Human-in-the-Loop Segmentation of Multi-species Coral Imagery

Supplementary Material

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

[1]QUT Centre for Robotics, Australia {sg.raine, f.maire, niko.suenderhauf, 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 'Human-in-the-Loop Segmentation of Multi-species Coral Imagery'. First, we extend two ablation studies in the main paper (Section 1): we investigate the impact of the value of k in the KNN classifier used for clustering deep pixel embeddings for different quantities of point labels, and we evaluate additional variations of the DINOv2 feature extractor used in our approach. Section 2 then outlines the process used for cleaning the UCSD Mosaics dataset. Finally, we include supplementary qualitative results in Section 3 and provide additional discussion of the results presented.

1. Extended Ablation Studies

This section outlines extensions of two of the ablation studies in the main paper: the value of k used by the K-Nearest Neighbor algorithm for clustering deep pixel features (Section 1.1), and the DINOv2 feature extractor used for generating per-pixel deep embeddings (Section 1.2).

1.1. Effect of k in KNN

In Fig. 5 of the main paper, we evaluate the impact of k in the KNN algorithm used for clustering pixels in the deep embedding space. We demonstrate that the best performance for 25 point labels is when $k = 1$. In this section, we perform a more comprehensive evaluation of values of k as the quantity of point labels is also varied (5, 10, 25, 100 and 300 point labels). Fig. 1 shows the results of this ablation.

For all values of point labels, the best performance is achieved when $k = 1$, *i.e.* when a nearest neighbor classifier is used. The effect is particularly pronounced for small values of point labels because there are fewer examples for the clustering algorithm, *i.e.* for 5 point labels in an image, it is likely that there is only one labeled point per class in the set. This means there is no benefit from taking the majority of three or five neighbors, as only one of the neighbors will be the correct class label.

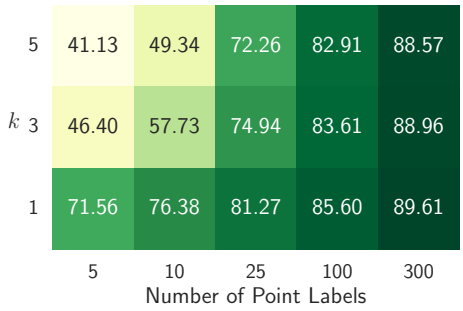

Figure 1. Effect of increasing the value of k on the pixel accuracy of point label propagation. For all quantities of point labels, the best pixel accuracy is achieved when $k = 1$, *i.e.* a nearest neighbor classifier is used for clustering deep pixel features.

1.2. DINOv2 Variations

In Section 5.2.1 of our main paper, we evaluate the impact of using the denoised version of DINOv2 described in [8], and establish that denoising the feature embeddings leads to an improvement in clustering pixel features. In this section, we also compare the variation of DINOv2 trained with registers [3], and the denoised version of DINOv2 trained with registers [3, 8]. The features are visualised by reducing the dimensions with Principal Components Analysis (PCA) into the RGB colour space in Fig. 2. This figure shows that training DINOv2 with registers reduces some of the feature artefacts, but the effect is not as pronounced as for the denoised features [8]. The features obtained through training DINOv2 with registers [3] as well as denoising the features [8] are not as clean as for the denoised features from the

