Image Labels Are All You Need for Coarse Seagrass Segmentation

Supplementary Material

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

Overview

This is Supplementary Material for the paper, 'Image Labels Are All You Need for Coarse Seagrass Segmentation'. We further explore the performance of our introduced ensemble of classifiers, SeaFeats and SeaCLIP. Section 1 supplements the results presented in the main paper with additional qualitative examples of failure cases and analysis for each case. We also provide additional implementation details in Section 2.

1. Additional Qualitative Results

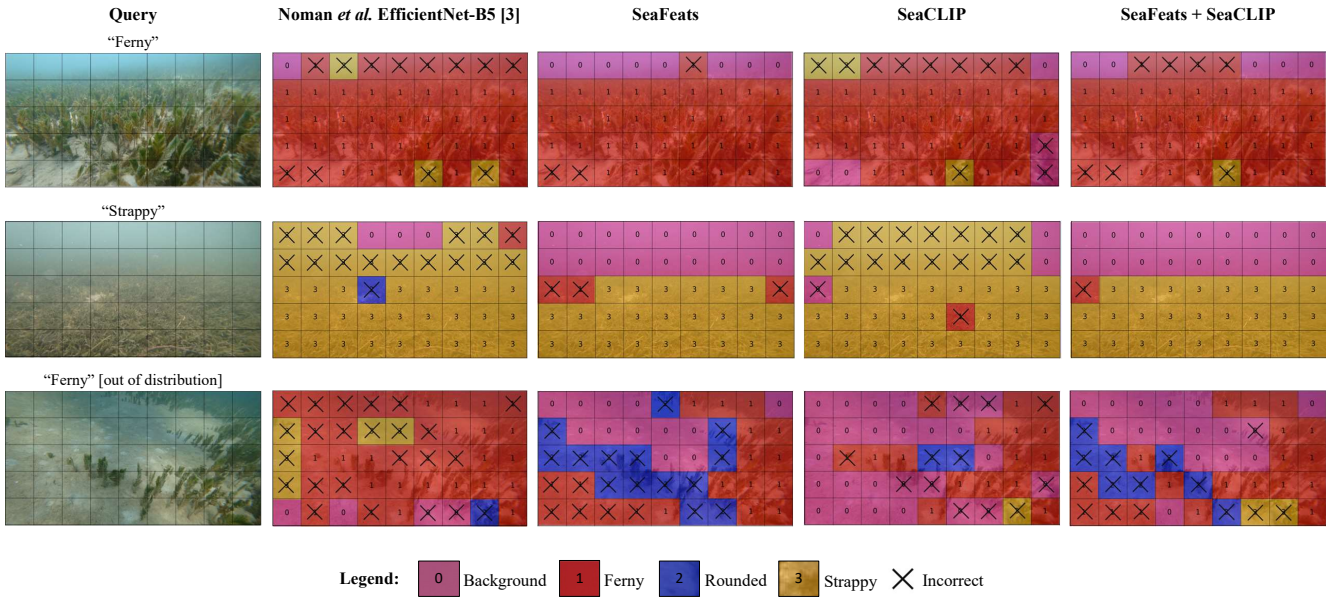
In Section 5.2.2 of our main paper, we present the output results when combining our classifiers, SeaFeats and SeaCLIP, in an ensemble. This combination exhibits superior performance than using either classifier individually, because the generally higher-performing SeaFeats model benefits from the conservative predictions of SeaCLIP to result in more robust performance overall. In this section, we further analyze the effect of combining SeaFeats and SeaCLIP, and we focus on failure cases which result in incorrect predictions. For each example, we also compare our qualitative predictions to the outputs of our re-implementation of the EfficientNet-B5 approach presented in [3].

There are a range of factors which may result in a failure case: edge patches which are subject to camera distortion, resulting in blurry, darkened or warped image patches (Supp. Fig. 1, first row); significant difference in scale or resolution of inference imagery; visually degraded images due to turbidity, lighting or algae (Supp. Fig. 1, second row); inference on images which contain out-of-distribution seagrass species or seagrasses at a different stage of growth than seen in the training set (Supp. Fig. 1, third row); or presence of unknown objects in inference images. These factors are largely caused by limitations of the training data – a larger, more varied dataset which encompasses a wider range of conditions, visual characteristics, seagrass appearance changes and image qualities would improve the ability of the models to generalize to previously unseen conditions.

Supp. Fig. 1 (first row) demonstrates the impact of camera distortion and blur for image patches at the edge of images. SeaFeats and SeaCLIP are more likely to incorrectly classify these patches than the clear image patches at the center of the image. SeaFeats classifies the majority of the patches correctly, however SeaCLIP is uncertain about multiple patches and misclassifies some edge patches as the 'Strappy' species of seagrass (yellow) and others as 'Background' (pink). This example also demonstrates a failure case for the ensemble of classifiers, such that the correct predictions from one model (SeaFeats) are overridden by the incorrect predictions of the other (SeaCLIP). Although there are examples when this occurs, in general the ensemble of classifiers results in improved and more robust performance, as demonstrated in the results section of the main paper.

Supp. Fig. 1 (second row) illustrates that images degraded due to turbidity, lighting and/or algae may result in incorrectly classified patches. In this example, the water column has high levels of turbidity, resulting in an image with a foggy appearance. Although the majority of patches in the image are correctly classified ('Strappy' class in yellow), all models incorrectly classify a few patches. This type of failure case could be mitigated by increasing the range of visual characteristics in the training dataset.

