Image Labels Are All You Need for Coarse Seagrass Segmentation

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

[1]QUT Centre for Robotics, Queensland University of Technology, Australia {sg.raine, f.maire, tobias.fischer}@qut.edu.au
[2]CSIRO Data61, Australia {scarlett.raine, brano.kusy}@csiro.au
[3]Image Analytics, Australia ross.g.marchant@gmail.com

Abstract

Seagrass meadows serve as critical carbon sinks, but estimating the amount of carbon they store requires knowledge of the seagrass species present. Underwater and surface vehicles equipped with machine learning algorithms can help to accurately estimate the composition and extent of seagrass meadows at scale. However, previous approaches for seagrass detection and classification have required supervision from patch-level labels. In this paper, we reframe seagrass classification as a weakly supervised coarse segmentation problem where image-level labels are used during training (25 times fewer labels compared to patch-level labeling) and patch-level outputs are obtained at inference time. To this end, we introduce SeaFeats, an architecture that uses unsupervised contrastive pre-training and feature similarity, and SeaCLIP, a model that showcases the effectiveness of large language models as a supervisory signal in domain-specific applications. We demonstrate that an ensemble of SeaFeats and SeaCLIP leads to highly robust performance. Our method outperforms previous approaches that require patch-level labels on the multi-species 'DeepSeagrass' dataset by 6.8% (absolute) for the class-weighted F1 score, and by 12.1% (absolute) for the seagrass presence/absence F1 score on the 'Global Wetlands' dataset. We also present two case studies for real-world deployment: outlier detection on the Global Wetlands dataset, and application of our method on imagery collected by the FloatyBoat autonomous surface vehicle.

1. Introduction

Blue Carbon refers to the capacity of coastal ecosystems to sequester significant amounts of carbon from the atmosphere [22]. Seagrass meadows play a vital role in this process, but the amount of carbon they store can vary depending on the seagrass species and habitat conditions [15]. Seagrass meadows also improve water quality, protect the coastline, and provide a source of food and nursery shelter for fish [23, 35]. Scientists require detailed data on seagrass meadow extent and composition to accurately estimate blue carbon sequestration [15], as well as for supporting the long-term ecosystem management and conservation of these critical ecosystems [22].

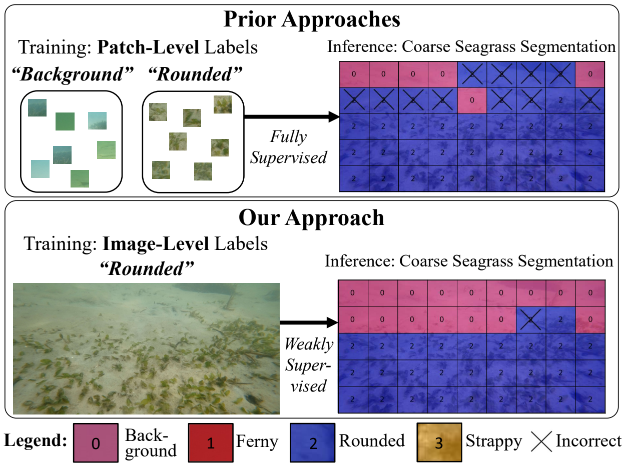


Figure 1. Top: Prior approaches for multi-species seagrass detection and classification relied on fully-supervised training with patch-level labels. Bottom: Our proposed approach is weakly supervised, requiring only image-level labels, while still achieving patch-level coarse seagrass segmentation at inference time. Refer to Section 4.2 for definition of classes 'Background', 'Ferny', 'Rounded' and 'Strappy'.

Autonomous and semi-autonomous underwater survey methods are critical for large-scale underwater imaging [14, 20, 26] and habitat surveys [6]. Underwater vehicles can validate remote geospatial and aerial sensing technology, analyze seagrass species distributions after disasters, and monitor climate change [23].

Machine learning enables automated, accurate processing of image data and allows ecologists to adapt survey transect paths in real-time [13, 20]. However, deep learning approaches require large amounts of labeled data for supervised training, but underwater image pixel-level labeling by domain experts is costly and time-consuming [25]. Seagrass also lacks distinct semantic features and has poorly defined boundaries, making it challenging to detect using typical object detection methods. An alternative to pixel-level labeling is grid-based patch labeling, which reduces annotation costs while enabling seagrass detection and classification [29, 35]. This is consistent with approaches for automated benthic habitat classification [1], which typically employ grid-based or point-grid labeling styles. We refer to class labels assigned to image patches as patch-level labels, while those assigned to the entire image are image-level labels (Fig. 1).

Earlier work in multi-species seagrass classification naively assigned each image patch the image-level label [35], which facilitates fast training but requires images that only contain seagrass – any regions of sand, water, or other objects introduce noise into the dataset. Other research explored a teacher-student framework to enable fast training on unlabeled image patches, however their teacher model requires patch-level labels [29].

In this paper, we propose a principled approach to coarse segmentation from image-level labels. Our framework is an ensemble of two complementary classifiers which leverage unsupervised pre-training and the general semantic knowledge of a large vision-language model to propose pseudo-labels during training.

Our paper presents the following key contributions:

1. We re-frame seagrass classification as a weakly supervised coarse segmentation task, such that only image-level labels are required for training, but patch-level classifications are obtained at inference time.
2. We present an ensemble of two novel methods for multi-species coarse segmentation: 'SeaFeats' uses a novel loss function and classifies patches into background and seagrass by comparing patch features to class templates. 'SeaCLIP' demonstrates that the CLIP large language model [33] can be effectively used as a supervisory signal in domain-specific applications.
3. We present exhaustive experimental trials, in which we outperform the state-of-the-art by 6.8% (absolute F1 score) on the DeepSeagrass dataset [35]. We also provide a variety of ablations that investigate the effectiveness of the various components in our model.
4. We perform two real-world deployment case studies: First, we contribute a labeled test set for the 'Global Wetlands' dataset [4] and demonstrate that SeaCLIP can be effectively used for outlier detection of fish. Second, given just 20 labeled images, we demonstrate generalization capability to underwater imagery collected by the FloatyBoat [26] autonomous surface vehicle.

