DGD-SAM：一种用于水下图像实例分割的动态引导SAM

尚毅涵，董兴辉*

（中国海洋大学海洋动力—物理环境与智能感知全国重点实验室，中国海洋大学信息科学与工程学部，山东青岛 266100）

摘　要： 随着深海探测与海洋资源开发需求的日益增长，水下视觉技术已成为机器人作业、海洋生物监测等领域的关键支撑。在众多的视觉任务中，水下图像实例分割因需同时实现目标的精确定位与像素级掩码预测而具有极高的挑战性。近年来，视觉基础模型，特别是Segment Anything Model（SAM），在通用场景下展现出卓越的零样本泛化能力，但在复杂的水下环境中，其表现仍不尽如人意。水下环境光线吸收、散射严重，导致图像伴随明显的色彩失真、对比度极低以及边缘模糊等退化现象，严重干扰了模型的特征提取。此外，SAM的分割性能高度依赖人工提供的显式提示信息（例如点、框和掩码），这种依赖不仅增加了人工成本，更限制了其在无人值守或复杂水下环境中的适用性。为了解决上述问题，本文提出了一种动态引导SAM（Dynamically Guided SAM，DGD-SAM）。DGD-SAM通过引入动态引导机制，结合特征聚合与多尺度增强模块，构建了完整的自动提示生成与精细化分割流程。首先，针对检测与分割任务特征分布不一致的问题，本文设计了自适应特征聚合模块。该模块通过引入通道注意力机制对特征依赖关系进行重新建模，在空间与通道维度上实现任务对齐，有效增强了模型对水下弱目标区域的感知灵敏度。其次，考虑到水下目标尺寸多变且背景干扰复杂的特性，构建了多尺度特征增强模块。该模块通过构建跨空间分辨率的特征金字塔，显著提升了模型在复杂场景下对各种尺度目标的捕捉能力。最后，在解码阶段，本文提出了动态引导解码器，先融合初始分割掩码与图像特征以生成动态引导信息，再通过提示与图像特征间的双向注意力交互实现精细掩码预测。实验结果显示，DGD-SAM在四个公开水下数据集LIACI、USIS10K、UIIS和UIIS10K以及两个陆地场景数据集COME15K-E和COME15K-H上均优于当前的先进方法，这表明本文方法不仅在水下场景中表现出色，在陆地场景中同样能够获得稳定且具有竞争力的分割性能，说明模型未过度依赖特定场景特征，具备良好的泛化能力和可扩展性。

关键词： SAM；视觉基础模型；图像分割；水下图像实例分割；动态引导解码器；提示生成

基金项目： 国家自然科学基金（No.42576200）；山东省重点研发计划（No.2024ZLGX06）

中图分类号： TP391　　**文献标识码：** AA　　**文章编号：** 0372-2112(XXXX)XX-0001-13

电子学报URL:http://www.ejournal.org.cn　　**DOI:**10.12263/DZXB.C251002.R1

DGD-SAM: A Dynamically-Guided SAM for Underwater Image Instance Segmentation

SHANG Yihan, DONG Xinghui*

(*State Key Laboratory of Physical Oceanography and the Faculty of Information Science and Engineering*, *Ocean University of China*, *Qingdao*, *Shandong* 266100, *China*)

Abstract: With the growing demand for deep-sea exploration and marine resource exploitation, underwater vision technologies have become a critical enabler for applications, such as robotic operations and marine biological monitoring. Among various vision tasks, Underwater Image Instance Segmentation (UIIS) is particularly challenging, as it requires both precise object localization and pixel-level mask generation. In recent years, vision foundation models, in particular, the Segment Anything Model (SAM), have demonstrated remarkable zero-shot generalization capabilities in generic scenes. However, their performance remains unsatisfactory in complex underwater environments. Severe light absorption and scattering in underwater environments lead to significant image degradation, including color distortion, extremely low contrast, and blurred boundaries, which substantially hinder effective feature extraction. Moreover, the segmentation performance of SAM heavily relies on manually provided explicit prompts (e.g., points, boxes, and masks). This dependency not only increases annotation costs but also limits its applicability in unattended or complex underwater scenarios. To address these challenges, we propose a Dynamically-Guided SAM (DGD-SAM). By introducing a dynamically-guided mechanism and integrating feature aggregation with a multi-scale feature enhancement module, DGD-SAM establishes a complete pipeline

收稿日期：2025-12-18；录用日期：2026-01-16；责任编辑：宋甲英

*通信作者：董兴辉

for automatic prompt generation and refined segmentation. First, to mitigate the feature distribution discrepancy between detection and segmentation tasks, an Adaptive Feature Aggregator (AFA) is designed. This module re-models inter-channel dependencies through a channel attention mechanism, achieving task alignment across both spatial and channel dimensions and effectively enhancing the model's sensitivity to weak underwater targets. Second, considering the large variation in underwater target scales and the complexity of background interference, a multi-scale feature enhancement module is constructed. By building a cross-resolution feature pyramid, this module significantly improves the model's ability to capture targets of various scales in complex scenes. During the decoding stage, a Dynamically-Guided Decoder (DGD) is proposed, which first integrates the initial segmentation mask with image features to generate dynamic guidance information, and then performs refined mask prediction through bidirectional attention interactions between the prompts and image features. Experimental results demonstrate that DGD-SAM consistently outperforms state-of-the-art methods on four public underwater data sets, including LIACI, USIS10K, UIIS, and UIIS10K, as well as two terrestrial scene data sets, i.e., COME15K-E and COME15K-H. These results indicate that the proposed method not only achieves superior performance in underwater environments but also maintains stable and competitive segmentation performance in terrestrial scenes, suggesting that the model does not overly rely on scene-specific characteristics and exhibits strong generalizability and scalability.

Keywords: segment anything model; vision foundation model; underwater image instance segmentation; image segmentation; dynamically-guided decoder; prompt generation

Foundation Item(s): National Natural Science Foundation of China (NSFC) (No.42576200); Key Research and Development Program of Shandong Province, China (No.2024ZLGX06)

0 引言

作为一项基本的计算机视觉任务,实例分割旨在准确识别和描绘图像中的每个独立物体。该技术已广泛应用于多个领域,例如自动驾驶[1]、医学图像分析[2]和遥感检测[3-4]。然而,水下图像与陆地图像不同,因为它们通常具有颜色偏移、模糊、低对比度和视觉畸变等特点。这些退化严重损害了图像质量,并对水下视觉任务(如水下目标检测、分割和识别)构成了挑战[5]。由于现有的实例分割方法通常是为处理在陆地环境所获取的图像而开发的,将它们直接应用于水下场景通常无法产生令人满意的结果。为解决水下图像实例分割问题,研究者们已开展了一系列探索。WaterMask[6]提出了一个面向水下实例分割的大规模数据集UIIS,并设计了专门的特征增强网络结构,包括差异相似图注意模块与多级特征细化模块,用于弥补水下图像细节损失和边界模糊的问题。该方法显著提升了传统实例分割模型在水下图像数据上的性能。随后,USIS-SAM[7]将视觉基础模型Segment Anything Model(SAM)[8]引入水下显著实例分割任务,提出显著特征提示生成器以自动生成提示,从而减少对人工输入的依赖。该方法同时发布了大规模数据集USIS10K,进一步推动了基于提示驱动的水下分割研究。

近年来,随着视觉基础模型的兴起,大规模预训练与可提示分割框架为下游任务带来了新的机遇。SAM[8]基于超大规模数据集SA-1B[8]进行预训练,在多个视觉任务中展现出卓越的泛化性与零样本分割能力。凭借提示驱动的设计理念,SAM能够根据用户输入的点、框或掩码等提示,在未知场景中快速生成分割结果,已被广泛应用于多种下游任务[9-10]。然而,SAM的性能高度依赖于提示信息的质量与准确性[11]。这种依赖性不仅增加了人工干预的成本,也限制了模型在无人干预和自动化分割场景中的应用潜力[12]。在水下任务中,这一问题尤为突出,由于水下影像退化严重,用户难以提供精确提示,导致SAM在该场景下的分割效果显著下降。

