Towards Dynamic Adaptation of Marine Surveys: Leveraging Fine-grained Segmentation from Sparse Labels

Scarlett Raine[1,2], Ross Marchant[1,2], Frederic Maire[1], Niko Sünderhauf [1], Brano Kusy[2]

Abstract— **There has been a decline in hard coral cover on the Great Barrier Reef, due to environmental stressors such as crown-of-thorns starfish (COTS) predation, cyclones and increased ocean temperatures due to climate change. Marine surveys are used to detect and monitor changes in reef ecosystems and provide intelligence to authorities to inform decision-making and mitigation strategies. Recent advances in broad-scale survey methods, including use of autonomous, remotely operated and towed underwater vehicles, have significantly increased the speed and extent of marine surveys. Real-time analysis of imagery from these systems enables on-the-fly adjustment of survey paths and sampling efforts during field operation. In particular, identification of various benthic classes can be used to guide the robots towards regions of interest, for example when searching for COTS. In this paper, we investigate benthic habitat identification to solve this part of the robot perception system. Our method involves training neural networks to perform fine-grained semantic segmentation from existing sparsely labelled underwater images. We use the XL CATLIN Seaview Survey as a basis for our experiments and establish a new train / test split and baseline accuracy for this dataset.**

I. INTRODUCTION

Marine surveys aim to identify and monitor reef status changes, and were traditionally completed manually by ecologists [1]. Recent advances in broad-scale marine survey methods have increased the range and accuracy of survey data [2], [3], [4], [5]. The efficiency of broad-scale marine surveys can be improved by dynamic adaptation of the survey trajectory and/or sampling effort, based on real-time analysis of imagery collected on-board [6].

Traditionally, ecologists collect photographs at intervals along transects and then scale each image to a 1m by 1m quadrat [7]. Photo-quadrats are analysed by randomly distributing points in the image and labelling into taxonomic or morphological classes [8]. Approximately 300,000 images can be collected per day using a towed underwater platform. These images must then be annotated by domain experts. Point labels are an efficient technique for labelling large quantities of images and offer a middle ground between the weak training signal of image-level labels and the intensive, time-consuming and costly method of densely labelling every pixel in each image [9]. Semantic segmentation is a dense prediction task in which every pixel in a query image is classified into one of a number of predefined classes [10]. Real-time semantic segmentation can inform guidance of a towed underwater platform towards areas with more interesting information (Fig. 1), e.g. larger population of Crown of Thorns starfish, the presence of bleached or otherwise damaged coral, or abundance of macro-algae [11].

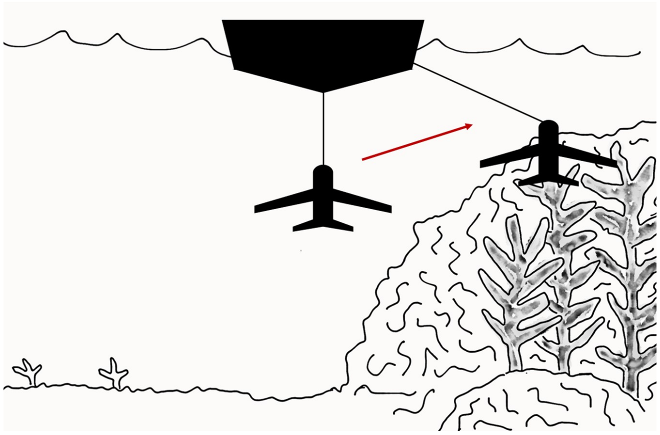


Fig. 1. Dynamic adaptation of towed underwater vehicle guidance based on segmentation model outputs

Semantic segmentation is performed by a neural network typically trained on densely labelled data in the form of image-mask pairs, where every pixel in the training image is assigned a class label. This paper compares methods and makes recommendations for achieving dense segmentation of benthic imagery using a neural network trained on sparse, randomly distributed point labels instead of masks.

The XL CATLIN Seaview Survey dataset (referred to as the 'CATLIN' dataset herein) is the largest publicly available database of marine images curated by domain experts and consists of 1.1 million standardised photo-quadrat images [12]. This work investigates semantic segmentation architectures trained on sparse point-labelled images from the CATLIN dataset. After reviewing existing work on segmentation of underwater imagery and approaches for training networks from sparsely labelled imagery in Section II, we review existing underwater datasets in Section III and outline our proposed train/test split for the CATLIN dataset. We present our fine-grained segmentation approach and methodology in Section IV. We evaluate and compare three architectures, analyse the impact of sampling strategies on training with imbalanced data and introduce a novel sampling strategy (Section V). The use of ground truth label propagation as a pre-processing step is analysed and a simple, computationally efficient method is introduced (Section V). Finally, we establish the baseline accuracy for

[1] QUT Centre for Robotics, Queensland University of Technology, Brisbane, Australia `sg.raine, r.marchant, f.maire@qut.edu.au`

[2] CSIRO Data61, Brisbane, Australia `scarlett.raine, ross.marchant, brano.kusy@csiro.au`

fine-grained segmentation of the CATLIN imagery (Section V) and discuss our findings in Section VI.

II. Related Work

Automating the analysis of underwater imagery is an active field of research at the intersection of the computer vision and marine biology communities. This section reviews advances in semantic segmentation of underwater imagery and recent approaches for training models from sparse or weak labels.

