。为定量分析这一依赖关系,本文在保持模型参数和其他推理配置不变的条件下,固定RPN的非极大值抑制(non-maximum suppression, NMS)前候选框数量为1000,并将NMS后每幅图像保留的候选框数量 N_r 分别设置为10、50、100、300和1000。对于每种设置,统计RPN候选区域的平均召回率(Proposal_AR)、最终检测框的平均召回率(Box_AR)及实例分割性能。

由表11可知,当 N_r 从10增加至50时,Proposal_AR由81.5%提高至83.5%,Box_AR由83.5%提高至84.9%, mAP_{mask} 由70.2%提高至71.0%。这

说明增加 RPN 候选框数量能够提高对真实目标的覆盖程度,使更多有效实例进入后续检测和 TAPM 分支,从而改善最终分割性能。

当 N_R 继续增加至 100、300 和 1000 时, Proposal_AR 分别达到 83.7%、83.9% 和 84.0%, mAP_{mask} 分别达到 71.1%、71.2% 和 71.3%, 增幅逐渐减小。与此同时, Box_AR 在 $N_R=100$ 时达到 85.5%, 此后基本保持稳定, 说明当候选框数量达到一定规模后, RPN 对真实目标的覆盖能力已趋于饱和, 继续增加候选框数量对最终检测和分割性能的改善有限。最终 Box_AR 略高于 Proposal_AR, 主要是因为 Box_head 进一步修正了 RPN 候选框的位置, 提高了部分候选框在较高 IoU 阈值下的定位精度。

综上, RPN 候选框召回率会影响能够进入 TAPM 的有效实例范围, 较低的候选框召回率将限制后续实例分割性能; 当 $N_R \geq 100$ 后, 各项指标逐渐稳定, 表明包含 TAPM 的模型在候选框召回率达到较高水平后具有一定稳定性。

表 11 不同 RPN 候选框数量对 TAPM 分割性能的影响
Table 11 Effect of Different Numbers of RPN Proposals on TAPM Segmentation Performance

Unit: %					
RPN 候选框数 N_R	Proposal_AR@ N_R	Box_AR@100	mAP_{mask}	AP_{mask}^{50}	AP_{mask}^{75}
10	81.5	83.5	70.2	90.8	76.4
50	83.5	84.9	71.0	92.0	77.1
100	83.7	85.5	71.1	92.2	77.2
300	83.9	85.4	71.2	92.4	77.3
1000	84.0	85.4	71.3	92.5	77.4

注: 加粗数值表示该列最优值; Proposal_AR@ N_R 表示在 IoU 阈值 0.50 ~ 0.95, 采用每幅图像前 N_R 个 RPN 候选框计算的平均召回率; Box_AR@100 则基于经过分类、边界框回归和非极大值抑制后的最终检测框计算

2.4.7 不同损失权重的影响

本文模型的训练目标由 RPN 阶段损失和 RoI 阶段损失共同构成。RPN 阶段负责生成候选区域, 其输出为后续实例分类、边界框回归和掩码预测提供位置基础; RoI 阶段则基于候选区域完成实例分类、边界框回归和掩码预测。因此, 两个阶段并非彼此

独立, 而是构成由候选区域生成到实例掩码预测的连续过程。为避免训练初期对任一阶段引入人为偏置, 本文默认对两个阶段采用相同的损失权重。

考虑到当前计算资源与实验规模的限制, 本文未对各损失项进行大规模组合搜索, 而是采用固定权重设置, 并以 RPN 阶段与 RoI 阶段损失的相对权重为切入点开展实验。为排除总体损失权重变化带来的干扰, 实验始终保持 $\lambda_1 + \lambda_2 = 2$, 仅调整两者的相对比例。令 $r = \lambda_2 / \lambda_1$, 其中, λ_1 和 λ_2 分别表示 RPN 阶段损失与 RoI 阶段损失的权重, 并设置 $r \in \{2, 1, 0.5\}$ 。实验结果如表 12 所示。

