arXiv:2303.00973v2 [cs.CV] 6 Sep 2023

粗略海草分割只需图像标签

Scarle tt Raine[1,2], Ross Marchant[3], Brano Kusy[2], Frederic Maire[1] 和 Tobias Fischer [1]

[1]QUT机器人中心，昆士兰科技大学，澳大利亚{sg.raine, f.maire, tobias.fischer}@qut.edu.au[2]CSIRO Data61，澳大利亚{scarlett.raine, brano.kusy}@csiro.au[3]Image Analytics，澳大利亚ross.g.marchant@gmail.com

摘要

海草草甸是重要的碳汇，但要估算其储存的碳量，必须了解其中存在的海草物种。搭载机器学习算法的水下航行器和水面航行器有助于大规模准确估算海草草甸的组成和范围。然而，以往的海草检测与分类方法需要图像块级标签提供监督。本文将海草分类重新定义为弱监督粗略分割问题：训练时使用图像级标签（所需标签数量比图像块级标注少 *25* 倍），推理时则获得图像块级输出。为此，我们提出 *SeaFeats*，一种结合无监督对比预训练与特征相似度的架构；同时提出 *SeaCLIP*，这一模型展示了大型语言模型作为监督信号在特定领域应用中的有效性。我们证明，*SeaFeats* 与 *SeaCLIP* 的集成能够带来极为稳健的性能。在多物种“*DeepSeagrass*”数据集上，我们的方法在类别加权*F1* 分数上比以往需要图像块级标签的方法高出 *6.8%*（绝对值）；在“*Global Wetlands*”数据集上，海草有无的 *F1* 分数高出 *12.1%*（绝对值）。我们还介绍了两个面向实际部署的案例研究：在 *Global Wetlands* 数据集上进行异常值检测，以及将我们的方法应用于 *FloatyBoat* 自主水面航行器采集的图像。

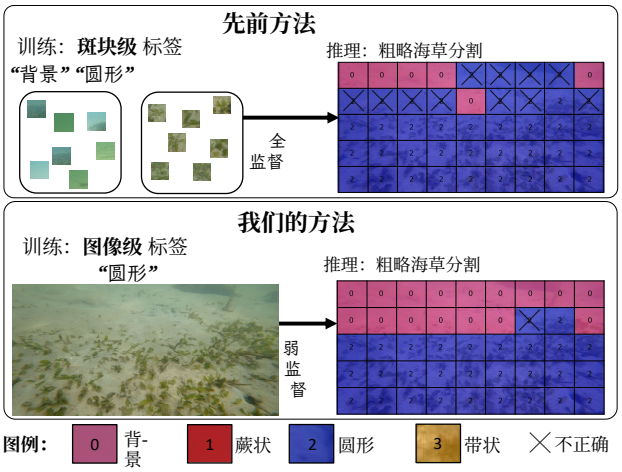


图 1. 上：先前用于多物种海草检测与分类的方法依赖带有图像块级标签的全监督训练。下：我们提出的方法采用弱监督，仅需图像级标签，同时仍能在推理时实现图像块级粗粒度海草分割。类“背景”“蕨状”“圆形”和“带状”的定义请参见第 4.2 节。

1. 引言

蓝碳 指沿海生态系统从大气中固存大量碳的能力 [22]。海草草甸在这一过程中发挥着至关重要的作用，但其储存的碳量可能因海草物种和栖息地条件而异 [15]。海草草甸还能改善水质、保护海岸线，并为鱼类提供食物来源和育幼栖息地 [23, 35]。科学家需要有关海草—草甸范围和组成的详细数据，以准确估算蓝碳固存 [15], ，并支持对这些关键生态系统的长期生态系统管理和保护 [22]。

自主和半自主水下调查方法对于大规模水下成像 [14,20, 26] 和栖息地调查 [6]至关重要。水下航行器可用于验证远程地理空间技术和航空感知技术、分析灾害发生后的海草物种分布，并监测气候变化 [23]。

机器学习能够自动、准确地处理图像数据，使生态学家能够实时调整调查路线 [13, 20]。然而，深度学习方法需要大量标注数据进行监督训练，但领域专家对水下图像进行像素级标注既昂贵又耗时 [25]。海草也缺乏鲜明的语义特征，且边界定义不清

，因此难以使用常见的目标检测方法进行检测。像素级标注的一种替代方法是基于网格的图像块标注，它能降低标注成本，同时支持海草检测与分类 [29, 35]。这与自动化底栖生境分类的方法一致 [1],，这类方法通常采用基于网格或点网格的标注方式。我们将分配给图像块的类别标签称为图像块级标签，而分配给整幅图像的类别标签称为图像级标签（图1）。

以往的多物种海草分类研究会直接将图像级标签赋给每个图像块 [35],，这有助于快速训练，但要求图像中只能包含海草——任何沙、水或其他物体区域都会给数据集引入噪声。其他研究探索了教师-学生框架，以便在未标注的图像块上快速训练，但其教师模型需要图像块级标签 [29]。

本文提出了一种基于图像级标签进行粗粒度分割的系统化方法。我们的框架集成了两个互补的分类器，利用无监督预训练和大型视觉语言模型的通用语义知识，在训练过程中生成伪标签。

本文的主要贡献如下：

1. 我们将海草分类重新定义为弱监督粗粒度分割任务，因此训练时只需要图像级标签，而推理时则可获得图像块级分类结果。 2. 我们提出了一种由两种新方法组成的集成，用于多物种粗粒度分割：SeaFeats采用一种新颖的损失函数，并通过比较图像块特征与类别模板，将图像块分类为背景或海草。SeaCLIP则证明，CLIP大型语言模型 [33] 可有效用作特定领域应用中的监督信号。 3. 我们进行了详尽的实验试验，在DeepSeagrass 数据集 [35]上以绝对F1分数计超越最先进水平6.8%。我们还提供了多种消融实验，以考察模型中各个组件的有效性。

4. 我们开展了两项真实世界部署案例研究：首先，我们为“Global Wetlands”数据集贡献了一个带标签的测试集 [4]，并证明 SeaCLIP 可有效用于鱼类异常值检测。其次，仅凭 20 张带标签的图像，我们证明了模型能够泛化至由 FloatyBoat [26] 自主水面航行器采集的水下图像。

我们公开代码，以促进未来对粗粒度海草分割的研究[1]。

[1] https://github.com/sgraine/bag-of-海草

2. 相关工作

珊瑚图像分类研究受到了研究人员的广泛关注 [7, 8, 13, 24, 25, 34]。相比之下，海草检测与分类尚未得到充分探索 [16, 30, 31, 36, 43],，仅有少数研究尝试进行多物种分割 [11, 29, 35, 37, 40]。我们在第 2.1 节详细介绍这些多物种方法。为解决多物种海草检测与分类问题，我们将其视为一个弱监督问题，并在第 2.2 节探讨该领域的现有方法。此外，我们还将在第 2.3 节讨论大型视觉语言模型的最新进展，以及它们在海洋调查领域中对弱监督和异常值检测的意义。

2.1. 多物种海草检测/分类

在大规模海草多物种制图方面，研究者已开发出多种方法，利用无人机 [37, 40] 或遥控飞行器 [11]采集的数据开展工作。然而，这些遥感方法在区分不同海草物种方面存在局限，并且只有在低风速条件、低潮时段以及一天中的特定时间才能有效运行 [37, 40]。

相比之下，Raine 等人 [35] 提出了水下图像原位多物种海草分类问题。具体而言， [35]他们贡献了“*DeepSeagrass*”数据集，其中包含从水下航行器视角拍摄的海草图像块，并提出一种方法，将每个图像块直接赋予其所属图像的标签。Noman等人 [29] 使用教师模型生成的伪真实标签训练学生模型。Noman等人。[29] 还贡献了“*ECU-MSS*”多物种海草数据集，但遗憾的是该数据集并未公开[2]，因此无法用于评估我们的方法。