A. Segmentation of Underwater Imagery

There are many existing approaches for classification of coral images [13], [14], [15], [16], [17], [18], [19], [20]. The CATLIN dataset [12] was used to train a VGG-16 convolutional neural network (CNN) to classify patches of CATLIN imagery [21]. The authors achieved a 97% agreement between the model's and an expert's estimations of benthic abundance, however the results of the classification task itself were not published [21]. There are limited approaches for pixel-wise segmentation of benthic imagery (such as [22]), and even fewer are trained from sparse point labels. Islam *et al.* developed a segmentation method and contributed a custom dataset with human-generated segmentation masks, however the segmentation classes were coarse-grained e.g. "background", "robot", "plant", "human" [23]. Pavoni *et al.* performed semantic segmentation of orthomosaic benthic images using a Bayesian CNN trained from human-generated polygons [24]. The work most similar to ours is called 'CoralSeg', which investigates segmentation of corals from sparse point labels [25]. The authors build on their earlier work in [26] and use a multi-level super-pixel approach to propagate the ground truth and then train on augmented dense masks, as discussed in Section II-C. A review of neural network methods for recognition of marine benthos highlighted the relative lack of approaches for fine-grained pixel-wise segmentation as opposed to frame level classification [27].

Coral species recognition is complicated by variation in the size, colour, texture and shape of corals within the same class, and by the difficulty in discerning boundaries between instances [13]. The visual traits of corals are particularly challenging due to the plasticity of forms, lack of shape definition, intricacy and density of growth and the varying scale of discriminative features [21]. The resolution of the labels introduces in-class variability as one label could describe several morphologically diverse species [21], highlighting the importance of the fine-grained segmentation case. Label resolution is especially critical for algae as there are approximately 630 species on the Great Barrier Reef alone, making functional grouping of species an important consideration [28]. Environmental and geographical conditions impact the morphology of coral reef benthos, as increased depth encourages growth structure to maximise light capture [21]. For these reasons, segmentation of fine-grained benthic classes is a difficult machine learning task which is under-explored in the existing literature.

B. Weakly Supervised Segmentation

Weakly supervised semantic segmentation refers to the task of generating a dense pixel-wise mask from a model which has been trained on weak labels [29]. These labels could take the form of bounding boxes, polygons, scribbles, point labels or whole image labels [29]. Bridging the gap between whole image labels and pixel-wise segmentation has attracted a large amount of research effort, however the gap between point based labels and segmentation has comparably fewer approaches.

Bearman *et al.* used a single point per object to train a segmentation architecture based on VGG-16 [9]. They added a convolution layer at the output corresponding to the number of classes and then a deconvolution layer to bilinearly upsample and obtain a mask [9]. They also incorporated an objectness prior and found an improvement in the boundaries of their object segmentation [9]. This approach cannot be applied to the context of coral segmentation because in this setting there is no background and foreground, all pixels in a photo-quadrat image must be designated a class. The nature of randomly placed points also means that there could be many classes present in the image which have not been labelled, nullifying the assumption that every "object" has a labelled point. Wang *et al.* used a U-Net architecture to segment cropland from satellite images [30]. Their method was trained using one randomly placed point label in each image and they determined that the architecture was able to achieve greater than 85% accuracy when more than 100 training examples were provided [30]. We extend this approach to enable training for the fine-grained multi-class and multi-label case and present results using this method.

C. Label Augmentation with Super-pixels

Friedman proposed a super-pixel method for using existing underwater photo-quadrat point-labelled images to train a pixel-wise segmentation network [31]. They assumed that homogeneous regions within an image belonged to a super-pixel and used the mean-shift segmentation and edge detection algorithm to segment images before consolidating super-pixels with point labels [31]. The labelled super-pixels were used to extract LBP, colour and shape descriptors which were used for classification with a support vector machine with a radial basis function [31]. Yu *et al.* proposed an iterative approach which used random point labels as input and a CNN to extract features before leveraging Latent Dirichlet Allocation (LDA) to generate additional labels to augment the original labels for a second round of training [32]. Yu *et al.* presented a method which used the Simple Linear Iterative Clustering (SLIC) algorithm to create super-pixels based on colour and therefore extra labels by propagation [33]. They implemented a coarse-to-fine approach to reduce computation time and only performed segmentation on the most likely areas of coral [33]. This approach considered five broad classes: Coral, Red Algae, Green Algae, Rock and Other [33]. Alonso *et al.* developed CoralSeg based on the DeepLabv3+ architecture [25]. The network was trained on augmented ground truth obtained using an iterative, multilevel approach to super-pixels [26]. This method generated super-pixels at a number of levels and assigned known labels to the super-pixels at those locations. The different levels were combined to “fill-in” the image, thus propagating the ground truth labels across the whole image [26]. The final number of super-pixels in the augmented ground truth mask poses a critical trade-off between maintaining the accuracy of the super-pixels and filling in the image with the propagated labels [26]. This method is computationally expensive as it relies on repeated generation of super-pixels. For an image of 1024x1024 pixels, the computation time was 113.56s. In this paper, we propose a simple, computationally efficient method of propagating point labels using the Delaunay triangulation of the points.

III. DATASETS

A. Existing Coral Datasets

There are a number of coral datasets (summarised in Table I) which have previously been used to assess and demonstrate the performance of deep learning approaches for underwater detection, classification or segmentation. The CATLIN dataset is the largest expert-labelled dataset with the largest number of fine-grained classes, and is therefore ideal for comparison of methods for fine-grained segmentation of underwater imagery.

B. Dataset Preparation

The CATLIN dataset consists of 1.1 million photo-quadrat images collected at nine locations worldwide [12]. A subset of 11,387 images were labelled using randomly placed points, yielding 858,257 labels [12]. Images were collected from a top-down viewpoint and scaled to 1m by 1m of the seafloor [12]. The labels are provided as a hierarchy, with a total of 228 specific benthic classes which are merged into 62 classes and 5 broad super-classes (Hard Coral, Soft Coral, Algae, Invertebrates and Other) [12]. An example CATLIN image with point labels is in Fig. 2.