Supp. Fig. 1 (third row) demonstrates that all models incorrectly classify image patches where the seagrass has a different visual appearance as compared to the training dataset. Here, the *Halophila spinulosa* seagrass is not as dense as in the training dataset (particularly in the center of the image), and the seagrass is at an earlier stage of growth. The distribution of training examples seen by the model needs to encompass all stages of seagrass growth and other factors including presence/absence of algae, season and weather



Supplementary Figure 1. Example cases where both the prior approach [3] and our proposed classifiers fail. Top row: Many failure cases occur around the edge of images due to camera distortion, blurring, or darker regions in these parts of the images. Middle row: Some images may be visually degraded due to turbidity, lighting, or algae, resulting in failure cases. Bottom row: When models are deployed on images from outside the distribution of the training data, the species of seagrass are more likely to be incorrectly assigned.

conditions to ensure that models can effectively transfer to a variety of inference images.

These failure cases demonstrate how the availability and variation of training data impacts on model performance. Future work could include human-in-the-loop training or bootstrapping to iteratively update models based on expert verification or correction of model inferences during deployment.

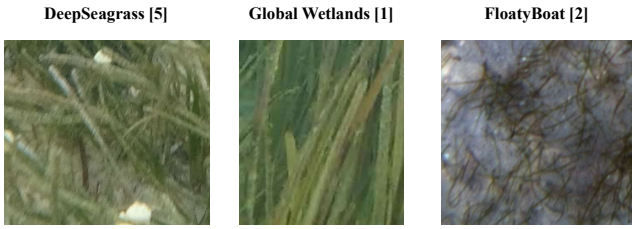
2. Additional Implementation Details

2.1. SeaCLIP

We train SeaCLIP on image patches pseudo-labeled by the CLIP large language model [4]. We use patches of size 520x578 pixels (Supp. Fig. 2), as per the DeepSeagrass dataset [5]. The query phrases used for obtaining the binary pseudo-labels are:

- 'Background': "a photo of sand", "a photo of water", "a photo of sand or water", "a blurry photo of water", "a blurry photo of sand"; and
- 'Seagrass': "a blurry photo of seagrass", "a photo containing some seagrass", "a photo of underwater plants", "a photo of underwater grass", "a photo of green, grass-like leaves underwater", "a photo of seagrass".

We assign the image-level seagrass species (i.e. 'Ferny', 'Rounded' or 'Strappy') to the patches pseudo-labeled by CLIP as 'Seagrass' at training time.



Supplementary Figure 2. Example patches from each of the datasets: DeepSeagrass [5] (left), Global Wetlands [1] (center) and FloatyBoat [2] (right). The patch size between the datasets differ, but are selected to maintain a similar scale in terms of seagrass appearance.

2.2. Inference on Global Wetlands Dataset

We use the Global Wetlands dataset [1] to evaluate the ability of our proposed models to generalize to unseen data and to assess the performance of the SeaCLIP model. We process the Global Wetlands dataset by splitting images into 50 (10x5) patches, resulting in a patch size of 192x216 pixels. We selected this grid size to maintain a similar scale for the seagrass within each image patch as for the DeepSeagrass patches (as seen in Supp. Fig. 2).

When evaluating the performance of CLIP as a zero-shot classifier on this dataset, we use the following prompts:

- 'Background': "a photo of sand", "a photo of blue water", "a photo of murky, green water", "a photo of sand or water", "a blurry photo of water", "a blurry photo of sand";
- 'Seagrass': "a blurry photo of seagrass", "a photo containing some seagrass", "a photo of underwater plants", "a photo of underwater grass", "a photo of green, grass-like leaves underwater", "a photo of seagrass"; and
- 'Fish': "a photo of fish", "a close-up photo of fish", "a blurry photo of fish", "a photo containing part of a fish", "a photo of fish scales".

When training SeaCLIP on Global Wetlands, the CLIP-generated pseudo-labels were created using the same prompts as above.

2.3. Fine-tuning for FloatyBoat Dataset

For evaluation of model generalization to the FloatyBoat [2] autonomous surface vehicle dataset, we take the SeaFeats model trained on DeepSeagrass and fine-tune it on 10 background images and 10 seagrass images for 10 epochs. We use a range of data augmentations to improve the ability of the model to generalize to the different camera characteristics and imagery viewpoint: we randomly apply augmentations which vary the color channels, linear contrast, Gaussian blur, brightness, hue and saturation. We additionally apply geometric augmentations including x and y scaling, and left/right flipping.

At inference time, our approach assumes a 6x4 grid for each FloatyBoat image, resulting in patches which are 468x455 pixels. Similarly to the Global Wetlands dataset, this grid size is selected so that the seagrass appears at a similar scale within each patch as the DeepSeagrass dataset (Supp. Fig. 2).

References

[1] 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 seagrass habitats. *Frontiers in Marine Science*, 8, 2021. 2

[2] Serena Mou, Dorian Tsai, and Matthew Dunbabin. Reconfigurable robots for scaling reef restoration. *arXiv preprint arXiv:2205.04612*, 2022. 2, 3

[3] 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. 1, 2

[4] 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

[5] 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. 2