为了解决这些方法的局限性，我们提出了一种用于粗略海草分割的图像级弱监督结构化框架，不会轻率地假定图像块级标签，也无需领域专家耗时进行图像块级标注。

2.2. 弱监督目标定位

由领域专家对水下图像进行像素级和图像块级标注，成本高昂且耗时极长。然而，水下识别主要依赖专家知识和人工标注的方法 [9]。为应对这一挑战，我们研究了弱监督目标检测与定位，其中模型仅使用图像级标签进行训练，以确定图像中实例的位置和类别 [3, 38, 45]。然而，我们的设置具有独特性，因为我们专注于粗粒度海草分割。海草沿着海底呈“地毯”状生长，

[2] 我们联系了作者，但他们无法分享该数据集。

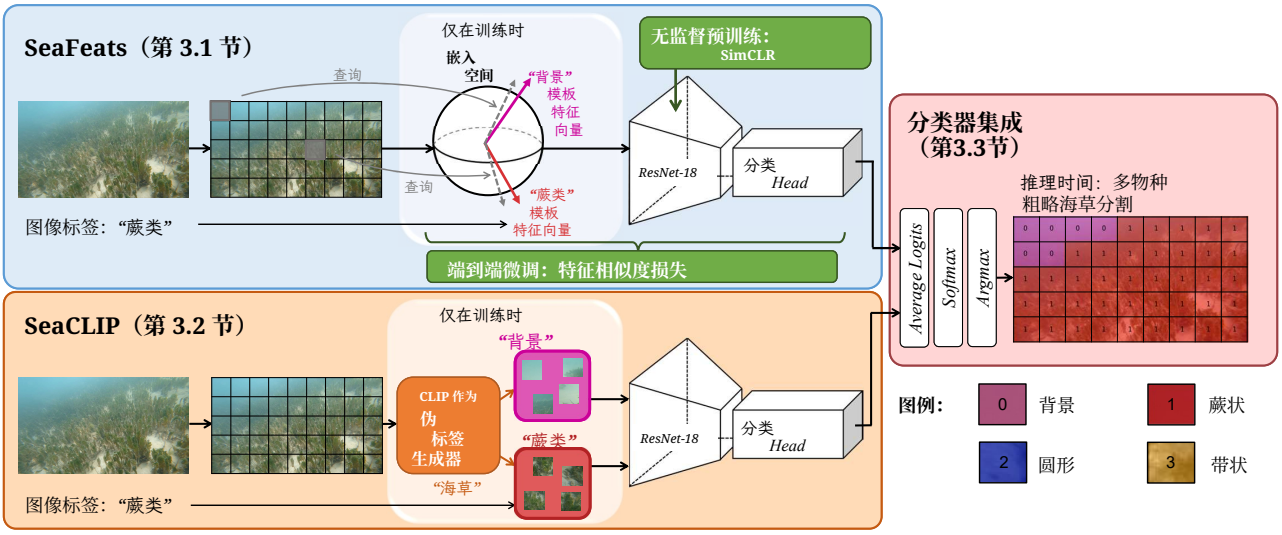


图 2. 所提算法示意图。我们提出了一种由两种互补方法组成的集成方案，用于海草粗略分割。上：SeaFeats 通过微调特征提取器（以无监督对比任务的权重初始化）并训练分类器来实现，方法是计算图像块级特征向量与各类别模板特征向量之间的余弦相似度（第 3.1 节）。下：SeaCLIP 使用预训练的大型语言模型 CLIP 预测的零样本伪标签进行训练 [33]，，并结合图像级领域特定海草标签（第 3.2 节）。右：在推理时，SeaFeats 和 SeaCLIP 并行组合为分类器集成，以预测查询图像的粗略分割掩膜（第 3.3 节）。

这意味着必须对推理图像中的所有区域进行分类，而边界框提议并不奏效，因为不存在明确定义的目标/背景边界。

Li 等。 [17] 通过自监督对比预训练实现了医学全切片图像分类和图像块级肿瘤定位——我们对这种预训练方法进行了改造，以区分不同类型的海草。

据我们所知，此前尚未有人尝试仅凭图像级标签对水下图像中的海草进行粗略分割。为此，我们将该问题构建为一项弱监督任务，并提出一种新颖的分类器集成，作为创新且有效的解决方案。

2.3. 视觉语言模型

近年来，大型视觉语言模型因其执行图像零样本分类的潜力而受到广泛关注。这些模型通常由两个编码器组成：一个用于处理文本描述，另一个用于处理视觉表征 [12, 19, 33]。通过联合训练这两个编码器，模型学会将文本描述符和基于图像的特征映射到共享嵌入空间中。其中一个著名的大型视觉语言模型是 CLIP，即“对比语言-图像预训练”（Contrastive Language-Image Pre-training）的缩写。CLIP 使用从互联网收集的 4 亿个图像-标题对进行训练 [33]。大量的训练数据，加上应用对比学习来建立共享嵌入空间，使 CLIP 具备了对视觉与语义概念的通用理解，从而能够应用于未见过的领域，并能够通过自然语言描述查询新颖类别 [33]。

近期许多方法都基于 CLIP，应用于零样本分布外检测 [5],指代图像分割 [21, 42] 和零样本分割 [18, 46] 等任务。然而，尚未有人尝试将 CLIP 用作水下图像中图块级检测与分类的训练信号。此外，CLIP 模型中蕴含的通用知识尚未与弱的领域特定标注结合使用。我们认为，将大型视觉语言模型引入生态学和海洋调查领域，是一项新颖且重要的贡献。

3. 方法

我们提出了两种仅依赖图像级标签进行粗略海草分割的新方法：SeaFeats 和 Sea-CLIP。为了消除对图像块级标注的监督需求，这两种方法都内置了在训练图像中提出伪标签的机制。首先，SeaFeats 利用深层特征的相似度，判断图像块在嵌入空间中更接近背景模板特征向量还是海草物种模板特征向量（第3.1节）。其次，SeaCLIP 展示了 CLIP 视觉语言模型可用作监督信号，以提出哪些图像块包含海草（第3.2节）。通过向 CLIP 查询描述海草的文本短语，SeaCLIP 尤其适用于模糊、远距离或图像

边缘图像块——在这些情况下，它会采取保守策略，将图像块分类为背景。因此，我们将两种方法并行组合为分类器集成，并发现这提高了方法的鲁棒性（第3.3节）。完整流程如图2所示。

在推理时，只需要 ResNet-18 分类器，即可实现轻量级部署，并达到与最先进水平相当的推理时间 [29, 35]。

3.1. SeaFeats – 特征相似度

受超维计算的启发 [27, 28, 44]，SeaFeats 通过为每个类别（背景和每种海草物种）生成高维模板特征向量，并测量每个图像块与这些模板向量的相似度，来提出伪标签（图2上；公式（1））。基于这些伪标签，我们随后训练一个粗略分割框架，该框架由 ResNet-18 编码器和分类头组成。我们使用自定义的特征相似度损失函数（第3.1.1节）对架构进行端到端微调。我们使用通过无监督预文本任务（第3.1.2节）学习的权重初始化编码器。在推理时，我们使用训练好的 SeaFeats ResNet-18 编码器和分类头，直接预测查询图像的粗略分割掩膜。

3.1.1 特征相似度损失函数

我们提出了一种新颖的损失函数，其灵感来自超维计算中的特征聚合和模板匹配 [27, 28, 44]。每个训练轮次结束后，我们从 ResNet-18 编码器的最后一层提取每个图像块的特征向量。对于每个类别，我们通过对图像级别中属于该类别的 L2归一化图像块特征取平均值，得到一个模板特征向量：

$$\bar{\mathbf{v}}_c = \frac{1}{N_c} \sum_x \mathbb{1}_{[x=c]} \frac{\mathbf{v}_x}{||\mathbf{v}_x||_2}, \tag{1}$$