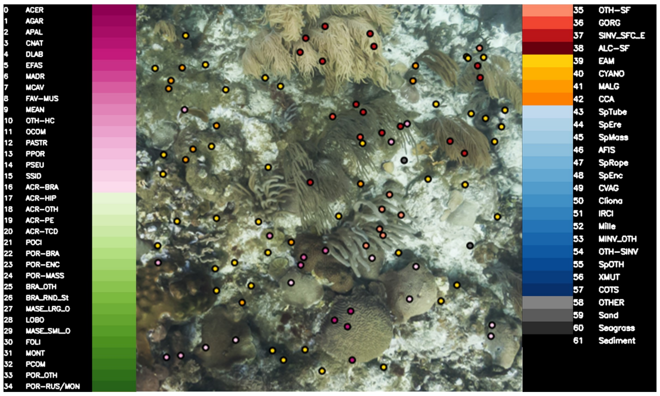


Fig. 2. Test image from CATLIN dataset with supplied point labels (best viewed in colour)

Datasets containing fine-grained classes often exhibit the long tail distribution problem, in which there are a large number of classes with a very small number of training examples. The number of samples in each of the 62 fine-grained classes in the CATLIN dataset is extremely imbalanced, with a class imbalance ratio of $\rho = 4,700$. The class imbalance ratio is the proportion of the number of samples in the largest class to the number of samples in the smallest class [42].

It was necessary to consider the long tail distribution of the dataset when forming a training/test dataset split. The CATLIN dataset was released with a random 20% training/test split, however this split did not consider class frequencies. The smallest class in the test dataset contained only three examples. Using this dataset split for evaluation of models would result in misrepresentation of performance.

The pronounced class imbalance prevents creation of a completely balanced test dataset. However, we propose a new training/test dataset split which considers the class frequencies and allows fairer evaluation of model results. We enforce a minimum of 20 data points per class in the test dataset and prescribe that 20% of the images collected at each of the nine geographical locations be used for testing. The details of our proposed test dataset are released to enable future work to directly compare with our baseline.

IV. METHOD

We use PyTorch [43] and Python 3.7 to design our stacked fully convolutional network. The number of filters at each consecutive layer is controlled by a single parameter, which we define as 1.05 for these experiments. Depth-wise separable convolutions [44] were introduced with the popular Xception architecture and consist of a ‘depth-wise’ convolution followed by a ‘point-wise’ convolution (Fig. 3). The point-wise convolution is a 1x1 convolution which projects the channels from the depth-wise convolution into a new channel space [44]. Separable convolutions significantly decrease the number of model parameters required [44]. In our architecture, the stacked depth-wise separable convolutions are followed by two fully convolutional layers with kernel size 1x1 (Fig. 4), designed to step down the filters to 64 and then the number of classes (in this case, 62).

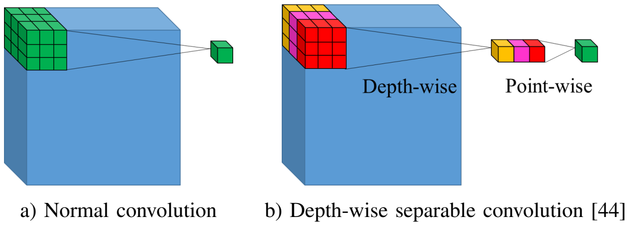


Fig. 3. Comparison of normal and depth-wise separable convolutions

The network was trained on a receptive field size of 51x51 pixels around each label, in batches containing 16 samples. The model was trained using the Categorical Cross-Entropy loss function, where only known pixels were used in the calculation. The Adam optimiser [45] was used with an initial learning rate of 0.0001. The learning rate was halved five times, each time after five epochs without an improvement in the validation loss. The training was stopped

TABLE I

SURVEY OF PUBLICLY AVAILABLE DATASETS FOR MARINE SPECIES

Paper or Dataset Name	Year	Viewpoint	Classes	Images	Annotations
BENTHOZ-2015 [34],[35]	2015	Top-down	CATAMI hierarchy (148 classes)	407,968 expert annotations of 9,874 images	Point-based
SUIM dataset [23]	2020	Oblique	8 coarse grained categories	1,525 train, 110 test pairs of images and masks	Segmentation masks
Moorea Labeled Corals (MLC) [13]	2008	Top-down	9 classes, 5 coral genera and 4 non-coral classes	400,000 human expert annotations on 2,055 coral reef survey images	Point-based
Pacific Labeled Corals (overlap with MLC)	2005-2012	Top-down	20 classes	251,988 expert annotations on 5,090 coral reef survey images	Point-based
CoralNet [36]	To present	Mainly top-down	User specified for every source	1.6 million in total, all labelled using crowd-sourcing	Point-based
XL Catlin Seaview Survey [12], [37]	2012-2018	Top-down	228 in total, 62 merged classes, 5 broad super-classes	1.1 million photo-quadrat images in total, 11,387 global labelled images, yielding 858,257 expert annotations	Point-based
EILAT [38]	2004	Top-down	8 classes	1,123 patches	Patch level
EILAT Mixx [25]		Top-down	10 classes	23 for training, 8 for testing	Point-based
Mosaics UCSD [39]	2017	Top-down	35 classes	4,193 for training, 729 for testing	Segmentation masks
RSMAS [17]	2017	Top-down	14 coral classes	766 patches	Patch level
StructureRSMAS [16]	2019	Mainly oblique, variety	14 coral structure types	409	Image level
LifeCLEF Coral Challenge 2019 [40]	2019	Top-down	13 classes	240 images with 6,670 substrates annotated	Bounding boxes and bounding polygons
Tasmania Coral Point Count Dataset [41]	2010	Top-down	36 classes	1,000 each with 50 points	Point-based

after the learning rate was dropped five times, or after a maximum of 300 epochs.

