Appendix: Guiding Labelling Effort for Efficient Learning With Georeferenced Images

APPENDIX A
SEAFLOOR IMAGERY DATASET

Camera equipped Autonomous Underwater Vehicles (AUVs) are routinely used in seafloor environmental monitoring applications. These mobile robotic platforms typically gather tens of thousands of seafloor images during their deployments and can observe several 10,000 m^2/h of seafloor [1]. Even though the cost of gathering data has been massively reduced through their introduction, annotating images for environmental monitoring applications is typically a manual task that requires significant expertise. Here we describe domain specific characteristics of seafloor imagery and describe the Seafloor dataset used in this study.

A.1 Characteristics of Seafloor Imagery

Subsea imaging surveys typically use Red Green Blue (RGB) colour or greyscale images, making their format compatible with modern CNNs. However, these datasets have properties that are not common in other domains:

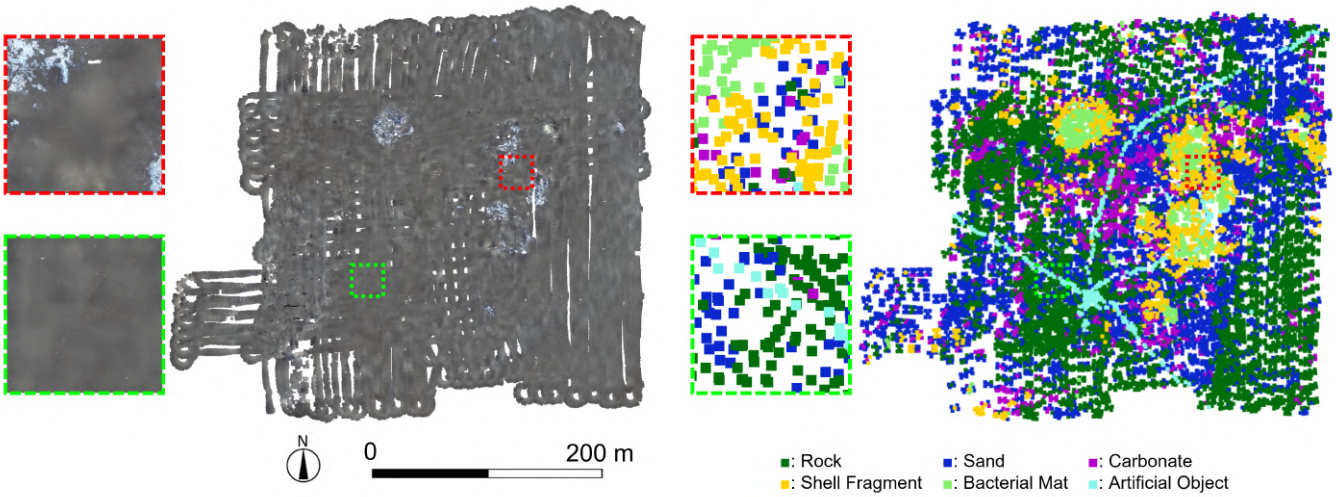
Colour and Geometry Distortion
: Different wavelengths of light attenuate at different rates in water, causing underwater images to look blue-green compared to the true colour of observed targets. The relatively low imaging altitudes (typically less than 10 m) and wide angle lenses often used to maximise area cover result in large relative range differences within an image due to terrain profiles and between images due to vehicle dynamics, which change the hue of images. Between datasets and platforms there are additional sources of variability, including different water column properties that affect the wavelength dependence of light attenuation, and the use of artificial light sources with different wavelength profiles. In addition to colour degradation, the variable range causes spatial inconsistencies that distort the shape and size of observed targets. There have been many studies investigating computational and physically grounded principles to compensate for these artefacts [3], [4].

Small Footprint
: Light rapidly attenuates in water, and so powerful artificial light sources are needed to obtain visual images in most applications. The range at which images can be obtained is limited to approximately 10 m for most setups, which constrains the footprint of a single frame to edge lengths of a similar magnitude. Since many patterns of interest (e.g. substrates, habitats, infrastructure) exist on far larger spatial scales, multiple images need to be taken along trajectories to capture these broader scale patterns.

Georeferencing
: Most images of the seafloor are gathered by robotic platforms or fixed observatories and georeference information is typically available. Since Global Navigation Satellite Systems (GNSS) cannot be used underwater, most mobile robotic platforms have navigational suites that fuse data from an AHRS, DVL and depth sensor with acoustic positioning systems such as a USBL. Georeferencing is typically achieved with a relative accuracy of approximately 1 % of distance travelled, and absolute accuracy of approximately 1 % of depth [5]. Stationary systems have similar absolute position accuracy.