其中，$\mathbf{v}_x$ 是提取出的图像块 x 的特征向量，$\mathbb{1}_{[x=c]}$ 是一个指示函数：如果图像块$\mathbf{v}_x$ 的图像标签 x 等于类别 c，则返回 1，否则返回 0。因此，对于每个类别 c，我们计算 $\bar{\mathbf{v}}_c$，即从类别 c的 N_c 个图像块样本中提取的归一化特征向量的平均值。需要注意的是，随着训练过程中对特征提取器进行微调，每个类别的模板特征向量也会在每个训练轮次动态更新。

训练期间，我们计算当前图像块的特征向量$\mathbf{v}_x$ 与对应于图像级标签 c的模板特征向量 $\bar{\mathbf{v}}_c$ 之间的余弦相似度：

$$\text{sim}(\mathbf{v}_x, \bar{\mathbf{v}}_c) = \frac{\mathbf{v}_x \cdot \bar{\mathbf{v}}_c}{||\mathbf{v}_x||\ ||\bar{\mathbf{v}}_c||}. \tag{2}$$

我们还会计算$\mathbf{v}_x$ 与背景模板特征向量 $\bar{\mathbf{v}}_0$之间的余弦相似度。如果该相似度高于与背景模板特征向量的相似度，则该图像块将被伪标注为背景，即使其图像级标签为海草。否则，该图像块的伪标签将与图像级标签一致，

$$p_x = \begin{cases} 0 & \text{sim}(\mathbf{v}_x, \bar{\mathbf{v}}_0) > \text{sim}(\mathbf{v}_x, \bar{\mathbf{v}}_c) \\ c & \text{otherwise} \end{cases} \tag{3}$$

其中，p_x 是图块 x 的伪标签，类别索引 0 对应背景类。我们使用模型输出的 Softmax值与伪标签之间的分类交叉熵来训练神经网络。

3.1.2 对比代理任务

由于我们的损失函数依赖于从架构的编码器中提取的特征，因此特征编码器的初始化对于获得高性能至关重要 [17]。我们采用无监督代理任务来获得预训练 ResNet-18 特征提取器，如 [17]所示。我们添加分类头（第4.1.1节），然后使用第3.1.1节所述的损失函数，对整个架构进行端到端微调。

具体而言，我们采用的代理任务使用 Sim- CLR [2] 对比学习框架。对于一个包含 B张图像的批次，SimCLR 使用数据增强变换为每张图像创建两个随机“视图”，因此每批次共包含 $2B$张图像（见第 4.1.1节）[2]。对于每张图像，其对应的两个增强视图被视为一个“正样本对”，而批次中的所有其他样本则为负样本 [2]。网络通过以下对比损失 [2]进行训练，以最大化样本表征 $\mathbf{v}_i$与其正视图 $\mathbf{v}j$之间的余弦相似度，并最小化其与负样本 $\{\mathbf{v}k\}_{k\in\{1,\ldots,2B\},\ k\neq i}$之间的相似度：

$$l_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{v}_i, \mathbf{v}_j)/\tau)}{\sum_{k=1}^{2B} \mathbb{1}_{[k\neq i]} \exp(\text{sim}(\mathbf{v}_i, \mathbf{v}_k)/\tau)}, \tag{4}$$

其中 τ是温度参数，我们按照建议将其设为 $\tau = 0.07$。

3.2. SeaCLIP – 大型视觉语言模型

我们利用预训练的大型视觉语言模型 CLIP [33] 作为零样本伪标签生成器。CLIP 使用图像和文本嵌入进行联合训练。尽管 CLIP 对海草物种的专业知识极为有限，但它具备关于沙、水和海草的一般语义知识，因此我们利用它来预测训练图像中的哪些图像块包含海草，哪些应标记为背景。请注意，CLIP 本身无法区分不同的海草物种。

我们使用原始的 CLIP 预训练模型，并遵循推荐的方法进行零样本分类 [33]。由于 CLIP 是在自然语言提示词上训练的，因此可以

利用这一点，过滤掉所有分布外对象或噪声图像块。我们从 CLIP 获取二元背景/通用海草伪标签（关于所用提示词的更多详情，请参见第4.1.2 节和补充材料），并将单个图像级物种特定海草标签分配给 CLIP 伪标注为通用海草的图像块。随后，我们端到端训练 SeaCLIP 分类器（第4.1.2节）。通过这种方式，SeaCLIP 将 CLIP 的通用语义知识与领域特定的专业知识相结合，为训练生成准确的图像块级伪标签。在推理时，直接使用训练好的 SeaCLIP ResNet-18 编码器和分类头来预测所有查询图像块的类别，无需耗费大量资源进行 CLIP 推理。

3.3. 分类器集成

我们发现，提出的两种方法 SeaFeats 和 SeaCLIP 是互补的分类器，将它们集成后能够提升性能（图2）。例如，我们观察到，SeaCLIP 更倾向于将视觉退化图像块分类为背景，而不是过度自信地错误分类其物种。与 SeaFeats 结合时，这一特性十分理想，因为它可以避免对视觉退化图像块进行错误的物种识别，从而提升我们方法的总体鲁棒性。我们通过对两个分类器的归一化输出 logit 取平均，然后应用 softmax 函数，将它们整合起来。我们发现，在应用 softmax 之前合并分类器的输出，可以保留 logit 的幅度，从而得到置信度更高的预测 [41]。在取平均之前，会对每个分类器的 logit 尺度进行归一化，以确保两个分类器在这一操作中具有相同的“权重”。

4. 实验设置

本节简要讨论实现细节（第4.1节）、评估数据集（第4.2节）以及所使用的评估指标（第4.3节）。

4.1. 实现

所有实验均使用 NVIDIA A100 GPU 进行。两个分类器均使用 PyTorch 实现 [32]。下文将讨论我们的方法的超参数和实现细节。

4.1.1 SeaFeats – 特征相似度

我们将架构设计为一个 ResNet-18 [10] 编码器，移除最后的全连接层，随后连接一个平均池化层，该层将分类头应用于推理图像中的图像块。在此之后， [35], 我们的分类头由一个包含 512 个节点并采用 ReLU 激活的全连接层、概率为 0.15 的丢弃层（用于防止过拟合）以及一个最终的全连接层组成，每个类别对应一个节点。我们使用 Adam 优化器训练网络，初始学习率为 0.00001，批次大小为 3 张图像，最多训练 150 个训练轮次。为缓解类别不平衡问题，我们根据每个类别的权重 $w = [1.0, 1.5, 1.2, 1.2]$ 对分类交叉熵进行加权，这些类别为 [“背景”、“蕨状”、“圆形”、“带状”]。

我们使用通过无监督对比预训练获得的权重初始化编码器。为此，我们采用 [39] 中的 SimCLR 实现。输入图像块的分辨率为 520x578，并将随机裁剪尺寸设为 132x132 像素。每个批次使用 132 个图像块，在 DeepSeagrass 数据集的 42,848 个未标注训练图像块上训练 200 个训练轮次；训练期间使用缩放和裁剪、水平翻转、垂直翻转、颜色抖动和高斯模糊来生成正样本。

4.1.2 SeaCLIP – 大型视觉语言模型

SeaCLIP 使用了 [33] 中介绍的预训练公开 CLIP 模型。为提升 CLIP 作为伪标签生成器的零样本性能，我们采用基础的提示工程，设计了多种形式的查询提示，例如“一张海草的照片”、“一张包含一些海草的模糊照片”或“一张水下草状叶片的照片”（更多详情请参阅补充材料）。如果这些提示中任意一个对应的概率最高，那么在我们的方法中，该图像块就会被伪标注为图像级海草物种。与 SeaFeats 模型类似，我们使用加权类别交叉熵和 Adam 优化器训练一个以 ImageNet 权重初始化的 ResNet-18 [10] 架构，批次大小为 32 个图像块。

