武汉大学学报(理学版)
J. Wuhan Univ. (Nat. Sci. Ed.) 1~15

DOI: 10.14188/j.1671-8836.2026.0124

基于语义动作一致性的水下异物动作图生成方法

聂骁阳[1],陶 卫[2,3],曹志勇[4],谢邦天[4†]

1. 帝国理工学院 电子电气工程系,伦敦,SW7 2AZ;
2. 上海交通大学 自动化与感知学院,上海 200240;
3. 海底科学与划界全国重点实验室,上海 200240;
4. 微特技术,湖北 宜昌 443000

收稿日期:2026-06-24 † 通信联系人 E-mail:15507209862@163.com
基金项目:国家自然科学基金资助(51975374);上海交通大学深蓝基金资助项目(SL2021ZD203)
第一作者:聂骁阳,男,硕士生,现从事 ARM Cortex-M 系列动态功耗协同架构、异物识别与环境感知等方面的研究。E-mail:Nefrid@163.com

摘 要:海洋牧场水下环境中常存在废弃网片、绳索、塑料和金属件等异物,这些异物不仅可能干扰养殖设施运行,还会影响水生生物活动空间和局部生态环境。针对不同类型目标需采取清除、避让或不处理等不同处置策略的需求,现有水下目标检测和语义分割方法主要学习目标的语义类别,相应动作通常在预测后通过预定义规则获得。由于动作类别依赖语义类别预测结果,模型训练主要关注目标类别识别准确性,难以对可能导致错误动作决策的预测结果进行针对性约束。针对这一问题,本文提出一种基于语义动作一致性的水下异物动作图生成方法,将动作预测直接纳入模型训练。模型通过语义分支和动作分支分别学习目标类别和动作类别,并利用语义与动作之间的对应关系约束两种预测结果保持一致,使目标识别同时受到动作预测结果的约束。在此基础上,引入风险惩罚损失,重点抑制需避让区域和无需处理区域被误判为可清除区域的高风险错误。实验结果表明,本文方法在 SeaClear 和 SUIM 数据集上的 Decision Accuracy 分别达到 89.8% 和 91.3%,Critical Error 分别降低至 4.3% 和 3.9%,验证了该方法在水下异物动作图生成和高风险误判控制方面的有效性。

关 键 词:海洋牧场;水下异物识别;动作图生成;水下图像

中图分类号:TP18; S951.2 **文献标志码**:A **文章编号**:1671-8836(XXXX)XX-0001-15

Semantic-Action Consistency Based Action Map Generation for Underwater Foreign Objects

NIE Xiaoyang[1], TAO Wei[2,3], CAO Zhiyong[4], XIE Bangtian[4†]

1. Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, United Kingdom;
2. School of Automation and Sensing, Shanghai Jiao Tong University, Shanghai 200240, China;
3. State Key Laboratory of Submarine Geoscience, Shanghai 200240, China;
4. Weite Technologies Co., Ltd., Yichang 443000, Hubei, China

Abstract: Marine ranching environments often contain foreign objects such as abandoned nets, ropes, plastics, and metal components, which may not only interfere with the operation of aquaculture facilities but also affect the activity space of aquatic organisms and the local ecosystem. As different types of underwater targets require different treatment strategies, including removal, avoidance, or no action, existing underwater object detection and semantic segmentation methods mainly focus on learning semantic categories, while the corresponding actions are typically derived from predictions using predefined rules. Since action categories depend on semantic predictions, model training primarily emphasizes semantic recognition accuracy, making it difficult to explicitly constrain prediction errors that may lead to incorrect action decisions. To address this issue, we propose a semantic-action consis-

引用格式:聂骁阳,陶卫,曹志勇,等. 基于语义动作一致性的水下异物动作图生成方法[J]. 武汉大学学报(理学版),XXXX,XX(XX):1-15. DOI:10.14188/j.1671-8836.2026.0124.

NIE Xiaoyang, TAO Wei, CAO Zhiyong, *et al*. Semantic-Action Consistency Based Action Map Generation for Underwater Foreign Objects [J]. *J Wuhan Univ (Nat Sci Ed)*, XXXX, XX(XX): 1-15. DOI: 10.14188/j.1671-8836.2026.0124(Ch).

tency based method for underwater foreign-object action map generation, which directly incorporates action prediction into model training. The model employs a semantic branch and an action branch to learn target categories and action categories, respectively, and exploits the correspondence between semantics and actions to enforce consistency between the two predictions, allowing semantic recognition to be jointly constrained by action prediction. Furthermore, a risk penalty loss is introduced to explicitly suppress high-risk errors in which regions requiring avoidance or no action are incorrectly predicted as removable. Experimental results show that the proposed method achieves Decision Accuracy values of 89.8% and 91.3% on the SeaClear and SUIM datasets, respectively, while reducing Critical Error to 4.3% and 3.9%. These results demonstrate the effectiveness of the proposed method in underwater foreign-object action map generation and high-risk error control.

Key words: marine ranching; underwater foreign object recognition; action map generation; underwater image

0 引 言

现代海洋牧场[1]是蓝色经济发展和海洋资源可持续利用的重要载体。随着海洋牧场规模不断扩大,水下养殖设施、人工鱼礁和生态修复区域的运行维护需求逐渐增加。在长期运行过程中,塑料、金属件、废弃网片和绳索等异物容易受水流作用发生漂移、缠绕和沉积,不仅可能干扰养殖设施正常运行,还可能影响水生生物活动空间和局部生态环境。因此,如何准确识别水下异物并为后续清除作业提供可靠的视觉信息,是海洋牧场智能化运维中的重要问题。

水下异物识别主要依赖水下图像理解[2-3]。准确识别水下图像中的目标,是开展水下环境感知和异物清除辅助的重要基础。然而,水下图像通常存在光照不足、颜色偏移、图像模糊和背景复杂等问题,这些因素会降低视觉模型的识别稳定性。同时,海洋牧场中的目标类型复杂,不同目标对应的处置需求存在明显差异。如塑料、瓶子、绳索和金属件等目标通常需要清除,鱼类、水草、礁石和养殖设施等目标需要避让,而水体和普通背景区域无需处理。因此,面向水下清除任务,视觉模型不仅需要识别目标类别,还需要进一步判断不同区域对应的动作类别。

现有水下目标检测和语义分割方法[4-5]主要关注目标类别识别。目标检测和语义分割方法能够分别提供目标位置与类别、实例级掩码以及像素级语义信息,为水下环境感知提供重要基础。水下异物识别方法[6-8]通常需要区分鱼类、水草、礁石、机器人、海底背景和海洋垃圾等多类目标,但其主要目标仍是获得准确的语义类别。对于异物清除任务而言,不同语义类别对应不同的处置策略,仅依赖语义预测结果难以直接给出清除、避让和无需处理等动作层面的判断。因此,如何在目标识别的基础上进一步建立语义类别与动作类别之间的联系,是水下异物清除任务需要解决的问题。

早期水下目标识别方法[8-9]主要依赖颜色、纹理、边缘和形状等人工特征。这类方法计算过程较为简单,但特征表达能力有限,在水下光照和背景发生明显变化时容易失效。随着深度学习的发展,卷积神经网络逐渐成为水下图像识别和分割任务的主要方法[10-11]。U-Net[12]采用编码器-解码器结构,并通过跳跃连接融合浅层空间信息和深层语义信息,能够较好地恢复目标边界。DeepLabV3+[13]通过空洞卷积和多尺度上下文建模增强语义表达能力,并利用解码结构改善分割边界。SegFormer[14]将层次化Transformer编码器与轻量级MLP解码器结合,进一步增强了大范围上下文信息的建模能力。

尽管上述方法能够有效提升水下目标识别和语义分割性能,但现有水下语义分割研究仍主要关注像素级类别预测[15]。在实际清除作业中,作业系统不仅需要判断目标“是什么”,还需要确定其是否应该清除、避让或无需处理。目前方法通常以语义类别作为最终输出,难以直接形成相应动作决策。若仅在推理阶段利用预定义规则将语义类别转换为动作,模型在训练过程中难以利用语义类别与动作类别之间已有的对应关系对预测结果进行约束。同时,当不同数据集具有不同的语义类别体系时,依赖人工规则进行后处理也会增加动作转换的复杂性。因此,仅依赖语义类别预测的方法难以满足面向水下清除任务的动作图生成需求。为此,有必要在目标识别过程中进一步考虑语义类别与动作类别之间的关联,使模型能够同时获得语义信息和动作信息。

从任务关系来看,细粒度语义识别与动作预测具有较强的相关性。二者均依赖水下图像中的视觉特征,同时语义类别与动作类别之间存在由清除需求确定的对应关系。因此,将语义识别与动作预测进行联合建模,可以利用两类监督信息之间的关联。多任务学习通常通过参数共享实现相关任务之间的信息共享。经典多任务学习方法多采用硬参数共享方式[16],即多个任务共享部分网络参数。

近年来,研究者进一步探索任务自适应参数共享方法[17],以提高不同任务之间参数共享的灵活性。随着深度神经网络的发展,多任务学习已广泛应用于视觉任务,包括水下目标检测与跟踪[18-19]、实例分割[20-21]以及语义分割[22-24]等。此类方法通常通过共享特征提取模块和多个任务分支实现不同任务之间的信息共享,但由于不同任务对视觉特征的关注重点不同,简单的特征共享仍难以充分利用任务之间的关联关系。

