计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.010133] [CSTR: 32024.14.csa.010133]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

面向海洋生物的多模态零样本语义分割①

周德龙[1], 张　宁[2], 程远志[1]

[1](青岛科技大学 信息科学技术学院, 青岛 266061)
[2](青岛市中医医院, 青岛 266033)
通信作者: 张　宁, E-mail: ruudcheung@163.com

摘　要: 海洋生物图像分割是智能化海洋监测的重要基础, 但在实际应用中仍面临跨模态语义偏差、多尺度融合效率低以及生物结构建模不足等挑战. 为此, 本文提出一种基于 CLIP 的多模态语义分割框架 Mseg, 能够在未见类别上实现有效分割. 该方法融合视觉图像与生物类别文本特征, 同时利用轻量级交叉注意力 (LCA) 机制和多层级图像特征融合策略引导图像与文本特征的交互, 从而生成语义增强的图像表征. 随后, 引入 BalanceITV 模块对两路特征进行动态加权融合, 实现主干融合视觉特征与语言引导特征的自适应平衡. 最后, 本文设计了基于海洋生物形态感知的不确定性建模方法, 在边界区域及复杂生物结构处提升了分割的精细度与鲁棒性. 实验结果表明, Mseg 在多个海洋生物零样本分割任务中均优于现有方法, 验证了其在复杂水下场景中的适应性与有效性.

关键词: 图像分割; 跨模态; 零样本; 不确定性估计

引用格式: 周德龙,张宁,程远志.面向海洋生物的多模态零样本语义分割.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/10133.html

Multimodal Zero-shot Semantic Segmentation for Marine Organisms

ZHOU De-Long[1], ZHANG Ning[2], CHENG Yuan-Zhi[1]

[1](School of Information Science and Technology, Qingdao University of Science & Technology, Qingdao 266061, China)
[2](Qingdao Hospital of Traditional Chinese Medicine, Qingdao 266033, China)

Abstract: Image segmentation of marine organisms is fundamental to intelligent ocean monitoring but remains challenging due to issues such as cross-modal semantic deviation, inefficient multi-scale fusion, and insufficient modeling of biological structures. To address these challenges, this study proposes Mseg, a CLIP-based multimodal semantic segmentation framework, to achieve effective segmentation of unseen categories. The method integrates visual image features with textual category descriptions, while employing a lightweight cross-attention (LCA) mechanism and a multi-level feature fusion strategy to guide the interaction between visual and textual representations, thereby generating semantically enriched image representations. Subsequently, a BalanceITV module is introduced to dynamically weight and adaptively balance the two streams of features, namely, the backbone visual features and the language-guided features. Moreover, an uncertainty modeling method on marine organism morphology perception is designed to enhance segmentation precision and robustness, particularly in boundary regions and areas with complex biological structures. Experiments on multiple marine organism datasets show that Mseg consistently outperforms existing methods in zero-shot segmentation tasks, demonstrating its strong adaptability and effectiveness in complex underwater environments.

Key words: image segmentation; cross-modal; zero-shot; uncertainty estimation

① 基金项目: 国家自然科学基金 (61971142)
收稿时间: 2025-09-23; 修改时间: 2025-10-14, 2025-11-11; 采用时间: 2025-11-24; csa 在线出版时间: 2026-03-09

随着海洋资源开发的不断加深与海洋生态系统研究的持续推进, 海洋生物图像分割技术在海洋科学中的重要性日益凸显. 准确地从图像中提取特定生物个体, 是实现物种识别、数量评估及行为分析等高层任务的基础环节. 传统的有监督方法高度依赖人工标注数据, 不仅成本高昂、扩展性差, 而且在多样化海洋场景下容易出现泛化失效的问题. 因此, 如何在缺乏标注样本的条件下利用语义关联实现零样本分割, 成为当前海洋智能感知领域亟需突破的关键问题.

然而, 零样本分割在海洋环境中仍面临 3 大挑战: ① 跨模态语义偏差: 视觉与语言特征空间分布差异大, 导致语义映射不稳定; ② 多尺度融合效率低: 海洋生物形态复杂、尺度差异显著, 传统静态融合难以兼顾局部细节与全局语义; ③ 生物结构建模不足: 现有模型缺乏对解剖形态的显式约束, 难以在模糊边界和复杂结构区域保持预测精度.

为此, 本文提出一种基于视觉语言联合建模的海洋生物图像分割框架 Mseg (multimodal segmentation guided by vision-language embedding), 以解决上述问题. 该框架以 CLIP (contrastive language-image pre-training) 为基础, 融合视觉与文本的多模态特征表示, 通过引入类别文本嵌入与跨尺度语义对齐机制, 增强模型对未见类别的理解与泛化能力. 在结构设计上, Mseg 采用双支路特征融合策略: 一方面, 通过视觉图像特征与文本嵌入的内积计算获取初步的语义融合特征图, 实现显式的跨模态语义对齐; 另一方面, 利用轻量级交叉注意力模块从特征层面引导视觉与文本信息的深层交互, 生成语义增强的特征融合图. 随后, 将两支路输出的融合特征共同传入 BalanceITV 模块进行动态平衡与自适应融合, 从而实现视觉特征与语言引导特征的协同优化. 最后, 本文设计了基于生物形态感知的不确定性建模方法, 在损失函数中引入结构置信度约束, 有效提升了模型在模糊边界与复杂结构区域的分割精度与鲁棒性.