4.2. 数据集

DeepSeagrass 数据集 [35] 由 1,701 张高分辨率（4624x2600 像素）训练图像组成。研究保留了 8 个与训练图像不重叠的地理区域，用于构建包含 335 张图像的测试数据集 [35]。根据图像级类别，将图像划分为 40 个（8x5）较小的图像块。在开展试验之前，我们对测试数据集进行了人工验证；在 13,378 个测试图像块中，我们修正了 591 个图像块的类别，并移除了 163 个模糊图像块。

DeepSeagrass 包含四个大类：“背景”，由海草总面积小于 1% 的图像块组成，例如水体、沙或图像中因过于模糊而无法辨别的区域；“蕨状”（*Halophilaspinulosa*）；“圆形”（卵叶喜盐草）；以及“带状”（齿叶针草、单脉二药草、丝粉草和穆勒氏鳗草） [35]。

我们提出使用**Global Wetlands** luderick 和海草数据集 [4] 来考虑部署机器人应用我们的方法的场景。当生态学家收集一组新的训练图像时，其中很可能包含生态学家不感兴趣的分布外

表 1. 多物种海草分类结果 – DeepSeagrass 数据集（请参阅第 4.3 节了解指标定义）

方法	标签	背景			蕨状			圆形			带状			总体	推理
		精确率	召回率	F1	精确率	召回率	F1	精确率	召回率	F1	精确率	召回率	F1	F1 分数	时间（秒）
零样本CLIP [33]	Nil	79.92	87.23	83.41	86.26	20.55	33.20	56.35	34.75	42.99	36.62	82.72	50.76	60.65	0.783
SimCLR [2] + Raine 等人. [35]	图块	99.64	83.45	90.83	84.63	53.70	65.70	51.24	96.58	66.96	60.90	93.41	73.73	77.16	**0.019**
Raine等人. ResNet-50 [35]	图块	99.76	70.97	82.94	79.45	**99.47**	88.34	86.10	97.58	91.48	86.01	97.87	91.56	87.41	0.022
Noman等人。 EfficientNet-B5 [29]	图块	**99.90**	73.11	84.43	88.00	99.01	93.18	86.56	**99.33**	92.50	76.26	**99.04**	86.17	88.52	0.047
我们的方法: SeaFeats+SeaCLIP	图像	97.47	**92.59**	**94.97**	**92.19**	98.89	**95.42**	**93.50**	95.92	**94.69**	**97.57**	95.06	**96.29**	**95.33**	0.025

对象。Global Wetlands 包含三个大类：海草、背景和鱼类（分布外）。我们还使用 Global Wetlands 来评估我们方法的领域泛化能力，并验证我们的 CLIP监督方法。

Global Wetlands 数据集采集于澳大利亚莫顿湾，使用远程水下摄像机拍摄。为进行评估，我们使用了 764 张“novel-test”划分图像；这些图像采集自不同于“train”和“test”划分的地理位置。我们从每张图像中创建 50 个（10x5）图像块，最终得到 38,200 个测试图像块，并将这些图像块手动标注为“鱼类”“海草”和“背景”（更多细节请参阅补充材料）。该数据集中的海草对应于 DeepSeagrass 数据集中的“带状”海草形态型。我们公开发布带标注的测试集，以促进未来在该数据集上的评估[3]。

4.2.1 FloatyBoat

我们还考虑了将我们的方法部署在自主水面航行器上的情况，并利用 FloatyBoat 自主水面航行器采集的水下图像，评估我们方法的领域泛化能力 [26] 。这些图像采集于澳大利亚蜥蜴岛。与 DeepSeagrass 数据集相比，这些图像表现出显著的领域偏移，具体体现在摄像机属性、视角（DeepSeagrass 从 45° 的倾斜角度拍摄，而 FloatyBoat 摄像机采用垂直俯视角度）、光照、浑浊度、地理位置以及海草的生长阶段和密度等方面（DeepSeagrass 包含茂密的海草，覆盖率为 >70%，而 FloatyBoat 海草草甸较为稀疏）。我们利用一条包含 1,148 张图像的调查样带来评估我们的方法。在应用我们的方法之前，我们会执行基本的自动颜色校正。

4.3. 评估指标

我们使用各类别的精确率、召回率、F1分数以及总体 F1 分数来报告我们方法的性能。我们主要使用 F1分数，因为它们兼顾了

[3]https://doi.org/10.5281/zenodo.7659203

表 2. 二元海草分类——DeepSeagrass F1分数（参见第 4.3 节了解指标定义）

方法	标签	背景	海草	总体
零样本CLIP [33]	Nil	83.41	87.74	85.90
SimCLR [2] +Raine 等。 [35]	图像块	90.85	94.55	93.17
Raine等。 ResNet-50 [35]	图像块	82.94	90.90	88.13
Noman等。 EfficientNet-B5 [29]	图像块	84.42	91.54	89.03
我们的方法: SeaFeats+SeaCLIP	图像	**95.01**	**96.72**	**96.04**

精确率和召回率的重要性：

$$\mathrm{F}_1 = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}. \tag{5}$$

我们注意到，先前方法经过调优，旨在获得较高的海草召回率，但这是以精确率为代价的，即许多背景图像块被错误地预测为海草。在海草密度估计中，准确预测海草是否存在非常重要，这样可以避免高估覆盖率和碳储量，而 F1 分数更能体现这一点。

我们评估了该方法在多物种应用场景中的表现，其中考虑了第4.2节所述的四个类别，以及用于估计海草是否存在的二元（背景/海草）分类场景。

5. 结果

在第5.1节中，我们将我们的方法与多种海草检测与分类的最先进方法进行比较 [29, 35]，并将我们的方法与两种广泛使用的分类方法 SimCLR 进行比较 [2] 和 CLIP [33]。在第5.2节中，我们进行消融实验，以展示特征提取器初始化和分类器集成的效果。最后，在第 5.3节中，我们通过两个案例研究探讨方法在实际部署中的应用：检测数据中不需要的对象这一异常值检测任务，以及在机器人平台上的部署。

表 3. SeaFeats 特征提取器初始化——DeepSeagrass F1 分数（指标定义请参阅第4.3 节）

方法	后景	蕨类	圆形	带状	总体
SeaFeats：随机初始化	86.48	84.54	86.10	82.16	85.05
SeaFeats：ImageNet 初始化	72.62	90.29	55.18	80.69	77.89
SeaFeats：SimCLR [2]初始化	**94.93**	**92.44**	**89.56**	**92.12**	**93.10**

表4. 分类器集成的效果——DeepSeagrass F1分数（参见第4.3 节的指标定义）

方法	后景	蕨类	圆形	带状	总体
SeaFeats	94.93	92.44	89.56	92.12	93.10
SeaCLIP	86.14	87.82	87.91	91.21	87.84
SeaFeats+SeaCLIP	**94.97**	**95.42**	**94.69**	**96.29**	**95.33**

5.1. 与最先进方法的比较

在表 1中，我们将我们的方法与两种最先进的方法进行比较：文献 [35] 中报告的 ResNet-50 分类器，以及文献 [29][4]中报告的 EfficientNet-B5 分类器。我们还与两种具有竞争力的基线方法进行比较：使用 CLIP [33]作为零样本分类器；以及使用 SimCLR [2] 对特征提取器进行预训练，然后以全监督方式，使用来自 [35] 的图像块训练线性分类头。

如表 1 所示，在多物种海草分类任务中，与所有先前方法和基线方法相比，我们的方法在各类别及总体 F1 分数方面均有显著优势。对于多物种分类，与 [29] 和 [35] 相比，我们的方法分别将总体 F1 分数提高了 6.8% 和 7.9%。在推理时间方面，我们的方法也与先前方法相当（~40帧/秒，见表 1）。

表2 显示，我们在二元海草存在/缺失问题上的表现也优于 [29]和 [35]，分别提升了 7.0% 和 7.9%。

5.2. 消融实验

5.2.1 特征提取器初始化的影响

我们发现，使用 Sim-CLR 无监督对比预训练任务训练特征提取器 [2]可显著提升自定义损失函数在微调架构时的性能；与使用 ImageNet 初始化和随机权重初始化相比，性能分别提升了 15.2% 和 8.1%（表3）。在 ImageNet 上训练的特征提取器无法很好地迁移到海草分类任务；而随机初始化的特征提取器能够捕捉空间特征，并将图像投影到判别式嵌入空间中，避免了 ImageNet 预训练任务所学到的无益偏差。