表 12 不同损失权重比例对模型分割性能的影响
Table 12 Effect of different loss-weight ratios on segmentation performance

Unit: %			
$r = \lambda_2 / \lambda_1$	mAP_{mask}	AP_{mask}^{50}	AP_{mask}^{75}
2	75.9	94.7	82.0
1	77.4	96.3	85.1
0.5	76.1	95.0	82.9

注: 加粗数值表示该列最优值

实验结果表明, 当 $r=2$ 时, RoI 阶段权重提高, 但分割精度并未随之提升。这是因为 RPN 阶段权重降低后, 候选区域的定位质量与目标覆盖能力可能受到影响, 而后续 RoI 预测难以完全弥补候选区域质量的不足, 导致 mAP_{mask} 和 AP_{mask}^{75} 分别下降至 75.9% 和 82.0%。当 $r=0.5$ 时, RPN 阶段权重提高, 但 RoI 阶段监督相对减弱。此时, 即使候选区域生成得到强化, 后续分类、边界框回归和掩码细化仍可能受到限制, 导致 mAP_{mask} 和 AP_{mask}^{75} 分别下降至 76.1% 和 82.9%。这说明仅侧重任一阶段均难以获得最优分割性能。总体而言, 在基于候选提示的 SAM 实例分割框架中, 高质量候选区域生成与精细掩码预测之间具有较强的相互依赖关系, 模型的整体性能取决于两个阶段的协同优化。

2.4.8 解码器冻结策略分析

考虑到 SAM 解码器主要在自然图像上完成预训练, 其直接应用于浮游动物显微图像时可能受到域偏移的影响, 本文进一步比较了冻结与解冻解码器两种设置。两组实验保持其他网络结构、训练配置及评价设置一致。解冻设置允许解码器参数参与

反向传播,并将其学习率设置为主学习率的0.1倍。

如表13所示,解冻解码器后, mAP_{mask} 和 AP_{mask}^{75} 分别提升0.2和0.4个百分点。这表明,解冻解码器仅带来了小幅且不一致的指标变化,未表现出全面的性能优势。这说明解码器微调可能对较严格交并比条件下的掩膜预测产生一定的正向作用,但其整体增益相对有限。

产生这一结果的可能原因与SAM解码器的功能和结构有关。相较于ViT-H图像编码器,SAM解码器的结构更加轻量,主要通过双向Transformer实现图像嵌入与提示嵌入之间的信息交互,并利用掩膜令牌生成分割掩膜及其质量预测结果。该解码器不直接处理原始显微图像,其预测结果在较大程度上取决于上游编码特征与提示信息的质量。本文利用DDF-ViT对编码特征进行领域适配,并利用MSFPM和TAPM增强提示特征与全局特征,使输入解码器的信息已经包含一定的显微图像领域特征。因此,自然图像与显微图像之间的部分表征差异能够由上游模块进行缓解,进一步微调解码器所带来的边际收益相对有限。

需要指出的是,本实验仅对预训练解码器进行参数微调,并未针对显微图像中的弱边界和细粒度形态改变其结构。因此,有限的微调增益并不意味着掩膜解码阶段不存在进一步优化空间。未来可在保留SAM预训练掩膜生成能力的基础上,引入边界感知特征交互、多尺度掩膜细化或领域适配解码结构,进一步增强模型对目标轮廓与局部形态细节的预测能力。

表13 SAM掩膜解码器冻结策略对分割性能的影响
Table 13 Effect of the SAM Mask Decoder Freezing Strategy on Segmentation Performance

Unit: %			
解码器状态	mAP_{mask}	AP_{mask}^{50}	AP_{mask}^{75}
冻结	77.4	96.3	85.1
解冻	77.6	96.0	85.5