We make our code available to foster future research on coarse seagrass segmentation[1].

2. Related Work

Research on coral image classification has received considerable attention from researchers [7, 8, 13, 24, 25, 34]. In contrast, seagrass detection and classification has not been explored as much [16, 30, 31, 36, 43], with only a few studies attempting multi-species segmentation [11, 29, 35, 37, 40]. We detail these multi-species approaches in Section 2.1. To address the problem of multi-species seagrass detection and classification, we frame it as a weakly supervised problem, and explore existing methods in this area in Section 2.2. In addition, we discuss recent advances in large vision-language models and their relevance for weak supervision and outlier detection in the field of marine surveys in Section 2.3.

2.1. Multi-species Seagrass Detection/Classification

In the context of multi-species mapping of seagrass at scale, numerous methods have been developed using data collected via unmanned aerial vehicles [37, 40] or remotely piloted aircraft [11]. However, these remote approaches have limitations in terms of their ability to discriminate between seagrass species, and can only operate effectively during low wind conditions, at low tide and at certain times of the day [37, 40].

In contrast, Raine *et al*. [35] posed the problem of in-situ multi-species seagrass classification of underwater images. Specifically, [35] contributed the *'DeepSeagrass'* dataset containing seagrass image patches taken from the viewpoint of an underwater vehicle, and presented a method where each patch is naively assigned the label of the parent image. Noman *et al*. [29] used pseudo ground-truth labels generated by a teacher model to train a student model. Noman *et al*. [29] also contributed the *'ECU-MSS'* multi-species seagrass dataset, which unfortunately is not publicly available[2] and therefore could not be used for evaluation of our methods.

To address the limitations of these approaches, we propose a structured framework for image-level weak supervision for coarse seagrass segmentation, without naively assuming the patch-level labels or requiring time-consuming patch-level labeling by domain experts.

2.2. Weakly Supervised Object Localization

Pixel-level and patch-level labeling of underwater imagery by domain experts is prohibitively costly and time-consuming. However, underwater recognition is dominated by methods which rely on expert knowledge and human labeling [9]. To address this challenge, we explore weakly supervised object detection and localization, where the model is trained only on the image-level label to determine the location and class of instances in images [3, 38, 45]. However, our setting is unique in that we focus on coarse seagrass segmentation. Seagrass grows as a 'carpet' along the seafloor,

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

[2] We contacted the authors, who are not able to share the dataset.

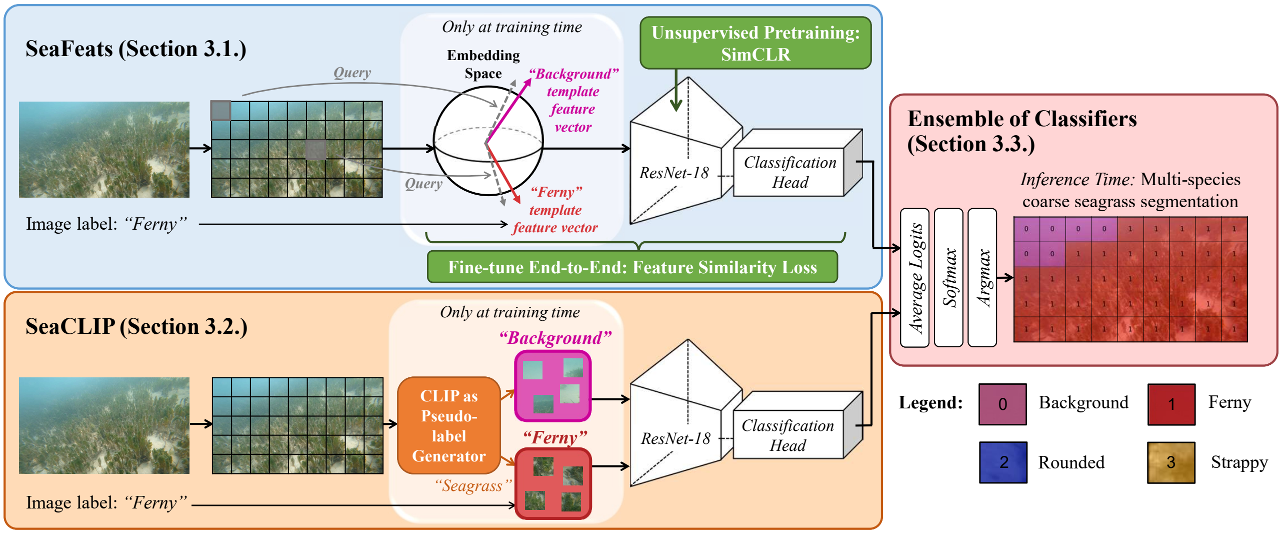


Figure 2. Proposed Algorithm Schematic. We propose an ensemble of two complementary methods for coarse seagrass segmentation. Top: SeaFeats is trained by fine-tuning a feature extractor (initialized with weights from an unsupervised contrastive task) and classifier by finding the cosine similarity between patch-level feature vectors and our per-class template feature vectors (Section 3.1). Bottom: SeaCLIP is trained using zero-shot pseudo-labels predicted by the pre-trained large language model CLIP [33], in combination with the image-level domain-specific seagrass labels (Section 3.2). Right: At inference time, SeaFeats and SeaCLIP are combined in parallel as an ensemble of classifiers and predict the coarse segmentation mask of the query image (Section 3.3).

meaning that all regions of the inference image must be classified, and bounding box proposal is ineffective as there is no clearly defined object/background boundary.