5.2.2 分类器集成

我们发现，特征相似度分类器（SeaFeats）与我们的 CLIP伪标签方法（SeaCLIP）集成后，性能更加稳健；与仅使用 SeaFeats 或仅使用 SeaCLIP 相比，绝对F1 分数分别提高了 2.2% 和 7.5%（表

4）。请注意，正如第 3.3 节所述，推理时会对每个模型归一化后的输出 l ogits 取平均

[4]来自 [29] 的源代码未公开，因此我们尽最大努力重新实现了他们的方法。

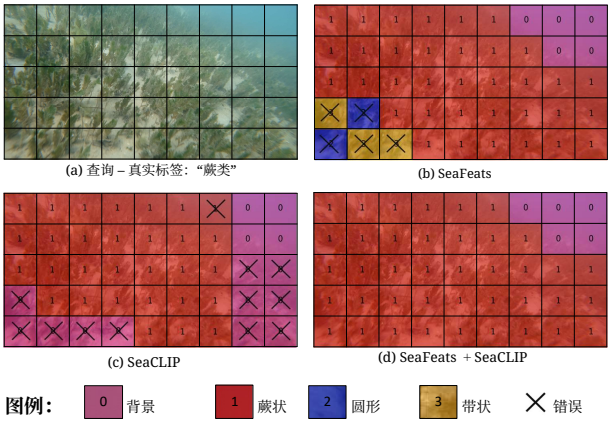


图3。对于 (a) 中所示的查询图像，我们的分类器集成结合 SeaFeats (b) 和 SeaCLIP (c) 的输出，得到 (d) 中的组合 Softmax 值。在 (b) 中，图像左下方模糊的图像块被错误地预测为带状和圆形海草。在 (c) 中，SeaCLIP 将这些图像块谨慎地预测为背景。结合这两个模型 (d) 可避免错误分类，并将这些图像块正确分类为蕨状海草。

时间，以提高对不确定分类的置信度。如图 3 所示，将 SeaCLIP 与总体表现更优的 SeaFeats 模型结合后，整体性能有所提升，这是因为 SeaCLIP 对视觉质量退化的图像块产生了影响。我们在补充材料中提供了更多定性结果。

5.3. 机器人部署案例研究

本节中，我们介绍两个关于将我们的方法部署于真实场景的案例研究。首先，我们考察“Global Wetlands”数据集，该数据集模拟了这样一种情形：生态学家收集一组新的图像，其中包含一些生态学家不感兴趣的物体。随后，我们通过将方法应用于 FloatyBoat 收集的调查样带，评估模型对新平台的泛化能力 [26] 自主水面航行器。

在我们的第一个案例研究中，我们展示了 CLIP 在标签更加有限的问题中的有效性：没有图像级标签，而且所有图像都包含多个类别。这种设置模拟了生态学家收集了一组海草图像的情形，不过这些图像可能包含分布外类别，而我们希望分类器能将这些类别与海草和背景地面图像块区分开来——

表5. Global Wetlands F1分数（参见第4.3 节了解指标定义）

方法	训练数据集	背景-地面	Fish	Sea-草	多类别 总体	二元 总体
零样本CLIP [33]	Global Wetlands	84.55	**60.73**	90.54	85.77	88.01
SeaCLIP	Global Wetlands	87.97	59.57	91.81	**87.86**	90.86
Raine等。 [35]	DeepSeagrass	30.04	n/a	75.04	n/a	63.21
Noman等。 [29]	DeepSeagrass	69.72	n/a	85.15	n/a	80.07
SeaFeats+SeaCLIP	DeepSeagrass	**89.93**	n/a	**93.58**	n/a	**92.16**

地面图像块（在本例中为“鱼类”类别）。我们利用 CLIP 预测哪些图像块属于各个类别，并将其作为训练 SeaCLIP 分类器的监督信号（有关所用提示词的详情，请参阅补充材料）。

SeaCLIP 能有效区分海草、鱼类和背景类别，在训练期间仅以 CLIP 作为监督信号，F1 分数便达到 87.9%（表 5 和图4）。还需注意，在这一场景中，各类别的范围更广、视觉差异也比多物种海草场景更加明显，因此零样本 CLIP 模型的准确率可达 85.8%。与使用 CLIP 相比，SeaCLIP 的性能进一步提升了 2.1%，且具有显著优势：其推理速度比零样本 CLIP 快 ~30 倍（表1）。

我们还评估了使用 DeepSeagrass 训练的方法在不同相机属性和视觉条件下，对 Global Wetlands 数据集的泛化能力。在海草存在/缺失问题上（表 [29]和 [35] 中的“二分类总体”），我们的方法优于上述两者，绝对 F1 分数分别高出 12.1% 和 28.9%；此外，尽管我们的集成方法仅使用 DeepSeagrass 进行训练，其表现仍比使用 Global Wetlands 数据集训练的 SeaCLIP 高出 1.3%。

在第二个案例研究中，我们评估了我们的方法在由自主水面航行器采集的图像上的性能 [26]。我们仅使用 10 张背景图像和 10 张海草图像对方法进行微调，训练 10 个训练轮次。我们证明了模型能够泛化到完全不同的平台和图像特征，并能够在每一帧中正确预测带状海草和背景图像块，如图 4 所示。需要指出的是，由于该数据集不提供真实标签，我们无法对方法进行定量评估，但图 4 及配套视频从定性角度证明了我们的方法在这一场景下表现良好。

6. 结论

在这项工作中，我们提出了一种用于海草粗粒度多物种分割的弱监督方法，仅需图像级标注。我们的方法在推理过程中，利用 SeaFeats

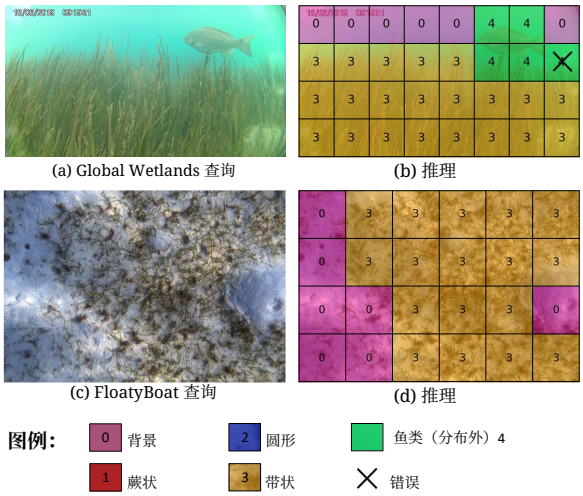


图 4模型对其他平台泛化能力的定性评估。来自 Global Wetlands 的测试图像样本（a），以及我们的 SeaCLIP 模型在（b）中的推理结果。由 FloatyBoat 上的俯视摄像机采集的一张横断面图像 [26] 见（c），微调后的 SeaFeats 对应的推理结果见（d）。

和 SeaCLIP 这两种互补方法，在图像块级别进行海草检测与分类。我们提出的集成方法（SeaFeats+SeaCLIP）在 DeepSeagrass 上取得了 95.3% 的类别加权F1分数，较此前最先进水平高出 6.8%。

我们展示了预训练大型视觉语言模型在生态学和海洋调查领域的相关性，并表明 CLIP 能够为弱监督训练提供有效的监督信号。我们的方法还可应用于其他场景，例如绘制精准农业中的杂草分布图，或对遥感影像进行粗粒度分割，以分析土地与植被覆盖。