为了应对上述挑战,本文提出了一种动态引导SAM(Dynamically Guided SAM, DGD-SAM)。该框架首次在SAM中引入动态引导机制,使模型能够在无需人工提示的情况下自主生成提示信息并实现精细化分割。为了适应水下环境,DGD-SAM使用了LoRA[13]和一个相对较小的数据集进行微调,仅需少量样本即可完成模型迁移。本文设计了一个动态引导解码器,利用图像编码器的多尺度特征与掩码解码器的区域响应进行信息交互。前者提供全局上下文与目标线索,后者强化边界与局部细节表达。通过这种跨层协同机制,模型能够动态生成引导信息以驱动分割过程,实现对不同尺度目标的自适应感知与精确定位。整个网络可在无需人工干预的情况下实现端到端的训练。

据作者所知,现有研究尚未从动态引导机制的角度对SAM进行系统性的探索。本文的主要贡献可以概括为以下三点:

(1)提出DGD-SAM框架。本文首次将动态引导机制引入SAM,用于水下图像实例分割任务。得益于SAM的强大泛化能力,所提出的网络仅需少量样

本即可通过微调来适应复杂水下场景。DGD-SAM摆脱了对用户输入提示的依赖,实现了自动化的分割过程。如图1所示,该方法在六个公开数据集上均显著优于当前最先进的基线模型。

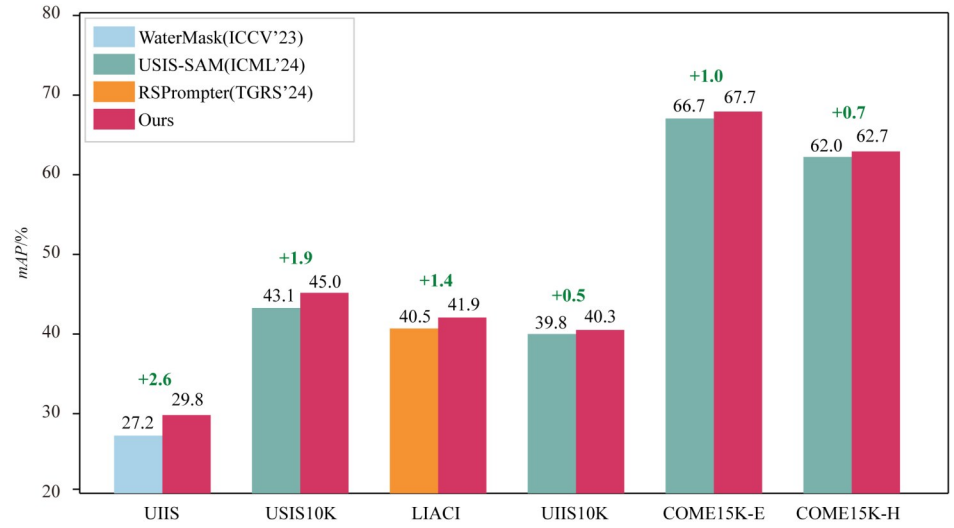


图1 本文所提出的DGD-SAM在六个公开数据集上与三种先进的实例分割方法的比较

Figure 1 Comparison of the proposed DGD-SAM with three state-of-the-art instance segmentation methods on six public datasets

(2)提出动态引导解码器结构。通过融合初始掩码与图像特征来生成动态提示信息,并利用多层注意力机制实现提示与图像特征的有效融合,显著提升了模型在弱特征与模糊边界条件下的鲁棒性与精度。

(3)在六个具有代表性的公开实例分割数据集上进行了全面实验验证与比较分析,为后续相关研究提供了统一的性能基准与可复现的评测参考。

1 相关工作

1.1 实例分割

实例分割要求模型不仅要区分不同的物体类别,还要区分同一类别内的不同实例[14]。早期的方法通常建立在两阶段框架之上,例如Mask R-CNN[15]、Cascade Mask R-CNN[16]和HTC[17]。这类方法通常遵循"候选区域生成—掩码预测"的两步范式,借助区域提议网络(RPN)[18]先定位目标,再通过精细的掩码分支预测每个实例的像素级区域。虽然这类方法在精度上表现优异,但其推理速度较慢、计算开销大,限制了其在实时或资源受限场景下的应用。为了提高效率和性能,单阶段实例分割方法被提出,例如YOLACT[19]、BlendMask[20]、CondInst[21]和SOLOv2[22]。YOLACT首次将掩码生成过程与检测解耦,通过学习一组全局原型掩码,并为每个实例预测组合系数,实现了高效推理。CondInst进一步利用动态卷积核为每个实例生成特定的掩码参数,从而避免了显式的RoI操作。SOLO系列方法则采用位置敏感的全卷积方式,将实例分割任务转化为像素级分类问题,大幅提高了并行性和推理速度。这些单阶段方法在保持较高精度的同时显著提升了效率。最近,Transformer的引入推动了实例分割进入新的阶段。MaskFormer[23]、Mask2Former[24]等基于Transformer的架构利用全局注意力机制统一了语义分割、实例分割和全景分割任务。此类方法通过掩码查询机制实现了端到端的分割预测,不再依赖显式的区域提议或锚点设计。随后的一系列改进[25-26]进一步优化了查询匹配策略与特征融合方式。然而,现有方法通常严重依赖于训练数据的分布,并且在应用于复杂或特定领域的图像(如水下场景、医学图像和遥感图像)时,往往会出现性能下降。这个问题阻碍了它们在实际应用中的部署,与此同时,基础视觉模型如SAM的出现,为通用实例分割提供了新的思路。它们通过大规模预训练展现出强大的跨领域能力,引发了人们对基础视觉模型日益增长的兴趣。

1.2 Segment Anything Model

SAM作为通用分割基础模型,旨在实现可提示的任意图像分割。作为计算机视觉领域的重要里程碑技术,SAM在大规模数据预训练的支持下,展现出了卓越的泛化能力和零样本分割性能。其核心设计理念是通过提示驱动机实现与类别无关的掩码生成,使用户能够通过点、框、文本等形式的提示来高效地获得目标分割结果。SAM的模型结构由三个主要组件构成:图像编码器、提示编码器和掩码解码器。图像编码器基于ViT(Vision Transformer)[27]结构,用于提

取全局视觉特征;提示编码器负责对输入提示(例如点或框)进行嵌入表示;掩码解码器则结合图像与提示特征,通过轻量级的注意力机制生成对应的分割掩码。这种模块化设计使SAM能够在多种交互式分割场景中灵活应用。得益于在超过10亿掩码(SA-1B数据集)上进行的预训练,SAM拥有极强的通用性和可迁移性。大量研究者将SAM应用于不同领域的图像分割任务,包括医学影像[28]、遥感场景[3]、水下图像[7]等。然而,尽管SAM展现出了卓越的泛化能力,其在特定领域中的应用仍面临若干挑战。首先,SAM的预训练数据主要来源于陆地上的自然场景,与水下等特定领域图像之间存在显著的分布差异,导致模型在这些领域的表现显著下降。其次,SAM严重依赖显式提示来实现目标定位,缺乏自动化分割能力,这限制了其在无人干预或大规模批处理任务中的使用。

1.3 提示学习

近年来,随着基础模型的兴起,一种新的学习模式,即预训练与提示逐渐成为主流[29]。与以往依赖任务特定微调的传统方法不同,这一模式通过在输入端设计合适的提示,将下游任务重新表述为与预训练阶段目标相一致的形式,从而实现任务迁移与知识复用。该思路最早来源于自然语言处理领域的大型语言模型(LLM),这一理念随后被扩展到多模态领域,其中对比语言-图像预训练(CLIP)[30]是具有代表性的工作。CLIP通过在大规模图文配对数据上进行对比学习,使视觉特征与语言语义在共享的嵌入空间中对齐,实现了图像与文本之间的跨模态理解。借助文本提示,CLIP能够在无需额外训练的情况下执行图像分类、检索等多种下游任务,开创了视觉领域提示学习的先河。