Class imbalance was mitigated using our proposed 'Midpoint Sampling' method (discussed in Section V-A.2). Our approach achieved a class-weighted overall precision of 52.7% on the CATLIN dataset.

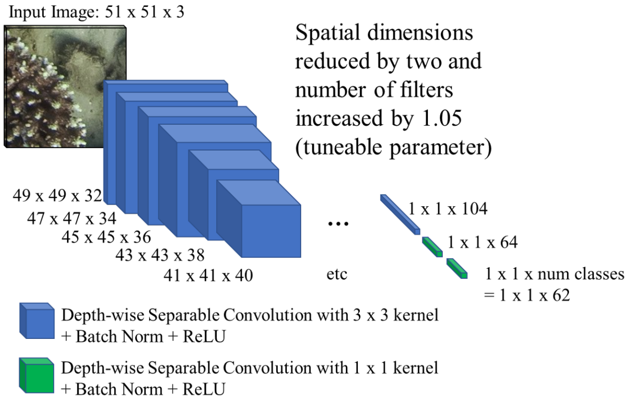

Input Image: 51 x 51 x 3
Spatial dimensions reduced by two and number of filters increased by 1.05 (tuneable parameter)
49 x 49 x 32
47 x 47 x 34
45 x 45 x 36
43 x 43 x 38
41 x 41 x 40
...
etc
1 x 1 x 104
1 x 1 x 64
1 x 1 x num classes = 1 x 1 x 62
Depth-wise Separable Convolution with 3 x 3 kernel + Batch Norm + ReLU
Depth-wise Separable Convolution with 1 x 1 kernel + Batch Norm + ReLU

Fig. 4. Architecture diagram for our stacked convolution model

V. RESULTS

Precision and recall were used for evaluating the performance of the models. The precision and recall were averaged across the 62 classes ('Macro-Averaged') and also averaged using class weights based on the number of samples for each class in the test dataset ('Weighted'). As the test dataset was not balanced, the weighted metrics provide a better evaluation of model performance. The best model was the stacked convolutional network with our novel 'midpoint-sampling' strategy, with a weighted precision of 57.2% and recall of 35.6% when evaluated on our proposed test split of the CATLIN dataset.

A. Ablation Studies

Ablation studies were performed to analyse the impact of the sampling strategy used with the stacked fully convolutional networks and the impact of the Delaunay triangulation propagation for the U-Net architecture.

1) Network Architectures: Three architectures were compared: the stacked network presented in Section IV, a two-headed stacked network and an encoder-decoder architecture. Both stacked networks were trained using mid-point sampling (Section V-A.2).

The two-headed network combines the fine-grained class and the super-class for each training point. The architecture was comprised of stacked depth-wise separable convolution layers and two separate "heads" to predict logits for the fine-grained and super-class (Fig. 5).

An encoder-decoder architecture based on the U-Net in [30] was trained using a sparse Focal loss [46]. Focal loss down-weights samples which are easily classified and focuses on more difficult examples [46]. Class weights for

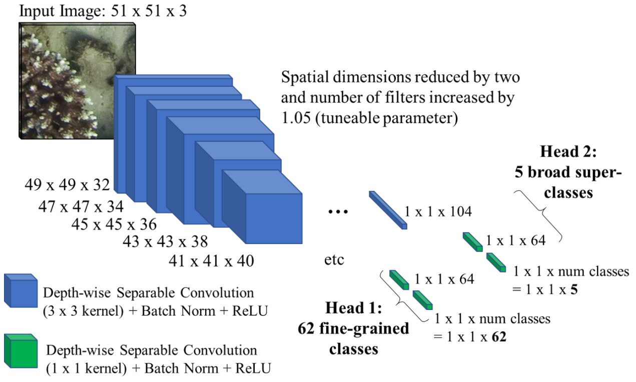

Input Image: 51 x 51 x 3
Spatial dimensions reduced by two and number of filters increased by 1.05 (tuneable parameter)
Head 2: 5 broad super-classes
49 x 49 x 32
47 x 47 x 34
45 x 45 x 36
43 x 43 x 38
41 x 41 x 40
...
etc
1 x 1 x 104
1 x 1 x 64
1 x 1 x num classes = 1 x 1 x 5
1 x 1 x 64
1 x 1 x num classes = 1 x 1 x 62
Head 1: 62 fine-grained classes
Depth-wise Separable Convolution (3 x 3 kernel) + Batch Norm + ReLU
Depth-wise Separable Convolution (1 x 1 kernel) + Batch Norm + ReLU

Fig. 5. Architecture diagram for our stacked convolution model with two heads

the focal loss were calculated by taking the maximum class frequency and dividing by the frequency of each class.

The inclusion of the broad super-classes in the two-headed network yielded similar results as the single-headed network (Table II), suggesting that the broad classes may not improve the model's ability to learn useful features for fine-grained classes. The inclusion of more specific classes as a second training signal may be more effective, and we intend to investigate this in future work.

TABLE II

EFFECT OF NETWORK ARCHITECTURE ON TEST PERFORMANCE

Architecture	Macro Prec./Rec.	Weighted Prec./Rec.
U-Net with focal loss	10.6% / 22.5%	46.5% / 11.1%
Two headed network	16.1% / 20.3%	55.3% / 32.5%
Stacked network	**18.7% / 25.0%**	**57.2% / 35.6%**

2) Sampling Strategy: The effect of sampling strategy when training the stacked network was investigated. A common method for mitigating class imbalance is to under-sample larger classes such that they contain the same number of samples as the smallest class [42]. Alternatively, the smallest classes could be over-sampled such that all classes have the same number of examples as the largest class [42].