我们的方法旨在满足机器人学应用场景的需求。实时推理将为自适应导航和调查带来令人振奋的新研究机会。例如，与沙质图像块等价值较低的区域相比，我们的方法能够以更高的细节程度调查海草密度较高或存在珍稀海草物种的区域。鉴于人们对调查区域的先验知识有限，这项能力尤为关键，因为遥感观测方法无法提供足够详细的信息，来指导水下或水面自主车辆。未来的其他研究方向还包括对多物种海草图像进行逐像素语义分割、估算海草密度，以及根据模型推断估算蓝碳储量。

致谢

这项工作由昆士兰科技大学与 CSIRO Data61 合作完成。S.R.、F.M. 和 T.F. 感谢昆士兰科技大学（昆士兰科技大学）通过机器人中心给予的持续支持。T.F. 感谢澳大利亚研究理事会桂冠学者奖学金 FL210100156 和英特尔实验室通过资助项目 RV3.290.Fischer 提供的资金支持。

参考文献

[1] Qimin Chen、Oscar Beijbom、Stephen Chan、Jessica Bouwmeester 和 David Kriegman。面向 CoralNet 的新型深度学习引擎。载于国际计算机视觉会议，第 3693–3702 页，2021年。2[2] Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey Hinton。用于视觉表征对比学习的简单框架。载于国际机器学习会议，第 1597–1607 页，2020年。4、6、7[3] Elijah Cole、Kimberly Wilber、Grant Van Horn、Xuan Yang、Marco Fornoni、Pietro Perona、Serge Belongie、Andrew Howard 和 Oisin Mac Aodha。论标签粒度与目标定位。载于欧洲计算机视觉会议，第 604–620 页，2022年。2[4] Ellen M Ditria、Rod M Connolly、Eric L Jinks 和 Sebastian Lopez-Marcano。用于在不受限海洋环境中自动识别和计数鱼类的标注视频片段，2021年。2、5[5] Sepideh Esmaeilpour、Bing Liu、Eric Robertson 和 Lei Shu。基于预训练模型 CLIP 的零样本分布外检测。载于AAAI人工智能会议，第 6568–6576 页，2022年。3[6] Renata Ferrari、Ezequiel M Marzinelli、Camila Rezende Ayroza、Alan Jordan、Will F Figueira、Maria Byrne、Hamish A Malcolm、Stefan B Williams 和 Peter D Steinberg。利用自主水下航行器对多个海洋保护区的底栖群落进行大规模评估。PLoS One，13(3)，2018年。1[7] Anabel Gómez-Ríos、Siham Tabik、Julián Luengo、ASM Shihavuddin 和 Francisco Herrera。利用纹理或结构图像，通过基于卷积神经网络的两级分类器识别珊瑚物种。Knowledge-Based S ystems，184：1–10，2019年。2[8] Manuel González-Rivero 等。利用人工智能监测珊瑚礁：一种可行且经济高效的方法。Remote Sensing，12(3)：489，2020年。2[9] Salma P González-Sabbagh 和 Antonio Robles-Kelly。水下计算机视觉综述。ACM Computing Surveys，2023年。2[10] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于IEEE计算机视觉与模式识别会议，第 770–778 页，2016年。5[11] Brandon Hobley、Riccardo Arosio、Geoffrey French、Julie Bremner、Tony Dolphin 和 Michal Mackiewicz。使用遥控驾驶航空器影像进行海岸监测海草半监督分割。Remote Sensing，13(9)：1741，2021年。2[12] Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。利用噪声文本监督扩展视觉及视觉语言表征学习。载于国际机器学习会议，第 4904–4916 页，2021年。3[13] Karim Koreitem、Florian Shkurti、Travis Manderson、Wei-Di Chang、Juan Camilo Gamboa Higuera 和 Gregory Dudek。野外单次示例引导的机器人视觉搜索。载于IEEE/RSJ智能机器人与系统国际会议，第 5800–5807 页，2020年。1、2。

[14] Brano Kusy、Jiajun Liu、Aninda Saha、Yang Li、Ross Marchant、Jeremy Oorloff、Lachlan Tychsen-Smith、David Ahmedt-Aristizabal、Brendan Do、Joey Crosswell 等。现场数据整理：边缘人工智能实现可执行应用的关键。载于移动计算与网络国际会议，第794–796页，2022年。1[15] Paul S Lavery、Miguel-Ángel Mateo、Oscar Serrano 和 Mohammad Rozai mi。海草栖息地碳储存的变化及其对蓝碳生态系统服务全球估算的影响。《PLoS One》，8(9)，2013年。1[16] NA Lestari、I Jaya 和 M Iqbal。使用深度学习 Mask R-CNN 算法分割海草（针叶藻）。载于IOP会议系列：地球与环境科学，2021年。2[17] Bin Li、Yin Li 和 Kevin W Eliceiri。用于全视野切片图像分类的双流多实例学习网络：结合自监督对比学习。载于IEEE计算机视觉与模式识别会议，第14318–14328页，2021年。3, 4[18] Boyi Li、Kilian Q Weinberger、Serge Belongie、Vladlen Koltun 和 Rene Ranftl。语言驱动的语义分割。载于国际学习表征会议，2022年。3[19] Junnan Li、Ramprasaath Selvaraju、Akhilesh Gotmare、Shafiq Joty、Caiming Xiong 和 Steven Chu Hong Hoi。先对齐，再融合：结合动量蒸馏的视觉与语言表征学习。神经信息处理系统进展，第9694–9705页，2021年。3[20] Yang Li、Jiajun Liu、Brano Kusy、Ross Marchant、Brendan Do、Torsten Merz、Joey Crosswell、Andy Steven、Lachlan Tychsen-Smith、David Ahmedt-Aristizabal 等。用于珊瑚礁调查的实时边缘人工智能系统。载于移动计算与网络国际会议，第 903 906, 2022 1 –页。Timo Lüddecke 和 [21] Alexander Ecker。使用文本与图像提示词进行图像分割。载于IEEE计算机视觉与模式识别会议，第7086–7096页，2022年。3 [22] Peter I Macreadie、Andrea Anton、John A Raven、Nicola Beaumont、Rod M Connolly、Daniel A Friess、Jeffrey J Kelleway、Hilary Kennedy、Tomohiro Kuwae、Paul S Lavery 等。蓝碳科学的未来。《自然·通讯》，10(1):1–13，2019年。1[23] Paul Maxwell、Rod Connolly、Chris Roelfsema、Dana Burfeind、James Udy、Kate O'Brien、Megan Saunders、Richard Barnes、Andrew D Olds、CJ Hendersen 等。摩顿湾昆达穆卡的海草：多样性、生态学与韧性。《摩顿湾昆达穆卡及流域：过去、现在与未来》，第 279 298, 2019 1 –页。 [24] Md Modasshir、Alberto Quattrini Li 和 Ioannis Rekleitis。MDNet：用于珊瑚分类的多块密集网络。载于MTS/IEEE OCEANS，第 1 6, 2018 2 –页。 [25] Md Modasshir 和 Ioannis Rekleitis。利用深度半监督学习方法增强珊瑚礁监测。载于IEEE机器人与自动化国际会议，第1874 1880 页，2020年 1, 2 – 。[26] Serena Mou、Dorian Tsai 和 Matthew Dunbabin。用于缩放珊瑚礁修复的可重构机器人。arXiv预印本 arXiv:2205.04612，2022年。1, 2, 6, 7, 8[27] Peer Neubert 和 Stefan Schubert。作为图像描述符系统化聚合框架的超维计算。载于IEEE计算机视觉与模式识别会议，第16938–16947页，2021年。4