Li *et al*. [17] performed medical whole slide image classification and patch-level tumor localization via self-supervised contrastive pre-training – we adapt such pre-training to distinguish different types of seagrass.

To the best of our knowledge, coarse seagrass segmentation in underwater imagery from image-level labels has not been attempted previously. To this end, we formulate this problem as a weakly supervised task, and propose our novel ensemble of classifiers as an innovative and effective solution.

2.3. Vision-Language Models

In recent years, large vision-language models have garnered significant interest for their potential to perform zero-shot classification of images. These models typically consist of two encoders: one for textual descriptions and another for visual representations [12, 19, 33]. By jointly training these encoders, the models learn to map the textual descriptors and image-based features into a shared embedding space. One notable large vision-language model, CLIP for 'Contrastive Language-Image Pre-training', was trained on 400 million image-caption pairs collected from the internet [33]. The substantial amount of training data, combined with the application of contrastive learning to establish a shared embedding space, results in CLIP possessing a general understanding of visual and semantic concepts, making it applicable to unseen domains and capable of being queried with natural language descriptions of novel classes [33].

Many recent approaches build on CLIP for tasks including zero-shot out-of-distribution detection [5], referring image segmentation [21, 42] and zero-shot segmentation [18, 46]. However, using CLIP as a training signal for patch-level detection and classification in underwater images has not yet been attempted. Furthermore, the general knowledge embedded in the CLIP model has not been leveraged in conjunction with weak, domain-specific annotations. We believe that introducing the use of large vision-language models to the fields of ecology and marine surveys is a novel and important contribution.

3. Method

We propose two novel approaches for coarse seagrass segmentation from image-level labels: SeaFeats and SeaCLIP. To eliminate the supervision required from patch-level annotations, both methods exhibit internal mechanisms for proposing pseudo-labels within training images. Firstly, SeaFeats uses the similarity of deep features to determine whether patches are closer in the embedding space to the background or seagrass species template feature vectors (Section 3.1). Secondly, SeaCLIP demonstrates that the CLIP vision-language model can be used as a supervisory signal for proposing which patches contain seagrass (Section 3.2). By querying CLIP with text phrases describing seagrass, SeaCLIP is particularly effective for blurry, distant or image edge patches – in these cases, it behaves conservatively and will classify patches as background. We therefore combine our two methods in parallel as an ensemble of classifiers, and find that this improves the robustness of our method (Section 3.3). Our full pipeline is shown in Fig. 2.

At inference time, only the ResNet-18 classifiers are required, enabling light-weight deployment and comparable inference times to the state-of-the-art [29, 35].

3.1. SeaFeats – Feature Similarity

Inspired by hyperdimensional computing [27, 28, 44], SeaFeats proposes pseudo-labels by generating high-dimensional template feature vectors for each of the classes (background and each of the seagrass species), and then measuring the similarity of each image patch to these template vectors (Fig. 2, top; Eq. (1)). Given these pseudo-labels, we then train a coarse segmentation framework based on a ResNet-18 encoder followed by a classification head. We fine-tune the architecture end-to-end using our custom feature similarity loss function (Section 3.1.1). We initialize the encoder with weights learned using an unsupervised pretext task (Section 3.1.2). At inference time, we take the trained SeaFeats ResNet-18 encoder and classification head and directly predict the output coarse segmentation mask of a query image.

3.1.1 Feature Similarity Loss Function

We propose a novel loss function inspired by feature aggregation and template matching in hyperdimensional computing [27, 28, 44]. After every epoch, we extract feature vectors from the final layer of the ResNet-18 encoder for each of the patches. We obtain a template feature vector for each class by averaging the L2 normalized patch features corresponding to each class at the image level:

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

where $\mathbf{v}_x$ is the extracted feature vector for patch x and $\mathbb{1}_{[x=c]}$ is an indicator function which returns 1 if the image label x of patch $\mathbf{v}_x$ is equal to class c, and 0 otherwise. Therefore, for each class c, we compute $\bar{\mathbf{v}}_c$, the average of the normalized feature vectors from a sample of N_c patches from class c. It is important to note that the per-class template feature vectors are dynamically updated every epoch as the feature extractor is fine-tuned during training.

During training, we calculate the cosine similarity between our current patch's feature vector $\mathbf{v}_x$ and the template feature vector $\bar{\mathbf{v}}_c$ corresponding to the image-level label c:

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

We also compute the cosine similarity between $\mathbf{v}_x$ and the background template feature vector $\bar{\mathbf{v}}_0$. If the similarity to the background template feature vector is higher, then the patch will be pseudo-labeled as background, despite the seagrass image-level label. Otherwise, the pseudo-label for the patch will correspond to the image-level label,

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

where p_x is the pseudo-label for patch x and the class index 0 corresponds to the background class. We use the categorical cross entropy between the model's output softmax values and the pseudo-label to train the neural network.

3.1.2 Contrastive Pretext Task

As our loss function relies on features extracted from the encoder of our architecture, the initialization of the feature encoder is important for high performance [17]. We use an unsupervised pretext task to obtain a pre-trained ResNet-18 feature extractor, as in [17]. We add our classification head (Section 4.1.1) and then fine-tune the architecture end-to-end with the loss function described in Section 3.1.1.

Specifically, the pretext task we consider uses the SimCLR [2] contrastive learning framework. For a batch of B images, SimCLR creates two random 'views' for each image using data augmentation transforms, giving a total of $2B$ images per batch (see Section 4.1.1) [2]. For each image, the corresponding two augmented views are considered as a 'positive pair', while all other samples in the batch are negative samples [2]. The network is trained to maximize the cosine similarity between the representation of a sample, $\mathbf{v}_i$, and its positive view $\mathbf{v}_j$, and minimize the similarity to the negative samples, $\{\mathbf{v}_k\}_{k \in \{1,\ldots,2B\}, k \neq i}$, using the following contrastive loss [2]:

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

where τ is a temperature parameter which we set to $\tau = 0.07$ as recommended.