注:加粗数值表示该列最优值

2.4.9 ZooMIIIS-SAM中不同结构的参数量

为充分利用SAM预训练模型的通用视觉表征与分割先验,并减少领域适配过程中需要更新的预训练参数,本文冻结SAM的预训练参数,仅训练新

增的SG-LoRA、FFEM、Bottleneck Adapter、MSSConv和TAPM等任务相关模块。其中,冻结范围包括SAM原始图像编码器参数、共享位置嵌入和解码器参数。SG-LoRA和瓶颈Adapter分别利用低秩映射与瓶颈结构进行参数适配;FFEM、MSSConv和TAPM则分别用于频域结构增强、多尺度特征融合与目标实例感知。ZooMIIIS-SAM各新增模块的参数构成如表14所示。SG-LoRA、FFEM、瓶颈Adapter和MSSConv在选定的12个编码层中配置,其单层参数量分别为0.1M、12.1M、0.8M和3.4M;TAPM在整体网络中部署一次,参数量为33.8M。

从参数构成来看,SG-LoRA和瓶颈Adapter的累计参数量分别约占模型总参数量的0.13%和1.08%,表明低秩映射与瓶颈结构能够以较少的附加参数完成特征适配。FFEM、MSSConv和TAPM的参数占比分别约为16.29%、4.58%和3.79%,其参数开销主要用于频域结构增强、多尺度特征融合和目标实例感知。其中,FFEM的参数量相对较大,主要源于对频域实部与虚部拼接特征进行通道映射,以增强显微图像中弱边界与结构细节的表达。总体而言,ZooMIIIS-SAM冻结SAM的预训练参数,仅训练新增的任务相关模块,从而在保留预训练视觉先验的基础上完成浮游动物显微图像领域适配。

尽管FFEM改善了高IoU阈值下的分割质量,但也是新增参数的主要来源。后续将研究分组卷积、低秩通道映射或幅值域调制,以降低其参数与部署开销。

2.4.10 推理效率与部署潜力分析

为评估ZooMIIIS-SAM在实际浮游动物监测中的

表14 ZooMIIIS-SAM中不同结构的参数量
Table 14 Number of parameters for each component of ZooMIIIS-SAM

结构	SG-LoRA	FFEM	Bottleneck Adapter	MSSConv	TAPM
单个参数量	0.1M	12.1M	0.8M	3.4M	33.8M
单个参数量占比	0.01%	1.36%	0.09%	0.38%	3.79%
累计参数量占比	0.13%	16.29%	1.08%	4.58%	3.79%

注:单个参数量占比表示单个模块参数量占模型总参数量的比例;累计参数量占比表示同类模块在网络中全部配置后的参数量之和占模型总参数量的比例

应用潜力,本文进一步比较了不同实例分割模型的推理延迟、每秒处理图像数(frames per second,FPS)和峰值显存占用。所有模型均在相同硬件环境下进行测试,输入尺寸统一为 512 x 512, Batch Size 设置为 2,并在预热后统计平均推理效率,结果如表 15 所示。

由表 15 可知,Mask2Former 具有较高的推理速度,其单图推理延迟为 33.8 ms,推理速度达到 29.6 FPS,但峰值显存占用为 9.73 GB, mAP_{mask} 为 71.9%。USIS-SAM 的峰值显存下降至 4.42 GB,但单图推理延迟达到 156.3 ms,推理速度为 6.4 FPS。相比之下,ZooMIIS-SAM 在取得最高 mAP_{mask} 的同时,单图推理延迟为 83.4 ms,推理速度达到 12.0 FPS,峰值显存占用仅为 4.04 GB。与 USIS-SAM 相比,

ZooMIIS-SAM 的 mAP_{mask} 提高 4.9 个百分点,推理延迟降低约 46.6%,推理速度提高 87.5%,峰值显存占用降低约 8.6%,表明本文方法在分割精度和推理效率之间取得了较好的平衡。

浮游动物样本通常在采集和显微成像后进行实验室离线分析,并不要求模型达到高帧率实时视频处理水平。在此应用模式下,ZooMIIS-SAM 能够以 12.0 FPS 处理 512 x 512 的显微图像,且峰值显存占用为 4.04 GB,可在实验室 GPU 工作站上完成批量样本的自动检测、物种识别和实例分割。因此,现有模型已具备用于浮游动物实际样本检测与辅助分析的能力,可为后续在海洋生态监测机构开展自动化样本分析提供技术支持。