[28] Peer Neubert、Stefan Schubert、Kenny Schlegel 和 Peter Protzel。向量语义表示作为视觉地点识别的描述符。载于机器人学：科学与系统，第1–11页，2021年。4[29] Md Kislu Noman、Syed Mohammed Shamsul Islam、Jumana Abu-Khalaf 和 Paul Lavery。使用半监督学习进行多物种海草检测。载于新西兰国际图像与视觉计算会议，第1–6页，2021年。2, 4, 6, 7, 8[30] Md Kislu Noman、Syed Mohammed Shamsul Islam、Jumana Abu-Khalaf 和 Paul Lavery。使用 Faster R-CNN 与 NASNet 从水下数字图像中检测海草。载于数字图像计算：技术与应用，第1–6页，2021年。2[31] SA Pamungkas、I Jaya 和 M Iqbal。使用 Mask R-CNN 方法从水下图像中分割针叶藻海草。载于IOP 会议系列：地球与环境科学，2021年。2[32] Adam Paszke 等。PyTorch：一种命令式、高性能深度学习库。载于神经信息处理系统进展，2019年。5[33] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议，第8748–8763页，2021年。2, 3, 4, 5, 6, 7, 8[34] Scarlett Raine、Ross Marchant、Brano Kusy、Frederic Maire 和 Tobias Fischer。用于水下图像多物种分割的点标签感知超像素。《IEEE机器人与自动化快报》，7(3):8291–8298，2022年。2[35] Scarlett Raine、Ross Marchant、Peyman Moghadam、Frederic Maire、Brett Kettle 和 Brano Kusy。从水下图像中检测与分类多物种海草。载于数字图像计算：技术与应用，第1–8页，2020年。1, 2, 4, 5, 6, 7, 8[36] Gereon Reus、Thomas Möller、Jonas Jäger、Stewart T Schultz、Claudia Kruschel、Julian Hasenauer、Viviane Wolff 和 Klaus Fricke-Neuderth。寻找海草：用于视觉覆盖度估算的深度学习。载于 MTS/IEEE OCEANS，第1–6页，2018年。2

[37] Alejandro Román、Antonio Tovar-Sánchez、Irene Olivé 和 Gabriel Navarro。使用搭载于无人机的多光谱相机监测海洋大型藻类。Frontiers in Marine Science，第1225页，2021年。2[38] Feifei Shao、Long Chen、Jian Shao、Wei Ji、Shaoning Xiao、Lu Ye、Yueting Zhuang 和 Jun Xiao。弱监督目标检测与定位的深度学习：综述。神经计算，496：192–207，2022年。2[39]Thalles Silva、Alessia Marcolini 和 But Yuhao。sthalles/SimCLR：PyTorch SimCLR，2021年2月。5[40] Satoru Tahara、Kenji Sudo、Takehisa Yamakita 和 Masahiro Nakaoka。利用无人机和深度学习技术绘制海草床物种级地图。PeerJ，10，2022年。2[41] Cedrique Rovile Njieutcheu Tassi、Jakob Gawlikowski、Auliya Unnisa Fitri 和 Rudolph Triebel。对概率取平均的逻辑值对神经网络集成的影响。发表于 AISafety，2022年。5[42] Zhaoqing Wang、Yu Lu、Qiang Li、Xunqiang Tao、Yandong Guo、Mingming Gong 和 Tongliang Liu。CRIS：由 CLIP 驱动的指代图像分割。发表于 IEEE计算机视觉与模式识别会议，第11686–11695页，2022年。3[43] Franz Weidmann、Jonas Jäger、Gereon Reus、Stewart T Schultz、Claudia Kruschel、Viviane Wolff 和 Klaus Fricke-Neuderth。深入观察海草草甸：用于视觉覆盖度估算的语义分割。发表于 MTS/IEEE OCEANS，第1–6页，2019年。2[44] Samuel Wilson、Tobias Fischer、Niko Sünderhauf 和 Feras Dayoub。用于分布外检测的超维特征融合。发表于 IEEE/CVF 冬季计算机视觉应用会议论文集，第2644–2654页，2023年。4 – 。[45] Dingwen Zhang、Junwei Han、Gong Cheng 和 Ming-Hsuan Yang。弱监督目标定位与检测：综述。IEEE模式分析与机器智能汇刊，44(9)：5866–5885，2021年。2[46] Ziqin Zhou、Bowen Zhang、Yinjie Lei、Lingqiao Liu 和 Yifan Liu。ZegCLIP：探索将 CLIP 应用于零样本语义分割。arXiv预印本 arXiv:2212 03588，2022年。3 。

arXiv:2303.00973v2 [cs.CV] 6 Sep 2023

粗略海草分割只需图像标签

补充材料

Scarlett Raine[1,2], Ross Marchant[3], Brano Kusy[2], Frederic Maire[1] 和 Tobias Fischer[1]

[1]QUT机器人中心，昆士兰科技大学，澳大利亚{sg.raine, f.maire, tobias.fischer}@qut.edu.au[2]CSIRO Data61，澳大利亚{scarlett.raine, brano.kusy}@csiro.au[3]Image Analytics，澳大利亚ross.g.marchant@gmail.com

概述

本文是论文《粗略海草分割只需图像标签》的补充材料。我们进一步探究了所提出的分类器集成、SeaFeats 和 SeaCLIP 的性能。第 1节通过更多失败案例的定性示例及对各案例的分析，对主论文中呈现的结果进行了补充。我们还在第2节提供了补充实现细节。

1. 补充定性结果

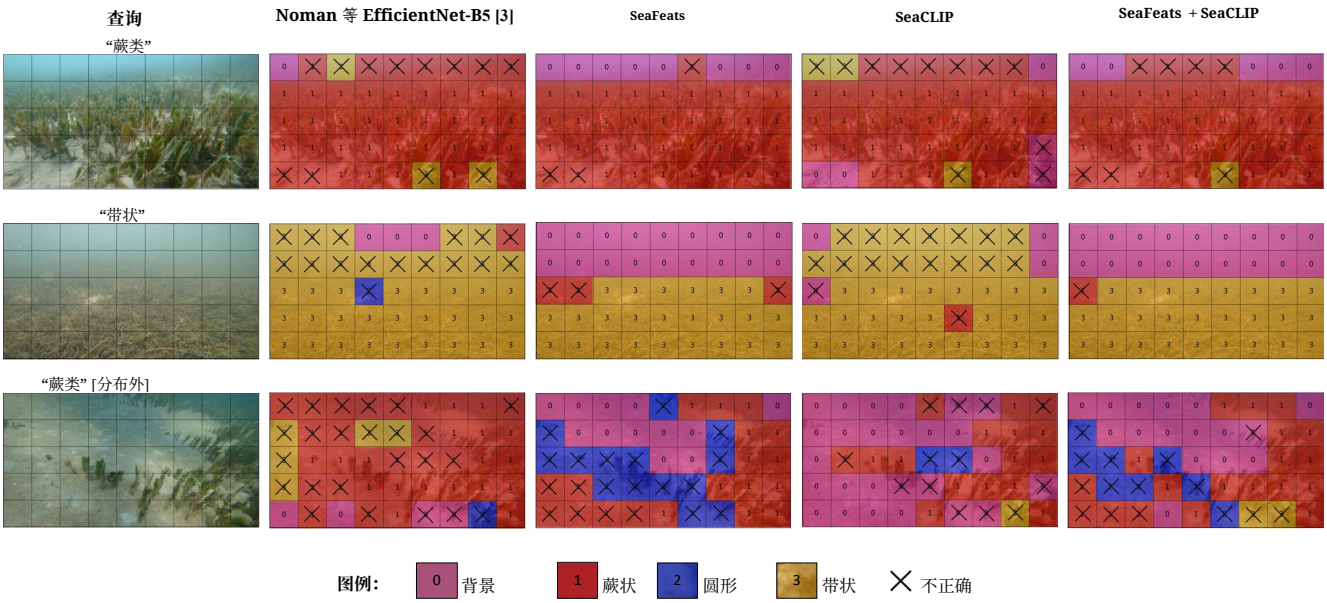
在我们主论文的第 5.2.2 节中，我们展示了将我们的分类器 SeaFeats 和 SeaCLIP 进行集成后得到的输出结果。与单独使用任一分类器相比，这种组合表现出更优的性能，因为总体表现更好的 SeaFeats 模型受益于 SeaCLIP 的保守预测，从而实现了更稳健的总体性能。在本节中，我们将进一步分析组合 SeaFeats 和 SeaCLIP 的影响，并重点关注导致预测错误的失败案例。对于每个示例，我们还会将定性预测结果与我们重新实现的 EfficientNet-B5 方法的输出进行比较，该方法介绍于 [3]。