为缓解硬参数共享对不同任务特征表达的限制,Misra等[25]提出Cross-Stitch Networks框架,通过可学习的cross-stitch单元融合不同任务网络的特征,使模型能够自适应调节任务之间的信息共享程度。与硬共享方法[26]相比,软共享方法能够保留更多任务特有信息,但网络结构和训练过程也更加复杂。除特征共享方式外,多任务学习还需要解决不同任务损失之间的平衡问题。Kendall等[27]提出基于同方差不确定性的损失加权方法,使模型能够自适应调整不同任务的损失权重;GradNorm[28]则从梯度角度平衡不同任务的训练速度,通过调整梯度幅值缓解多任务联合训练中的优化不均衡问题。

上述多任务学习方法主要关注不同任务之间的特征共享、损失权重平衡和联合优化,而较少显式利用不同任务输出之间具有确定结构关系的先验信息。在水下异物清除任务中,语义识别和动作预测并非两个相互独立的任务,而是具有由清除需求决定的明确对应关系。细粒度语义类别为动作图生成提供目标信息,动作类别则直接反映不同区域的清除需求。因此,仅通过共享视觉特征进行联合学习,仍难以充分利用语义类别与动作之间的结构关系。

针对上述问题,本文提出一种基于语义动作一致性的水下异物动作图生成方法,将动作预测直接纳入模型训练。与此同时,本文构建统一的动作标签空间,将不同水下数据集中的原始语义类别分别映射为Remove、Avoid和No action三类动作标签。其中,Remove表示可清除异物区域,包括塑料、废弃网片、绳索和金属件等;Avoid表示需避让对象区域,包括鱼类、水草、礁石和养殖设施等;No action表示无需处理区域,如水体、普通背景。模型通过语义分支和动作分支分别学习目标类别和动作类别,并利用语义到动作的映射矩阵将语义预测转换为语义引导的动作概率。在此基础上,本文设计语义动作一致性,通过语义动作一致性减小两种动作概率分布之间的差异,促使其在训练过程中趋于一致,使目标识别同时受到动作预测结果的约束。进一步地,考虑到不同误判类型在水下清除任务中的风险存在差异,本文引入风险惩罚损失,重点抑制Avoid或No action被误判为Remove的高风险错误。

本文的主要贡献如下:第一,提出一种面向水下异物处置的像素级动作图生成方法,将动作预测直接纳入模型训练,使模型能够直接输出符合清除任务需求的动作区域。第二,提出语义动作一致性,通过语义到动作的映射关系连接语义分支和动作分支,使语义预测与动作预测保持一致,并利用动作预测结果进一步约束目标识别。第三,设计风险惩罚损失,根据不同误判类型的风险差异重点约束Avoid和No action被误判为Remove的高风险错误,从而降低视觉识别层面的误清除风险。

1 本文方法

1.1 问题定义

现有语义分割方法在语义预测后经预定义规则获得动作,难以在训练阶段直接约束动作决策。为此,本文将水下异物识别由语义类别预测进一步扩展为面向清除任务的动作图生成任务。假设I为一张水下RGB图像,$I\in\mathbb{R}^{H\times W\times 3}$。其中,$H$和$W$分别表示图像的高度和宽度,3表示RGB三个颜色通道。

Y_{a}为模型需要预测的动作图:

$$Y_{\mathrm{a}}=[\,y_{i,j}^{\mathrm{a}}\,]_{H\times W},y_{i,j}^{\mathrm{a}}\in\{1,\cdots,C_{\mathrm{a}}\}^{H\times W} \tag{1}$$

其中,Y_{a}表示像素级动作标签图,$y_{\mathrm{a},ij}$表示位置(i,j)处的动作标签,C_{a}表示动作类别数量。$C_{\mathrm{a}}=3$,分别对应Remove、Avoid和No action。

不同水下数据集具有不同的语义类别体系。为统一不同数据集的动作表示,本文将各数据集中的语义标签映射到统一的动作标签空间,使模型能够在一致的动作图生成目标下进行学习。

原始语义标签图为:

$$Y_{\mathrm{s}}=[\,y_{i,j}^{\mathrm{s}}\,]_{H\times W},y_{i,j}^{\mathrm{s}}\in\{1,\cdots,C_{\mathrm{s}}\}^{H\times W} \tag{2}$$

其中,C_{s}表示对应数据集的语义类别数量,$y_{\mathrm{s},ij}$表示位置(i,j)处的语义标签。由于不同数据集具有不同的语义类别体系,C_{s}的取值由具体数据集决定。本文定义语义到动作的映射函数$M(\cdot)$,将各数据集的原始语义类别统一映射为Remove、Avoid和No action三类动作标签:

$$Y_{\mathrm{a}}=M(Y_{\mathrm{s}}) \tag{3}$$

其中,$M(\cdot)$对Y_{s}中的语义标签进行逐像素映射,即$y_{\mathrm{a},ij}=M\left(y_{\mathrm{s},ij}\right)$,经过映射后,$Y_{\mathrm{a}}$中的每个像素均属于Remove、Avoid和No action三类动作之一。

为实现统一动作空间下语义识别与动作预测的联合学习,本文提出水下动作图生成网络(Underwater Cleanup Decision Network,UCD-Net)。

1.2 UCD-Net整体框架

UCD-Net以水下图像作为输入,并直接输出像素级动作图。如图1所示,本文提出的UCD-Net采用共享编码器和双分支预测结构。共享特征编码器提取水下图像特征,双分支预测中的语义预测分支和动作预测分支可分别得到语义概率图和动作概率图。在训练阶段,利用语义动作一致性约束建立两种预测结果之间的联系。

具体步骤:输入图像首先经过共享特征编码器得到视觉特征,随后将该特征分别送入语义预测分支和动作预测分支,生成语义概率图和动作概率图。为利用语义类别与动作类别之间的对应关系,本文先通过映射矩阵$\boldsymbol{A}$将语义分支的预测结果转换为对应的动作概率;然后,利用一致性损失约束该动作概率,与动作分支直接预测的动作概率保持一致,使语义监督信息能够参与动作预测过程,并在训练阶段加强语义类别与动作类别之间的关联。最终,模型通过动作预测分支输出Remove、Avoid和No action三类像素级动作图,为水下异物清除任务提供动作决策依据。

给定输入图像$\boldsymbol{I}$,共享特征编码器首先提取视觉特征:

$$F=E(I) \tag{4}$$

其中,$E(\cdot)$为共享特征编码器,F为提取到的图像特征。

语义预测分支根据共享特征预测语义类别概率图:

$$P_{\mathrm{s}}=H_{\mathrm{s}}(F) \tag{5}$$

其中,$H_{\mathrm{s}}(\cdot)$表示语义预测头,$P_{\mathrm{s}}\in\mathbb{R}^{C_{\mathrm{s}}\times H\times W}$表示语义类别概率图,$C_{\mathrm{s}}$表示语义类别数量。

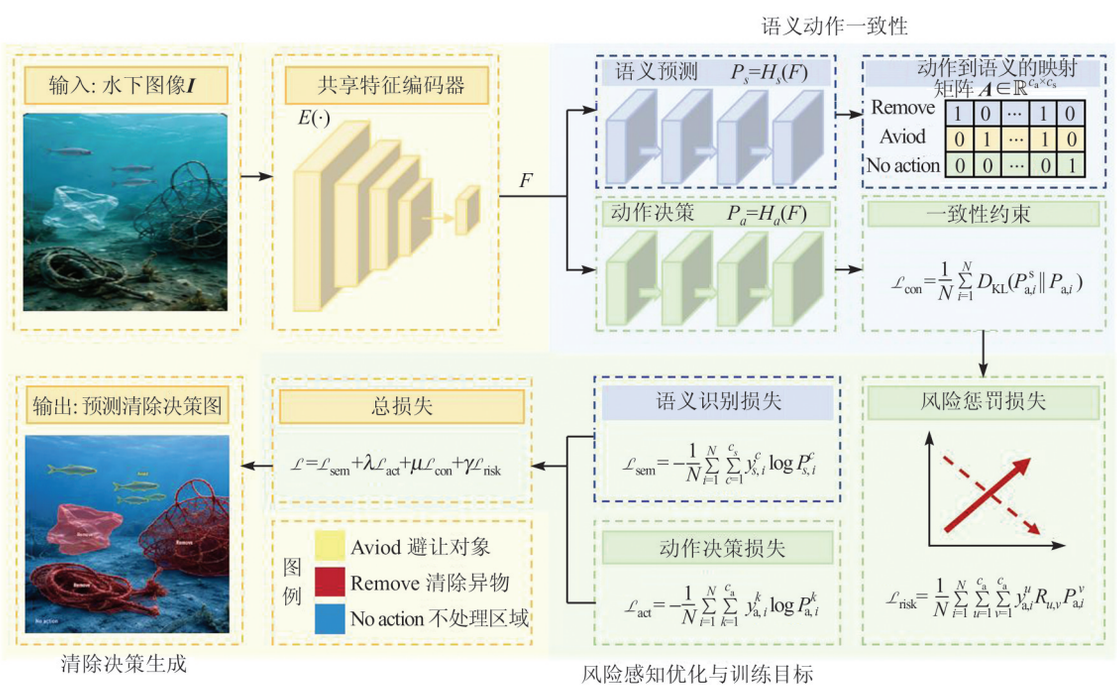


图1 UCD-Net整体框架图

Fig. 1 The overall framework of UCD-Net

动作预测分支根据共享特征预测动作概率图:

$$P_{\mathrm{a}}=H_{\mathrm{a}}(F) \tag{6}$$

其中,$H_{\mathrm{a}}(\cdot)$表示动作预测头,$P_{\mathrm{a}}\in\mathbb{R}^{C_{\mathrm{a}}\times H\times W}$表示动作类别概率图。