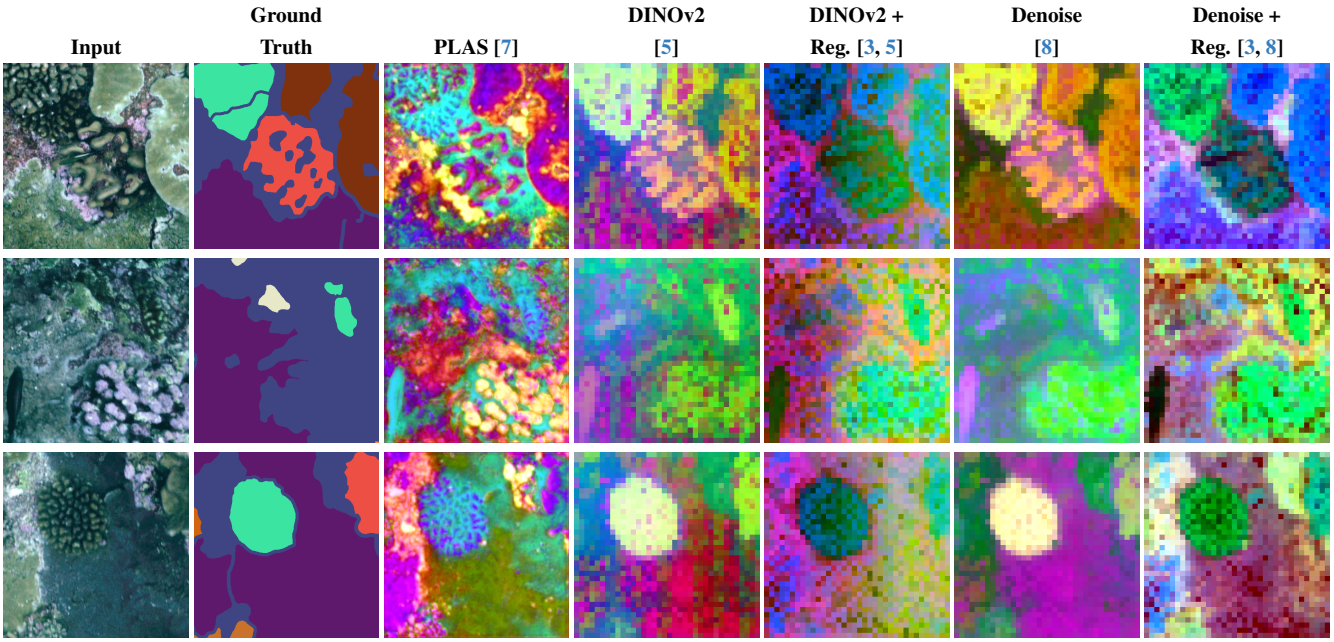

Figure 2. Comparison of Point Label Aware Superpixels [7] features, DINOv2 raw features [5], DINOv2 trained with registers features [3], denoised DINOv2 features [8], and denoised DINOv2 features trained with registers [3, 8] for UCSD Mosaics coral images. For the transformer approaches, features for every 14x14 pixel patch in the original image have been upsampled with bilinear interpolation. All features are reduced to RGB for visualisation with Principal Components Analysis (PCA). Pixels with similar RGB colors are similar in the deep embedding space. The CNN features used by Point Label Aware Superpixels (PLAS) [7] do not effectively group pixels into meaningful segments. The denoising model clearly reduces the position embedding artefacts, resulting in smoother, cleaner features and therefore improved clustering performance.

original DINOv2. This is reflected in the quantitative results for this ablation, shown in Table 1, which shows that the highest performance across the three metrics is for the denoised DINOv2 model.

2. Dataset Details

In Section 4.2 of the main paper, we describe the UCSD Mosaics dataset, which is used for development and evaluation of the point label propagation approach. This dataset is the only multi-species coral image dataset where the images are accompanied by pixel-wise ground truth masks. It was originally collected and contributed by [4], and has been used extensively in the coral segmentation literature [1, 2, 6, 7]. We noticed a small number of ground truth masks in the dataset are corrupted, so we excluded these from the dataset. Fig. 3 demonstrates the issue with the ground truth masks. The dataset was carefully inspected and 219 images were removed from the training set, resulting in 3,974 images and another 32 were removed from the test split, yielding 696 images. Although it is unlikely that this small quantity of images would significantly impact the reported results, we re-ran the comparison approaches [6, 7] on the cleaned version of the dataset for accurate evaluation. The specific details for the images in the cleaned version of the dataset can be found at https://github.com/sgraine/HIL-coral-segmentation.

3. Additional Qualitative Results

In Fig. 6 of the main paper, we show a selection of example images and compare our point propagation approach which leverages DINOv2, KNN and our smart point selection regime with prior approaches Fast Multi-level Superpixel Segmentation [6] and Point Label Aware Superpixels [7]. In this section, we provide a more comprehensive version of the figure, which shows the augmented ground truth masks from each of the approaches and each of the four example images (Fig. 4).

This figure highlights that grid-based sparse labels improve the coverage over randomly placed sparse labels. Row 6 shows that for the Fast MSS approach [6], one of the beige segments is entirely missed by the randomly placed points but captured by the grid points.