可能导致失败的因素有多种：边缘图像块容易受到相机畸变影响，从而产生模糊、变暗或变形的图像块（补充图 1，第一行）；推理图像的尺度或分辨率存在显著差异；由于浑浊度、光照或藻类导致图像视觉质量下降（补充图 1，第二行）；对包含分布外海草物种，或处于不同于训练集所见生长阶段的海草的图像进行推理（补充图 1，第三行）；或者推理图像中存在未知物体。这些因素主要是由训练数据的局限性造成的——若数据集能够涵盖更广泛的条件范围、视觉特征、海草外观变化和图像质量，将提高模型泛化到此前未见条件的能力。

补充图 1（第一行）展示了相机畸变和模糊对图像边缘处图像块的影响。与图像中心的清晰图像块相比，SeaFeats 和 SeaCLIP 更有可能错误地分类这些图像块。然而，SeaFeats 能够正确分类大多数图像块；SeaCLIP 则对多个图像块的类别不确定，并将一些边缘图像块误分类为“带状”海草物种（黄色），将另一些误分类为“背景”（粉色）。该示例还展示了分类器集成的一个失败案例：一个模型（SeaFeats）的正确预测被另一个模型（SeaCLIP）的错误预测覆盖。尽管存在这种情况，但总体而言，正如主论文结果部分所示，分类器集成能够带来更优且更稳健的性能。

补充图 1 （第二行）说明，因浑浊度、光照和/或藻类而质量下降的图像可能导致图像块被错误分类。在此示例中，水体具有较高的浑浊度，导致图像呈现雾状外观。尽管图像中的大多数图像块都被正确分类（黄色表示“Strappy”类别），但所有模型都错误地分类了少数图像块。通过增加训练数据集中视觉特征的范围，可以减轻此类失败情况。

补充图 1 （第三行）表明，所有模型都会错误地分类海草视觉外观与训练数据集不同的图像块。这里的 *Halophila spinulosa* 海草不如训练数据集中的海草茂密（尤其是在图像中央），且处于较早的生长阶段。模型所见训练样本的分布需要涵盖海草生长的所有阶段，以及其他因素，包括藻类存在/缺失、季节和天气



补充图1。先前方法 [3] 和我们提出的分类器均失败的示例案例。顶行：由于图像边缘区域存在相机畸变、模糊或较暗区域，许多失败案例发生在图像边缘附近。中行：由于浑浊度、光照或藻类的影响，一些图像的视觉质量可能下降，从而导致失败案例。底行：当模型应用于训练数据分布之外的图像时，海草物种更有可能被错误分类。

条件，才能确保模型能够有效迁移到各种推理图像。

这些失败案例表明，训练数据的可用性和多样性会影响模型性能。未来的工作可以采用人在回路训练或自举方法，根据专家在部署过程中对模型推断的验证或纠正，迭代更新模型。

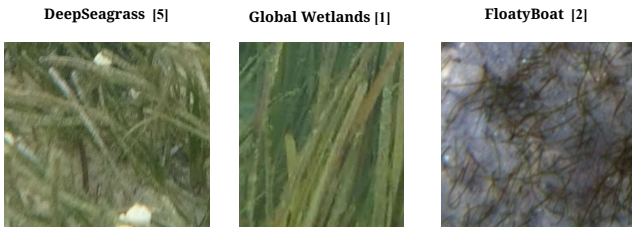
2. 补充实现细节

2.1. SeaCLIP

我们使用 CLIP 大型语言模型对图像块进行伪标注，并据此训练 SeaCLIP [4]。我们使用大小为 520x578 像素的图像块（补充图 2），与 DeepSeagrass 数据集 [5] 保持一致。用于获取二元伪标签的查询短语如下：

- ‘背景’：“一张沙的照片”、“一张水的照片”、“一张沙或水的照片”、“一张模糊的水的照片”、“一张模糊的沙的照片”；以及
- ‘海草’：“一张模糊的海草照片”、“一张含有一些海草的照片”、“一张水下植物的照片”、“一张水下草的照片”、“一张水下绿色、类似草的叶片的照片”、“一张海草的照片”。

我们将图像级海草物种（即“蕨状”“圆形”或“带状”）分配给由



补充图2。来自各数据集的示例图像块：DeepSeagrass [5]（左）、Global Wetlands [1]（中）和 FloatyBoat [2]（右）。不同数据集之间的图像块大小有所不同，但均经过选择，以保持海草外观在尺度上的相似性。

CLIP 在训练时伪标注为“海草”的图像块。

2.2. 在 Global Wetlands 数据集上的推理

我们使用 Global Wetlands 数据集 [1] 来评估所提出模型对未见数据的泛化能力，并评估 SeaCLIP 模型的性能。我们将 Global Wetlands 数据集中的图像划分为 50 个（10x5）图像块，每个图像块的大小为 192x216 像素。我们选择这一网格大小，以保持每个图像块中的海草与 DeepSeagrass 图像块中的海草具有相似的尺度（如补充图 2所示）。

在评估 CLIP 作为零样本分类器在此数据集上的性能时，我们使用了以下提示词：

•‘背景’：“一张沙的照片”、“一张蓝色水体的照片”、“一张浑浊的绿色水体的照片”、“一张沙或水的照片”、“一张模糊的水的照片”、“一张模糊的沙的照片”；

• ‘海草’：“一张模糊的海草照片”、“一张包含一些海草的照片”、“一张水下植物的照片”、“一张水下草的照片”、“一张水下绿色、类似草的叶片照片”、“一张海草照片”；以及

• ‘鱼类’：“一张鱼类照片”、“一张鱼类特写照片”、“一张模糊的鱼类照片”、“一张包含部分鱼的照片”、“一张鱼鳞照片”。

在 Global Wetlands 上训练 SeaCLIP 时，CLIP 生成的伪标签使用了与上述相同的提示词。

2.3. FloatyBoat 数据集微调

为评估模型对 Floaty- Boat [2] 自主水面航行器数据集的泛化能力，我们采用在 DeepSeagrass 上训练的 SeaFeats 模型，并使用 10 张背景图像和 10 张海草图像对其进行微调，共训练 10 轮次。我们使用多种数据增强来提升模型对不同相机特性和影像视角的泛化能力：随机应用会改变颜色通道、线性对比度、高斯模糊、亮度、色调和饱和度的增强。此外，我们还应用几何增强，包括 x 和 y 方向缩放以及左右翻转。

推理时，我们的方法假设每张 FloatyBoat 图像均划分为 6x4 网格，从而得到大小为 468x455 像素的图像块。与 Global Wetlands 数据集类似，我们选择这一网格大小，使每个图像块中的海草呈现的尺度与 DeepSeagrass 数据集中的相近（补充图 2）。

参考文献

[1] Ellen M Ditria、Rod M Connolly、Eric L Jinks 和 Sebastian Lopez-Marcano。用于在不受限制的海草栖息地中自动识别和计数鱼类的带注释视频片段。*Frontiers in Marine Science*，8，2021年。12[2] Serena Mou、Dorian Tsai 和 Matthew Dunbabin。用于缩放珊瑚礁修复的可重构机器人。*arXiv* 预印本 *arXiv:2205.04612*，2022年。12，13[3] Md Kislu Noman、Syed Mohammed Shamsul Islam、Ju mana Abu-Khalaf 和 Paul Lavery。使用半监督学习进行多物种海草检测。载于新西兰国际图像与视觉计算会议，第 1–6 页，2021年。11，12[4] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议，第 8748–8763 页，2021年。12[5] Scarlett Raine、Ross Marchant、Peyman Moghadam、Frederic Maire、Brett Kettle 和 Brano Kusy。基于水下图像的多物种海草检测与分类。载于数字图像计算：技术与应用，第 1 8 页，2020 年 12–。