受CLIP启发,Kirillov等人[8]首次将提示机制引入视觉分割任务,提出了SAM。SAM将提示概念从文本扩展至视觉域,使得用户可以通过点击、框选或输入文本等方式对模型进行显式引导,从而实现对任意物体的掩码分割。SAM在自然场景中展现出了极强的零样本泛化能力,并成为构建下游任务的基础组件。然而,尽管SAM在自然图像分割中表现卓越,其图像生成过程仍高度依赖人工交互。这种依赖限制了其在自动化或无人干预场景中的应用。此外,手动提示难以覆盖复杂场景中的全部目标对象,也导致了模型性能的受限。

2 本文方法

为减轻SAM对人工提示的依赖,并解决外部提示在水下环境中易失效的问题,本文提出了一种动态引导SAM(Dynamically Guided SAM, DGD-SAM)。得益于SAM卓越的泛化能力,该网络可适用于水下实例分割任务。DGD-SAM的架构如图2所示,整个网络主要由图像编码器(Image Encoder)、自适应特征聚合模块(Adaptive Feature Aggregator, AFA)、多尺度特征增强模块(Multi-Scale Feature Enhancement Module, MFEM),以及动态引导解码器(Dynamically Guided Decoder, DGD)组成。该结构旨在在保留SAM全局特征提取能力的同时,增强对局部细节与多尺度信息的感知,从而在复杂场景下获得更加精确的目标分割结果。

2.1 图像编码器

图像编码器采用了基于Transformer的结构,与SAM的主干网络保持一致,并加载SAM预训练的ViT-B权重,用于从输入图像中提取全局语义特征。为了在保持模型主干网络权重冻结的情况下高效地适配下游任务,在图像编码器中引入了LoRA[13](Low-Rank Adaptation)微调机制。与传统的全参数微调不同,LoRA通过在自注意力模块的线性投影层中插入低秩可学习矩阵,实现了参数高效化的模型适配方式。具体而言,LoRA Adapter仅插入于图像编码器中各全局Transformer Block的自注意力模块内,对其中的$Q/K/V$线性投影层进行低秩分解。设原始$Q/K/V$投影层的权重矩阵为$\boldsymbol{W}$,引入一个低秩分解形式的参数更新项$\Delta\boldsymbol{W}$,即

$$\boldsymbol{W}'=\boldsymbol{W}+\Delta\boldsymbol{W},\ \Delta\boldsymbol{W}=\frac{\alpha}{r}\boldsymbol{BA} \tag{1}$$

其中,$\boldsymbol{A}\in\mathbb{R}^{r\times d_{\text{proj}}}$,$B\in\mathbb{R}^{d_{\text{in}}\times r}$,且$r=16\ll\min(d_{\text{in}},d_{\text{proj}})$。我们将LoRA的缩放系数$\alpha$设置为32,并在LoRA分支中施加dropout(概率为0.05),以增强模型的泛化能力并缓解过拟合。假设线性投影层的输入序列为X,输出序列为Y,LoRA在冻结原始权重$\boldsymbol{W}$的基础上引入低秩更新项,将线性变换表示为

$$Y=X\left(\boldsymbol{W}+\frac{\alpha}{r}\boldsymbol{BA}\right) \tag{2}$$

这样在训练时仅优化低秩矩阵$\boldsymbol{A}$和$\boldsymbol{B}$,其余参数保持冻结,从而大幅降低显存与计算开销。

2.2 自适应特征聚合模块

SAM预训练的分割主干网络在大规模图像数据上学习到较强的语义表征能力,其特征更多地关注于区域内部一致性与边界细节的精确分割。然而,在实例检测与目标定位任务中,RPN(Region Proposal Network)对目标的形态、位置及上下文变化更为敏感,二者在特征分布与关注重点上存在一定的不匹配。若直接将SAM的特征用于RPN,会导致特征表达偏向分割而非检测,从而影响候选区域生成与后续掩码预

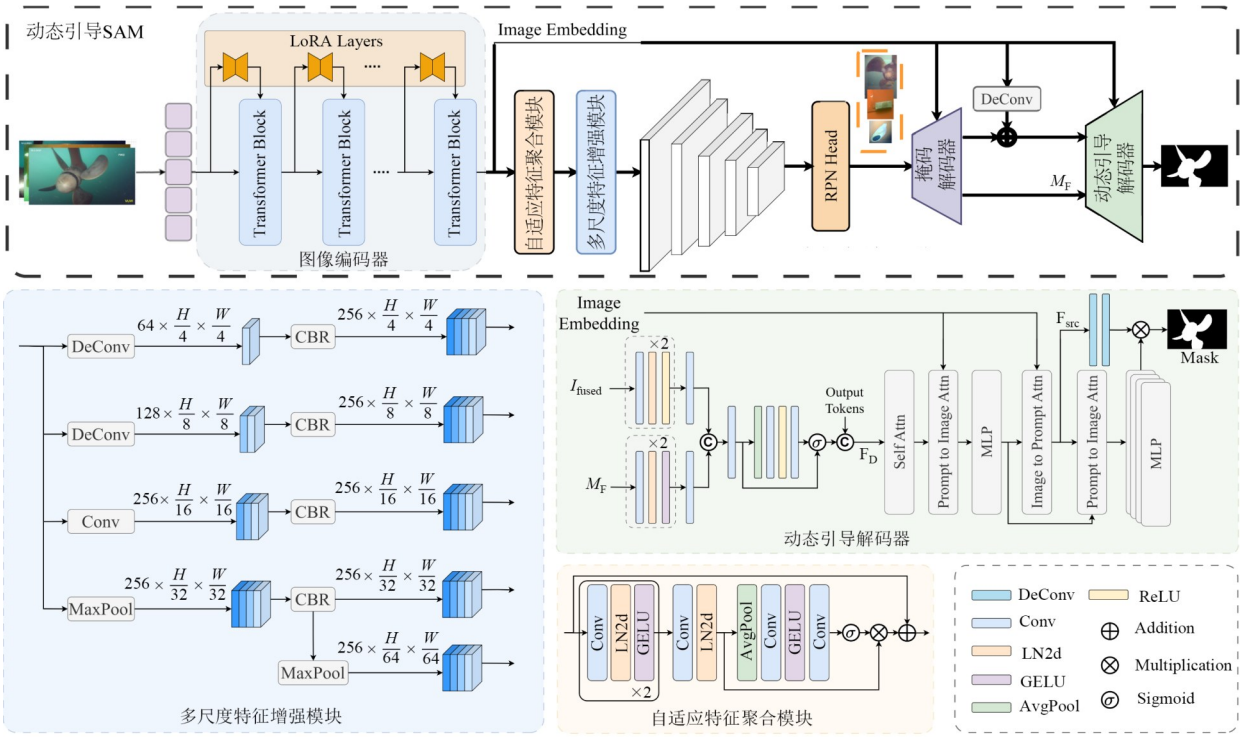


图2 动态引导SAM的模型架构

Figure 2 Architecture of the proposed Dynamically-Guided SAM

测的准确性。

为此,本文设计了自适应特征聚合模块(Adaptive Feature Aggregator,AFA),作为一个轻量级可插拔的任务对齐器,在不改变空间分辨率的前提下,对来自SAM主干网络的单尺度特征进行通道层面的再加权与语义对齐。AFA通过对特征通道间的依赖关系进行建模,使网络能够根据任务需求动态调整特征响应,从而提升特征对目标区域的敏感度与判别性。具体而言,AFA先由两层卷积模块(Conv-LN-GELU)对输入特征F_{in}先进行初步变换与语义重投影,再通过第三个卷积模块映射到输出通道,得到F_{out}。为实现任务自适应性,AFA在卷积后引入通道注意力机制,通过全局平均池化对通道统计量进行建模,并利用逐通道权重调整特征响应强度。该通道注意力可表示为

$$W_{\text{c}}=\sigma(\text{MLP}(\text{AvgPool}(F_{\text{in}}))) \tag{3}$$

其中,W_{c}为通道权重,σ表示Sigmoid激活函数。为了保留原始特征的全局语义信息并稳定训练,AFA模块在卷积与通道注意力处理后引入了残差连接

$$F_{\text{AFA}}=F_{\text{out}}\odot W_{\text{c}}+F_{\text{in}} \tag{4}$$

其中,$\odot$表示逐元素乘法。该模块输出的特征不仅保留了原始特征的全局语义一致性,同时在通道维度上对任务相关信息进行了显式强化,为后续的多尺度特征增强模块(MFEM)提供更具区分性的输入。