We determined that in the case of fine-grained segmentation of benthic classes, it is undesirable to enforce balance between the classes because it creates an unrealistic representation of species abundance, resulting in lower accuracy (Table III). Training without sampling also resulted in lower precision, because in this case the model ignored 61.3% of the classes at inference time and instead favoured larger classes such as the Epilithic Algal Matrix (EAM), which comprises 49.0% of the training dataset.

Our novel method, which we call 'Mid-point Sampling', is a combination of under- and over-sampling and enables the user to specify how balanced the classes should appear via a single parameter. This parameter is the desired upper and lower bound of the class frequency distribution i.e. if 0.2 was chosen, the smallest 20% would be over-sampled and the largest 20% would be under-sampled to bring the class frequencies with a middle range of 60% (Fig. 6). This parameter is tuned by the user based on the class imbalance ratio of the dataset. Our novel sampling strategy is a simple way of preserving the natural distribution of species, while encouraging the model to learn features of all classes in the training dataset. This method achieved a weighted precision of 57.2% (Table III).

TABLE III

EFFECT OF SAMPLING STRATEGY ON TEST PERFORMANCE

Sampling Strategy	Macro Prec./Rec.	Weighted Prec./Rec.
No Sampling	14.3% / 8.6%	45.7% / **53.5%**
Balanced under-sampling	5.7% / 15.1%	41.7% / 16.3%
Midpoint sampling	**18.7% / 25.0%**	**57.2%** / 35.6%

3) Propagation of Point Labels using Delaunay Triangulation: Extending the U-Net architecture discussed in Section V-A.1, the impact of label propagation was investigated. Delaunay triangulation is a method of joining points into triangles such that triangles do not overlap and no point is inside the circumcircle of any triangle [47]. Following triangulation of the random point labels, triangles were selected only if all three vertices belonged to the same class. The class label was propagated to pixels inside the triangle (Fig. 7). Propagation of point labels for a 1200x1200 pixel image took 0.465 seconds on an Intel Core i7-6500. The number of training points increased from 660,939 labelled pixels in the training dataset to 932,612,517, however there still existed a similar distribution of class frequencies even after propagation, resulting in comparable overall weighted accuracies for both models (Table IV).

TABLE IV

EFFECT OF LABEL PROPAGATION ON TEST PERFORMANCE

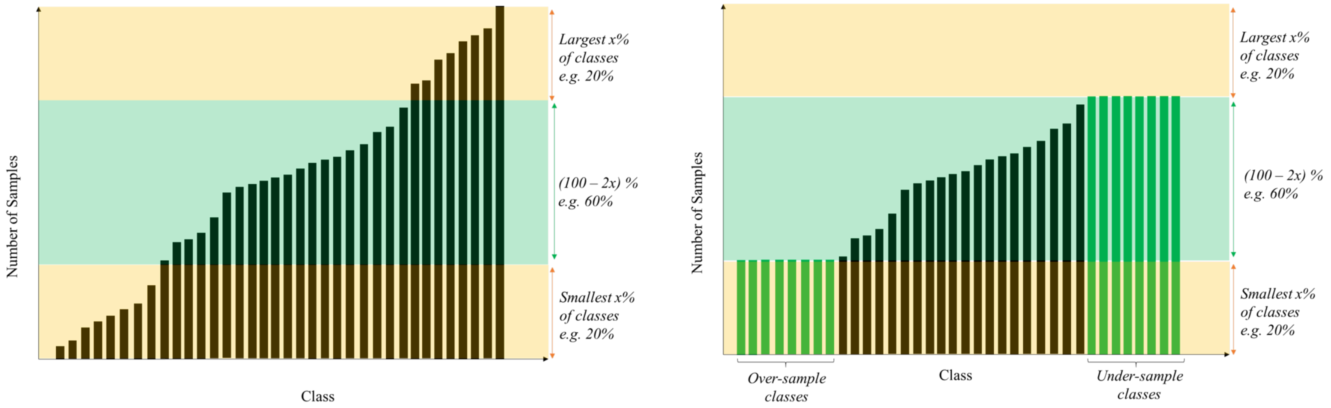
Architecture	Macro Prec./Rec.	Weighted Prec./Rec.
U-Net with label propagation and focal loss	8.1% / 16.9%	45.3% / 5.4%
U-Net with focal loss	**10.6% / 22.5%**	**46.5% / 11.1%**

B. Inference on Images

The trained models were applied to images in our proposed test split of the CATLIN dataset. Pixels were classified by assigning the label corresponding to the maximum value of the Softmax function when applied to the output of the final layer of the model. The pixel classifications were visualised using a colour-coded mask, where labels on the left of each image (in pink and green) correspond with fine-grained hard coral classes, classes in shades of red correspond to soft corals, blue denotes invertebrates and shades of grey represent sub-classes in the 'other' super-class (e.g. sand). The visualisations in Fig. 8 correspond with outputs for the test image in Fig. 2.

VI. DISCUSSION

There are benefits and drawbacks for the architectures compared: the stacked network enables inference on any size and shape of input image regardless of aspect ratio or resolution. This is beneficial for photo-quadrat imagery



a) Example class frequencies

b) Class frequencies after mid-point sampling with x=0.2

Fig. 6. Our proposed 'mid-point sampling' method

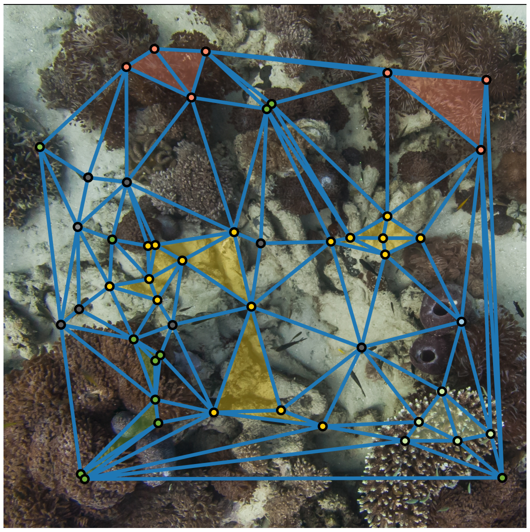


Fig. 7. Our proposed label propagation method using Delaunay triangulation. Pixels inside the shaded regions are labelled as the class of the vertices.

datasets, which often contain a range of image dimensions due to scaling [12], and for compatibility with a range of cameras.