3.2. SeaCLIP – Large Vision-Language Model

We leverage the pre-trained large vision-language model, CLIP [33] as a zero-shot pseudo-label generator. CLIP has been trained jointly on image and text embeddings. While CLIP has extremely limited specialist knowledge in seagrass species, it has general semantic knowledge of sand, water and seagrass, which we leverage to predict which patches in training images contain seagrass and which should be labeled as background. Note that CLIP by itself cannot be used to distinguish between different seagrass species.

We use the original CLIP pre-trained model and follow the recommended method for zero-shot classification [33]. As CLIP is trained on natural language prompts, it can be leveraged in this way to filter out any out-of-distribution objects or noisy image patches. We obtain binary background / generic seagrass pseudo-labels from CLIP (see Section 4.1.2 and the Supplementary Material for further details on the prompts we use) and assign the single image-level species-specific seagrass label to the patches pseudo-labeled as generic seagrass by CLIP. We then train our SeaCLIP classifier (Section 4.1.2) end-to-end. In this way, SeaCLIP combines CLIP's general semantic knowledge with domain-specific expertise to obtain accurate patch-level pseudo-labels for training. At inference time, the trained SeaCLIP ResNet-18 encoder and classification head are used directly to predict the class of all query image patches, without needing the expensive inference of CLIP.

3.3. Ensemble of Classifiers

We find that our two proposed methods, SeaFeats and SeaCLIP, are complementary classifiers that result in improved performance when combined in an ensemble (Fig. 2). For example, we observe that SeaCLIP is more likely to classify visually degraded patches as background, rather than overconfidently misclassifying the species. When combined with SeaFeats, this is a desirable property as it prevents incorrect species identification of visually degraded patches, resulting in improved overall robustness of our method. We integrate the two classifiers by averaging the normalized output logits and then applying the softmax function. We find that combining the classifier outputs prior to applying softmax preserves the magnitudes of the logits, thereby resulting in more confident predictions [41]. The scale of the logits from each classifier is normalized prior to averaging to ensure that both classifiers are 'weighted' equally in this operation.

4. Experimental Setup

In this section, we briefly discuss implementation details (Section 4.1), evaluation datasets (Section 4.2), and evaluation metrics used (Section 4.3).

4.1. Implementation

All experiments are conducted with an NVIDIA A100 GPU. Both classifiers were implemented in PyTorch [32]. In the following, we discuss the hyperparameters and implementation details of our approach.

4.1.1 SeaFeats – Feature Similarity

We design our architecture as a ResNet-18 [10] encoder with the final fully connected layers removed, followed by an average pooling layer which applies the classification head to patches in an inference image. Following [35], our classification head is comprised of a fully connected layer with 512 nodes and ReLU activation, dropout with probability of 0.15 to prevent over-fitting, and a final fully connected layer with one node per class. We train our network with the Adam optimizer with an initial learning rate of 0.00001, a batch size of 3 images and a maximum of 150 epochs. To mitigate class imbalance, we weight the categorical cross entropy with per-class weights of $w = [1.0, 1.5, 1.2, 1.2]$ for classes ['Background', 'Ferny', 'Rounded', 'Strappy'].

We initialize our encoder with weights obtained from unsupervised contrastive pre-training. To this end, we use the SimCLR implementation at [39]. We feed patches of resolution 520x578, and set the random cropping operation to 132x132 pixels. We use 132 patches per batch, train for 200 epochs on 42,848 unlabeled training patches from the DeepSeagrass dataset, and use resizing and cropping, horizontal flipping, vertical flipping, color jitter and Gaussian blur to create the positive samples during training.

4.1.2 SeaCLIP – Large Vision-Language Model

SeaCLIP leverages the pre-trained and publicly available CLIP model reported in [33]. To improve the zero-shot performance of CLIP as a pseudo-label generator, we employ basic prompt engineering, where we design multiple query prompts in various forms, *e.g.* "a photo of seagrass", "a blurry photo containing some seagrass" or "a photo of grass-like leaves underwater" (refer to the Supplementary Material for further details). If any one of these prompts has the highest associated probability, then the patch would be pseudo-labeled as the image-level seagrass species in our method. Similarly to our SeaFeats model, we train a ResNet-18 [10] architecture initialized with ImageNet weights using weighted categorical cross entropy and the Adam optimizer with a batch size of 32 patches.

4.2. Datasets

The **DeepSeagrass** dataset [35] consists of 1,701 high resolution (4624x2600 pixels) training images. Eight geographic areas that are not overlapping with the training images were reserved for a test dataset containing 335 images [35]. Images were divided into 40 (8x5) smaller patches based on the image-level class. Prior to performing trials, we manually validated the test dataset and out of the 13,378 test patches, we corrected the class of 591 patches and removed 163 ambiguous patches.

DeepSeagrass considers four broad classes: 'Background', consisting of patches with less than 1% total seagrass, *e.g.* water column, sand or regions of the image which are too blurry to distinguish; 'Ferny' (*Halophila spinulosa*); 'Rounded' (*Halophila ovalis*); and 'Strappy' (*Cymodocea serrulata*, *Halodule uninervis*, *Syringodium isoetifolium* and *Zostera muelleri*) [35].

We propose to use the **Global Wetlands** luderick and seagrass dataset [4] to consider the scenario of a robot deployment of our methods. When an ecologist collects a new set of training images, they will likely contain out-of-distribution

Table 1. Multi-Species Seagrass Classification Results – DeepSeagrass Dataset (Refer to Section 4.3 for Metric Definitions)

Method	**Labels**	**Background**			**Ferny**			**Rounded**			**Strappy**			**Overall**	**Inference**
		Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1	**F1 Score**	**Time (s)**
Zero-shot 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 *et al.* [35]	Patch	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 *et al.* ResNet-50 [35]	Patch	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 *et al.* EfficientNet-B5 [29]	Patch	**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
Ours: SeaFeats+SeaCLIP	Image	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

objects which the ecologist is not interested in. Global Wetlands contains three broad classes: seagrass, background and fish (out-of-distribution). We also use Global Wetlands to evaluate the domain generalization of our approach and for validating our CLIP-supervised method.