2.3 多尺度特征增强模块

为了应对复杂场景中目标尺度变化大的问题,本文设计了多尺度特征增强模块(MFEM),用于在不同空间分辨率下提取互补的语义与细节信息,从而提升网络对多尺度目标的鲁棒性与定位精度。该模块以AFA输出的特征为输入,通过上采样与下采样路径构建多尺度特征集,使得网络能够在统一的语义空间中融合不同感受野的信息。具体而言,MFEM首先对输入特征$F_{\text{AFA}}\in\mathbb{R}^{C\times H\times W}$进行多尺度变换,生成四个层级的特征表示$F_i\in\mathbb{R}^{C\times\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}},i=1,2,3,4$。每个尺度的特征均通过一个卷积模块进行通道对齐与特征增强,该卷积模块由卷积层、批归一化层和ReLU激活函数组成,从而得到统一维度的多尺度输出$O_i=\text{CBR}(F_i),i=1,2,3,4$。为了进一步扩大特征金字塔的覆盖范围,MFEM通过步长为2的最大池化继续向下生成更低分辨率的层级特征$O_5=\text{MaxPool}_{2\times2}(O_4)$,最终输出的多尺度特征集合可表示为

$$F_{\text{MSEM}}=\{O_1,O_2,O_3,O_4,O_5\} \tag{5}$$

2.4 动态引导解码器

在分割阶段,本文提出的动态引导解码器旨在实现从候选框提示到最终精细分割的动态语义引导。该模块通过结合RPN的定位能力与SAM的强大分割先验,在分割过程中逐步优化目标区域的边界与语义

一致性。训练过程中,动态引导解码器权重均采用随机初始化。

2.4.1 初始提示引导阶段

首先,RPN Head根据多尺度特征生成候选框,这些初始候选框被视作提示,输入至SAM的掩码解码器以获得初始分割结果。在这一阶段,模型利用RPN的目标定位能力,将全局特征约束在候选区域内,从而实现对潜在目标的粗粒度分割。输出的初始掩码不仅提供了目标的空间先验,也为后续的特征融合与精细分割提供了引导。

2.4.2 动态特征引导阶段

在获得初始分割掩码后,我们进一步将该掩码M_F与融合了初始提示的特征嵌入I_{fused}先经过通道对齐与卷积融合,生成动态提示特征F_D,该特征不仅包含初始提示的位置信息,还融合了图像的上下文语义,为后续注意力模块提供了动态、可学习的引导信号。随后,动态引导解码器(DGD)进入与SAM掩码解码器类似的注意力交互阶段。该阶段由自注意力层(Self-Attn)、提示到图像的注意力层(Prompt-to-Image Attention)和图像到提示的注意力层(Image-to-Prompt Attention)三部分组成,用以实现提示特征与图像特征之间的多层交互融合。

在自注意力层中,F_D经过线性投影得到查询(Q)、键(K)和值(V):

$$Q,K,V=F_DW_Q,F_DW_K,F_DW_V \tag{6}$$

然后,利用标准多头注意力来计算融合特征内部的全局依赖:

$$\text{SelfAttn}(F_D)=\text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{7}$$

提示到图像的注意力层的作用是使提示特征从图像嵌入中吸收上下文信息,从而具备更强的语义表达能力。假设将经过自注意力后的提示特征记为T_{P},图像嵌入记为E_{I},则

$$Q,K,V=T_{\text{P}}W_Q^{\text{P}},E_{\text{I}}W_K^{\text{I}},E_{\text{I}}W_V^{\text{I}} \tag{8}$$

得到的注意力输出为

$$F_{\text{P}\to\text{I}}=\text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{9}$$

通过上述计算方式,提示信息在语义空间上融合了图像上下文,从而为后续的交互提供更精确的引导信号。随后执行反向的信息交互,将提示语义注入图像特征,即

$$Q,K,V=E_{\text{I}}W_Q^{\text{I}},F_{\text{P}\to\text{I}}W_K^{\text{P}},F_{\text{P}\to\text{I}}W_V^{\text{P}} \tag{10}$$

$$F_{\text{I}\to\text{P}}=\text{Softmax}(\frac{QK^T}{\sqrt{d}})V \tag{11}$$

该过程将提示语义反向注入图像特征中,促使图像嵌入在语义空间上与提示特征对齐,从而实现图像与提示的双向融合。经过多层注意力交互后,输出的特征嵌入$F_{\text{I}\to\text{P}}$经过反卷积来恢复空间分辨率:

$$F_{\text{src}}=\text{DeConv}(\text{DeConv}(F_{\text{I}\to\text{P}})) \tag{12}$$

提示特征$F_{\text{p}\to\text{I}}$再经过一层提示到图像的注意力层和多层感知机得到输出Token,并将输出Token与F_{src}进行逐元素乘法以生成最终分割结果。

2.5 损失函数

除了区域建议网络的损失函数(包括目标分类损失和边界框回归损失)之外,本文还引入了两个基于像素级交叉熵的损失函数,用于分别监督通过掩码解码器和动态引导解码器生成的分割掩码。整体损失函数可表示为

$$L=\lambda_1\cdot L_{\text{RPN}}+\lambda_2\cdot L_{\text{MD}}+\lambda_3\cdot L_{\text{DGD}} \tag{13}$$

其中,L_{MD}为SAM掩码解码器输出的分割掩码的监督项,通过像素级交叉熵损失约束基础分割结果。L_{DGD}为动态引导解码器(Dynamically Guided Decoder)对应的掩码损失,用于进一步细化目标边界与结构细节。在实验中,所有损失权重均设置为1(即$\lambda_1=\lambda_2=\lambda_3=1$)。该设置在实验中表现稳定,无需复杂的权重调节即可在六个公开数据集上获得最佳性能。

3 实验

本节旨在验证本文提出的DGD-SAM在水下图像实例分割任务中的有效性与优越性。接下来,将首先介绍实验设置,包括数据集、评价指标及实现细节。随后,通过与多种主流方法的对比实验、消融实验以及定性结果分析,全面评估模型在不同数据集与场景下的性能表现。

3.1 实验设置

为确保实验结果的公平性与可复现性,本节详细介绍了实验所使用的数据集、评价指标及实现细节。所有实验均在统一的训练策略与硬件环境下进行,模型训练与评测过程保持一致。

3.1.1 数据集

我们在四个公开可用的水下实例分割数据集上评估了所提出的方法,包括LIACI[31]、UIIS[6]、USIS10K[7]、UIIS10K[32]。为验证所提出方法在不同数据分布与应用场景下的可扩展性,本文进一步在地面场景数据集COME15K-E[33]与COME15K-H[33]上对所提方法进行了实验评估。对于每个数据集,我们始终遵循原始的划分方法,将数据集分为训练集、验证集以及测试集。对于具有实例级标注的数据集,我们直接使用它们作为标注数据。对于只包含语义标注的数据集,我们使用连通区域分析法将语义分割掩码