表 15 不同实例分割模型的推理效率对比

Table 15 Comparison of inference efficiency among different instance segmentation models

模型	输入尺寸	Batch Size	mAP_{mask}	推理延迟(ms/ image)	推理速度 (FPS)	峰值显存 (GB)
Mask2Former	512 x 512	2	71.9	33.8	29.6	9.73
USIS-SAM	512 x 512	2	72.5	156.3	6.4	4.42
ZooMIIS-SAM	512 x 512	2	77.4	83.4	12.0	4.04

注:加粗数值表示该列最优值

需要指出的是,ZooMIIS-SAM 的总参数量仍然较大,在边缘设备或资源受限平台上的部署能力仍有提升空间。后续将采用知识蒸馏,将模型学习到的领域特征和实例分割能力迁移至轻量化学生网络;同时针对 FFEM 等开销较大的模块研究结构化剪枝、低秩分解和通道压缩,并结合混合精度推理、算子融合及推理引擎优化,进一步降低模型参数量、推理延迟和显存占用,在保持分割性能的同时提升轻量化部署能力。

2.4.11 TAPM 轻量化分析

为分析 TAPM 中完整动态矩阵设计的必要性,本文在相同 Baseline 和训练配置下,将 TAPM 分别替换为通道级调制和低秩矩阵调制,结果如表 16 所示。

通道级调制仅增加 0.3 M 参数,其 mAP_{mask} 提高 0.3 个百分点,但 AP_{mask}^{75} 未发生变化,说明独立调整各通道响应难以充分建模实例特征之间的关联。低秩矩阵通过受限子空间建立跨通道交互,使 mAP_{mask}

提高至 70.7%,但受秩约束影响,其对复杂实例特征关系的表达仍然有限。采用完整 TAPM, mAP_{mask} 和 AP_{mask}^{75} 分别达到 71.3% 和 77.4%。与低秩矩阵相比,虽然其新增参数量增加 16.9 M,但 AP_{mask}^{75} 进一步提高 1.7 个百分点,表明完整跨通道调制更有利于建模目标主体、细长附肢与弱边界之间的实例相关关系,从而改善严格重叠条件下的掩码质量。低秩矩阵在参数效率方面具有优势,而完整动态矩阵在严格重叠条件下方面取得了更好的性能。

2.5 泛化性实验

为验证 ZooMIIS-SAM 的跨场景泛化能力,本文在公开的 USIS10K 数据集上进行了对比实验,实验结果如表 17 所示。结果表明,ZooMIIS-SAM 在所有对比方法中取得了最佳性能。这表明本文所采用的双域特征增强和实例感知机制不仅适用于浮游动物显微图像,在自然水下场景中也表现出一定的跨场景适应能力。

与 ZooMIIS5K 上的实验结果相比,ZooMIIS-SAM

表 16 TAPM 与不同轻量化调制方式的对比

Table 16 Comparison of TAPM with different lightweight modulation methods

调制方式	动态权重形式	新增参数量/M	mAP_{mask}	AP_{mask}^{50}	AP_{mask}^{75}
Baseline (MSSConv)	—	0	68.6	90.6	74.5
Baseline (MSSConv) + 通道级调制	$w_{ins} \in \mathbf{R}^C$	0.3	68.9	91.4	74.5
Baseline (MSSConv) + 低秩矩阵	$\mathbf{W}_{ins} = \mathbf{U}_{ins} \mathbf{V}_{ins}^T, \mathbf{U}_{ins}, \mathbf{V}_{ins} \in \mathbf{R}^{C \times 64}$	16.9	70.7	92.8	75.7
Baseline (MSSConv) + TAPM	$\mathbf{W}_{ins} \in \mathbf{R}^{C \times C}$	33.8	71.3	92.5	77.4