Imbalanced Class Distribution
: Seafloor substrates and habitats can change over spatial scales larger than the extents observed during most robotic imaging surveys. Furthermore, many types of benthic communities, geological features and infrastructure are sparsely distributed, making seafloor datasets susceptible to skewed class membership [6], [7].

Imbalanced class distributions, colour and geometry distortions can degrade learning performance [8], [9]. The problem of small footprints can potentially be solved if pixel-order accurate georeferencing can be achieved, as artefact-free photomosaics can be generated and cropped to form image patches for processing. However, for seafloor imaging applications position estimates contain non-negligible uncertainty compared to the resolution and footprint of obtained imagery. Although techniques such as simultaneous localisation and mapping are available [10], the need for artificial strobes and the limited energy available on robotic platforms limits the relative overlap that can be achieved between images. This makes generating pixel order accurate photomosaics more challenging to obtain than with satellite and aerial drone imagery, which typically have lower resolution, larger image footprints with greater overlap and



(a) Mosaiced imagery of the Seafloor dataset

(b) Ground truth classes for the Seafloor dataset

Fig. A1: Seafloor dataset consisting of ~63k image patches and ~19k ground truth annotations. The data covers approximately 12 ha and has an average depth of 780 m. The lines formed by 'Artificial Object' show the routing of exposed cables connecting various bits of observatory infrastructure. The light green 'Bacterial Mats' form discrete patches around active methane gas venting from the seafloor and are sparsely distributed around the site. These are often surrounded by 'Shell Fragments' and 'Carbonates', which are distributed over background substrates of 'Sediments' and 'Rocks'. Examples of representative images in each class can be found in [2].

TABLE A1: Description of the Seafloor Dataset

No. of Image Patches	62,875
No. of Annotations	18,740
Resolution [mm/pixel]	10
Imaged Area [m^2]	118,000
No. of Classes	6
Latitude [°N]	44.5683 to 44.5715
Longitude [°W]	125.1455 to 125.1506
Lat. × Lon. Edge Lengths [m]	360 × 410
Seafloor Depth [m]	765 - 785
Location	Southern Hydrate Ridge

accurate position information. These points favour the use of single image frames for automated interpretation of underwater imagery since these contain fewer artefacts.

A.2 Dataset Description

The Seafloor dataset analysed in this work is of the Southern Hydrate Ridge, a gas hydrate field that is also the site of a seafloor cabled observatory [11] located 100 km offshore of Oregon, USA at a depth of ~780 m. Dataset characteristics are given in TABLE A1 and a mosaic generated from the data is given in Fig. A1. The dataset consists of 12,575 images that were collected using the *SeaXerocks* mapping system [1] mounted on the AUV AE2000f of the Institute of Industrial Science, University of Tokyo, Japan, during the Schmidt Ocean Institute's FK180731 #Adaptive Robotics campaign in August 2018 [2]. The georeferenced position where each image was captured is determined using vehicle navigation data, which consists of an Attitude and Heading Reference System (AHRS), a Doppler Velocity Log (DVL), a depth sensor and an Ultra-Short BaseLine (USBL) acoustic positioning system [5]. These are processed together with the seafloor images using a visual Simultaneous Localisation and Mapping (SLAM) pipeline [10], with an estimated relative position error of less than 1 m. The images are colour-corrected, undistorted and resampled to a constant spatial resolution of 10 mm/pixel to minimise the impact of altitude variation between observations. Five 224 × 224 pixels regions are cropped from each image (four corners and centre, partially overlapping) to form 62,875 images patches, which includes 18,740 patches that are annotated by human experts [2]. The unsupervised learning step of Fig. 1 in the main text uses all available image patches since annotations are not needed for LGA training. This workflow matches real seafloor survey scenarios, where a set of completely unknown images are collected during each deployment. The annotated patches randomly sample approximately 30 % of the entire dataset, which is sufficient to consider the distribution of class annotations as representative of the full dataset's distribution.