转换为实例级标注。

3.1.2 评估指标

为了评估所提出方法的性能,本文采用了广泛认可的COCO平均精度(*mAP*)度量[34]。该指标经常用于客观评价物体检测和实例分割方法的有效性。在本研究中,使用*mAP*、AP_{50}和AP_{75}行评估。*mAP*指的是在IoU阈值从0.5到0.95(步长为0.05)以及所有类别上的平均指标。*AP*值越大,表示预测的实例掩码越准确、实例分割性能越好。AP_{50}代表IoU阈值0.50下的计算,而AP_{75}则体现了更严格的指标,对应于IoU阈值0.75下的计算。因此AP_{75}值越高,表明实例掩码越准确。

3.1.3 实现细节

在本文的实验中,除非另有说明,均基于预训练的SAM模型[8],并使用ViT-B[27]作为主干网络。整个方法基于PyTorch与MMDetection[35]框架实现。在训练过程中,模型共训练50个epoch,采用AdamW优化器,初始学习率设为0.0002,权重衰减系数设为0.05。为实现高效且稳定的优化,本文设计了一种两阶段学习率调度策略。具体而言,在训练初期的前50个iteration中采用线性预热策略,将学习率从较小的初始值逐步提升至目标值;随后采用余弦退火策略,在整个训练过程中逐步衰减学习率,最终降至基础学习率的0.001倍。

此外,为提高训练效率并减少显存占用,本文还引入了自动混合精度(Automatic Mixed Precision, AMP)训练机制。所有实验均在批大小为4的设置下进行。除非特别说明,所有实验均在一张NVIDIA L40 GPU上完成。

3.2 实验结果

本文在四个公开可用的水下实例分割数据集与两个地面实例分割数据集上,对所提出的DGD-SAM模型进行了系统评估,并与当前主流方法进行了全面比较。对比方法包括两阶段实例分割框架、单阶段方法,以及近年来兴起的基于视觉基础模型SAM的方法。相关结果报告如下。

3.2.1 LIACI数据集

为了全面评估所提出的DGD-SAM在LIACI数据集[31]上的性能,本文将其与八种当前最先进的实例分割方法进行了对比,实验结果如表1所示。可以看出,本文提出的DGD-SAM在所有评估指标上均取得了最优结果。具体来说,DGD-SAM在*mAP*、AP_{50}和AP_{75}上分别较次优方法提升了1.4、0.2和4.9个百分点。这些结果表明,本文所提出的方法相比于多种架构的实例分割模型,均表现出优异的分割性能与泛化能力。

表1 本文的方法在LIACI数据集上与八种实例分割方法的比较

Table 1 Comparison of our method with eight instance segmentation methods on LIACI

方法	出处	骨干网络	评估指标		
			mAP	AP_{50}	AP_{75}
SOLOv2[22]	TPAMI21	ResNet-50	34.8	54.7	37.5
QueryInst[36]	CVPR21	ResNet-50	36.0	51.0	38.9
FastInst[37]	CVPR23	ResNet-50	38.6	55.0	41.0
HTC[17]	CVPR19	ResNet-50	36.4	55.2	39.2
BoxInst[38]	CVPR21	ResNet-50	27.7	50.0	26.5
USIS-SAM[7]	ICML24	ViT-Huge	27.8	44.6	28.8
Efficient-SAM[39]	CVPR24	ViT-Small	33.3	51.4	33.9
RSPrompter[3]	TGRS24	ViT-Base	<u>40.5</u>	<u>58.7</u>	<u>42.3</u>
DGD-SAM	本文	ViT-Base	**41.9**	**58.9**	**47.2**

注:表中以粗体和下划线分别标示最优和次优结果。后续表格中亦采用相同的标注方式。

3.2.2 USIS10K数据集

在USIS10K[7]数据集上,本文遵循原始基准设置,实验结果如表2所示。与对比方法相比,DGD-SAM在所有评价指标上均取得显著提升。具体而言,相比于次优的方法,DGD-SAM在*mAP*、AP_{50}和AP_{75}上分别提升了1.9、0.9和2.4。值得注意的是,尽管本文的方法仅采用了更轻量的ViT-Base主干网络,但其性能仍然优于使用ViT-Huge主干网络的对比方法,这体现了模型的高效性与竞争力。

表2 本文方法与10种实例分割方法在USIS10K上的比较

Table 2 Comparison with ten instance segmentation methods on the USIS10K dataset

方法	出处	骨干网络	评估指标		
			mAP	AP_{50}	AP_{75}
RDPNet[40]	TIP21	ResNet-50	37.9	55.3	42.7
WaterMask[6]	ICCV23	ResNet-50	37.7	54.0	42.5
OQTR[41]	TMM22	ResNet-50	19.7	30.6	21.9
S4Net[42]	CVPR19	ResNet-50	23.9	43.5	24.4
CondInst[21]	ECCV20	ResNet-101	38.9	55.8	43.1
Mask RCNN[15]	ICCV17	ResNet-101	39.6	58.7	45.0
SAM+BBox[8]	ICCV23	ViT-Huge	26.4	38.9	29.0
SAM+Mask[8]	ICCV23	ViT-Huge	38.5	56.3	44.0
RSPrompter[3]	TGRS24	ViT-Huge	40.2	55.3	44.8
USIS-SAM[7]	ICML24	ViT-Huge	<u>43.1</u>	<u>59.0</u>	<u>48.5</u>
DGD-SAM	本文	ViT-Base	**45.0**	**59.9**	**50.9**

3.2.3 UIIS和UIIS10K

此外,本文还在UIIS[6]和UIIS10K[7]两个水下实例分割数据集上进一步评估了DGD-SAM的泛化性能,并将其与已有的主流方法进行了比较。从表3可以看出,在UIIS10K数据集上,DGD-SAM在*mAP*、AP_{50}和AP_{75}指标上分别较次优方法提升了0.5、0.7和2.3。

而在UIIS数据集上，提升幅度分别达到2.6、3.5和3.4。

表3 UIIS和UIIS10K上的实例分割性能比较

Table 3 Instance segmentation performance on the UIIS and UIIS10K datasets

方法	UIIS			UIIS10K		
	mAP	AP_{50}	AP_{75}	*mAP*	AP_{50}	AP_{75}
WaterMask[6]	27.2	43.7	29.3	37.4	51.6	41.7
YOLACT[19]	18.5	36.2	17.8	35.0	50.9	37.9
PointRend[43]	25.9	43.4	27.6	37.3	53.0	41.3
RSPrompter[3]	25.1	40.3	26.2	33.2	44.9	36.0
USIS-SAM[7]	26.3	41.3	28.7	39.8	52.0	40.6
DGD-SAM	**29.8**	**47.2**	**32.7**	**40.3**	**53.7**	**44.0**

综上所述，本文的实验结果充分表明了DGD-SAM在不同类型的水下数据集和多样化场景中均表现出更强的分割能力与更高的鲁棒性，体现了其在水下实例分割任务中的广泛适应性与稳定性。

3.2.4 COME15K-E和COME15K-H

为验证所提出方法在不同数据分布与应用场景下的可扩展性，本文进一步在地面场景数据集COME15K-E[33]和COME15K-H[33]上对所提方法进行了实验评估，实验结果如表4所示，DGD-SAM在*mAP*、AP_{50}和AP_{75}指标上分别较次优方法提升了1.0、2.5和2.4。而在COME15K-H数据集上，提升幅度分别为0.7、2.2和1.6。实验结果表明，本文方法在地面场景中依然能够获得稳定且具有竞争力的分割性能，表明模型并未过度依赖特定场景特征，具备良好的可扩展性。

表4 COME15K-E和COME15K-H上的实例分割性能比较

Table 4 Instance segmentation performance on the COME15K-E and COME15K-H datasets