Global Wetlands was collected in Moreton Bay, Australia, using remote underwater video cameras. For evaluation, we use the 764 'novel-test' split images which were collected from a different geographic location than the 'train' and 'test' splits. We create 50 (10x5) patches from each image, resulting in 38,200 test image patches, and manually label the patches into 'fish', 'seagrass' and 'background' (refer to the Supplementary Material for further details). The seagrass in this dataset corresponds to the 'Strappy' seagrass morphotype in the DeepSeagrass dataset. We publicly release the labeled test set to facilitate future evaluation on this dataset[3].

4.2.1 FloatyBoat

We also consider the deployment of our methods onboard an autonomous surface vehicle and evaluate the domain generalization of our approach on underwater imagery collected by the FloatyBoat autonomous surface vehicle [26] at Lizard Island, Australia. This imagery exhibits a profound domain shift from the DeepSeagrass dataset, with different camera properties, viewpoint (DeepSeagrass is taken from the oblique angle at 45°, while the FloatyBoat camera is vertically top-down), lighting, turbidity, geographical location and seagrass stage of growth and density (DeepSeagrass contains dense seagrass with $>70\%$ cover, while the FloatyBoat seagrass meadow is sparse). We utilize a survey transect of 1,148 images to evaluate our approach. Prior to applying our method, we perform basic automatic color correction.

4.3. Evaluation Metrics

We report the performance of our method using the per-class precision, recall, F1 scores, and the overall F1 score. We predominantly use the F1 scores because they balance the importance of precision and recall:

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{5}$$

Table 2. Binary Seagrass Classification – DeepSeagrass F1 Scores (Refer to Section 4.3 for Metric Definitions)

Method	**Labels**	**Background**	**Seagrass**	**Overall**
Zero-shot CLIP [33]	Nil	83.41	87.74	85.90
SimCLR [2] + Raine *et al.* [35]	Patch	90.85	94.55	93.17
Raine *et al.* ResNet-50 [35]	Patch	82.94	90.90	88.13
Noman *et al.* EfficientNet-B5 [29]	Patch	84.42	91.54	89.03
Ours: SeaFeats+SeaCLIP	Image	**95.01**	**96.72**	**96.04**

We note that prior approaches have been tuned for high seagrass recall values, however this is at the expense of precision, *i.e.* many background patches are incorrectly predicted as seagrass. In the context of seagrass density estimation, it is important to accurately predict the presence of seagrass to prevent overestimation of coverage and carbon stock, and this is better captured by the F1 scores.

We evaluate the performance of our method in the multi-species use case, in which we consider the four categories described in Section 4.2, as well as the binary (background/seagrass) setting for estimation of seagrass presence/absence.

5. Results

In Section 5.1, we compare our method to the multi-seagrass detection and classification state-of-the-art [29, 35] and also compare our approach to two widely-used classification methods, SimCLR [2] and CLIP [33]. In Section 5.2, we perform ablation studies to demonstrate the effect of feature extractor initialization and our ensemble of classifiers. Finally, in Section 5.3, we consider two case studies for realistic deployment of our methods: outlier detection of unwanted objects in data and deployment on robot platforms.

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

Table 3. SeaFeats Feature Extractor Initialization – DeepSeagrass F1 Scores (Refer to Section 4.3 for Metric Definitions)

Method	Back.	Fern.	Roun.	Strap.	Overall
SeaFeats: Random init	86.48	84.54	86.10	82.16	85.05
SeaFeats: ImageNet init	72.62	90.29	55.18	80.69	77.89
SeaFeats: SimCLR [2] init	**94.93**	**92.44**	**89.56**	**92.12**	**93.10**

5.1. Comparison to State-of-the-art Methods

In Table 1, we compare our method to two state-of-the-art approaches: the ResNet-50 classifier in [35] and the EfficientNet-B5 classifier reported in [29][4]. We also compare to two competitive baseline approaches: using CLIP [33] as a zero-shot classifier; and using SimCLR [2] to pre-train the feature extractor, and then train a linear classification head in a fully supervised manner on the patches from [35].

As shown in Table 1, our approach outperforms all prior approaches and baseline methods for the per-class and overall F1 scores by a large margin in the case of multi-species seagrass classification. Our method improves on the overall F1 score by 6.8% and 7.9% for the multi-species case when compared to [29] and [35] respectively. Our method is also comparable to prior approaches in terms of inference time (~40 fps, as seen in Table 1).

Table 2 shows that we also outperform [29] and [35] on the binary seagrass presence/absence problem, with an improvement of 7.0% and 7.9% respectively.

5.2. Ablation Studies

5.2.1 Effect of Feature Extractor Initialization

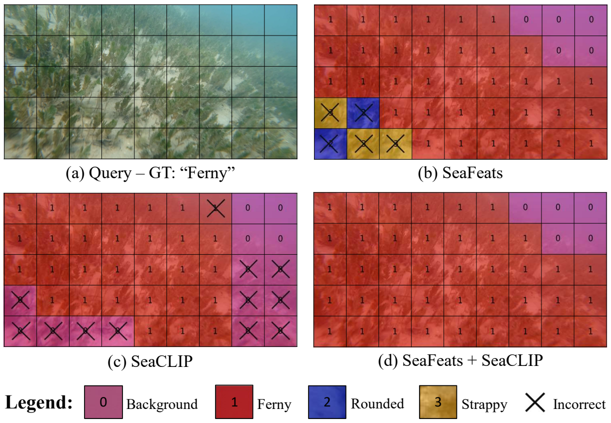
We find that training our feature extractor using the SimCLR unsupervised contrastive pretext task [2] significantly improves the performance of our custom loss function for fine-tuning our architecture, with an improvement of 15.2% and 8.1% when compared to initialization with ImageNet and random weights respectively (Table 3). The feature extractor trained on ImageNet does not transfer well to the seagrass classification task, whereas using the randomly initialized feature extractor captures spatial features and projects the images into a discriminative embedding space without the unhelpful bias learnt from the ImageNet pre-training task.