UCD-Net在训练阶段同时学习语义类别和动作类别。语义分支提供目标理解能力。动作分支提供任务决策能力。一致性将两类预测结果联系起来,从而形成统一的语义动作学习过程。

1.3 语义动作一致性

仅使用语义监督时,动作通常由语义预测结果经预定义规则获得,模型训练主要依据语义类别误差进行优化,难以直接约束可能导致错误动作决策的预测结果。仅使用动作监督时,多种不同目标被归入相同的动作类别,模型难以充分利用原始语义标签提供的目标类别信息,从而影响对复杂水下目标的区分。为同时利用目标类别信息和动作信息,本文设计语义动作一致性学习机制。

为了建立语义类别和动作之间的对应关系，本文构建语义到动作的映射矩阵：

$$\boldsymbol{A}\in\{0,1\}^{C_a\times C_s} \tag{7}$$

当语义类别 c 对应动作类别 k 时，矩阵元素 $\boldsymbol{A}_{k,c}=1$，否则 $\boldsymbol{A}_{k,c}=0$。该矩阵将语义类别空间转换为动作类别空间。

对于第 i 个像素，语义预测分支输出的概率向量为 $\boldsymbol{P}_{s,i}$。通过映射矩阵 $\boldsymbol{A}$，可以得到由语义预测推导出的动作概率：

$$\boldsymbol{P}_{a,i}^{s}=\boldsymbol{A}\boldsymbol{P}_{s,i} \tag{8}$$

其中，$\boldsymbol{P}_{a,i}^{s}$ 表示语义引导的动作概率，由语义预测结果经过动作映射得到。

为减少语义预测与动作预测之间的冲突，本文对语义分支推导得到的动作概率 $\boldsymbol{P}_{a}^{s}$ 与动作分支直接输出的动作概率 $\boldsymbol{P}_{a}$ 施加一致性约束。当两者的预测概率分布不一致时，通过一致性损失对其差异进行惩罚，差异越大，对应的损失越大，从而在训练过程中促使两个分支的动作预测结果逐渐一致。其损失函数定义为：

$$\mathcal{L}_{con}=\frac{1}{N}\sum_{i=1}^{N}D_{KL}\left(\boldsymbol{P}_{a,i}^{s}\|\boldsymbol{P}_{a,i}\right) \tag{9}$$

其中，$D_{KL}(\cdot)$ 表示 KL 散度，N 表示像素数量。

如图 2 所示，语义概率 $\boldsymbol{P}_{s}$ 首先通过映射矩阵 $\boldsymbol{A}$ 被转换为语义引导动作概率 $\boldsymbol{P}_{a}^{s}$。动作分支直接输出动作概率 $\boldsymbol{P}_{a}$。一致性通过 KL 散度度量 $\boldsymbol{P}_{a}^{s}$ 与 $\boldsymbol{P}_{a}$ 之间的分布差异。该约束使语义预测结果能够为动作决策提供结构引导，并减少两条分支之间的预测冲突。

动作标签由语义标签经过固定映射得到，因此动作监督并非独立人工标注。本文方法的目的不是利用两个独立标签源进行传统多任务学习，而是将细粒度语义标签重构为面向清除需求的低维动作标签空间，并通过双路径预测和一致性对动作空间进行结构正则化。

1.4 风险感知优化与训练目标

水下清除任务中不同动作误判的实际代价并不相同。为使模型在训练过程中重点抑制可能导致误清除的高风险错误，本文引入风险惩罚损失，并通过动作风险矩阵描述不同误判类型的风险差异。

本文按照 Remove、Avoid 和 No action 的类别顺序定义动作风险矩阵：

$$\boldsymbol{R}=\begin{bmatrix}0&1&1\\3&0&1\\2&1&0\end{bmatrix} \tag{10}$$

其中，矩阵的行和列均按照 Remove、Avoid 和 No action 的顺序排列，$\boldsymbol{R}_{u,v}$ 表示真实动作类别为 u、预测动作类别为 v 时对应的风险权重。对角线上的元素设为 0，表示预测正确时不施加额外风险惩罚。Avoid→Remove 的风险权重设为 3，以重点抑制需避让对象被误判为可清除区域的情况；No action→Remove 的风险权重设为 2，用于降低无需处理区域被错误清除的风险；其余误判的风险权重设为 1。该设置使模型在训练过程中对高风险误判施加更强惩罚，同时保留对其他错误类型的基本约束。

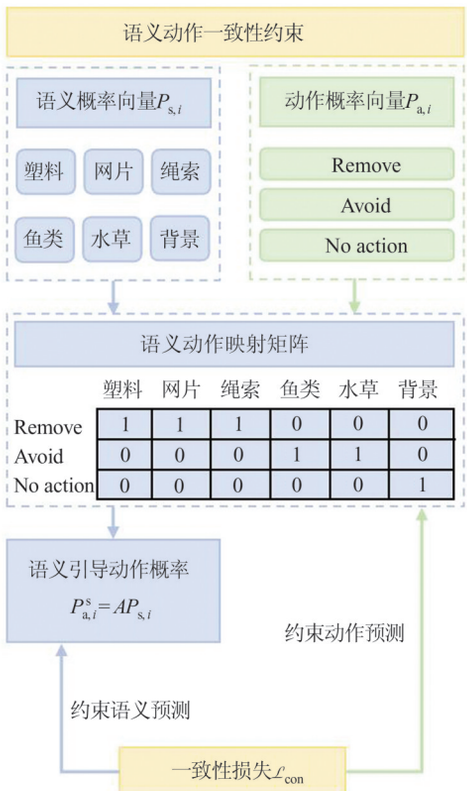


图 2 语义动作一致性结构图

Fig. 2 Structure of semantic-action consistency

基于风险矩阵 $\boldsymbol{R}$，风险惩罚损失定义为：

$$\mathcal{L}_{risk}=\frac{1}{N}\sum_{i=1}^{N}\sum_{u=1}^{C_a}\sum_{v=1}^{C_a}y_{a,i}^{u}\boldsymbol{R}_{u,v}\boldsymbol{P}_{a,i}^{v} \tag{11}$$

其中，$y_{a,i}^{u}$ 表示第 i 个像素真实动作标签在类别 u 上的 one-hot 编码分量，$\boldsymbol{P}_{a,i}^{v}$ 表示模型将该像素预测为动作类别 v 的概率。

该损失直接约束高风险预测概率。当真实类别为 Avoid，而模型对 Remove 给出较高概率时，损失值会明显增加。该设计能够降低需避让对象被误清除的风险。

UCD-Net 的整体训练目标由语义识别损失（$\mathcal{L}_{sem}$）、动作决策损失（$\mathcal{L}_{act}$）、语义动作一致性损失（$\mathcal{L}_{con}$）和风险惩罚损失（$\mathcal{L}_{risk}$）组成：

$$\mathcal{L}=\mathcal{L}_{sem}+\lambda\mathcal{L}_{act}+\mu\mathcal{L}_{con}+\gamma\mathcal{L}_{risk} \tag{12}$$

其中，λ、μ 和 γ 分别表示动作决策损失、语义动作一

致性损失和风险惩罚损失的权重系数。

语义识别损失定义为:

$$\mathcal{L}_{\text{sem}}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C_s}y_{\text{s},i}^{c}\log \boldsymbol{P}_{\text{s},i}^{c} \tag{13}$$

动作决策损失定义为:

$$\mathcal{L}_{\text{act}}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C_a}y_{\text{a},i}^{k}\log \boldsymbol{P}_{\text{a},i}^{k} \tag{14}$$

其中,$\mathcal{L}_{\text{sem}}$保留原始语义类别监督,$\mathcal{L}_{\text{act}}$直接优化动作预测,$\mathcal{L}_{\text{con}}$建立语义类别与动作类别之间的结构一致性,$\mathcal{L}_{\text{risk}}$抑制高风险错误。四个部分共同服务于动作图生成目标。

1.5 动作图生成

模型输入一张水下图像,并直接输出动作概率图P_{a}。最终动作图由动作决策获得:

$$\hat{Y}_{\text{a},i}=\operatorname{argmax}_k \boldsymbol{P}_{\text{a},i}^{k} \tag{15}$$

其中,$\hat{Y}_{\text{a},i}$表示第i个像素的预测动作类别,$\boldsymbol{P}_{\text{a},i}^{k}$表示动作决策分支将第$i$个像素预测为第$k$类动作的概率,argmax表示选取概率最大值所对应的动作类别。若预测类别为Remove,则该区域被判定为可清除异物区域;若预测类别为Avoid,则该区域被判定为需避让对象;若预测类别为No action,则该区域被判定为不需要处理区域。