方法	COME15K-E			COME15K-H		
	mAP	AP_{50}	AP_{75}	*mAP*	AP_{50}	AP_{75}
Mask RCNN[15]	48.8	71.2	58.6	42.2	65.7	50.8
YOLACT[19]	48.1	70.7	56.2	41.4	66.0	48.5
RDPNet[40]	49.8	72.2	59.5	42.1	65.2	49.7
RSPrompter[3]	65.0	81.6	70.8	59.9	78.6	65.0
USIS-SAM[7]	66.7	81.7	70.6	62.0	78.7	65.5
DGD-SAM	**67.7**	**84.2**	**73.0**	**62.7**	**80.9**	**67.1**

3.2.5 计算复杂度分析

在计算复杂度方面，本文进一步将所提出的DGD-SAM与九种实例分割方法进行了比较，如表5所示。可以看出，基于ViT-Base[27]主干网络的DGD-SAM在保持适当推理速度的同时，具有中等数量的参数和FLOPs。尽管DGD-SAM的FLOPs值高于一些轻量级模型（例如FastInst和Efficient-SAM），但其分割性能更优。如表1所示，在LIACI数据集[31]上，本文所提出的DGD-SAM方法在*mAP*指标上分别比FastInst和Efficient-SAM高出3.3和8.6。与USIS-SAM相比，DGD-SAM将参数量和FLOPs减少了一半以上，同时将*mAP*从27.8提升到41.9。

这些结果表明，本文所提出的DGD-SAM在计算复杂度与分割精度之间实现了良好的平衡。

表5 DGD-SAM与九种实例分割方法的复杂度与推理速度比较

Table 5 Complexity and inference speed comparison of DGD-SAM and nine instance segmentation methods

方法	骨干网络	参数量(M)	计算量(G)	推理速度(FPS)
Mask R-CNN[15]	ResNet-50	44.4	253.0	13.8
BoxInst[38]	ResNet-50	35.1	372.0	22.0
CondInst[21]	ResNet-50	34.2	331.0	22.7
FastInst[37]	ResNet-50	**34.1**	58.2	28.9
SOLOv2[22]	ResNet-50	46.6	239.0	13.4
QueryInst[36]	ResNet-50	172.0	170.0	14.5
Efficient-SAM[39]	ViT-Small	52.0	**32.5**	**29.9**
RSPrompter[3]	ViT-Base	117.5	114.5	15.5
USIS-SAM[7]	ViT-Huge	696.8	824.4	5.6
DGD-SAM	ViT-Base	87.9	115.4	16.2

3.3 消融实验

为验证所设计的各模块的有效性，在主干网络与训练策略保持一致的条件下，本文对动态引导解码器（DGD）、自适应特征聚合模块（AFA）以及训练中的超参数进行了系统的消融实验。所有消融实验均在USIS10K数据集上进行。

表6展示了不同解码器结构的性能对比，包括原始SAM的掩码解码器（原始SAM）、仅保留初始提示分割阶段的模型（仅Stage1）以及本文提出的动态引导解码器（DGD）。可以观察到，无论使用哪种解码器结构，使用本文的动态引导解码器都会产生最佳结果。因此，动态引导解码器对于本文所提出的DGD-SAM是有用的。

表6 动态引导解码器消融实验

Table 6 Ablation study on the dynamically guided decoder

方法	评估指标		
	mAP	AP_{50}	AP_{75}
原始SAM	44.6	59.1	50.8
仅Stage1	43.8	58.6	49.4
DGD	**45.0**	**59.9**	**50.9**

自适应特征聚合模块相关实验结果如表7所示。可以观察到，在去除整个AFA模块（w/o AFA）后，模型整体性能明显下降，说明RPN与SAM主干网络输出特征之间确实存在分布差异，AFA在特征对齐中起到了关键作用。当仅去掉注意力机制（w/

AFA w/o Attn)时,性能虽略优于完全去除AFA的情况,但仍低于完整版本,表明AFA内部的通道注意力在抑制冗余语义、突出目标相关特征方面具有显著贡献。

表7 自适应特征聚合模块消融实验

Table 7 Ablation study on the adaptive feature aggregator

方法	评估指标		
	mAP	AP_{50}	AP_{75}
w/ AFA w/o Attn	44.5	59.0	50.7
w/o AFA	44.1	59.3	50.5
w/ AFA(完整版本)	**45.0**	**59.9**	**50.9**

为分析关键训练超参数对模型性能的影响,本文进一步开展了超参数消融实验,重点考察了学习率、权重衰减系数以及线性预热策略的作用。除被消融的超参数外,其余训练配置均采用本文的默认设置,以确保实验结果的可比性。具体而言,我们分别在不同学习率设置下对模型进行训练,以评估模型对学习率变化的敏感性。同时,我们测试了两种权重衰减系数设置,以分析正则化强度对模型收敛与泛化性能的影响。此外,为验证线性预热策略在训练初期的作用,我们移除了该策略并与默认设置进行了对比实验。实验结果如表8—10所示,消融分析结果表明,合理的学习率与权重衰减设置对于模型性能具有重要影响,而线性预热策略能够在一定程度上提升训练稳定性并改善最终性能。

表8 不同学习率设置下的消融实验结果

Table 8 Ablation results under different learning rates

学习率	评估指标		
	mAP	AP_{50}	AP_{75}
2×10^{-5}	39.5	54.4	44.2
5×10^{-5}	42.5	57.3	48.8
1×10^{-4}	44.7	59.5	50.7
2×10^{-4}**(默认)**	**45.0**	**59.9**	**50.9**

表9 不同权重衰减系数下的消融实验结果

Table 9 Ablation results under different weight decay settings

权重衰减系数	评估指标		
	mAP	AP_{50}	AP_{75}
5×10^{-1}	39.6	53.8	45.7
1×10^{-1}	44.0	58.6	49.8
5×10^{-2}**(默认)**	**45.0**	**59.9**	**50.9**

表10 线性预热策略对模型性能的影响分析

Table 10 Analysis of the effect of the linear warm-up strategy on model performance

线性预热策略	评估指标		
	mAP	AP_{50}	AP_{75}
无	44.2	59.2	50.0
有(默认)	**45.0**	**59.9**	**50.9**

4 定性结果分析

图3展示了本文方法与USIS-SAM、RSPrompter在UIIS数据集上的分割结果对比。从图中可以直观地看出,本文方法在多种典型水下场景下均取得了更为精确且边界清晰的分割效果,尤其在背景复杂、光照不佳等场景下表现更为稳定。此外,图4给出了本文方法在地面场景数据集COME15K-E上的分割结果可视化。可以观察到,尽管COME15K-E与水下数据集在成像条件、目标外观及背景结构等方面存在显著差异,本文方法仍能够准确定位显著目标区域,并生成

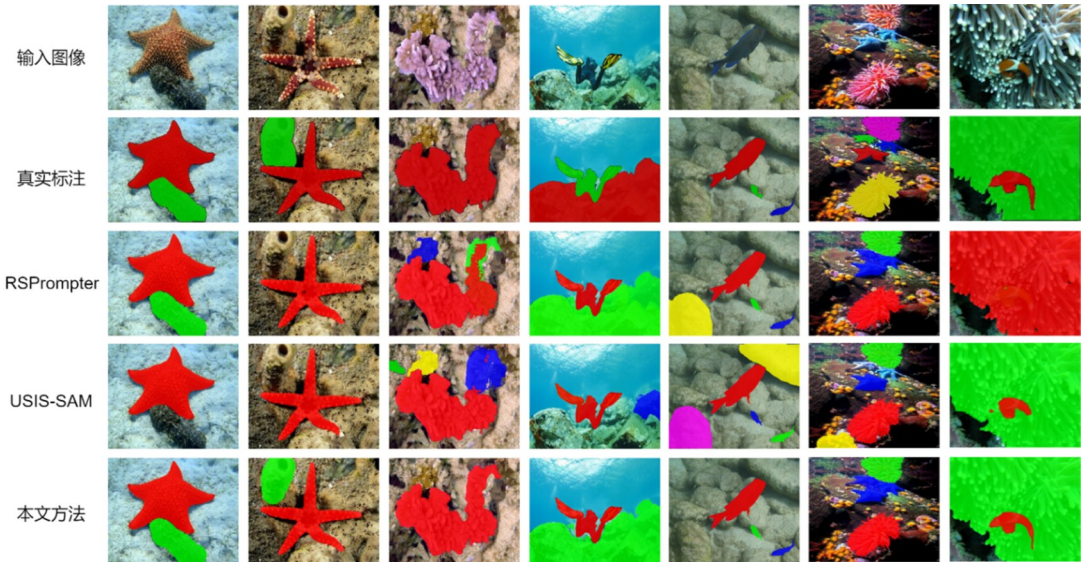


图3 本文方法与USIS-SAM和RSPrompter在UIIS数据集上的分割结果比较

Figure 3 Comparison of segmentation results of the proposed method, USIS-SAM, and RSPrompter on the UIIS dataset