注:加粗数值表示该列最优值;“—”表示未采用实例调制

在 USIS10K 上的分割性能有所下降,可能的原因是两个数据集在成像环境和目标特征方面存在较大差异。ZooMIIS5K 主要由实验室显微成像条件下采集的浮游动物图像构成,而 USIS10K 包含多种自然水下场景,图像更容易受到水体散射、颜色偏移等因素影响。同时,USIS10K 中的目标在尺度、形态和空间分布方面具有更大的变化,增加了特征提取与实例分割的难度。

表 17 不同模型在 USIS10K 数据集上的泛化性能对比
Table 17 Generalization performance comparison of different models on the USIS10K dataset

Unit: %			
模型	mAP_{mask}	AP_{mask}^{50}	AP_{mask}^{75}
Mask R-CNN	39.2	57.8	44.7
YOLACT	38.2	57.1	42.1
SOLO	30.4	50.9	31.8
Mask2Former	37.8	52.4	41.2
WaterMask	38.7	54.9	43.2
RSPrompter	40.2	55.3	44.8
USIS-SAM	43.9	59.6	50.0
ZooMIIS-SAM	49.6	65.8	56.1

注:加粗数值表示该列最优值

2.6 可视化实验

1)编码器特征可视化。为验证本文所提出编码器的特征提取能力,本文分别在自然图像和浮游动物显微图像上采用主成分分析(principal component analysis, PCA)(Shlens 等, 2014)进行特征可视化,并将其与原始 SAM 编码器进行对比。结果如图 6 所示,在自然图像中,本文编码器仍能够保留较为清晰的目标结构,说明所引入的特征增强模块并未破坏

SAM 原有的通用表征能力。然而,在浮游动物显微图像中,受悬浮颗粒、碎片以及低对比度背景的影响,原始 SAM 编码器提取的特征容易出现前景与背景混杂的现象,尤其对于边界较弱、半透明的浮游动物目标,其目标响应更加不稳定。相比之下, ZooMIIS-SAM 能够更有效地抑制背景干扰,并增强目标实例响应,表明本文编码器具有更强的显微图像特征表达能力。

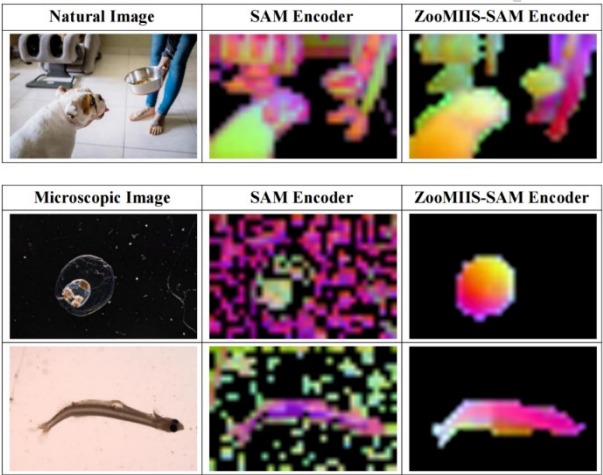


图 6 编码器特征可视化分析
Figure 6 Encoder feature visualization analysis

2)TAPM 模块的掩码响应热力图可视化。为进一步直观分析 TAPM 的作用,本文对 Baseline 模型与引入 TAPM 后模型的掩码响应热力图进行可视化,结果如图 7 所示。相比 Baseline,引入 TAPM 后模型的响应更加集中于浮游动物目标实例,而背景区域中的无关响应明显减少。该结果表明,TAPM 能够利用候选实例信息增强目标实例感知能力,还能抑制背景干扰,从而提升模型在复杂显微场景下的实例分割效果。