5.2.2 Ensemble of Classifiers

We find that the performance of the feature similarity classifier (SeaFeats) is more robust when combined in an ensemble with our CLIP pseudo-labeling approach (SeaCLIP), with an improvement of 2.2% and 7.5% in the absolute F1 score as compared to only SeaFeats or only SeaCLIP, respectively (Table 4). Recall that as detailed in Section 3.3, the normalized output logits from each model are averaged at inference time to improve confidence for uncertain classifications. As seen in Fig. 3, the combination of SeaCLIP with our generally higher-performing SeaFeats model results in overall improved performance due to the influence of SeaCLIP on visually degraded patches. We provide additional qualitative results in the Supplementary Material.

Table 4. Effect of Ensemble of Classifiers – DeepSeagrass F1 Scores (See Section 4.3 for Metric Definitions)

Method	Back.	Fern.	Roun.	Strap.	Overall
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**


(a) Query – GT: "Ferny"

(b) SeaFeats

(c) SeaCLIP

(d) SeaFeats + SeaCLIP

Figure 3. For the query image shown in (a), our ensemble of classifiers combines the output of SeaFeats (b) and SeaCLIP (c) to give the combined softmax values in (d). In (b), the blurry patches on the bottom left of the image have been incorrectly predicted as Strappy and Rounded seagrass. In (c), SeaCLIP has conservatively predicted these patches as background. Combining both models (d) prevents incorrect classification and the patches are correctly classified as Ferny seagrass.

5.3. Case Studies for Robot Deployment

In this section, we present two case studies for real-world deployment of our method. First, we consider the 'Global Wetlands' dataset which emulates the scenario where an ecologist collects a new set of images which contain some objects the ecologist is not interested in. We then evaluate the generalization capability of our model to new platforms by applying our method on a survey transect collected by the FloatyBoat [26] autonomous surface vehicle.

In our first case study, we show the effectiveness of CLIP in a problem where the labels are even more limited: image-level labels are not available and all images contain multiple classes. This setting mimics the scenario where an ecologist has collected a set of seagrass images, however the images may contain out-of-distribution classes which we want our classifier to separate from the seagrass and back-

[4] The source code from [29] was not made publicly available and we re-implemented their method to the best of our ability.

Table 5. Global Wetlands F1 Scores (See Section 4.3 for Metric Definitions)

Method	Training Dataset	Back-ground	Fish	Sea-grass	Multi-class Overall	Binary Overall
Zero-shot 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 *et al.* [35]	DeepSeagrass	30.04	n/a	75.04	n/a	63.21
Noman *et al.* [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**

ground patches (in this case, the 'fish' class). We leverage CLIP to predict which image patches belong to each class, acting as the supervisory signal for training our SeaCLIP classifier (refer to the Supplementary Material for details on the prompts we use).

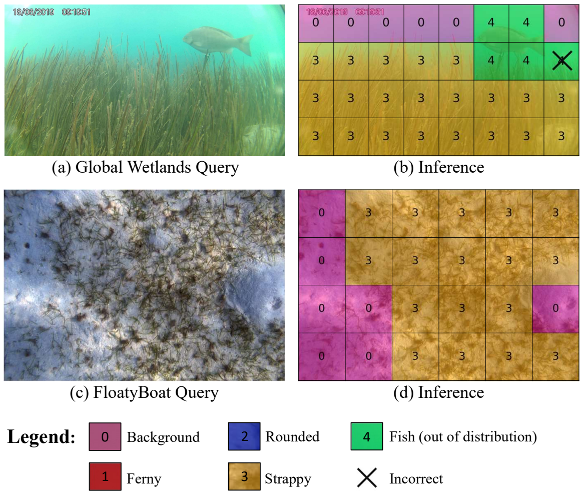
SeaCLIP effectively separates the seagrass, fish, and background classes, achieving 87.9% F1 score with only CLIP as supervision during training (Table 5 and Fig. 4). It is also important to note that in this scenario, where the classes are broader and more visually distinct than in the multi-species seagrass case, the zero-shot CLIP model is able to achieve 85.8% accuracy. SeaCLIP further improves performance by 2.1% compared to using CLIP and has the distinct advantage of performing inference ~30 times faster than zero-shot CLIP (Table 1).

We also evaluate the capacity of our DeepSeagrass-trained method to generalize across the different camera properties and visual conditions of the Global Wetlands dataset. Our ensemble method outperforms [29] and [35] for the seagrass presence/absence problem ("Binary Overall" in Table 5) by 12.1% and 28.9% respectively for the absolute F1 score, and even outperforms SeaCLIP trained on the Global Wetlands dataset by 1.3%, even though our ensemble was only trained on DeepSeagrass.

In our second case study, we evaluate the performance of our method on images collected by an autonomous surface vehicle [26]. We fine-tune our method on only 10 background images and 10 seagrass images for 10 epochs. We demonstrate that our model is able to generalize to a completely different platform and image characteristics, and is able to predict Strappy seagrass and background patches correctly in each frame, as seen in Fig. 4. We note that we could not quantitatively evaluate our method on this dataset as ground truth labels are not available, but both Fig. 4 and the accompanying video demonstrate qualitatively that our method performs well in this setting.