轮廓完整、边界清晰的分割掩码。这表明所提出方法在不同数据分布与应用场景下具有良好的泛化能力。最后，为揭示模型内部的决策过程，本文进一步对中间结果进行了可视化分析。具体而言，我们对模型中的掩码解码器以及动态引导解码器输出的前景概率图进行了可视化展示，如图5所示。可以观察到，动态引导解码器在融合初始掩码与图像语义信息后，对前景区域表现出了更高的置信度，并有效抑制背景干扰。上述结果表明，动态引导解码器能够实现从粗定位到精细分割的逐步优化。

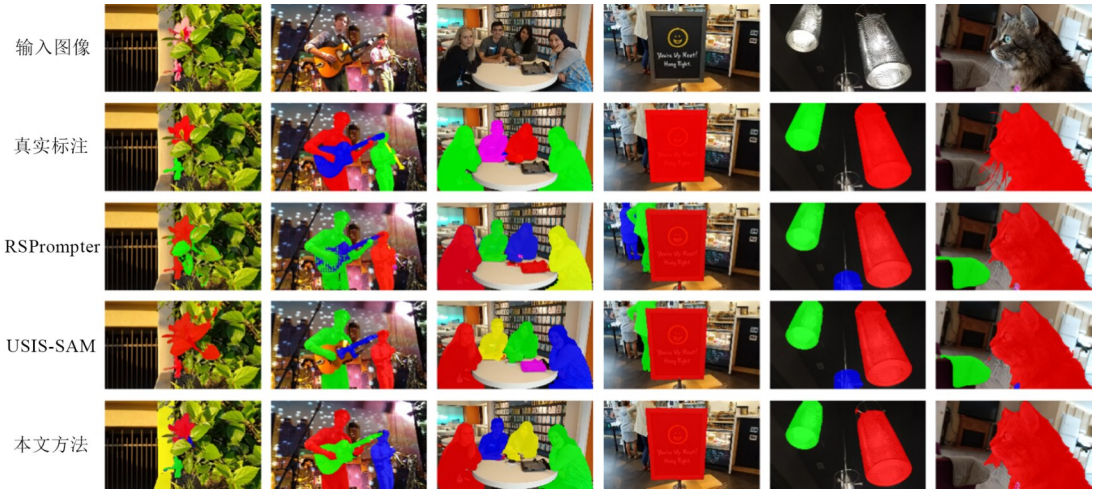


图4 本文方法在COME15K-E数据集上的分割结果可视化

Figure 4 Segmentation result visualization on the COME15K-E dataset

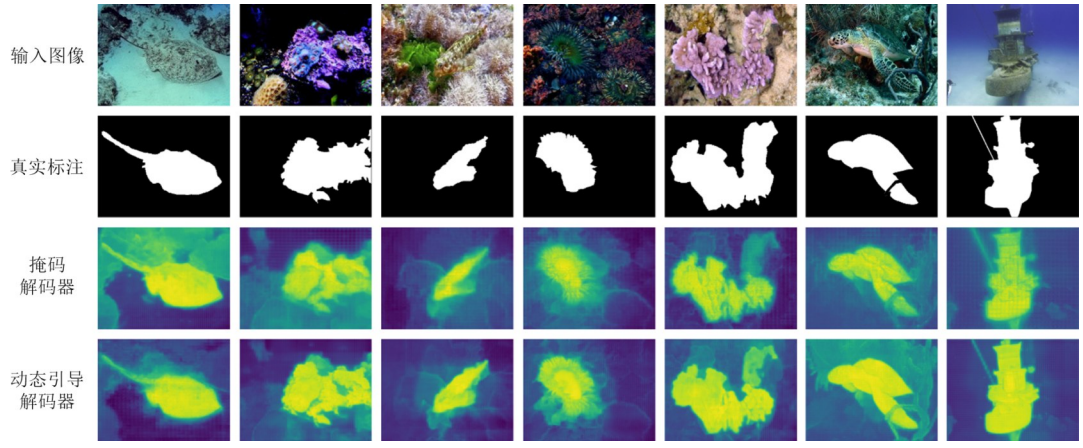


图5 本文方法的掩码解码器与动态引导解码器预测的前景概率

Figure 5 Foreground probability maps from the mask decoder and dynamically-guided decoder

5 结论

本文提出了一种用于水下图像实例分割的动态引导 Segment Anything Model（Dynamically Guided SAM）。具体而言，针对水下场景中目标边界模糊、尺度差异大以及背景复杂等问题，本文对SAM进行了改进。首先，设计了一个自适应特征聚合模块，用于在不改变空间分辨率的前提下，对SAM主干提取的单尺度特征进行通道级重标定与语义对齐，从而缓解分割主干与RPN特征关注点不一致的问题。其次，所提出的多尺度增强模块能够从单尺度特征中构建多层语义金字塔，有效补充不同尺度下的上下文信息，增强模型对小目标与复杂背景的表征能力。最后，所设计的动态引导解码器（DGD）通过融合初始掩码与图像特征来生成动态提示信息，并利用多层注意力机制实现提示与图像特征的有效融合从而获得更加精确的目标掩码。实验结果表明，本文方法在UIIS等四个水下实例分割数据集与两个地面实例分割数据集上的表现均显著优于USIS-SAM与其他主流方法。作者认为，其优越性主要是由于引入了任务语义对齐、多尺度增强与动态提示引导机制，实现了对SAM在水下实例分割任务中的有效适配。尽管如

此,本文方法仍存在一定的局限性。一方面,引入多尺度增强模块与动态引导解码器在一定程度上增加了模型的计算开销,对实时性要求较高的应用场景仍有改进空间;另一方面,本文主要在水下及部分地面数据集上进行了验证,对于更复杂或极端成像条件下的泛化能力仍有待进一步评估。未来的工作将从模型结构轻量化、推理效率优化以及更广泛场景下的泛化能力提升等方面展开研究,并探索将该方法扩展至多模态感知等更复杂的任务中。

参考文献

[1] Zhou Dingfu, Fang Jin, Song Xibin, et al. Joint 3D instance segmentation and object detection for autonomous driving[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA, 2020: 1836-1846.

[2] Zhou Sihang, Nie D, Adeli E, et al. Semantic instance segmentation with discriminative deep supervision for medical images[J]. Medical Image Analysis, 2022, 82: 102626.

[3] Chen Keyan, Liu Chenyang, Chen Hao, 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: 4701117.

[4] Su Hao, Wei Shunjun, Liu Shan, et al. HQ-ISNet: High-quality instance segmentation for remote sensing imagery [J]. Remote Sensing, 2020, 12(6): 989.

[5] 牛玉贞, 张凌昕, 兰杰, 等. 基于分频式生成对抗网络的非成对水下图像增强[J]. 电子学报, 2025, 53(2): 527-544.
Niu Yuzhen, Zhang Lingxin, Lan Jie, et al. FD-GAN: Frequency-decomposed generative adversarial network for unpaired underwater image enhancement[J]. Acta Electronica Sinica, 2025, 53(2): 527-544. (in Chinese)

[6] Lian Shijie, Li Hua, Cong Runmin, et al. WaterMask: Instance segmentation for underwater imagery[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France, Piscataway: IEEE, 2023: 1305-1315.