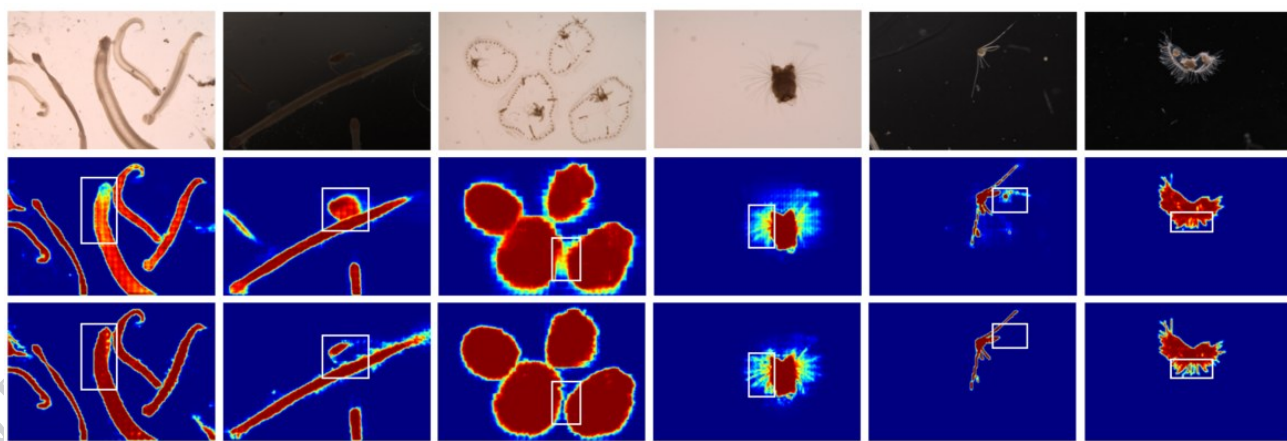
3 结 论

本文面向浮游动物显微图像自动实例分割任务,针对目标边界弱、背景干扰强、密集实例辨别困难以及像素级标注数据不足等问题,构建了包含48类浮游动物的显微图像实例分割数据集 ZooMIIS5K,并提出基于SAM的自动实例分割框架 ZooMIIS-SAM。该框架通过 DDF-ViT 对 SAM 编码器进行空间域与频域联合适配,增强目标结构及边界特征;通过 MSFPM 融合多层、多尺度特征并自动生成实例相关的软提示,减少对人工提示和密集网格提示的依赖;通过 TAPM 将候选实例信息引入全局图像特征,提升模型对目标实例的辨别能力。

实验结果表明,ZooMIIS-SAM 在 ZooMIIS5K 数据集上取得了 77.4% mAP_{mask} 、96.3% AP_{mask}^{50} 和 85.1% AP_{mask}^{75} ,整体性能优于对比方法。消融实验及典型困难场景分析验证了 DDF-ViT 和 TAPM 的有效性,表明空间域与频域特征增强有助于改善弱边界表征,实例感知机制能够提升不同实例密度场景下的分割性能。MSFPM 则利用检测分支定位的实例

区域自动生成软提示,替代 SAM 对人工交互式提示的依赖,实现端到端自动实例分割。可视化结果也证明其能够更稳定地聚焦于浮游动物目标实例,并有效抑制背景干扰。USIS10K 上的泛化实验也表明,该方法具备一定的跨场景适应能力。此外,ZooMIIS-SAM 在 512 x 512 输入下达到 12.0 FPS,峰值显存占用为 4.04 GB,能够满足实验室环境下浮游动物样本的离线自动分析需求。

未来将进一步采集低频及稀有物种样本,并结合类别均衡采样和针对性数据增强,缓解长尾分布对少样本类别识别与分割稳定性的影响;研究边界感知交互、多尺度掩码细化及解码器领域适配方法,提高透明目标、细长附肢和复杂轮廓的分割质量。同时,将针对 FFEM 等参数开销较大的结构开展低秩分解、通道压缩、模型剪枝和知识蒸馏研究,并结合混合精度与推理引擎优化进一步提升部署效率。在此基础上,将推动模型在海洋生态监测机构中的实际应用,并进一步探索开放集识别和未知物种发现,以增强模型在真实、动态海洋监测场景中的实用性与适应能力。



第二行为 Baseline 模型的掩码响应热力图,第三行为在 Baseline 基础上引入 TAPM 后的掩码响应热力图。白色框标出了典型响应差异区域。

图7 TAPM 模块的掩码响应热力图可视化

Figure 7 Visualization of mask response heatmaps with TAPM

参考文献(References)

Ratnarajah L, Abu-Alhaija R, Atkinson A, Batten S, Bax N J, Bernard K S, et al. Monitoring and modelling marine zooplankton in a changing climate[J]. Nature Communications, 2023, 14(1): 564.