1 研究背景及相关工作

1.1 海洋图像语义分割

海洋图像语义分割是智能水下感知[1]的重要组成部分, 广泛应用于生物监测、种群估计与生态分析等任务. 与陆地图像相比, 海洋图像具有光照变化大、背景干扰强、目标结构复杂等特点, 对分割精度和模型鲁棒性提出了更高要求.

近年来, 许多先进分割模型在自然图像和遥感图像上取得了显著进展, 例如 HRNet[2]通过保持高分辨率特征并进行多尺度融合, 有效提升了边界识别能力; SegFormer[3]则基于 Transformer[4]设计轻量主干, 在提升上下文建模能力的同时降低了计算开销; 而 Mask2Former[5]将多种分割任务统一为 mask 表达的框架, 在开放场景下展现出更好的泛化能力.

这些方法虽然在结构建模和全局感知方面表现出色, 但面对水下模糊、类间形态相似和数据标签稀缺等挑战, 仍存在小目标识别能力不足、边界模糊预测不稳等问题. 针对这些瓶颈, Mseg 融合多模态语义引导[6]、跨尺度注意力融合与结构感知机制[7], 不仅提升了对目标边界与纹理的刻画能力, 也增强了模型在弱监督[8]或零样本[9]下的分割性能, 尤其适用于多类相似物种共存的复杂水下环境.

1.2 视觉语言模型与多模态语义对齐

近年来, 多模态学习逐渐成为计算机视觉研究热点. 视觉语言模型 (VLM)[10]通过联合建模图像与语言模态, 在语义理解、开放集识别等任务中表现出色. CLIP[11]是该领域的代表性工作, 其通过图文对比学习将图像和文本嵌入到统一语义空间, 从而具备零样本预测能力.

后续研究如 LSeg[12]将 CLIP 的语言引导能力应用于分割任务中, 借助文本嵌入提升模型对类语义的理解, 提升了小样本和弱监督条件下的性能. DenseCLIP[13]和 MaskCLIP+[14]等则通过更细粒度的特征匹配与注意力机制优化视觉文本融合效果, 逐步向高密度像素级预测任务拓展.

然而, 这些方法大多在自然图像语料上训练, 缺乏对海洋语义的领域适应性; 此外, 其融合策略多为直接拼接或静态融合, 缺乏针对不同语义尺度的动态匹配能力, 难以处理同类物种间语义重叠严重的问题. Mseg 在此基础上引入层级式语义引导机制, 通过构建多尺度视觉与文本交互模块, 提升了模型对生物类属性、上下文关系的建模能力, 尤其适用于种间相似度高、语义模糊的海洋生物场景.

1.3 不确定性建模研究

在海洋图像语义分割任务中, 目标生物常具有边界模糊、结构复杂及类间外观相似等特点, 加之水下图像普遍存在光照不均、水体悬浮物干扰与分辨率退

化等噪声因素, 易导致模型预测结果出现不稳定与模糊现象. 因此, 如何对预测结果进行可信度评估, 并聚焦模型最不确定的区域进行强化学习, 成为提升水下分割鲁棒性的关键途径. 近年来, 不确定性建模[15]在医学图像分割、遥感场景解析等对细节敏感的任务中取得显著成果, 并逐步应用于海洋场景语义分割.

现有方法主要包括两类: 一类为预测阶段的不确定性估计, 如 Monte Carlo dropout (MC Dropout)[16]、Test-time augmentation (TTA)[17]和 Deep ensembles[18], 通过多次前向传播获取输出分布的方差, 进而评估模型对各区域预测的稳定性; 另一类在训练阶段引入不确定性感知机制, 在损失函数中对边界模糊区或低置信区域赋予更高权重, 引导模型关注易错区域.

在海洋图像中, 由于鱼类、贝类、软体动物等生物具备明显的解剖学结构先验[19], 如脊柱曲率、鳍基位置、体型轮廓等, 结合结构感知的不确定性建模策略能有效约束模型的预测空间.

基于此, 本文在 Mseg 模型的损失构建中引入了解剖学结构感知的不确定性估计机制, 用以动态衡量模型在结构边界处的预测置信度, 并引导模型聚焦于形态边界模糊及生物体态变化剧烈的高风险区域, 从而显著提升分割结果的鲁棒性与局部结构刻画能力.

2 方法介绍

本文提出的模型 Mseg 在统一的多模态空间中联合建模图像与文本语义, 从而实现海洋生物的零样本语义分割. 整体框架如图 1 所示, 以下将对各组成部分进行详细介绍.