[7] Lian Shijie, Zhang Ziyi, Li Hua, et al. Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset[PP/OL]. V1. arXiv (2024-06-10)[2025-12-19]. https://arxiv.org/abs/2406.06039.

[8] Kirillov A, Mintun E, Ravi N, et al. Segment anything[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 3992-4003.

[9] Zhang Xin, Liu Yu, Lin Yuming, et al. UV-SAM: Adapting segment anything model for urban village identification [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(20): 22520-22528.

[10] Wu Junde, Wang Ziyue, Hong Mingxuan, et al. Medical SAM adapter: Adapting segment anything model for medical image segmentation[J]. Medical Image Analysis, 2025, 102: 103547.

[11] Huang Jiaxing, Jiang Kai, Zhang Jingyi, et al. Learning to prompt segment anything models[PP/OL]. V1. arXiv (2024-01-09) [2025-12-18]. https://doi. org/10.48550/arXiv.2401.04651.

[12] Chen Tianrun, Zhu Lanyun, Ding Chaotao, et al. SAM fails to segment anything? : SAM-adapter: Adapting SAM in underperformed scenes: Camouflage, shadow, medical image segmentation, and more[PP/OL]. V3. arXiv (2023-05-02) [2025-12-18]. https://doi. org/10.48550/arXiv.2304.09148.

[13] Hu E J, Shen Yelong, Wallis P, et al. Lora: Low-rank adaptation of large language models[PP/OL]. V2. arXiv (2021-10-16)[2025-12-18]. https://arxiv.org/abs/2106.09685.

[14] 梁新宇, 林洗坤, 权冀川, 等. 基于深度学习的图像实例分割技术研究进展[J]. 电子学报, 2020, 48(12): 2476-2486.
Liang Xinyu, Lin Xikun, Quan Jichuan, et al. Research on the progress of image instance segmentation based on deep learning[J]. Acta Electronica Sinica, 2020, 48(12): 2476-2486. (in Chinese)

[15] He Kaiming, Gkioxari G, Dollár P, et al. Mask R-CNN [C]//2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2980-2988.

[16] Cai Zhaowei, Vasconcelos N. Cascade R-CNN: High quality object detection and instance segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(5): 1483-1498.

[17] Chen Kai, Pang Jiangmiao, Wang Jiaqi, et al. Hybrid task cascade for instance segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA, 2019: 4969-4978.

[18] Ren Shaoqing, He Kaiming, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[19] Bolya D, Zhou Chong, Xiao Fanyi, et al. YOLACT: Real-time instance segmentation[C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 9156-9165.

[20] Chen Hao, Sun Kunyang, Tian Zhi, et al. BlendMask: Top-down meets bottom-up for instance segmentation [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 8570-8578.

[21] Tian Zhi, Shen Chunhua, Chen Hao. Conditional convolutions for instance segmentation[C]//Computer Vision - ECCV 2020. Cham: Springer, 2020: 282-298.

[22] Wang Xinlong, Zhang Rufeng, Kong Tao, et al. Solov2: Dynamic and fast instance segmentation[J]. Advances in Neural Information Processing Systems, 2020, 33: 17721-17732.

[23] Cheng Bowen, Schwing A G, Kirillov A. Per-pixel classification is not all you need for semantic segmentation[J]. Advances in Neural Information Processing Systems, 2021, 34: 17864-17875.

[24] Cheng Bowen, Misra I, Schwing A G, et al. Masked-attention mask transformer for universal image segmentation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 1280-1289.

[25] Li Feng, Zhang Hao, Xu Huaizhe, et al. Mask DINO: Towards a unified transformer-based framework for object detection and segmentation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 3041-3050.

[26] Zhang Wenwei, Pang Jiangmiao, Chen Kai, et al. K-net: Towards unified image segmentation[J]. Advances in Neural Information Processing Systems, 2021, 34: 10326-10338.

[27] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[PP/OL]. V2. arXiv (2021-06-03) [2025-12-18]. https://doi.org/10.48550/arXiv.2010.11929.

[28] Li K, Rajpurkar P. Adapting segment anything models to medical imaging via fine-tuning without domain pretraining[C/OL]//AAAI 2024 Spring Symposium on Clinical Foundation Models. Openreview, 2024. https://openreview.net/forum?id=Fxi7pRmnYJ.

[29] Jia Menglin, Tang Luming, Chen Bochun, et al. Visual prompt tuning[C]//Computer Vision - ECCV 2022. Cham: Springer, 2022: 709-727.

[30] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]// International Conference on Machine Learning. [S. l.]: PmLR, 2021: 8748-8763.

[31] Waszak M, Cardaillac A, Elvesæter B, et al. Semantic segmentation in underwater ship inspections: Benchmark and data set[J]. IEEE Journal of Oceanic Engineering, 2023, 48(2): 462-473.

[32] Li Hua, Lian Shijie, Li Zhiyuan, et al. UWSAM: Segment anything model guided underwater instance segmentation and a large-scale benchmark dataset[PP/OL]. V1. arXiv (2025-05-21) [2025-12-18]. https://arxiv.org/html/2505.15581v1.

[33] Zhang Jing, Fan Dengping, Dai Yuchao, et al. RGB-D saliency detection *via* cascaded mutual information minimization[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 4318-4327.

[34] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[M]//Computer Vision - ECCV 2014. ChamSpringer International Publishing2014: 740-755.

[35] Chen Kai, Wang Jiaqi, Pang Jiangmiao, et al. MMDetection: Open MMLab detection toolbox and benchmark[PP/OL]. V1. arXiv (2019-06-17)[2025-12-18].https://doi.org/10.48550/arXiv.1906.07155.

[36] Fang Yuxin, Yang Shusheng, Wang Xinggang, et al. Instances as queries[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 6890-6899.

[37] He Junjie, Li Pengyu, Geng Yifeng, et al. FastInst: A simple query-based model for real-time instance segmentation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 23663-23672.

[38] Tian Zhi, Shen Chunhua, Wang Xinlong, et al. BoxInst: High-performance instance segmentation with box annotations[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 5439-5448.

[39] Xiong Yunyang, Varadarajan B, Wu Lemeng, et al. EfficientSAM: Leveraged masked image pretraining for efficient segment anything[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 16111-16121.

[40] Wu Yuhuan, Liu Yun, Zhang Le, et al. Regularized densely-connected pyramid network for salient instance segmentation[J]. IEEE Transactions on Image Processing, 2021, 30: 3897-3907.

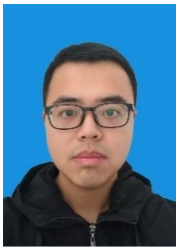
[41] Pei Jialun, Cheng Tianyang, Tang He, et al. Transformer-based efficient salient instance segmentation networks

with orientative query[J]. IEEE Transactions on Multimedia, 2023, 25: 1964-1978.

[42] Fan Ruochen, Cheng Mingming, Hou Qibin, et al. S4Net: Single stage salient-instance segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 6096-6105.

[43] Kirillov A, Wu Yuxin, He Kaiming, et al. PointRend: Image segmentation as rendering[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 9796-9805.

作者简介



尚毅涵 男,于2021年在哈尔滨理工大学获得学士学位。现于中国海洋大学攻读计算机技术专业硕士学位。主要研究方向为计算机视觉、图像分割和密集预测。

E-mail: shangyihan@stu.ouc.edu.cn

董兴辉 男,于2014年在英国赫瑞-瓦特大学获得博士学位。现任中国海洋大学教授、人工智能学院副院长。主要研究方向为计算机视觉、智能检测、纹理分析和视觉感知。

E-mail: xinghui.dong@ouc.edu.cn