[DOI: 10.1038/s41467-023-36241-5]

Chiba S, Batten S, Martin C S, Ivory S, Miloslavich P, Weatherdon L V. Zooplankton monitoring to contribute towards addressing global biodiversity conservation challenges [J]. Journal of plankton research, 2018, 40(5): 509-518.[DOI: 10.1093/plankt/fby030]

Yuan D, Li Z, Wang L, Qi Y, Qiao Z, Zhang J. ZS2Net: Frequency-aware semantic segmentation for zooplankton microscopic image

© 中国图象图形学报版权所有

- [J]. Expert Systems with Applications, 2026: 132072. [DOI: 10.1016/j.eswa.2026.132072]
- Bergum S, Saad A, Stahl A. Automatic in-situ instance and semantic segmentation of planktonic organisms using Mask R-CNN, " Global Oceans 2020: Singapore - U.S. Gulf Coast, Biloxi, MS, USA, 2020, pp. 1-8. [DOI: 10.1109/IEEECONF38699.2020.9389377].
- Sosa-Trejo D, Bandera A, González M, Hernández-León S. Vision-based techniques for automatic marine plankton classification [J]. Artificial Intelligence Review, 2023, 56 (11) : 12853-12884. [DOI: 10.1007/s10462-023-10456-w]
- Maracani A, Pastore V P, Natale L, Rosasco L, Odone F. In-domain versus out-of-domain transfer learning in plankton image classification [J]. Scientific Reports, 2023, 13(1) : 10443. [DOI: 10.1038/s41598-023-37627-7]
- He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn [C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969. [DOI: 10.48550/arXiv.1703.06870]
- Bolya D, Zhou C, Xiao F, Lee Y J. Yolact: Real-time instance segmentation [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9157-9166. [DOI: 10.1109/ICCV.2019.00925]
- Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment anything [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 4015-4026. [DOI: 10.48550/arXiv.2304.02643]
- Li Z W, Guo H N, Qi Y P, Yuan D K. 2026. Zooplankton classification based on multiscale feature fusion and attention guidance. Journal of Image and Graphics, 31(5):1583-1594. (李忠伟, 郭浩宁, 齐衍萍, 袁德坤. 2026. 多尺度特征融合与注意力引导的浮游动物分类. 中国图象图形学报, 31(5):1583-1594). [DOI: 10.11834/jig.250352]
- Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images [J]. Nature communications, 2024, 15(1) : 654. [DOI: 10.1038/s41467-024-44824-z]
- Wu J, Wang Z, Hong M, Ji W, Fu H, Xu Y, et al. Medical sam adapter: Adapting segment anything model for medical image segmentation [J]. Medical image analysis, 2025, 102: 103547. [DOI: 10.1016/j.media.2025.103547]
- Chen K, Liu C, Chen H, Zhang H, Li W, Zou Z, et al. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model [J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-17. [DOI: 10.1109/TGRS.2024.3356074]
- Lian S, Zhang Z, Li H, Li W, Yang L T, Kwong S, et al. Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset [J]. arXiv preprint arXiv:2406.06039, 2024. [DOI: 10.48550/arXiv.2406.06039]
- Bureš J, Eerola T, Lensu L, Kälviäinen H, Zemčík P. Plankton recognition in images with varying size [C]//International Conference on Pattern Recognition. Cham: Springer International Publishing, 2021: 110-120. [DOI: 10.1007/978-3-030-68780-9_11]
- Yuan C, He Z, Ning C, Wang W, Zhao J, Yuan G, et al. Research on in situ observation method of plankton based on convolutional neural network [J]. Journal of Marine Science and Engineering, 2024, 12(10) : 1702. [DOI: 10.3390/jmse12101702]
- Overbeek M. Declumping the Clumps: Instance Segmentation of Multiple Plankton in ZooScan Images [D]. 2025.
- Chen T, Zhu L, Deng C, Cao R, Wang Y, Zhang S, et al. Sam-adapter: Adapting segment anything in underperformed scenes [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 3367-3375. [DOI: 10.1109/ICCVW60793.2023.00361]
- Zhang K, Liu D. Customized segment anything model for medical image segmentation [J]. arXiv preprint arXiv:2304.13785, 2023. [DOI: 10.48550/arXiv.2304.13785]
- LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11) : 2278-2324. [DOI: 10.1109/5.726791]
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv:2010.11929, 2020. [DOI: 10.48550/arXiv.2010.11929]
- Wang X, Kong T, Shen C, Jiang Y, Li L. Solo: Segmenting objects by locations [C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 649-665. [DOI: 10.1007/978-3-030-58523-5_38]
- Wang X, Zhang R, Kong T, Li L, Shen C. SOLOv2: Dynamic and Fast Instance Segmentation [J]. arXiv preprint arXiv:2003.10152, 2020. [DOI: 10.48550/arXiv.2003.10152]
- Cheng B, Misra I, Schwing A G, Kirillov A, Girdhar R. Masked-attention mask transformer for universal image segmentation [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 1290-1299. [DOI: 10.1109/CVPR52688.2022.00135]
- Lian S, Li H, Cong R, Li S, Zhang W, Kwong S. Watermask: Instance segmentation for underwater imagery [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 1305-1315. [DOI: 10.1109/ICCV51070.2023.00126]
- Shlens J. A tutorial on principal component analysis [J]. arXiv preprint arXiv:1404.1100, 2014. [DOI: 10.48550/arXiv.1404.1100]
- Chen K, Wang J, Pang J, Cao Y, Xiong Y, Li X, et al. MMDetection: Open mmlab detection toolbox and benchmark [J]. arXiv preprint arXiv:1906.07155, 2019. [DOI: 10.48550/arXiv.1906.07155]
- Cheng B, Girshick R, Dollár P, Berg A C, Kirillov A. Boundary IoU: Improving object-centric image segmentation evaluation [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 15329-15337. [DOI: 10.48550/arXiv.2103.16562]