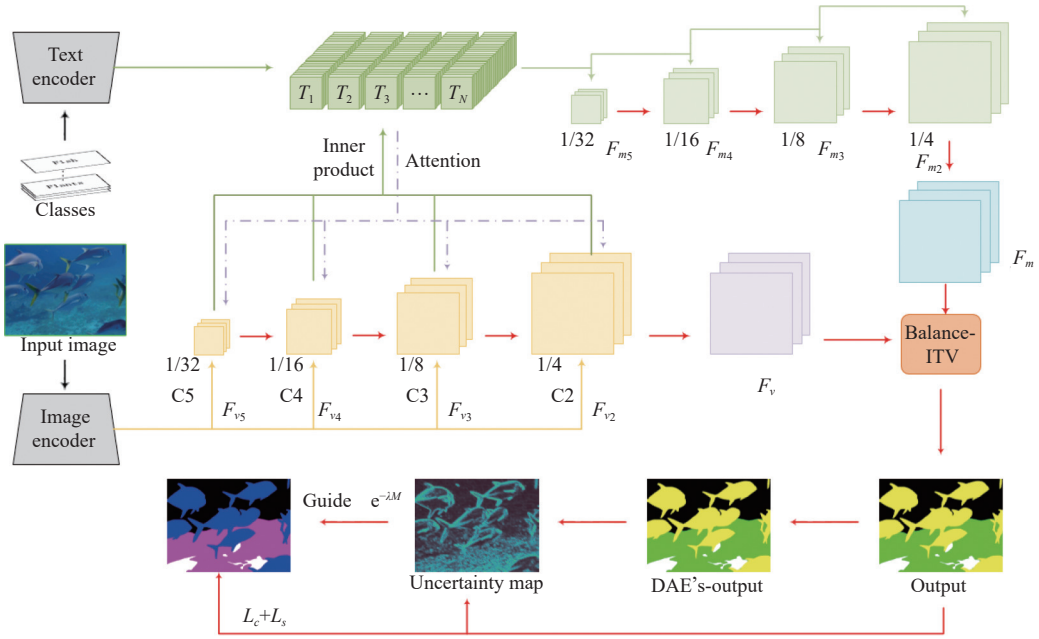

图 1 Mseg 框架总览图

2.1 特征提取和跨模态相似性融合

本工作设计了一种双支路特征融合策略, 通过前期特征提取与两种跨模态相似性融合机制协同实现, 其整体流程如下: 我们首先采用 CLIP 预训练的图像与文本编码器分别作为图像和类目语义信息提取模块. 对于图像通道, 可选择基于 ResNet 或 ViT 的 CLIP 编码器, 提取来自 C2–C5 阶段的多尺度视觉特征, 依次被定义为F_{v2}、F_{v3}、F_{v4}、F_{v5}. 其中, 它们的尺寸大小依次为原始图像的 1/4、1/8、1/16、1/32, 特征维度依次为c_2、c_3、c_4、c_5. 同时, 文本编码器接收所有生物类目的文本标签, 并将其编码为一组固定维度的嵌入向量$\{T_1, T_2, \cdots, T_N\} \in R^{1\times1\times C}$, 构成统一的语义表示空间.

在完成图像编码器与文本编码器的特征提取后, 由于图像编码器在多尺度阶段提取的特征图通道数不一致, 为保证后续与文本特征的融合一致性, 我们在各尺度特征上引入1×1卷积将通道数统一映射为C. 接下来模型采用并联的两种跨模态融合方式, 以生成不同的特征融合图.

第 1 种融合方式基于内积计算实现图像与文本特征的显式语义对齐. 我们将图像空间位置(i,j)处的特征向量$I_{ij}\in R^{1\times1\times C}$与文本类别嵌入$T_k\in R^{1\times1\times C}$做内积, 定义为:

$$f_{ijk}=I_{ij}\cdot T_k \tag{1}$$

上面计算得到的$f_{ijk}\in R$便可以表示空间位置(i,j)与类别k的匹配度, 接下来我们依次将I_{ij}与N个文本类别嵌入计算便可以得到三维匹配张量$F_{ij}=\left[f_{ij1},f_{ij2},\cdots,f_{ijN}\right]\in R^{1\times1\times N}$, 表示图像像素点$(i,j)$与所有类别文本的相关性, 然后针对不同尺度的图像采用上述计算方式, 便可以得到不同尺度的融合特征图像, 分别记为F_{m2}、F_{m3}、F_{m4}、F_{m5}, 最后通过空洞卷积融合成最终图像F_m.

第 2 种融合方式基于交叉注意力机制实现. 我们设计了一种轻量级交叉注意力 (lightweight cross-attention, LCA) 融合模块, 用于在不同尺度上高效对齐视觉特征与文本语义, 并实现两者间的互补增强. 该模块在保持计算开销较低的同时, 强化了语义指导信息对视觉感知的调控作用, 提升了分割结果的语义一致性和边界清晰度.

该模块的核心思想是: 将图像主干网络提取的多尺度特征作为查询(*Query*), 文本特征作为键/值 (*Key*/*Value*) 输入, 构建一个跨模态注意力映射, 突出图像中的语义显著区域, 同时压制背景干扰.

设视觉主干在某尺度层提取的图像特征表示为:

$$F_v\in R^{C\times H\times W} \tag{2}$$

其中, C表示通道数, $H\times W$为特征图空间尺寸. 文本编码器输出的嵌入向量为:

$$T\in R^{D} \tag{3}$$

其中, D为文本向量维度. 为了使文本与图像特征可以进行交叉注意力计算, 先将文本特征线性映射到图像特征维度下:

$$T'=W_tT+b_t,\ T'\in R^{C} \tag{4}$$

其中, $W_t\in R^{C\times D}$, $b_t\in R^{C}$为可学习参数. 然后, 将其广播复制为与图像特征空间匹配的尺寸:

$$T_b=Repeat(T',H,W)\in R^{C\times H\times W} \tag{5}$$

接下来我们引入图像引导的文本注意力机制, 即图像特征作为*Query*, 文本嵌入作为*Key*和*Value*进行计算, 构造跨模态注意力权重矩阵. 首先进行线性变换:

$$Q=W_qF_v,\ K=W_kT_b,\ V=W_vT_b \tag{6}$$

其中, $W_q,W_k,W_v\in R^{C'\times C}$, $Q,K,V\in R^{C'\times H\times W}$, 然后按空间位置展开为:

$$Q',K',V'\in R^{(H\cdot W)\times C'}$$

注意力权重计算如下 (缩放点积注意力):

$$A=Softmax\left(\frac{Q'K'^{\mathrm{T}}}{\sqrt{C'}}\right)\in R^{(H\cdot W)\times(H\cdot W)} \tag{7}$$