6. Conclusions

In this work, we proposed a weakly supervised method for coarse multi-species segmentation of seagrass, which requires only image-level annotations. Our approach performs seagrass detection and classification at the patch-level during inference using two complementary methods, SeaFeats and SeaCLIP. Our proposed ensemble (SeaFeats+SeaCLIP) achieves a class-weighted F1 score of 95.3% on DeepSeagrass, outperforming the prior state-of-the-art by 6.8%.



(a) Global Wetlands Query

(b) Inference

(c) FloatyBoat Query

(d) Inference

Figure 4. Qualitative evaluation of model generalization to other platforms. Sample test image from Global Wetlands (a) and inference from our SeaCLIP model in (b). A transect image collected by top-down camera onboard FloatyBoat [26] is shown in (c), with corresponding inference from the fine-tuned SeaFeats in (d).

We demonstrated the relevance of pre-trained large vision-language models to the field of ecology and marine surveys, and showed that CLIP provides an effective supervisory signal for weakly supervised training. Our method could be adapted to other applications such as mapping weeds in precision agriculture or for coarse segmentation of remotely sensed imagery for land and vegetation cover analysis.

Our method has been designed with a robotics use-case in mind. The real-time inference will create exciting new research opportunities in adaptive navigation and surveys. For instance, our method could facilitate the survey of high-density seagrass areas or areas with rare seagrass species at a higher level of detail compared to less interesting areas such as sand patches. This capability is particularly critical given the limited prior knowledge about survey areas, as remote observation methods do not provide sufficient detail to guide underwater or surface autonomous vehicles. Other future works could involve pixel-wise semantic segmentation of multi-species seagrass images, seagrass density estimation, and blue carbon stock estimation from our model inferences.

Acknowledgments

This work was done in collaboration between QUT and CSIRO Data61. S.R., F.M., and T.F. acknowledge continued support from the Queensland University of Technology (QUT) through the Centre for Robotics. T.F. acknowledges funding from ARC Laureate Fellowship FL210100156 and Intel Labs via grant RV3.290.Fischer.

References

[1] Qimin Chen, Oscar Beijbom, Stephen Chan, Jessica Bouwmeester, and David Kriegman. A new deep learning engine for CoralNet. In *Int. Conf. Comput. Vis.*, pages 3693–3702, 2021. 2

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Int. Conf. Mach. Learn.*, pages 1597–1607, 2020. 4, 6, 7

[3] Elijah Cole, Kimberly Wilber, Grant Van Horn, Xuan Yang, Marco Fornoni, Pietro Perona, Serge Belongie, Andrew Howard, and Oisin Mac Aodha. On label granularity and object localization. In *Eur. Conf. Comput. Vis.*, pages 604–620, 2022. 2

[4] Ellen M Ditria, Rod M Connolly, Eric L Jinks, and Sebastian Lopez-Marcano. Annotated video footage for automated identification and counting of fish in unconstrained marine environments, 2021. 2, 5

[5] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model CLIP. In *AAAI Conf. on Art. Int.*, pages 6568–6576, 2022. 3

[6] Renata Ferrari, Ezequiel M Marzinelli, Camila Rezende Ayroza, Alan Jordan, Will F Figueira, Maria Byrne, Hamish A Malcolm, Stefan B Williams, and Peter D Steinberg. Large-scale assessment of benthic communities across multiple marine protected areas using an autonomous underwater vehicle. *PLoS One*, 13(3), 2018. 1

[7] Anabel Gómez-Ríos, Siham Tabik, Julián Luengo, ASM Shihavuddin, and Francisco Herrera. Coral species identification with texture or structure images using a two-level classifier based on convolutional neural networks. *Knowledge-Based Systems*, 184:1–10, 2019. 2

[8] Manuel González-Rivero et al. Monitoring of coral reefs using artificial intelligence: A feasible and cost-effective approach. *Remote Sensing*, 12(3):489, 2020. 2

[9] Salma P González-Sabbagh and Antonio Robles-Kelly. A survey on underwater computer vision. *ACM Computing Surveys*, 2023. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 5

[11] Brandon Hobley, Riccardo Arosio, Geoffrey French, Julie Bremner, Tony Dolphin, and Michal Mackiewicz. Semi-supervised segmentation for coastal monitoring seagrass using RPA imagery. *Remote Sensing*, 13(9):1741, 2021. 2

[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Int. Conf. Mach. Learn.*, pages 4904–4916, 2021. 3

[13] Karim Koreitem, Florian Shkurti, Travis Manderson, Wei-Di Chang, Juan Camilo Gamboa Higuera, and Gregory Dudek. One-shot informed robotic visual search in the wild. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, pages 5800–5807, 2020. 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, et al. In-situ data curation: a key to actionable AI at the edge. In *Int. Conf. Mob. Comput. and Net.*, pages 794–796, 2022. 1

[15] Paul S Lavery, Miguel-Ángel Mateo, Oscar Serrano, and Mohammad Rozaimi. Variability in the carbon storage of seagrass habitats and its implications for global estimates of blue carbon ecosystem service. *PloS one*, 8(9), 2013. 1

[16] NA Lestari, I Jaya, and M Iqbal. Segmentation of seagrass (enhalus acoroides) using deep learning mask R-CNN algorithm. In *IOP Conf. Series: Earth and Env. Science*, 2021. 2

[17] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14318–14328, 2021. 3, 4

[18] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *Int. Conf. Learn. Represent.*, 2022. 3

[19] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Adv. in Neur. Inf. Proc. Sys.*, pages 9694–9705, 2021. 3

[20] Yang Li, Jiajun Liu, Brano Kusy, Ross Marchant, Brendan Do, Torsten Merz, Joey Crosswell, Andy Steven, Lachlan Tychsen-Smith, David Ahmedt-Aristizabal, et al. A real-time edge-AI system for reef surveys. In *Int. Conf. Mob. Comput. and Net.*, pages 903–906, 2022. 1