- Tang S, Ran H, Yang S, Wang Z, Li W, Li H, et al. A frequency selection network for medical image segmentation [J]. Heliyon, 2024, 10(16): e35698. [DOI: 10.1016/j.heliyon.2024.e35698]
- Gupta A, Dollár P, Girshick R. LVIS: A dataset for large vocabulary instance segmentation [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 5351-5360. [DOI: 10.1109/CVPR.2019.00550]
- Wang T, Li Y, Kang B, Li J, Liew J H, Tang S, et al. The devil is in classification: A simple framework for long-tail instance segmentation [C]//Computer Vision - ECCV 2020. Cham: Springer, 2020: 728-744. [DOI: 10.1007/978-3-030-58568-6_43]
- Spearman C. The proof and measurement of association between two things[J]. The American Journal of Psychology, 1904, 15(1): 72-101. [DOI: 10.2307/1412159]

作者简介

鲍君龙,男,硕士研究生,主要研究方向为计算机视觉及应用。E-mail:bao001126@163.com

乔钺,男,博士研究生,主要研究方向为显微图像处理、多模态融合与鲁棒控制。E-mail:qz00519@163.com

许向阳,女,硕士研究生,主要研究方向为计算机视觉及应用。E-mail:xyx030203@163.com

齐衍萍,女,正高级工程师。主要研究方向为海洋生物多样性调查、海洋生态灾害监测等。E-mail:qiyanning@ncs.mnr.gov.cn

刘昕,女,副教授。主要研究方向为数据挖掘、区块链等。E-mail:lx@upc.edu.cn