与DeepLabV3+[13]、SegFormer[14]、Mask R-CNN[29]和YOLO-seg[30]等主要面向语义类别或实例区域预测的方法相比,本文进一步将动作预测直接纳入水下异物识别模型的训练过程。现有方法通常在语义预测后通过预定义规则获得动作,动作结果依赖语义预测且不直接参与训练,因此难以针对可能导致错误动作决策的预测结果进行约束。本文在训练阶段直接学习Remove、Avoid和No action三类动作,并通过语义动作一致性使语义预测与动作预测相互约束,同时利用风险信息重点抑制高风险动作误判。

1.6 算法流程

算法1给出了UCD-Net的水下异物动作图生成流程。

算法1 UCD-Net水下异物识别动作图生成算法

输入:水下图像(I),原始语义标签(Y_{s}),动作标签(Y_{a}),语义到动作映射矩阵($\boldsymbol{A}$),动作风险矩阵($\boldsymbol{R}$),损失权重(λ、μ、γ)

输出:水下异物动作图($\hat{Y}_{\text{a}}$),其中每个像素对应Remove、Avoid或No action

将$\boldsymbol{I}$输入共享特征编码器,提取图像特征(F)

将F分别输入语义预测分支和动作预测分支,得到语义概率图($\boldsymbol{P}_{\text{s}}$)和动作概率图($\boldsymbol{P}_{\text{a}}$)

利用$\boldsymbol{A}$将$\boldsymbol{P}_{\text{s}}$转换为语义引导的动作概率图($\boldsymbol{P}_{\text{a}}^{s}$)

根据$\boldsymbol{P}_{\text{s}}$与$\boldsymbol{Y}_{\text{s}}$计算语义识别损失$\mathcal{L}_{\text{sem}}$

根据$\boldsymbol{P}_{\text{a}}$与$\boldsymbol{Y}_{\text{a}}$计算动作决策损失$\mathcal{L}_{\text{act}}$

根据$\boldsymbol{P}_{\text{a}}^{s}$与$\boldsymbol{P}_{\text{a}}$计算语义动作一致性损失$\mathcal{L}_{\text{con}}$

根据风险矩阵$\boldsymbol{R}$计算风险惩罚损失$\mathcal{L}_{\text{risk}}$,重点约束Avoid或No action被误判为Remove的情况

计算总损失:$\mathcal{L}=\mathcal{L}_{\text{sem}}+\lambda\mathcal{L}_{\text{act}}+\mu\mathcal{L}_{\text{con}}+\gamma\mathcal{L}_{\text{risk}}$

利用$\mathcal{L}$更新模型参数

训练完成后,将待测试图像输入模型,获得$\boldsymbol{P}_{\text{a}}$

对每个像素选取概率最大的动作类别,生成$\hat{Y}_{\text{a}}$

输出$\hat{Y}_{\text{a}}$,其中Remove表示可清除区域,Avoid表示需避让区域,No action表示不处理区域

2 实验与结果分析

2.1 实验数据集

实验使用两个公开水下视觉数据集,分别为SeaClear Marine Debris(以下简称SeaClear)数据集[31]和SUIM数据集[32]。SeaClear数据集主要包含废弃瓶罐、绳索、塑料制品、金属制品、水生动物、海底植被和机器人部件等水下目标。根据本文面向水下清除任务定义的动作标签,废弃瓶罐、绳索、塑料制品和金属制品等需要清除的目标被映射为Remove,水生动物、海底植被和机器人部件等需要避让的目标被映射为Avoid。SUIM数据集主要面向水下场景语义分割任务,数据中包含水体、鱼类、水草、机器人、礁石和海底等语义类别。该数据集适合验证模型对水下背景、需避让对象和普通环境区域的理解能力。

SeaClear数据集共包含8 610张水下图像。参考原数据集论文的训练集与测试集划分方式,本文将其中6 888张图像作为训练和验证数据,1 722张图像作为测试数据,并进一步从训练和验证数据中划分861张图像作为验证集。最终,SeaClear数据集的训练集、验证集和测试集的数据分别为6 027、861和1 722,对应比例为7∶1∶2。SUIM数据集共包含1 635张图像,其中官方提供1 525张训练和验证图像以及110张独立测试图像。本文将官方训练和验证数据按照8∶2的比例进行划分,最终得到1 220张训练图像、305张验证图像和110张测试图像。

本文根据水下清除作业需求,将两个数据集中的原始语义类别统一映射为Remove、Avoid和No action三类动作标签。SeaClear数据集包含40个目标类别,并将背景作为独立语义类别,因此$C_{\text{s}}^{\text{SeaClear}}=41$;其中,废弃瓶罐、绳索、塑料制品、金属

制品等可清除异物映射为Remove,水生动物、海底植被、机器人部件等需保护或避让目标相关类别映射为Avoid,水体、普通背景映射为No action。SUIM数据集包含8个语义类别,$C_s^{\mathrm{SUIM}}=8$。具体映射关系如表1所示。通过该映射方式,不同数据集可以在统一动作生成空间下进行训练和评价。

表1 不同数据集的语义类别到动作类别映射关系

Table 1 Mapping from semantic categories to action categories across different datasets

数据集	原始语义类别	动作类别
SeaClear	废弃瓶罐、绳索、塑料制品、金属制品等可清除异物	Remove
	水生动物、海底植被、机器人部件等需保护或避让目标	Avoid
	水体、普通背景	No action
SUIM	Wrecks and ruins (WR)	Remove
	Human divers (HD)、Aquatic plants and sea-grass (PF)、Robots (RO)、Reefs and invertebrates (RI)、Fish and vertebrates (FV)	Avoid
	Background/waterbody (BW)、Sea-floor and rocks (SR)	No action

2.2 实验设置与评价指标

为保证实验对比的公平性,本文在相同的数据划分、输入尺寸和训练策略下训练所有对比方法与本文方法。实验中,所有输入图像统一缩放至512×512。本文采用SegFormer[14]中的Mix Transformer-B0(MiT-B0)作为共享特征编码器,并采用轻量级解码头分别构建语义预测分支和动作预测分支。语义分支用于预测原始语义类别,动作分支用于预测Remove、Avoid和No action三类像素级动作区域。所有消融变体均采用相同的backbone和解码结构,仅改变损失项或监督方式,以减少网络结构差异对实验结果的影响。模型采用AdamW优化器进行训练,初始学习率和权重衰减均设置为1.0×10^{-4},batch size设置为8,训练轮数设置为100 epochs。训练过程中采用随机翻转、随机缩放、颜色扰动和模糊扰动等数据增强方式,以提高模型对水下光照变化、图像浑浊和目标尺度变化的适应能力。

为评估模型的计算复杂度与推理效率,本文还统计了参数量、FLOPs和推理速度。以512×512输入图像为基准,UCD-Net的参数量约为4.2×10^6,单次前向推理的计算量约为9.1 G FLOPs(浮点运算次数)。在单张NVIDIA RTX 3090 GPU上,以batch size为1进行测试时,模型平均推理速度约为58.9 FPS(每秒帧数)。FPS统计不包括图像读取和结果保存时间。上述结果表明,UCD-Net具有较低的参数规模和计算开销,并能够保持较高的推理效率。

本文从分割精度、动作图整体预测正确性、关键动作类别识别性能和误清除风险四个方面评价模型性能。其中,分割精度用于衡量预测区域与真实区域之间的重合程度,动作图整体预测正确性用于评价所有有效像素上的动作分类结果,关键动作类别识别性能用于评价不同动作类别的识别能力,误清除风险用于衡量需避让区域被误判为可清除区域的情况。

所有指标均以像素为统计单位,并在测试集有效标注像素集合Ω上计算,无效标注像素不参与统计。设第i个像素的真实动作标签为y_i,预测动作标签为$\hat{y}_i$,且$y_i,\hat{y}_i\in\{1,\cdots,C_a\}$。对于任意动作类别$c\in\{1,\cdots,C_a\}$,$\mathrm{TP}_c$、$\mathrm{FP}_c$和$\mathrm{FN}_c$分别表示类别$c$的真正例、假正例和假负例。

mIoU用于衡量预测动作区域与真实动作区域之间的平均重叠程度,定义为:

$$\mathrm{mIoU}=\frac{1}{|C_a|}\sum_{c\in C_a}\frac{\mathrm{TP}_c}{\mathrm{TP}_c+\mathrm{FP}_c+\mathrm{FN}_c}$$

Decision Accuracy用于评价模型对所有有效像素动作类别的整体预测正确率,定义为:

$$\text{Decision Accuracy}=\frac{\sum_{i\in\Omega}\mathrm{I}(\widehat{y_i}=y_i)}{|\Omega|}\times100\%$$

其中,I(·)为指示函数,当$\widehat{y_i}=y_i$,即第i个像素的预测动作类别与真实动作类别一致时,$\mathrm{I}(\widehat{y_i}=y_i)=1$,否则取0。

Remove $F1$和Avoid $F1$分别用于评价模型对可清除异物区域和需避让对象区域的识别效果。对于任意动作类别c,Precision、Recall和$F1$定义为:

$$\mathrm{Precision}_c=\frac{\mathrm{TP}_c}{\mathrm{TP}_c+\mathrm{FP}_c}\times100\%$$

$$\mathrm{Recall}_c=\frac{\mathrm{TP}_c}{\mathrm{TP}_c+\mathrm{FN}_c}\times100\%$$

$$F1_c=\frac{2\cdot\mathrm{Precision}_c\cdot\mathrm{Recall}_c}{\mathrm{Precision}_c+\mathrm{Recall}_c}$$