The modes of failure for Fast MSS and the Point Label Aware Superpixel approach in the 5 pixel case can be observed in Fig. 4. Fast MSS fails to produce useful segments because only segments containing a point label are used in

Table 1. Effect of DINOv2 Feature Extractor Variations (Refer to Section 4.3 of the Main Paper for Metric Definitions)

Method	PA 5 / 10 / 25 / 300	mPA 5 / 10 / 25 / 300	mIoU 5 / 10 / 25 / 300
DINOv2 [5]	68.58 / 73.32 / 76.94 / 88.10	60.23 / 68.04 / 70.97 / 85.58	50.28 / 55.76 / 61.61 / 83.79
DINOv2 with Registers [3, 5]	68.49 / 73.12 / 76.65 / 87.41	59.79 / 67.48 / 72.44 / 84.84	49.80 / 55.96 / 61.46 / 82.68
Denoised DINOv2 [5, 8]	**71.57** / **76.38** / **80.71** / **89.61**	**61.46** / 69.87 / **75.91** / **86.45**	**52.60** / **59.48** / **67.97** / **85.00**
Denoised DINOv2 with Registers [3, 5, 8]	70.15 / 75.41 / 78.88 / 88.16	61.85 / **70.75** / 75.81 / 85.42	52.36 / 59.47 / 67.28 / 83.68

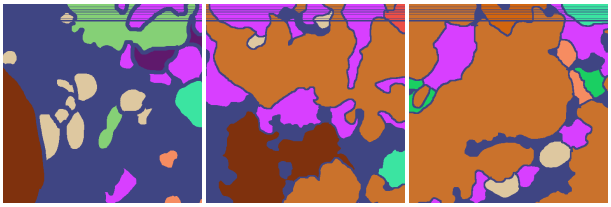

Figure 3. Some of the ground truth masks in the UCSD Mosaics dataset exhibited corruption, as seen at the top of these examples.

the augmented ground truth mask. In the case that segments do not contain any points (which occurs frequently in this setting), the unknown/unlabeled class is used, meaning that the majority of the mask is this class. In the case of the Point Label Aware Superpixel approach, any superpixel segment that does not contain a point label is labeled based on feature similarity with the segments which do have an associated label. This results in significant over-prediction of classes. In addition, the point label aware superpixel approach relies on sufficient points for the conflict loss function to force the boundaries of superpixels to neatly conform to species [7].

Our DINOv2 and KNN approach effectively produces augmented ground truth masks, even in the extremely sparse label setting. However, one limitation of our approach is that spatially small species segments can be missed when there are very few point labels available (as seen in row 1 of Fig. 4, the orange segment is not included in the augmented ground truth). One avenue for future work would be to incorporate mechanisms which place more emphasis on species which are spatially small and prevent model bias towards species with larger instances.

References

[1] Iñigo Alonso and Ana C Murillo. Semantic segmentation from sparse labeling using multi-level superpixels. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5785–5792, 2018. 2

[2] Iñigo Alonso, Matan Yuval, Gal Eyal, Tali Treibitz, and Ana C Murillo. CoralSeg: Learning coral segmentation from sparse annotations. *Journal of Field Robotics*, 36(8):1456–1477, 2019. 2

[3] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 1, 2, 3

[4] Hannah M Murphy and Gregory P Jenkins. Observational methods used in marine spatial monitoring of fishes and associated habitats: A review. *Marine and Freshwater Research*, 61(2):236–252, 2010. 2

[5] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3

[6] Jordan P Pierce, Yuri Rzhanov, Kim Lowell, and Jennifer A Dijkstra. Reducing annotation times: Semantic segmentation of coral reef survey images. In *Global Oceans*, pages 1–9, 2020. 2, 4

[7] Scarlett Raine, Ross Marchant, Brano Kusy, Frederic Maire, and Tobias Fischer. Point label aware superpixels for multi-species segmentation of underwater imagery. *IEEE Robotics and Automation Letters*, 7(3):8291–8298, 2022. 2, 3, 4

[8] Jiawei Yang, Katie Z Luo, Jiefeng Li, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers. *arXiv preprint arXiv:2401.02957*, 2024. 1, 2, 3, 4