输出特征为:

$$F_{\mathrm{attn}}=Reshape(AV')\in R^{C'\times H\times W} \tag{8}$$

最后通过一个残差连接与融合映射得到注意力增强特征:

$$F_{\mathrm{out}}=ReLU\left(BN\left(W_o\left[F_{\mathrm{attn}}\|F_v\right]\right)\right)+F_v \tag{9}$$

其中, $[\cdot\|\cdot]$表示通道维度拼接; $W_o\in R^{C\times(C'+C)}$; BN表示批归一化. 为了捕捉不同感受野下的注意力增强结果, 我们在C_2、C_3、C_4、C_5这 4 个尺度分别执行上述注意力交叉计算, 并对结果执行上采样对齐:

$$F_i'=LCAFM(F_{v(i)},T),\ i=2,3,4,5 \tag{10}$$

将所有图像调整至最大输入分辨率后进行融合:

$$F_{\mathrm{cross}}=\varphi\left(F_2',F_3',F_4',F_5'\right) \tag{11}$$

其中, $\varphi(\cdot)$表示注意力聚合融合函数.

2.2 双支路特征动态平衡融合

本文提出了一种 BalanceITV 模块来解决多模态特征融合过程中, 缺乏动态调节机制, 从而容易导致单一路径的特征主导融合结果, 削弱另一路径特征有效表达的问题, 具体设计如图 2 所示. 该模块通过残差增强、跨模态注意力与自适应加权机制, 实现了主干视觉特征与语言引导特征的动态平衡融合. 我们将来自图像-文本内积路径的特征记作为$F^{(I)}$, 来自轻量级交叉注意力路径的特征为$F^{(A)}$.

首先, 两种特征图像经过上采样操作将图像大小变为原始图像大小, 然后模块引入残差增强机制, 在对齐通道维度后, 将一个模态的特征作为残差信息输入另一个模态, 通过跨模态注意力机制计算其最相关的补充信息, 例如对$F^{(I)}$的残差建模可表示为:

$$Res\left(F^{(A)}\right)=\sum_{i=1}^{H\times W}\beta_i\cdot F_i^{(A)} \tag{12}$$

其中, 注意力权重β_i通过 *Softmax* 归一化相似性函数计

算得到:

$$\beta_i = \frac{\exp\left(\left(F_i^{(A)}\right)^{\mathrm{T}} W F^{(I)}\right)}{\sum_{k=1}^{H\times W} \exp\left(\left(F_k^{(A)}\right)^{\mathrm{T}} W F^{(I)}\right)} \tag{13}$$

该机制使得$Res\left(F^{(A)}\right)$能为$F^{(I)}$提供最相关的补充信息. 同理, 可以计算$Res\left(F^{(I)}\right)$并作用于$F^{(A)}$, 从而得到双向残差增强特征:

$$\hat{F}^{(I)} = AttentionBRFF\left(F^{(I)}, Res\left(F^{(A)}\right)\right) \tag{14}$$

$$\hat{F}^{(A)} = AttentionBRFF\left(F^{(A)}, Res\left(F^{(I)}\right)\right) \tag{15}$$

在残差增强后, 得到两组更新的特征$\hat{F}^{(I)}$和$\hat{F}^{(A)}$. 该模块会通过自适应加权机制整合两路信息, 动态控制它们对最终特征表示的贡献.

首先, 计算权重向量$\alpha \in R^N$:

$$\alpha = Softmax\left(W \cdot \left[Pool\left(\hat{F}^{(I)}\right); Pool\left(\hat{F}^{(A)}\right)\right] + b\right) \tag{16}$$

其中, $Pool(\cdot)$表示对空间维度的平均池化操作, 用于提取全局语义; W和b为可学习参数. 最终融合特征表示为:

$$F^{\text{fuse}} = \alpha \odot \hat{F}^{(I)} + (1-\alpha) \odot \hat{F}^{(A)} \tag{17}$$

其中, $\odot$表示逐元素加权. 这样, 该模块可以在不同任务和不同类别下, 动态调整图像-文本内积特征与交叉注意力特征的贡献比例.

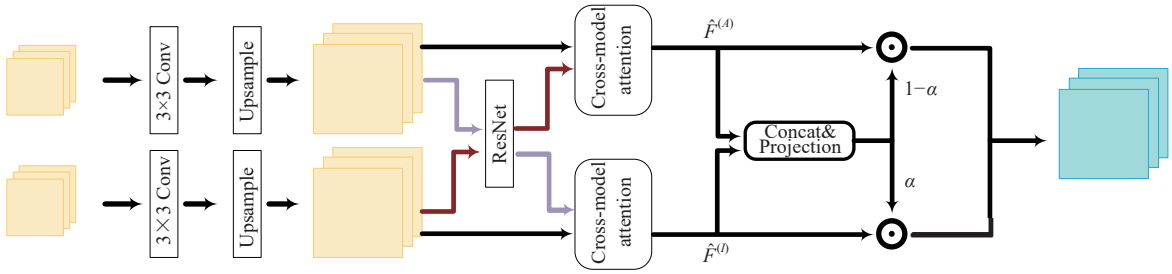

图 2 BalanceITV 模块展示图

2.3 海洋生物形态感知的不确定性方法

本研究设计了一种基于自编码器的形态感知先验的不确定性方法, 在海洋生物语义分割任务中引入形态感知先验知识, 从而实现面对复杂的生物体形态和背景干扰时, 仍然确保模型学习到海洋生物的全局形态特征, 而不仅依赖于局部像素的信息的能力. 具体来说, 我们通过引入去噪自编码器 (DAE) 将分割掩码编码到一个非线性潜在空间中, 从而学习到海洋生物的形态感知表示先验. 这个先验能够捕获海洋生物的全局形态特征, 使得模型能够调整不合理的分割预测, 得到更符合实际形态的分割结果.