其中,TP_c、FP_c和FN_c分别表示动作类别c的真正例、假正例和假负例。当c分别取Remove和Avoid

时,所得$F1$分别记为Remove $F1$和Avoid $F1$。

在水下异物动作图生成任务中,Critical Error是衡量模型安全性的重要指标。该指标专门用于衡量需避让对象或不处理区域被误识别为可清除区域的高风险错误。模型将可清除异物误判为Avoid或No action,会降低清除完整性;而模型将鱼类、水草、礁石、养殖设施或普通背景误判为Remove,则可能产生视觉层面的误清除风险。因此,本文重点统计Avoid和No action被误判为Remove的高风险错误。Critical Error定义为:

$$\text{Critical Error}=\frac{\sum_{i=1}^{N}\left(y_i\in\{\text{Avoid},\text{No action}\}\right)\cdot\left(\hat{y}_i=\text{Remove}\right)}{\sum_{i=1}^{N}\left(y_i\in\{\text{Avoid},\text{No action}\}\right)}\times 100\%$$

其中,Critical Error的分母为测试集中真实类别属于Avoid或No action的有效像素总数,分子为其中被误判为Remove的像素数。Critical Error越小,说明模型在视觉识别层面的误清除风险越低。

为增强实验结果的可信度,本文在固定数据划分下进行多次独立实验。SeaClear和SUIM数据集均按照训练集、验证集和测试集进行划分,验证集用于模型选择,测试集仅用于最终性能评价。所有对比方法和本文方法均使用相同的数据划分、输入尺寸、训练轮数、优化器和数据增强策略。为降低随机初始化和数据加载顺序对结果的影响,对每种方法进行3次独立训练。最终实验结果以均值和标准差的形式报告。

2.3 对比实验

本文选取语义分割方法和实例分割方法作为对比模型。语义分割方法采用U-Net[12]、DeepLabV3+[13]和SegFormer[14]。实例分割方法采用Mask R-CNN[29]和YOLO-seg[30]。本文先使用模型预测原始语义类别或实例类别,再根据人工规则将预测结果映射为Remove、Avoid和No action三类动作标签。因此,这类方法记为"+Rule"。

Mask R-CNN和YOLO-seg更适合具有实例级标注的数据集。SUIM数据集主要提供语义分割标注,因此本文在SUIM上主要与语义分割方法进行比较。

2.3.1 SeaClear数据集

SeaClear数据集主要用于评价模型对水下异物和可清除目标的识别能力。实验结果如表2所示。

从表2可以看出,Mask R-CNN和YOLO-seg在SeaClear数据集上取得了较好的Remove $F1$,说明实例分割方法对可清除异物区域具有较好的识别能力。然而,这类方法通常需要根据语义或目标类别预测结果,通过预定义规则进一步获得动作类别。当语义预测出现偏差时,错误可能继续传递至动作判断,且不同动作误判带来的风险未被直接纳入模型训练过程。

相比之下,本文方法在mIoU、Decision Accuracy、Remove $F1$和Avoid $F1$上分别达到79.9%、89.8%、86.7%和85.4%,均优于其他对比方法,同时Critical Error降低至4.3%±0.3%,为所有方法中最低。结果表明,本文方法不仅能够提高动作区域的整体分割精度和像素级决策准确率,还能够同时提升对可清除区域和需避让区域的识别能力,并有效减少高风险动作误判。进一步说明,语义动作一致性有助于利用语义信息改善动作预测,而风险惩罚损失能够重点抑制需避让对象被误判为Remove的情况,从而提升动作图生成的整体可靠性。

表2 不同方法在SeaClear数据集上的动作图生成性能对比

Table 2 Comparison of action map generation performance of different methods on the SeaClear dataset

%

方法	mIoU ↑	Decision Acc ↑	Remove $F1$ ↑	Avoid $F1$ ↑	Critical Error ↓
U-Net[12]+Rule	66.9±0.3	80.6±0.3	78.4±0.3	76.8±0.4	9.6±0.2
DeepLabV3+[13]+Rule	69.9±0.2	82.4±0.1	80.6±0.2	78.9±0.3	8.9±0.1
SegFormer[14]+ Rule	74.0±0.2	85.3±0.2	83.2±0.3	81.6±0.2	7.1±0.2
Mask R-CNN[29]+Rule	73.2±0.3	84.6±0.5	84.5±0.1	80.8±0.4	8.0±0.1
YOLO-seg[30]+Rule	75.8±0.3	86.1±0.3	85.0±0.2	82.3±0.2	7.4±0.1
本文方法	79.9±0.2	89.8±0.1	86.7±0.3	85.4±0.2	4.3±0.3

如图3所示,原始图像包含鱼类、珊瑚、潜水员、废弃绳索和金属结构等多类目标。场景背景较为复杂,且不同目标对应的处理动作存在明显差异。由语义标注图可以看出,本文方法能够区分水体背景、需避让目标和可清除异物。其中,鱼类、珊瑚、礁石和潜水员被划分为需避让区域,废弃绳索和金

属件被划分为可清除区域。该结果表明,本文方法能够在复杂水下生态场景中有效区分可清除目标和需避让对象,并生成面向水下清除任务的动作图。

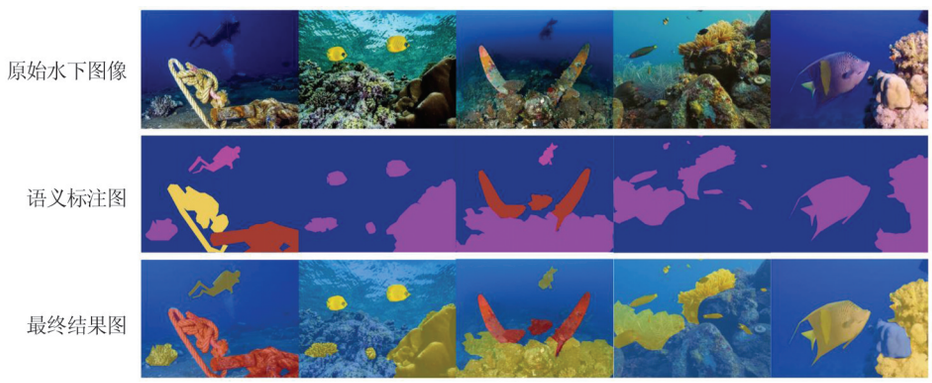


图3 SeaClear数据集可视化结果图

Fig. 3 Visualization results on the SeaClear dataset

2.3.2 SUIM数据集

SUIM数据集主要用于评价模型对复杂水下背景和需避让对象的理解能力。与SeaClear数据集相比,SUIM数据集中可清除异物类别较少。因此,本文在SUIM数据集上重点关注Avoid类别识别、No action区域识别和Critical Error。实验结果如表3所示。

从表3可以看出,SegFormer+Rule的整体表现优于U-Net+Rule和DeepLabV3++Rule。这说明全局上下文信息有助于复杂水下场景理解。本文方法的Decision Accuracy比SegFormer+Rule的高,Critical Error从6.8%±0.3%降低到3.9%±0.1%。该结果说明,直接学习动作决策比单纯依赖语义分割后处理更加稳定。

同时,本文方法在Avoid $F1$上取得较好结果。这表明模型能够较好地区分需避让对象和不处理区域。该能力对于水下清除任务具有实际意义,因为需避让对象的误判通常会带来更高的作业风险。

表3 不同方法在SUIM数据集上的动作图生成性能对比

Table 3 Comparison of action map generation performance of different methods on the SUIM dataset

%

方法	mIoU ↑	Decision Acc ↑	Avoid $F1$(%) ↑	Critical Error ↓
U-Net[12]+Rule	63.7±0.2	82.6±0.4	80.1±0.3	8.7±0.2
DeepLabV3+[13]+Rule	68.5±0.1	85.1±0.5	83.4±0.4	8.3±0.2
SegFormer[14]+Rule	73.1±0.6	88.4±0.4	86.8±0.4	6.8±0.3
本文方法	76.2±0.1	91.3±0.3	88.6±0.2	3.9±0.1

如图4所示,原始图像中包含水体、海底、鱼类、水草、礁石和机器人等类别,同时存在水体浑浊、对比度较低、目标边界模糊和背景干扰较强等问题。图4表明,本文方法能够将鱼类、水草、礁石和机器人等对象标记为Avoid区域,并将水体和普通海底背景标记为No action区域。因此,本文方法在复杂水下背景中能够保持较好的动作区域识别稳定性,并降低需避让对象被误判为Remove的风险。

综合SeaClear和SUIM两个数据集的实验结果可以看出,本文方法在不同水下场景和语义类别体系下均保持了较好的动作图生成性能,并能够有效降低高风险动作误判。

2.4 不同动作类别的识别性能

本文以SeaClear数据集为例,分别统计Remove、Avoid和No action三类动作的Precision、Recall、$F1$和类别IoU,结果如表4所示。其中,IoU表示单个动作类别预测区域与真实区域之间的交并比,Mean行中的IoU为三类动作IoU的平均值,即前文使用的mIoU。