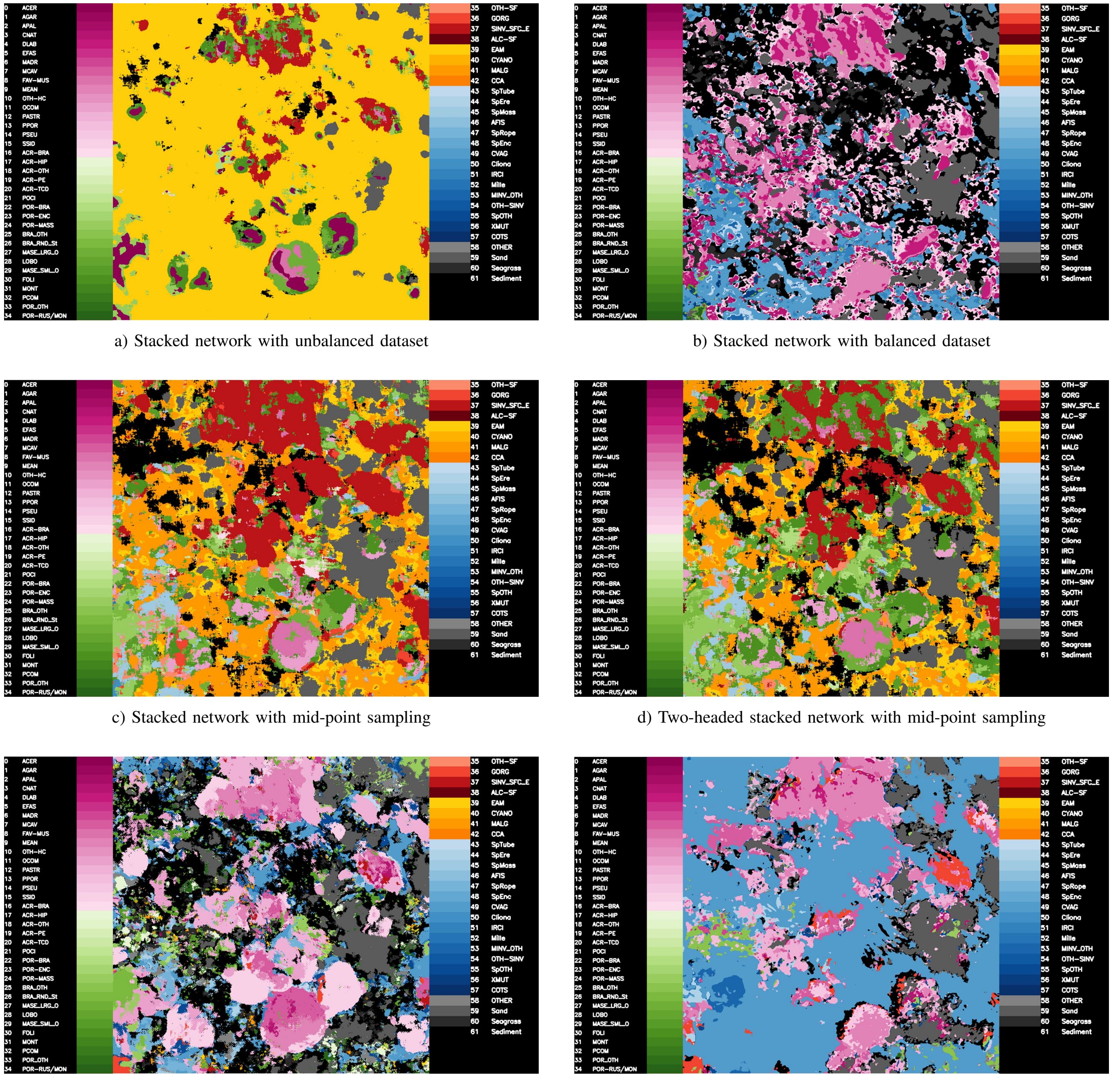
The U-Net approach enabled ground truth propagation using the Delaunay triangulation of the point labels. While the number of training samples was significantly increased, the frequency of classes was comparable, resulting in similar precision as the U-Net without propagation (Table IV). Limiting the Delaunay propagation to under-represented classes and constraining the size of triangles used for propagation will be investigated in future work. It is likely that the stacked networks outperformed the U-Net architectures because considering each point as an individual training sample allows for re-sampling to mitigate the long-tail distribution problem. The increased depth of the stacked network when compared to the U-Nets may have resulted in learning superior feature representations. Future work will include trials of additional encoder-decoder architectures to further explore this finding.

We find that our novel mid-point sampling approach is effective for the marine application because it does not remove all variation in species abundance in the dataset. When classes are balanced during training, the model is rewarded by predicting rare classes more frequently, resulting in higher recall but lower precision for those classes. We determined that enforcing balance between the classes results in over-prediction of the rare classes leading to unrealistic inference masks. This can be seen in Fig. 8, where balanced models (sub-figures b, e and f) all output a large number of pixels classified as invertebrates (depicted in blue), which is not correct based on the corresponding point labels provided in Fig. 2. The 15 invertebrate classes are rare in the CATLIN dataset, comprising only 4%. The unbalanced model (Fig. 8) exhibits the opposite problem and classifies the majority of the image as the class 'EAM', which forms 49.0% of the training dataset. In the case of fine-grained segmentation, there exists a critical balance between encouraging the model to predict all of the classes while maintaining a realistic representation of rare class frequency.