不确定性的作用是识别模型预测中的可靠目标区域, 从而提高分割结果的可信度. 我们的方法直接利用形态感知表示网络来估计不确定性, 仅需一次推理步骤. 具体来说, DAE 利用模型分割预测结果 (p_{m_i}) 生成合理的生物形态分割预测 ($\hat{p}_{m_i} = DAE\left(p_{m_i}\right)$), 然后计算该预测与主模型的输出之间的逐像素差异, 从而得出不确定性. 具体计算如下:

$$M_i = \|\hat{p}_{m_i} - p_{m_i}\|^2 \tag{18}$$

通过这种方式, 我们能够精准识别海洋生物分割中的不确定区域, 尤其是那些形态复杂或被背景干扰的区域. 不确定性公式与传统的基于样本方差的不确定性估计有关. 具体来说, 对于给定的输入x_i及其对应的多个模型预测p_{i_n}, 样本方差估计定义如下:

$$var(p_i) = \frac{1}{N-1}\sum_{n=1}^{N}(p_{i_n} - \overline{p}_i)^2 \tag{19}$$

其中, $\overline{p}_i$代表样本均值, 并被定义为: $\overline{p}_i = \frac{1}{N}\sum_{n=1}^{N}(p_{i_n})$. 参数$N$表示预测样本的数量, 当$N$被设为 2 时, 样本均值$\overline{p}_i$减少为$(p_{i_1} + p_{i_2})/2$, 导致方差估计采用以下形式:

$$\begin{aligned} var(p_i) &= \left(p_{i_1} - \frac{p_{i_1} + p_{i_2}}{2}\right)^2 + \left(p_{i_2} - \frac{p_{i_1} + p_{i_2}}{2}\right)^2 \\ &= \left(\frac{p_{i_1} - p_{i_2}}{2}\right)^2 + \left(\frac{p_{i_2} - p_{i_1}}{2}\right)^2 \end{aligned} \tag{20}$$

$$var(p_i) = \frac{1}{2}(p_{i_1} - p_{i_2})^2 \tag{21}$$

最终得到的方差计算公式与前面的不确定性计算公式只有一个常数比例的差异, 因此可以认为两者在数学意义上等价. 最后可通过由不确定性计算得到的不确定性图来获取可靠的目标区域, 具体为: $e^{-\lambda M_i}$, 其中, λ是一个不确定性加权因子, 在这里我们将它设定为 1. 利用可靠的目标区域最终组合成一致性损失为:

$$L_c\left(p_{m_i},GT\right)=\frac{\sum_x e^{-\lambda M_{i,x}}\|p_{m_{i,x}}-GT\|^2}{\sum_x e^{-\lambda M_{i,x}}} \tag{22}$$

其中, x表示一个像素. 除此之外, 我们还将一致性损失L_c和监督损失L_s作为学习目标进行联合优化, 其中L_s是交叉熵损失和 Dice 损失的组合.

3 实验

3.1 实验数据集

本文使用的是 SUIM 数据集[20]和 UIIS10K 数据集[21]. SUIM 是一个公开的水下图像分割数据集, 由孟加拉国达卡大学的研究团队发布, 主要面向水下机器人视觉与智能监测任务. 该数据集包含约 1 525 张水下场景图像, 涵盖 8 个类别, 包括鱼类、植物、岩石、沙地、珊瑚、潜水员、船只以及杂物等, 提供逐像素语义分割标注, 能够支持对复杂水下场景的目标检测与分割研究. UIIS10K 是一个大规模的水下图像分割数据集, 由 UWSAM 团队构建, 旨在推动水下目标检测与分割技术的发展. 该数据集包含 10 048 张真实水下图像, 并覆盖 10 个主要类别, 如鱼类、海星、甲壳类、软体动物等, 每张图像均配有高质量的实例级标注, 适用于水下多目标实例分割和语义识别等任务.

值得注意的是, 为了更好地模拟和展示模型零样本/少样本分割能力, 我们将数据集进行了特殊划分. 将 SUIM 分成 4 组, 每组中有 2 个类别, 训练时其中 3 组会被用来训练, 1 组用来测试, 从而人为模拟零样本分割实验, 最后会进行 4 折交叉验证, 实验结果取平均值. UIIS10K 无法被平均分组, 于是我们采取的策略是每次随机分成 4 组, 其中 1 组只有 1 个类别, 其余 3 组混合为 3 个类别, 同样进行 4 次实验, 并对实验结果取平均值.

3.2 实验细节及评价指标

训练过程中, 我们将输入图像统一裁剪为 256×256 像素的随机采样块, 以适配主干网络的输入需求. 为增强模型的泛化能力, 我们在训练中采用了多种数据增强技术, 包括随机缩放、随机水平翻转、颜色抖动、高斯噪声扰动以及对比度随机调整. 进行本文实验的训练与测试环境均为 Ubuntu 20.04 操作系统, 使用一块 NVIDIA GeForce RTX 4090 显卡, 编程语言为 Python3.8, 使用 PyTorch 深度学习框架完成整个模型的训练与测试过程. 优化器采用 Adam, 其参数设置为 β_1=0.9, β_2=0.99, 权重衰减率为 1×10^{-5}. 此外, 在训练阶段引入了标准的数据增强与随机丢弃策略以防止过拟合. 本文采用平均交并比 mIoU (mean intersection over union) 和前景-背景交并比 FB-IoU (foreground-background intersection over union) 作为评估指标. 其中, mIoU 用于计算所有类别的平均交并比, 以衡量整体分割精度; 而 FB-IoU 则分别计算前景与背景的交并比, 并取其平均值, 同时忽略具体的目标类别, 从而综合反映模型在前景与背景区分方面的性能.