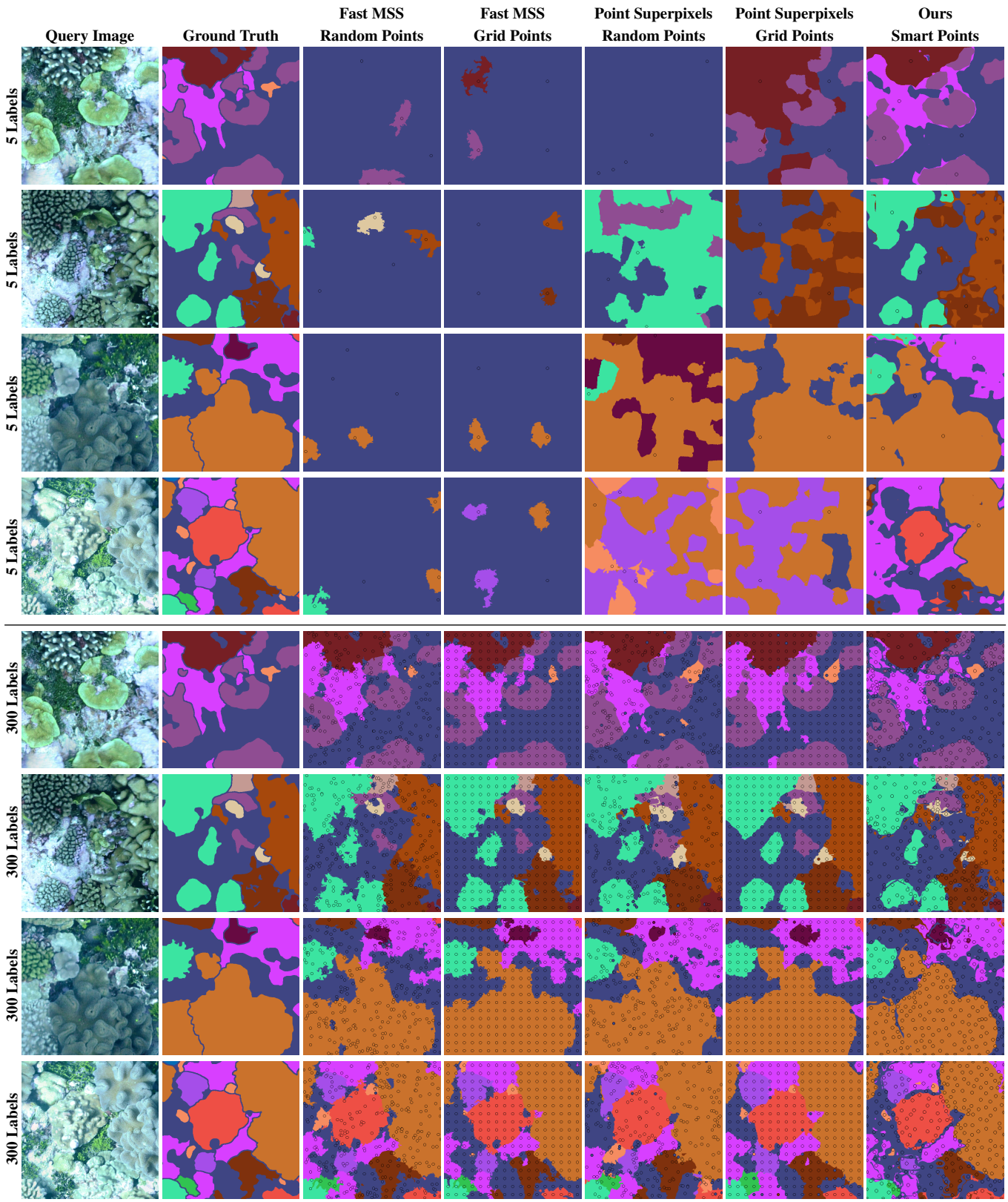

Figure 4. Additional Qualitative Results. Comparison between the Fast MSS approach [6], the point label aware superpixel approach [7] and our approach, based on denoised DINOv2 features [8], K-Nearest Neighbors and our Human-in-the-Loop labeling regime. The same four examples are shown for all approaches. The top section shows point propagation for 5 labels, and the bottom section demonstrates point propagation when there are 300 labels available. The pixels used in the point label propagation are shown as black circles within the output augmented ground truth masks.