The inclusion of super-class information in the two-headed stacked network resulted in similar performance as the single-headed stacked network. The CATLIN dataset also has a label set consisting of 228 classes. Future work will involve inclusion of these labels alongside the 62 class labelset to determine whether this additional representation improves segmentation accuracy.

VII. Conclusion

This work has implemented a neural network for segmentation of fine-grained benthic classes in underwater marine imagery. We define a new train/test data split for the 62 class version of the CATLIN dataset and establish baseline accuracy. Our approach was a deep stacked fully convolutional network which used depth-wise separable convolutions, trained using our novel midpoint sampling method. We achieved an overall precision of 57.2% when weighted based on the frequency of classes in our proposed CATLIN test dataset. The approaches presented could be used in training dense segmentation networks for fine-grained segmentation of sparsely labelled datasets in any field.



a) Stacked network with unbalanced dataset

b) Stacked network with balanced dataset

c) Stacked network with mid-point sampling

d) Two-headed stacked network with mid-point sampling

e) U-Net with focal loss

f) U-Net with focal loss and Delaunay triangulation ground truth propagation

Fig. 8. Comparison of model inferences on test image (best viewed in colour)

ACKNOWLEDGMENT

This work was done in collaboration between QUT and CSIRO Data61. S.R., R.M., F.M. and N.S. acknowledge continued support from the Queensland University of Technology (QUT) through the Centre for Robotics.

REFERENCES

[1] D. O. Obura, G. Aeby, N. Amornthammarong, W. Appeltans, N. Bax, J. Bishop, R. E. Brainard, S. Chan, P. Fletcher, T. A. Gordon *et al.*, “Coral reef monitoring, reef assessment technologies, and ecosystem-based management,” *Frontiers in Marine Science*, vol. 6, p. 580, 2019.

[2] D. Sward, J. Monk, and N. Barrett, “A systematic review of remotely operated vehicle surveys for visually assessing fish assemblages,” *Frontiers in Marine Science*, vol. 6, p. 134, 2019.

[3] A. W. Bicknell, B. J. Godley, E. V. Sheehan, S. C. Votier, and M. J. Witt, “Camera technology for monitoring marine biodiversity and human impact,” *Frontiers in Ecology and the Environment*, vol. 14, no. 8, pp. 424–432, 2016.

[4] L. Xu, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, “Deep learning for marine species recognition,” in *Handbook of Deep Learning Applications*. Springer, 2019, pp. 129–145.

[5] J. Monk, N. Barrett, T. Bridge, A. Carroll, A. Friedman, A. Jordan, G. Kendrick, V. Lucieer *et al.*, “Marine sampling field manual for auv’s (autonomous underwater vehicles).” 2018.

[6] S. Raine, R. Marchant, P. Moghadam, F. Maire, B. Kettle, and B. Kusy, “Multi-species seagrass detection and classification from underwater images,” in *2020 Digital Image Computing: Techniques and Applications (DICTA)*, 2020, pp. 1–8.

[7] S. English, C. Wilkinson, V. Baker *et al.*, “Survey manual for tropical

marine resources," 1997.

[8] H. M. Murphy and G. P. Jenkins, "Observational methods used in marine spatial monitoring of fishes and associated habitats: a review," *Marine and Freshwater Research*, vol. 61, no. 2, pp. 236–252, 2010.

[9] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *European conference on computer vision*. Springer, 2016, pp. 549–565.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[11] G. Diaz-Pulido, "Supplementary report to the final report of the coral reef expert group: S2. practical taxonomy for rimrep coral reef monitoring—macroalgae," 2020.

[12] M. González-Rivero, A. Rodriguez-Ramirez, O. Beijbom, P. Dalton, E. V. Kennedy, B. P. Neal, J. Vercelloni, P. Bongaerts, A. Ganase, D. E. Bryant *et al.*, "Seaview survey photo-quadrat and image classification dataset," 2019.

[13] O. Beijbom, P. J. Edmunds, D. I. Kline, B. G. Mitchell, and D. Kriegman, "Automated annotation of coral reef survey images," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 1170–1177.

[14] A. Mahmood, M. Bennamoun, S. An, F. Sohel, F. Boussaid, R. Hovey, G. Kendrick, and R. B. Fisher, "Coral classification with hybrid feature representations," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 519–523.

[15] M. Modasshir, A. Q. Li, and I. Rekleitis, "Mdnet: Multi-patch dense network for coral classification," in *OCEANS 2018 MTS/IEEE Charleston*. IEEE, 2018, pp. 1–6.

[16] A. Gómez-Ríos, S. Tabik, J. Luengo, A. Shihavuddin, and F. Herrera, "Coral species identification with texture or structure images using a two-level classifier based on convolutional neural networks," *Knowledge-Based Systems*, vol. 184, p. 104891, 2019.

[17] A. Shihavuddin, N. Gracias, R. Garcia, A. C. Gleason, and B. Gintert, "Image-based coral reef classification and thematic mapping," *Remote Sensing*, vol. 5, no. 4, pp. 1809–1841, 2013.

[18] M. H. Shakoor and R. Boostani, "A novel advanced local binary pattern for image-based coral reef classification," *Multimedia Tools and Applications*, vol. 77, no. 2, pp. 2561–2591, 2018.

[19] N. A. B. Mary and D. Dejey, "Classification of coral reef submarine images and videos using a novel z with tilted z local binary pattern," *Wireless Personal Communications*, vol. 98, no. 3, pp. 2427–2459, 2018.