3.3 实验结果和分析

3.3.1 对比实验

我们使用最先进的几个零样本分割方法 (包括 SPNet[22], ZegFormer[23], MaskCLIP+[14], TAS[24], ZS3Net[25], ZegCLIP[26]) 与本文方法对比. 不同方法数据集整体对比结果如表 1 和表 2 所示, 表中加粗表示最优结果, 下划线为次优结果. 可视化对比图如图 3 所示.

表 1 不同方法在 SUIM 数据集上的评估表现 (%)

方法	mIoU					FB-IoU
	实验 1	实验 2	实验 3	实验 4	Avg	
SPNet	60.63	61.37	58.52	56.53	59.26	68.08
ZegFormer	57.97	70.73	58.72	58.30	61.43	69.60
MaskCLIP+	<u>69.44</u>	59.79	58.43	60.08	61.94	71.71
TAS	56.38	63.50	<u>69.41</u>	59.39	62.17	70.92
ZS3Net	68.05	59.40	67.03	63.86	64.59	<u>72.95</u>
ZegCLIP	61.90	<u>71.67</u>	67.15	**64.45**	66.29	71.89
Mseg (Ours)	**70.21**	**72.24**	**69.81**	<u>64.06</u>	**69.08**	**75.61**

表 2 不同方法在 UIIS10K 数据集上的评估表现 (%)

方法	mIoU					FB-IoU
	实验 1	实验 2	实验 3	实验 4	Avg	
SPNet	56.40	62.23	52.88	60.83	58.09	59.20
ZegFormer	55.71	54.01	58.93	58.60	56.81	62.35
MaskCLIP+	56.40	<u>63.23</u>	53.88	61.83	58.84	60.42
TAS	58.35	62.59	<u>62.67</u>	57.23	60.21	65.71
ZS3Net	60.18	61.13	62.04	59.46	60.70	63.55
ZegCLIP	<u>60.65</u>	62.79	57.04	<u>64.12</u>	<u>61.15</u>	<u>67.09</u>
Mseg (Ours)	**64.11**	**64.38**	**64.73**	**64.53**	**64.44**	**68.65**

结合图 3 的可视化结果可以发现, Mseg 在鱼类、珊瑚等边界复杂或形态细碎的生物类别中表现尤为突出. 相比其他方法, 其预测轮廓更平滑且与真实结构更

贴合, 特别是在细长鱼鳍、珊瑚分枝及多目标重叠区域中, 模型能够准确保持物体完整性并有效抑制背景误分割现象. 这一视觉改善与 mIoU 的量化提升趋势一致, 说明 Mseg 不仅在全局语义对齐方面具有优势, 同时在局部结构精细度与边界辨识能力上也取得显著增强.

对于 SUIM 数据集, Mseg 在 4 次独立实验中有 3 次取得最高分, 且在平均结果上依然保持领先, 尽管在其中一次实验中表现略逊于最佳方法, 但差距极小, 并未影响整体最优的平均表现; 而在 UIIS10K 数据集上, Mseg 的优势更加显著, 不仅在 4 次实验中均取得最佳成绩, 在平均结果上也显著超过了所有对比方法. 这种在两个数据集上均保持领先的趋势表明, 该模型在多样化任务环境中展现出持续一致的竞争力和优越表现, 并具备较强的泛化能力和适应性, 验证了 Mseg 在零样本分割任务中的有效性与鲁棒性, 为其在更大规模、更复杂的真实水下场景中的推广与应用提供了坚实依据和发展潜力, 展示了该方法在未来自动化海洋监测任务中的广阔前景与研究价值.