从表4可以看出,本文方法对三类动作均取得了较好的识别效果。其中,Remove类别的Precision、Recall和$F1$分别达到87.5%、85.9%和86.7%,IoU达到76.5%,表明模型能够较准确地识别可清除异物区域。Avoid类别的$F1$和IoU分别达到85.4%和74.5%,说明模型能够较好地区分需避让对象。No action类别取得了93.0%的Precision、95.0%的Recall和94.0%的$F1$,其IoU达到88.7%。这是由于No action主要对应水体和普通

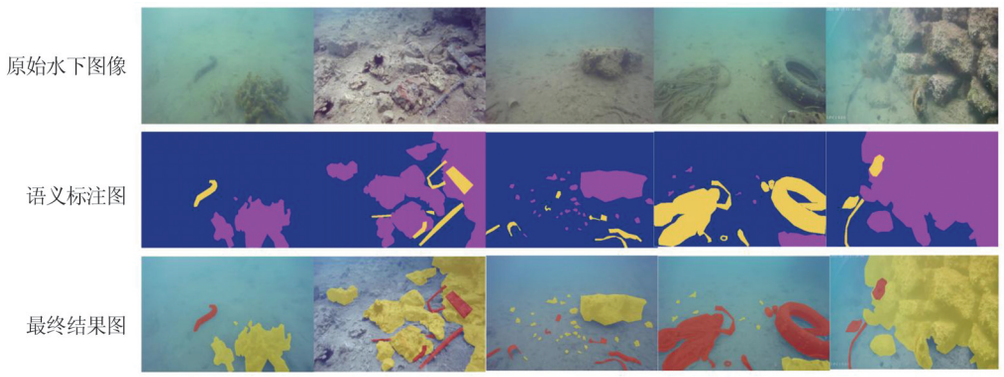


图4 SUIM数据集可视化结果图

Fig. 4 Visualization results on the SUIM dataset

表4 本文方法在SeaClear数据集上不同动作类别的识别性能

Table 4 Recognition performance of the proposed method for different action categories on the SeaClear dataset

%

决策类别	含义	Precision ↑	Recall ↑	$F1$ ↑	IoU ↑
Remove	可清除异物	87.5 ± 0.3	85.9 ± 0.4	86.7 ± 0.3	76.5 ± 0.4
Avoid	需避让对象	86.2 ± 0.3	84.6 ± 0.3	85.4 ± 0.2	74.5 ± 0.3
No action	不处理区域	93.0 ± 0.2	95.0 ± 0.2	94.0 ± 0.2	88.7 ± 0.2
Mean	平均结果	88.9 ± 0.3	88.5 ± 0.3	88.7 ± 0.2	79.9 ± 0.2

背景等大面积区域,其视觉分布相对连续,因此具有较高的识别稳定性。综合三类动作,本文方法的mIoU达到79.9%。其中,Remove和Avoid直接关系到后续清除与避让操作,其较高的$F1$表明模型能够在保持背景区域稳定识别的同时,有效区分可清除异物和需避让目标。

2.5 消融实验

为验证本文方法中各监督信号与损失项的作用,本文设计消融实验,分别分析语义监督、动作监督、语义动作一致性和风险惩罚损失对动作图生成性能的影响。

Semantic-only+Rule表示模型仅学习原始语义类别,并通过人工规则将语义预测结果转换为动作标签。Action-only表示模型仅学习Remove、Avoid和No action三类动作标签。Semantic+Action表示模型同时学习语义和动作类别。Semantic+Action+Consistency表示在双监督基础上加入语义动作一致性。本文方法加入风险惩罚损失,以降低误清除风险。实验结果如表5和表6所示。

从表5可以看出,在SeaClear数据集上,随着上述监督信号与损失项的逐步加入,模型性能整体呈提升趋势。与Semantic-only+Rule相比,Action-only的Decision Accuracy有所提高,说明直接利用动作监督有助于提升动作区域的预测准确性。进一步加入语义监督和语义动作一致性后,mIoU和Decision Accuracy继续提高,Critical Error明显降低,说明一致性约束能够减少语义预测与动作预测之间的冲突。加入风险惩罚损失后,Critical Error

表5 SeaClear数据集消融实验结果

Table 5 Ablation study results on the SeaClear dataset

%

方法变体	$\mathcal{L}_{sem}$	$\mathcal{L}_{act}$	$\mathcal{L}_{con}$	$\mathcal{L}_{risk}$	mIoU ↑	Decision Acc ↑	Remove Recall ↑	Remove $F1$ ↑	Avoid $F1$ ↑	Critical Error ↓
Semantic-only+Rule	√	-	-	-	74.0 ± 0.2	85.3 ± 0.2	82.1 ± 0.3	83.2 ± 0.3	81.6 ± 0.2	7.1 ± 0.2
Action-only	-	√	-	-	76.5 ± 0.3	87.1 ± 0.2	84.6 ± 0.4	84.9 ± 0.3	82.7 ± 0.3	6.4 ± 0.2
Semantic + Action	√	√	-	-	77.4 ± 0.2	87.8 ± 0.2	85.1 ± 0.3	85.5 ± 0.2	83.4 ± 0.2	5.9 ± 0.2
Semantic+Action+Consistency	√	√	√	-	78.9 ± 0.2	88.9 ± 0.1	85.7 ± 0.3	86.2 ± 0.2	84.5 ± 0.2	5.1 ± 0.2
本文方法	√	√	√	√	79.9 ± 0.2	89.8 ± 0.1	85.9 ± 0.4	86.7 ± 0.3	85.4 ± 0.2	4.3 ± 0.3

进一步降低到4.3%±0.3%，而Remove Recall仅由85.7%±0.3%提高至85.9%±0.4%，说明风险惩罚损失在降低误清除风险的同时，并未明显牺牲可清除异物的召回能力。

从表6可以看出，在SUIM数据集上，各监督与损失项的加入也呈现出与SeaClear数据集相似的性能提升趋势。在共有指标mIoU、Decision Accuracy、Avoid $F1$和Critical Error上，随着语义监督、动作监督、语义动作一致性和风险惩罚损失的逐步引入，模型整体性能得到进一步提升。语义动作一致性和风险惩罚损失均能降低Critical Error，并提升Avoid $F1$。由于SUIM数据集中可清除异物类别较少，Remove Recall不作为参考指标。总体来看，本文方法在SeaClear和SUIM两个数据集上均能稳定降低高风险误判，并在基本保持Remove召回能力的同时提升动作图生成的安全性。

表6 SUIM数据集消融实验结果

Table 6 Ablation study results on the SUIM dataset

%

方法变体	$\mathcal{L}_{sem}$	$\mathcal{L}_{act}$	$\mathcal{L}_{con}$	$\mathcal{L}_{risk}$	mIoU ↑	Decision Acc ↑	Avoid $F1$ ↑	Critical Error ↓
Semantic-only+Rule	√	-	-	-	73.1±0.6	88.4±0.4	86.8±0.4	6.8±0.3
Action-only	—	√	-	-	73.8±0.4	89.0±0.3	87.1±0.3	6.1±0.2
Semantic+Action	√	√	-	-	74.6±0.3	89.7±0.3	87.6±0.3	5.4±0.2
Semantic+Action+Consistency	√	√	√	-	75.4±0.2	90.5±0.2	88.1±0.2	4.6±0.2
本文方法	√	√	√	√	76.2±0.1	91.3±0.3	88.6±0.2	3.9±0.1

注：SUIM数据集中Remove类别样本较少，因此本文在SUIM数据集上主要关注Avoid F1和Critical Error。

2.6 高风险误判与失败案例分析

为进一步分析模型在清除任务中的高风险错误，本文统计Avoid和No action被误判为Remove的比例。其中，Avoid → Remove表示需避让区域被错误判定为可清除区域，No action → Remove表示无需处理区域被错误判定为可清除区域。除对比方法外，本文还选取2.5节中定义的部分消融配置，以分析语义动作一致性和风险惩罚损失对高风险误判的影响，结果如表7所示。

从表7可以看出，DeepLabV3++Rule和SegFormer+Rule的Avoid → Remove与No action → Remove均相对较高。仅采用语义监督和动作监督时，Critical Error降低至5.9%；进一步加入语义动作一致性后，该指标降低至5.1%；在此基础上加入风险惩罚损失后，Critical Error进一步降低至4.3%。同时，Avoid → Remove和No action → Remove分别降低至3.9%和4.5%。结果表明，语义动作一致性有助于减少语义预测与动作预测之间的不一致，而风险惩罚损失能够进一步抑制需避让区域和无需处理区域被错误预测为Remove的情况。

表7 高风险误判统计结果

Table 7 Statistical results of high-risk misclassifications

%

方法	Avoid → Remove ↓	No action → Remove ↓	Critical Error ↓
DeepLabV3+[13]+Rule	8.5±0.2	9.1±0.1	8.9±0.1
SegFormer[14]+Rule	6.8±0.2	7.2±0.2	7.1±0.2
Semantic+Action	5.5±0.2	6.1±0.2	5.9±0.2
Semantic+Action+Consistency	4.7±0.2	5.3±0.2	5.1±0.2
本文方法	3.9±0.2	4.5±0.3	4.3±0.3

为进一步分析高风险误判产生的具体场景，本文给出典型失败案例，如图5所示。图5展示了原始水下图像、语义标注图、最终预测结果以及主要错误区域。根据本文定义的语义到动作映射关系，语义标注可转换为对应的动作真值，用于判断模型预测错误。