[21] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7086–7096, 2022. 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, et al. The future of blue carbon science. *Nature Communications*, 10(1):1–13, 2019. 1

[23] Paul Maxwell, Rod Connolly, Chris Roelfsema, Dana Burfeind, James Udy, Kate O'Brien, Megan Saunders, Richard Barnes, Andrew D Olds, CJ Hendersen, et al. Seagrasses of moreton bay quandamooka: Diversity, ecology and resilience. *Moreton Bay Quandamooka & Catchment: Past, Present, and Future*, pages 279–298, 2019. 1

[24] Md Modasshir, Alberto Quattrini Li, and Ioannis Rekleitis. MDNet: Multi-patch dense network for coral classification. In *MTS/IEEE Oceans*, pages 1–6, 2018. 2

[25] Md Modasshir and Ioannis Rekleitis. Enhancing coral reef monitoring utilizing a deep semi-supervised learning approach. In *IEEE Int. Conf. Robot. Autom.*, pages 1874–1880, 2020. 1, 2

[26] Serena Mou, Dorian Tsai, and Matthew Dunbabin. Reconfigurable robots for scaling reef restoration. *arXiv preprint arXiv:2205.04612*, 2022. 1, 2, 6, 7, 8

[27] Peer Neubert and Stefan Schubert. Hyperdimensional computing as a framework for systematic aggregation of image descriptors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16938–16947, 2021. 4

[28] Peer Neubert, Stefan Schubert, Kenny Schlegel, and Peter Protzel. Vector semantic representations as descriptors for visual place recognition. In *Robotics: Science and Systems*, pages 1–11, 2021. 4

[29] Md Kislu Noman, Syed Mohammed Shamsul Islam, Jumana Abu-Khalaf, and Paul Lavery. Multi-species seagrass detection using semi-supervised learning. In *Int. Conf. Image and Vis. Comp. New Zealand*, pages 1–6, 2021. 2, 4, 6, 7, 8

[30] Md Kislu Noman, Syed Mohammed Shamsul Islam, Jumana Abu-Khalaf, and Paul Lavery. Seagrass detection from underwater digital images using Faster R-CNN with NASNet. In *Digital Image Computing: Techniques and Applications*, pages 1–6, 2021. 2

[31] SA Pamungkas, I Jaya, and M Iqbal. Segmentation of enhalus acoroides seagrass from underwater images using the Mask R-CNN method. In *IOP Conf. Series: Earth and Env. Science*, 2021. 2

[32] Adam Paszke et al. PyTorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inform. Process. Syst.*, 2019. 5

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, pages 8748–8763, 2021. 2, 3, 4, 5, 6, 7, 8

[34] Scarlett Raine, Ross Marchant, Brano Kusy, Frederic Maire, and Tobias Fischer. Point label aware superpixels for multi-species segmentation of underwater imagery. *IEEE Robot. Autom. Lett.*, 7(3):8291–8298, 2022. 2

[35] Scarlett Raine, Ross Marchant, Peyman Moghadam, Frederic Maire, Brett Kettle, and Brano Kusy. Multi-species seagrass detection and classification from underwater images. In *Digital Image Computing: Techniques and Applications*, pages 1–8, 2020. 1, 2, 4, 5, 6, 7, 8

[36] Gereon Reus, Thomas Möller, Jonas Jäger, Stewart T Schultz, Claudia Kruschel, Julian Hasenauer, Viviane Wolff, and Klaus Fricke-Neuderth. Looking for seagrass: Deep learning for visual coverage estimation. In *MTS/IEEE OCEANS*, pages 1–6, 2018. 2

[37] Alejandro Román, Antonio Tovar-Sánchez, Irene Olivé, and Gabriel Navarro. Using a UAV-mounted multispectral camera for the monitoring of marine macrophytes. *Frontiers in Marine Science*, page 1225, 2021. 2

[38] Feifei Shao, Long Chen, Jian Shao, Wei Ji, Shaoning Xiao, Lu Ye, Yueting Zhuang, and Jun Xiao. Deep learning for weakly-supervised object detection and localization: A survey. *Neurocomputing*, 496:192–207, 2022. 2

[39] Thalles Silva, Alessia Marcolini, and But Yuhao. sthalles/SimCLR: Pytorch SimCLR, Feb. 2021. 5

[40] Satoru Tahara, Kenji Sudo, Takehisa Yamakita, and Masahiro Nakaoka. Species level mapping of a seagrass bed using an unmanned aerial vehicle and deep learning technique. *PeerJ*, 10, 2022. 2

[41] Cedrique Rovile Njieutcheu Tassi, Jakob Gawlikowski, Auliya Unnisa Fitri, and Rudolph Triebel. The impact of averaging logits over probabilities on ensembles of neural networks. In *AISafety*, 2022. 5

[42] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. CRIS: CLIP-driven referring image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11686–11695, 2022. 3

[43] Franz Weidmann, Jonas Jäger, Gereon Reus, Stewart T Schultz, Claudia Kruschel, Viviane Wolff, and Klaus Fricke-Neuderth. A closer look at seagrass meadows: Semantic segmentation for visual coverage estimation. In *MTS/IEEE OCEANS*, pages 1–6, 2019. 2

[44] Samuel Wilson, Tobias Fischer, Niko Sünderhauf, and Feras Dayoub. Hyperdimensional feature fusion for out-of-distribution detection. In *Proceedings of the IEEE/CVF Winter Conf. on Applications of Comput. Vis.*, pages 2644–2654, 2023. 4

[45] Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: A survey. *IEEE Trans. on Pattern. Anal. and Mach. Int.*, 44(9):5866–5885, 2021. 2

[46] Ziqin Zhou, Bowen Zhang, Yinjie Lei, Lingqiao Liu, and Yifan Liu. ZegCLIP: Towards adapting CLIP for zero-shot semantic segmentation. *arXiv preprint arXiv:2212.03588*, 2022. 3