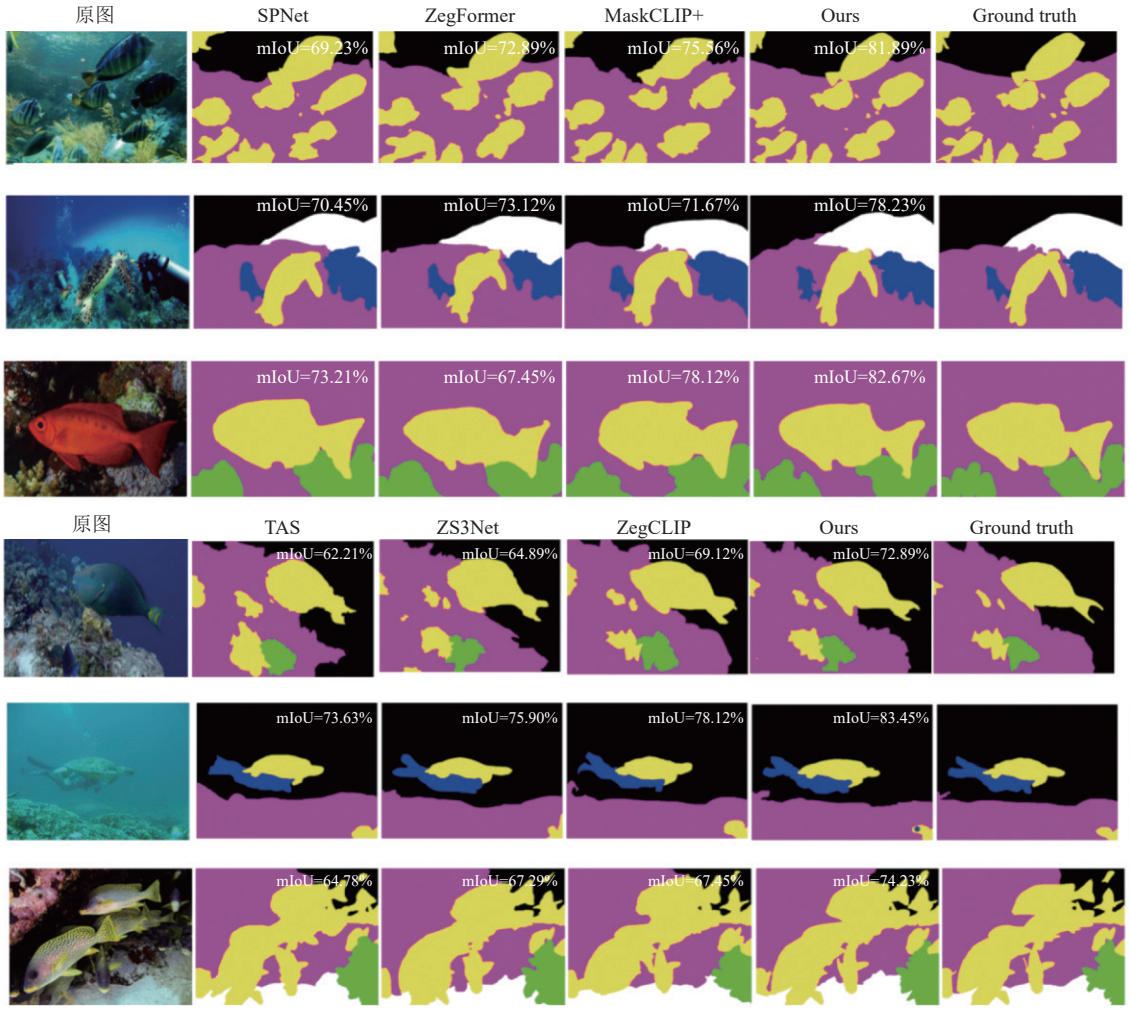

图 3 不同方法在 SUIM 数据集图像上的可视化结果

3.3.2 消融实验

(1) 为了进一步分析轻量级交叉注意力模块的作用, 我们设计了两组对照实验并分别进行了 10 次独立运行. 第 1 组对照将交叉注意力替换为简单的拼接操作, 并通过 1×1 卷积进行融合, 以评估显式交互机制相较于直接特征整合的优势. 第 2 组对照采用参数规模更大的多头交叉注意力 (multi-head cross-attention, MHCA) 替代轻量化实现, 以验证轻量化设计是否会带来性能损失.

如图 4 所示, 3 种融合方式在多次运行中表现出明显差异: Concat+1×1 Conv 整体性能最低且波动较大, 说明其难以有效建模跨模态语义; 多头交叉注意力

(MHCA) 虽具备较强表达能力, 但在性能上仅与轻量级交叉注意力 (LCA) 接近. 相比之下, 轻量级交叉注意力模块在保证较低参数开销的同时实现了稳定且优越的分割效果, 验证了轻量化设计的有效性.

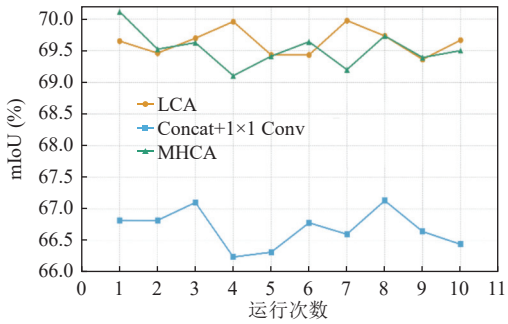

图 4 不同设置下的实验效果

(2) 为了评估我们提出的不确定性方法中超参数λ的影响, 设计了系统性的消融实验. 该参数用于调节最终一致性损失函数中的不确定性建模权重, 从而影响模型在边界细节与全局语义一致性之间的平衡. 具体地, 我们将λ分别设置为{0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5}, 并在每种设定下进行了多次实验, 最终报告平均 mIoU 结果, 以分析性能随参数变化的趋势.

如图 5 所示, 超参数λ在不确定性一致性损失中起到关键作用.

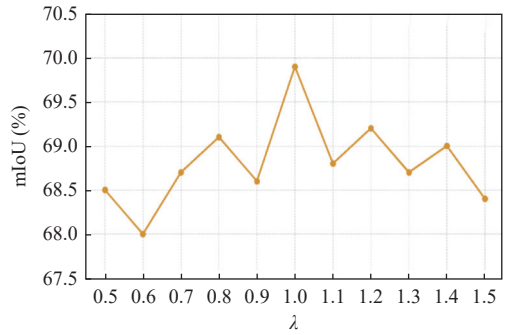

图 5 λ大小对模型的影响

从第 2.3 节中可知, λ控制了不确定性映射$M_{i,x}$的加权强度, 从而影响不同像素在损失计算中的贡献. 当λ过小时, 不确定性区域与高置信度区域差异不明显, 模型在复杂边界和细粒度结构处表现受限; 当λ过大时, 加权因子对高不确定性区域的惩罚过强, 模型容易过拟合局部噪声, 削弱全局语义一致性. 综合实验结果可见, $\lambda = 1.0$ 时模型在全局与局部建模之间达成了良好平衡, 取得了最高的 mIoU.