[20] M. Sotoodeh, M. R. Moosavi, and R. Boostani, "A structural based feature extraction for detecting the relation of hidden substructures in coral reef images," *Multimedia Tools and Applications*, vol. 78, no. 24, pp. 34 513–34 539, 2019.

[21] M. González-Rivero, O. Beijbom, A. Rodriguez-Ramirez, D. E. Bryant, A. Ganase, Y. Gonzalez-Marrero, A. Herrera-Reveles, E. V. Kennedy, C. J. Kim, S. Lopez-Marcano *et al.*, "Monitoring of coral reefs using artificial intelligence: A feasible and cost-effective approach," *Remote Sensing*, vol. 12, no. 3, p. 489, 2020.

[22] A. King, S. M. Bhandarkar, and B. M. Hopkinson, "A comparison of deep learning methods for semantic segmentation of coral reef survey images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1394–1402.

[23] M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, and J. Sattar, "Semantic segmentation of underwater imagery: Dataset and benchmark," *arXiv preprint arXiv:2004.01241*, 2020.

[24] G. Pavoni, M. Corsini, M. Callieri, M. Palma, and R. Scopigno, "Semantic segmentation of benthic communities from ortho-mosaic maps." *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2019.

[25] I. Alonso, M. Yuval, G. Eyal, T. Treibitz, and A. C. Murillo, "Coralseg: Learning coral segmentation from sparse annotations," *Journal of Field Robotics*, vol. 36, no. 8, pp. 1456–1477, 2019.

[26] I. Alonso and A. C. Murillo, "Semantic segmentation from sparse labeling using multi-level superpixels," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 5785–5792.

[27] A. Raphael, Z. Dubinsky, D. Iluz, and N. S. Netanyahu, "Neural network recognition of marine benthos and corals," *Diversity*, vol. 12, no. 1, p. 29, 2020.

[28] G. Diaz-Pulido and L. McCook, "State of the reef report 2008: Macroalgae (seaweeds)," 2008.

[29] M. Zhang, Y. Zhou, J. Zhao, Y. Man, B. Liu, and R. Yao, "A survey of semi-and weakly supervised semantic segmentation of images," *Artificial Intelligence Review*, pp. 1–30, 2019.

[30] S. Wang, W. Chen, S. M. Xie, G. Azzari, and D. B. Lobell, "Weakly supervised deep learning for segmentation of remote sensing imagery," *Remote Sensing*, vol. 12, no. 2, p. 207, 2020.

[31] A. L. Friedman, "Automated interpretation of benthic stereo imagery," 2013.

[32] X. Yu, B. Ouyang, J. C. Principe, S. Farrington, J. Reed, and Y. Li, "Weakly supervised learning of point-level annotation for coral image segmentation," in *OCEANS 2019 MTS/IEEE SEATTLE*. IEEE, 2019, pp. 1–7.

[33] X. Yu, Y. Ma, S. Farrington, J. Reed, B. Ouyang, and J. C. Principe, "Fast segmentation for large and sparsely labeled coral images," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–6.

[34] I. M. O. S. (IMOS), "IMOS - AUV SIRIUS, CAMPAIGN: GREAT BARRIER REEF, FEBRUARY 2011." 2011. [Online]. Available: https://catalogue-imos.aodn.org.au/geonetwork/srv/eng/catalog.search/metadata/ae70eb18-b1f0-4012-8d62-b03daf99f7f2

[35] M. Bewley, A. Friedman, R. Ferrari, N. Hill, R. Hovey, N. Barrett, E. M. Marzinelli, O. Pizarro, W. Figueira, L. Meyer *et al.*, "Australian sea-floor survey data, with images and expert annotations," *Scientific data*, vol. 2, no. 1, pp. 1–13, 2015.

[36] O. Beijbom, P. J. Edmunds, C. Roelfsema, J. Smith, D. I. Kline, B. P. Neal, M. J. Dunlap, V. Moriarty, T.-Y. Fan, C.-J. Tan *et al.*, "Towards automated annotation of benthic survey images: Variability of human experts and operational modes of automation," *PloS one*, vol. 10, no. 7, p. e0130312, 2015.

[37] M. González-Rivero, O. Beijbom, A. Rodriguez-Ramirez, T. Holtrop, Y. González-Marrero, A. Ganase, C. Roelfsema, S. Phinn, and O. Hoegh-Guldberg, "Scaling up ecological measurements of coral reefs using semi-automated field image collection and analysis," *Remote Sensing*, vol. 8, no. 1, p. 30, 2016.

[38] Y. Loya, "The coral reefs of eilat—past, present and future: three decades of coral community structure studies," in *Coral health and disease*. Springer, 2004, pp. 1–34.

[39] C. B. Edwards, Y. Eynaud, G. J. Williams, N. E. Pedersen, B. J. Zgliczynski, A. C. Gleason, J. E. Smith, and S. A. Sandin, "Large-area imaging reveals biologically driven non-random spatial patterns of corals at a remote reef," *Coral Reefs*, vol. 36, no. 4, pp. 1291–1305, 2017.

[40] J. Chamberlain, A. Campello, J. Wright, L. Clift, A. Clark, and A. Seco De Herrera, "Overview of imageclefcoral 2019 task." CEUR Workshop Proceedings, 2019.

[41] M. Bewley, B. Douillard, N. Nourani-Vatani, A. Friedman, O. Pizarro, and S. Williams, "Automated species detection: An experimental approach to kelp detection from sea-floor auv images," in *Proc Australas Conf Rob Autom*, vol. 2012, 2012.

[42] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[43] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.

[44] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[46] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[47] D.-T. Lee and B. J. Schachter, "Two algorithms for constructing a delaunay triangulation," *International Journal of Computer & Information Sciences*, vol. 9, no. 3, pp. 219–242, 1980.