从图5可以看出，模型主要存在两类失败情况。第一，在水体浑浊和低能见度场景下，可清除异物与周围背景的视觉差异减弱，部分Remove区域可能被误判为Avoid或No action，造成可清除异物漏检。第二，在礁石、鱼类等需避让目标密集或边界复杂的区域，局部纹理和形态可能与可清除异物相似，从而产生Avoid → Remove的高风险误判。

上述结果说明，本文方法虽然能够显著降低整体高风险误判，但在低能见度、复杂边界及外观相似目标场景下仍存在一定局限。后续可进一步结

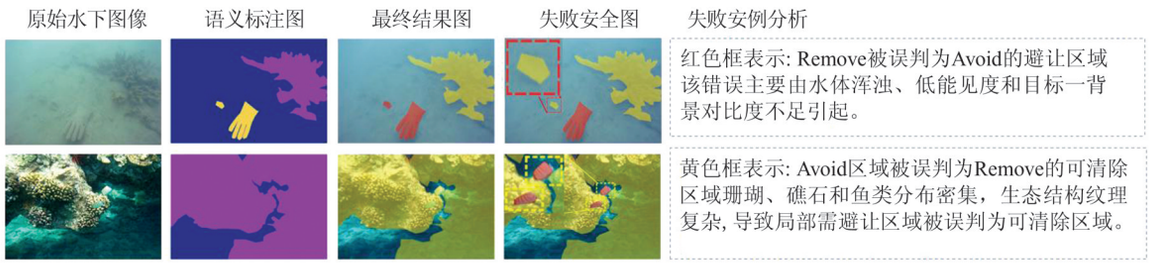


图5 失败案例分析图

Fig. 5 Analysis of failure cases

合边界信息、时序信息以及真实海洋牧场场景数据,提高模型在复杂水下环境中的稳定性。

2.7 参数敏感性分析

为分析语义动作一致性损失权重μ和风险惩罚损失权重γ对模型性能的影响,本文在SeaClear和SUIM数据集上开展参数敏感性实验。表8和表9以SeaClear为例报告具体结果,图6和图7进一步给出两个数据集上的变化趋势。本文主要采用Decision Accuracy、Remove $F1$、Avoid $F1$和Critical Error评价参数变化对模型性能的影响;在分析风险惩罚损失权重γ时,进一步报告Remove Recall,以考察风险约束对可清除异物召回能力的影响。

为单独分析语义动作一致性损失权重μ对模型性能的影响,本文固定动作决策损失权重$\lambda=1.0$,并暂不引入风险惩罚损失,即设置$\gamma=0$。在此基础上调整μ的取值,其他训练设置保持不变,实验结果如表8所示。

由表8可以看出,当$\mu=0$时,模型不使用语义动作一致性损失,仅通过语义监督和动作监督进行联合学习,此时Decision Accuracy为87.8%,Critical Error为5.9%。随着μ增大,Decision Accuracy、Remove $F1$和Avoid $F1$整体呈上升趋势,同时Critical Error逐渐降低,说明适当引入语义动作一致性有助于利用语义类别与动作类别之间的对应关系,减少两个预测分支之间的不一致。当$\mu=0.5$时,Decision Accuracy、Remove $F1$和Avoid $F1$分别达到88.9%、86.2%和84.5%,Critical Error降低至5.1%,取得较好的综合性能。当μ继续增大至0.7和1.0时,各项识别指标出现一定下降,表明过大的语义动作一致性权重可能使模型过度强调两种预测分布的一致性,从而影响两个预测分支对各自任务信息的学习。因此,本文最终设置$\mu=0.5$。

表8 不同μ取值下的参数敏感性分析

Table 8 Parameter sensitivity analysis under different values of μ

%

μ	Decision Acc ↑	Remove $F1$ ↑	Avoid $F1$ ↑	Critical Error ↓
0	87.8 ± 0.2	85.5 ± 0.2	83.4 ± 0.2	5.9 ± 0.2
0.1	88.2 ± 0.2	85.7 ± 0.3	83.8 ± 0.2	5.7 ± 0.2
0.3	88.6 ± 0.1	86.0 ± 0.2	84.2 ± 0.2	5.4 ± 0.2
0.5	88.9 ± 0.1	86.2 ± 0.2	84.5 ± 0.2	5.1 ± 0.2
0.7	88.7 ± 0.2	86.0 ± 0.3	84.3 ± 0.3	5.2 ± 0.2
1.0	88.3 ± 0.3	85.6 ± 0.3	83.9 ± 0.3	5.5 ± 0.2

表9 不同γ取值下的参数敏感性分析

Table 9 Parameter sensitivity analysis under different values of γ

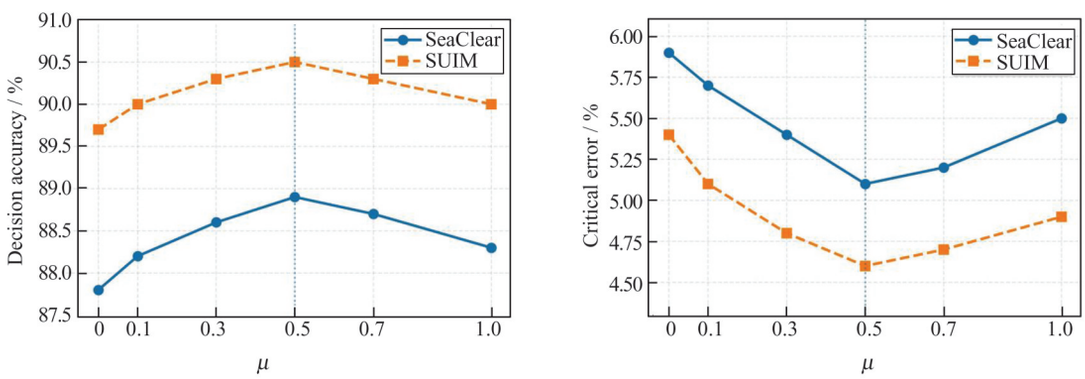
%

γ	Decision Acc ↑	Remove Recall ↑	Remove $F1$ ↑	Avoid $F1$ ↑	Critical Error ↓
0	88.9 ± 0.1	85.7 ± 0.3	86.2 ± 0.2	84.5 ± 0.2	5.1 ± 0.2
0.1	89.2 ± 0.2	85.8 ± 0.3	86.4 ± 0.3	84.8 ± 0.2	4.8 ± 0.2
0.3	89.5 ± 0.2	86.0 ± 0.4	86.6 ± 0.2	85.1 ± 0.3	4.5 ± 0.2
0.5	89.8 ± 0.1	85.9 ± 0.4	86.7 ± 0.3	85.4 ± 0.2	4.3 ± 0.3
0.7	89.4 ± 0.2	85.1 ± 0.3	86.0 ± 0.3	85.3 ± 0.2	4.1 ± 0.2
1.0	88.8 ± 0.3	83.9 ± 0.4	85.1 ± 0.4	85.0 ± 0.3	3.9 ± 0.2

根据表8的实验结果,本文将语义动作一致性损失权重固定为$\mu=0.5$,并保持$\lambda=1.0$,进一步调整风险惩罚损失权重γ,以分析其对动作图生成性能和高风险误判的影响,结果如表9所示。

由表9可以看出,当$\gamma=0$时,模型不使用风险惩罚损失,Critical Error相对较高。随着γ增大,Critical Error持续下降,说明风险惩罚损失能够有效抑制Avoid和No action被误判为Remove的高风险错误。当$\gamma=0.5$时,模型取得最高Decision Accuracy和Remove $F1$,同时Critical Error明显低于不使用风险惩罚损失的情况。当γ继续增大到0.7或1.0时,Critical Error仍然下降,但Remove $F1$明显降低。这说明过大的风险权重会使模型对Remove类别预测过于保守,从而增加可清除异物漏检。因此,本文选择$\gamma=0.5$作为清除完整性和误清除风险之间的折中设置。

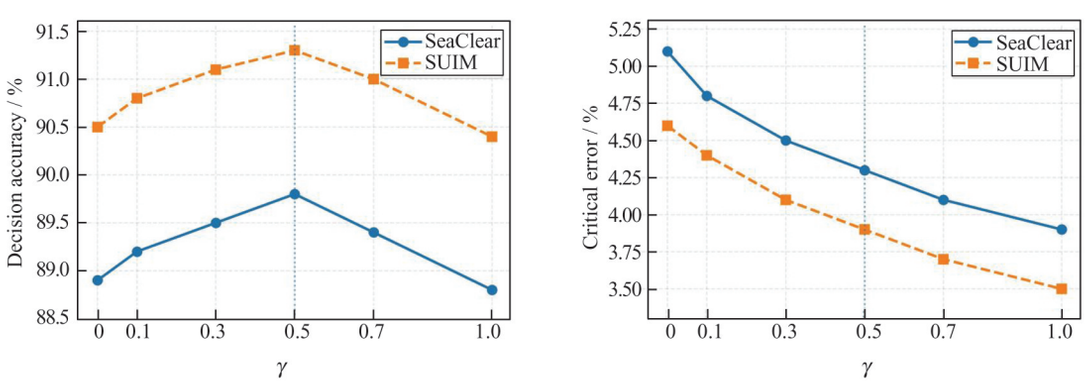
图6和图7进一步展示了μ和γ变化时各项指标的趋势。总体来看,μ主要影响语义识别与动作图生成之间的结构一致性,适当的一致性约束能够提升动作区域预测稳定性;γ主要影响模型对高风险误判的抑制能力,适当的风险惩罚能够降低Avoid或No action被误判为Remove的比例。综合Decision Accuracy、Remove $F1$、Avoid $F1$和Critical Error,本文最终设置$\mu=0.5$,$\gamma=0.5$。