4 结论与展望

本文针对海洋生物图像分割中语义对齐不足与边界模糊等问题, 提出了一种基于视觉语言联合建模的零样本分割方法 Mseg. 该方法融合图像与文本语义, 引入轻量级交叉注意力与动态平衡模块增强模态交互, 并结合不确定性建模提升结构感知能力. 模型无需额外标注即可识别未见类别, 展现出良好的泛化性与应用潜力. 未来将进一步探索跨域适应与多模态预训练, 以推动智能化海洋感知的发展.

参考文献

1 Misra R, Trivedi S, Minu RI. Deep learning-based framework for enhancing underwater visual perception. Proceedings of the 2025 International Conference on Multi-agent Systems for Collaborative Intelligence (ICMSCI). Erode: IEEE, 2025. 976–982. [doi: 10.1109/icmsci62561.2025.10894484]

2 Sun K, Xiao B, Liu D, *et al*. Deep high-resolution representation learning for human pose estimation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5686–5696. [doi: 10.1109/CVPR.2019.00584]

3 Xie EZ, Wang WH, Yu ZD, *et al*. SegFormer: Simple and efficient design for semantic segmentation with Transformers. Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021. 924.

4 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 6000–6010.

5 Cheng BW, Misra I, Schwing AG, *et al*. Masked-attention mask Transformer for universal image segmentation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 1280–1289. [doi: 10.1109/CVPR52688.2022.00135]

6 Gao SN, Yang XB, Jiang L, *et al*. Global feature-based multimodal semantic segmentation. Pattern Recognition, 2024, 151: 110340. [doi: 10.1016/j.patcog.2024.110340]

7 Zhou YJ, Yang JF, Cao HZ, *et al*. Learning multi-modal scale-aware attentions for efficient and robust road segmentation. Unmanned Systems, 2024, 12(2): 201–213. [doi: 10.1142/S2301385024410048]

8 Zheng W, Zhang ZX. Weakly-supervised object localization by cutting background with deep reinforcement learning. Proceedings of the 15th Pacific Rim International Conference on Artificial Intelligence. Nanjing: Springer, 2018. 210–218. [doi: 10.1007/978-3-319-97310-4_24]

9 Long Y, Liu L, Shen FM, *et al*. Zero-shot learning using synthesised unseen visual data with diffusion regularisation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(10): 2498–2512. [doi: 10.1109/TPAMI.2017.2762295]

10 Kim W, Son B, Kim I. ViLT: Vision-and-language Transformer without convolution or region supervision. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 5583–5594.

11 Radford A, Kim JW, Hallacy C, *et al*. Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 8748–8763.

12 Zhang ZC, Ke W, Zhu Y, *et al*. Language-driven visual consensus for zero-shot semantic segmentation. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(4): 3185–3195. [doi: 10.1109/TCSVT.2024.3504816]

13 Rao YM, Zhao WL, Chen GY, *et al*. DenseCLIP: Language-guided dense prediction with context-aware prompting. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 18061–18070. [doi: 10.1109/CVPR52688.2022.01755]

14 Zhou C, Loy CC, Dai B. Extract free dense labels from CLIP. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 696–712. [doi: 10.1007/978-3-031-19815-1_40]

15 Boas SEM, Jimenez MIN, Merks RMH, *et al*. A global sensitivity analysis approach for morphogenesis models. BMC Systems Biology, 2015, 9(1): 85. [doi: 10.1186/s12918-015-0222-7]

16 Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning. New York: JMLR.org, 2016. 1050–1059.

17 Labach A, Valaee S. Regularizing neural networks by stochastically training layer ensembles. Proceedings of the 30th International Workshop on Machine Learning for Signal Processing. Espoo: IEEE, 2020. 1–6. [doi: 10.1109/MLSP49062.2020.9231761]

18 Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 6405–6416.

19 Sukesh AV, Dolz J, Lombaert H. Anatomically-aware uncertainty for semi-supervised image segmentation. Medical Image Analysis, 2024, 91: 103011. [doi: 10.1016/j.media.2023.103011]

20 Islam MJ, Edge C, Xiao Y, *et al*. Semantic segmentation of underwater imagery: Dataset and benchmark. Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas: IEEE, 2020. [doi: 10.1109/IROS45743.2020.9340821]

21 Lian S, Zhang Z, Li H, *et al*. Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset. arXiv:2406.06039, 2024.

22 Xian YQ, Choudhury S, He Y, *et al*. Semantic projection network for zero- and few-label semantic segmentation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 8248–8257. [doi: 10.1109/CVPR.2019.00845]

23 Ding J, Xue N, Xia GS, *et al*. Decoupling zero-shot semantic segmentation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2021. 11573–11582. [doi: 10.1109/CVPR52688.2022.01129]

24 Suo YC, Zhu LC, Yang Y. Text augmented spatial aware zero-shot referring image segmentation. Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: ACL, 2023. 1032–1043.

25 Bucher M, Vu TH, Cord M, *et al*. Zero-shot semantic segmentation. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: NeurIPS, 2019. 43.

26 Zhou ZQ, Lei YJ, Zhang BW, *et al*. ZegCLIP: Towards adapting CLIP for zero-shot semantic segmentation. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 11175–11185.

(校对责编: 张重毅)