(a) 清除决策准确率 (b) 高风险误判率

图6 一致性损失权重μ的敏感性分析

Fig. 6 Sensitivity analysis of the consistency loss weight μ



(a) 清除决策准确率 (b) 高风险误判率

图7 风险惩罚损失权重γ的敏感性分析

Fig. 7 Sensitivity analysis of the risk penalty loss weight γ

3 结 语

本文面向海洋牧场水下环境维护需求,提出了一种水下异物像素级动作图生成方法,将动作预测直接纳入模型训练,并构建Remove、Avoid和No action三类统一动作标签。模型通过语义分支和动作分支分别学习目标类别和动作类别,并利用语义动作一致性使两种预测结果保持一致,使目标识别同时受到动作预测结果的约束。在此基础上,风险惩罚损失进一步重点抑制需避让区域和不处理区域被误判为可清除区域的高风险错误。实验结果表明,本文方法在SeaClear和SUIM两个公开水下视觉数据集上取得了较好的像素级动作识别效果。未来工作将进一步引入真实海洋牧场数据,并结合

水下机器人平台开展在线测试,以提升方法在实际作业场景中的适用性。

参考文献:

[1] 杨红生,章守宇,张秀梅,等.中国现代化海洋牧场建设的战略思考[J].水产学报,2019,**43**(4):1255-1262. DOI:10.11964/jfc.20190211670.
YANG H S, ZHANG S Y, ZHANG X M, *et al*. Strategic thinking on the construction of modern marine ranching in China[J]. *Journal of Fisheries of China*, 2019, **43**(4): 1255-1262. DOI:10.11964/jfc.20190211670 (Ch).

[2] LI C Y, GUO C L, REN W Q, *et al*. An underwater image enhancement benchmark dataset and beyond [J]. *IEEE Transactions on Image Processing*, 2020, **29**: 4376-4389. DOI:10.1109/TIP.2019.2955241.

[3] RAVEENDRAN S, PATIL M D, BIRAJDAR G K. Underwater image enhancement: A comprehensive review, recent trends, challenges and applications[J]. *Artificial Intelligence Review*, 2021, **54**(7): 5413-5467. DOI:10.1007/s10462-021-10025-z.

[4] ZHAO Z Q, ZHENG P, XU S T, *et al*. Object detection with deep learning: A review[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, **30**(11): 3212-3232. DOI:10.1109/TNNLS.2018.2876865.

[5] XIAO Y Z, TIAN Z Q, YU J C, *et al*. A review of object detection based on deep learning [J]. *Multimedia Tools and Applications*, 2020, **79**(33/34): 23729-23791. DOI:10.1007/s11042-020-08976-6.

[6] ELMEZAIN M, SAAD-SAOUD L, SULTAN A, *et al*. Advancing underwater vision: A survey of deep learning models for underwater object recognition and tracking[J]. *IEEE Access*, 2025, **13**: 17830-17867. DOI: 10.1109/ACCESS.2025.3534098.

[7] YANG H B, LIU P, HU Y Z, *et al*. Research on underwater object recognition based on YOLOv3[J]. *Microsystem Technologies*, 2021, **27**(4): 1837-1844. DOI: 10.1007/s00542-019-04694-8.

[8] SUN X, SHI J Y, LIU L P, *et al*. Transferring deep knowledge for object recognition in low-quality underwater videos[J]. *Neurocomputing*, 2018, **275**(C): 897-908. DOI: 10.1016/j.neucom.2017.09.044.

[9] WANG X H, OUYANG J H, LI D Y, *et al*. Underwater object recognition based on deep encoding-decoding network[J]. *Journal of Ocean University of China*, 2019, **18**(2): 376-382. DOI:10.1007/s11802-019-3858-x.

[10] HAN F L, YAO J Z, ZHU H T, *et al*. Marine organism detection and classification from underwater vision based on the deep CNN method[J]. *Mathematical Problems in Engineering*, 2020, **2020**(1): 3937580. DOI: 10.1155/2020/3937580.

[11] HAN F L, YAO J Z, ZHU H T, *et al*. Underwater image processing and object detection based on deep CNN method [J]. *Journal of Sensors*, 2020, **2020**: 6707328. DOI:10.1155/2020/6707328.

[12] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation [C]//Medical Image Computing and Computer-Assisted Intervention — MICCAI 2015. Cham: Springer, 2015: 234-241. DOI:10.1007/978-3-319-24574-4_28.

[13] CHEN L C, ZHU Y K, PAPANDREOU G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Computer Vision — ECCV 2018. Cham: Springer, 2018: 833-851. DOI: 10.1007/978-3-030-01234-2_49.

[14] XIE E Z, WANG W H, YU Z D, *et al*. SegFormer: Simple and efficient design for semantic segmentation with transformers [C]//Advances in Neural Information Processing Systems. 2021, 34: 12077-12090. DOI: 10.5555/3540261.3541185.

[15] CHEN Z, WANG Y, TIAN W, *et al*. Underwater sonar image segmentation combining pixel-level and region-level information [J]. *Computers and Electrical Engineering*, 2022, **100**: 107853. DOI: 10.1016/j.compeleceng.2022.107853.

[16] CARUANA R. Multitask learning[J]. *Machine Learning*, 1997, **28**(1): 41-75. DOI:10.1023/A:1007379606734.

[17] WALLINGFORD M, LI H, ACHILLE A, *et al*. Task adaptive parameter sharing for multi-task learning[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2022: 7551-7560. DOI:10.1109/CVPR52688.2022.00741.

[18] CHEN L, HUANG Y Z, DONG J Y, *et al*. Underwater optical object detection in the era of artificial intelligence: Current, challenge, and future[J]. *ACM Computing Surveys*, 2025, **58**(3): 1-34. DOI:10.1145/3759243.

[19] ZHANG L Y, LI C Y, SUN H F. Object detection/tracking toward underwater photographs by remotely operated vehicles (ROVs)[J]. *Future Generation Computer Systems*, 2022, **126**: 163-168. DOI:10.1016/j.future.2021.07.011.

[20] BOLYA D, ZHOU C, XIAO F Y, *et al*. YOLACT: Real-time instance segmentation [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE Press, 2019: 9156-9165. DOI:10.1109/iccv.2019.00925.

[21] HAFIZ A M, BHAT G M. A survey on instance segmentation: State of the art[J]. *International Journal of Multimedia Information Retrieval*, 2020, **9**(3): 171-189. DOI: 10.1007/s13735-020-00195-x.

[22] STRUDEL R, GARCIA R, LAPTEV I, *et al*. Segmenter: Transformer for semantic segmentation [C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE Press, 2021: 7242-7252. DOI:10.1109/ICCV48922.2021.00717.

[23] GUO Y M, LIU Y, GEORGIOU T, *et al*. A review of semantic segmentation using deep neural networks[J]. *International Journal of Multimedia Information Retrieval*, 2018, **7**(2): 87-93. DOI:10.1007/s13735-017-0141-z.

[24] YU C Q, WANG J B, PENG C, *et al*. BiSeNet: bilateral segmentation network for real-time semantic segmentation [C]//Computer Vision — ECCV 2018. Cham: Springer, 2018: 334-349. DOI:10.1007/978-3-030-01261-8_20.

[25] MISRA I, SHRIVASTAVA A, GUPTA A, *et al*. Cross-stitch networks for multi-task learning[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2016: 3994-4003. DOI:10.1109/CVPR.2016.433.

[26] LIU S K, JOHNS E, DAVISON A J. End-to-end multi-task learning with attention[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2019: 1871-1880. DOI:10.1109/CVPR.2019.00197.

[27] KENDALL A, GAL Y, CIPOLLA R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2018: 7482-7491. DOI:10.1109/CVPR.2018.00781.

[28] CHEN Z, BADRINARAYANAN V, LEE C Y, *et al*. GradNorm: gradient normalization for adaptive loss balancing in deep multitask networks[C]// Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018:794-803. DOI: 10.48550/arXiv.1711.02257.

[29] HE K M, GKIOXARI G, DOLLAR P, *et al*. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision (ICCV). New York: IEEE Press, 2017: 2980-2988. DOI:10.1109/iccv.2017.322.

[30] JOCHER G, QIU J. Ultralytics YOLO[CP/OL]. 2023 [2026-08-12]. https://github.com/ultralytics/ultralytics.

[31] ĐURAŠ A, WOLF B J, ILIOUDI A, *et al*. A dataset for detection and segmentation of underwater marine debris in shallow waters[J]. *Scientific Data*, 2024, **11**: 921. DOI:10.1038/s41597-024-03759-2.

[32] ISLAM M J, EDGE C, XIAO Y Y, *et al*. Semantic segmentation of underwater imagery: Dataset and benchmark [C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2020: 1769-1776. DOI:10.1109/IROS45743.2020.9340821.

□