The Seafloor datasets can be accessed via SQUIDLE+ (http://soi.squidle.org) as (Campaign: fk180731[ID:53], deployment: 20180804_093404_20180804_143258_20180805_123456_20180809_083837_ae2000f_sx3[ID:711]). The expert annotations for the images can be accessed at SHR_AE2000_3000samples[ID:80] and SHR_AE2000_1000samples[ID:74] in uos-oplab-fk180731[ID:9] datasets. The colour correction and undistortion methods used to pre-process images in this work can be found on https://github.com/ocean-perception/oplab_pipeline/tree/master/correct_images.

APPENDIX B
AERIAL IMAGERY DATASET

Experiments are performed for land cover classification of three different aerial image datasets to assess the versatility

TABLE B1: Description of Aerial Imagery Dataset

	Mountain	Island	Urban
No. of Image Patches	46,200	15,128	47,961
Resolution [m/pixel]	2.0	2.0	1.0
Imaged Area [km^2]	9,520	3,120	2,470
No. of Classes	6	4	6
Latitude [°N]	65.01 to 66.09	56.91 to 58.00	59.14 to 59.60
Longitude [°E]	14.79 to 16.53	17.96 to 19.35	17.45 to 18.36
Lat. $\times$ Lon. Edge Lengths [km]	120×80	150×84	50×50
Location	Vindelfjällen	Gotland	Stockholm

of the proposed method across environmental monitoring application domains.

B.1 Dataset Description

Aerial image datasets from three different regions (Mountain, Island and Urban) of Sweden are used to test the versatility of our method. TABLE B1 shows details of each datasets. The Mountain dataset consists of images of the area surrounding the Vindelfjällen Nature Reserve, which is one of the largest protected areas in Europe (see Fig. B1). Six classes are observed in this area, where 'Wetland' and 'Other Non-vegetated' (corresponding to alpine peaks) are unique to this dataset in our experiments. The region also has areas of 'Water'. The Island dataset is of Gotland island, which consists of four classes, including large regions of farmland ('Arable' class), as shown in Fig. B2. The Urban dataset consists of images around the city of Stockholm (see Fig. B3). This dataset consists of six classes, where the 'Artificial' class is used to describe the city and other built up areas, where this class is unique to this dataset in our experiments. The dataset also contains some 'Arable' and 'Water' regions. All datasets have 'Coniferous', 'Deciduous' and 'Other Vegetated' areas, although their appearances and distribution patterns differ between the datasets.

The dataset images are cropped from ESRI World Imagery. Each image is rescaled and cropped to 227 $\times$ 227 pixels patches. The datasets have different spatial resolutions, 2.0 m/pixel for Mountain and Island and 1.0 m/pixel for Urban, where it is often the case that higher resolution data is available near populated areas. The physical sizes of the image patches are 454 $\times$ 454 m (Mountain and Island) and 227 $\times$ 227 m (Urban), respectively.

The ground truth annotations used are based on the National Land Cover Database (NMD) published by the Swedish Environmental Protection Agency, which assigns land cover classes to every 10 $\times$ 10 m region of the country. In our experiments, we use the majority land cover class in each image patch as the ground truth class, and some detailed classes are merged as they cannot be distinguished using only RGB colour channels (e.g. six types of coniferous forest classes in NMD are dealt with as a single 'Coniferous' class in this experiment).

REFERENCES

[1] B. Thornton, A. Bodenmann, O. Pizarro, S. B. Williams, A. Friedman, R. Nakajima, K. Takai, K. Motoki, T.-o. Watsuji, and H. Hirayama, "Biometric assessment of deep-sea vent megabenthic communities using multi-resolution 3d image reconstructions," *Deep Sea Research Part I: Oceanographic Research Papers*, vol. 116, pp. 200–219, 2016.

[2] T. Yamada, A. Prügel-Bennett, and B. Thornton, "Learning features from georeferenced seafloor imagery with location guided autoencoders," *Journal of Field Robotics*, 2020.

[3] M. Bryson, M. Johnson-Roberson, O. Pizarro, and S. B. Williams, "True color correction of autonomous underwater vehicle imagery," *Journal of Field Robotics*, vol. 33, no. 6, pp. 853–874, 2016.

[4] A. Bodenmann, B. Thornton, and T. Ura, "Generation of high-resolution three-dimensional reconstructions of the seafloor in color using a single camera and structured light," *Journal of Field Robotics*, vol. 34, no. 5, pp. 833–851, 2017.

[5] L. Paull, S. Saeedi, M. Seto, and H. Li, "Auv navigation and localization: A review," *IEEE Journal of Oceanic Engineering*, vol. 39, no. 1, pp. 131–149, 2014.

[6] 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, p. 150057, 2015.

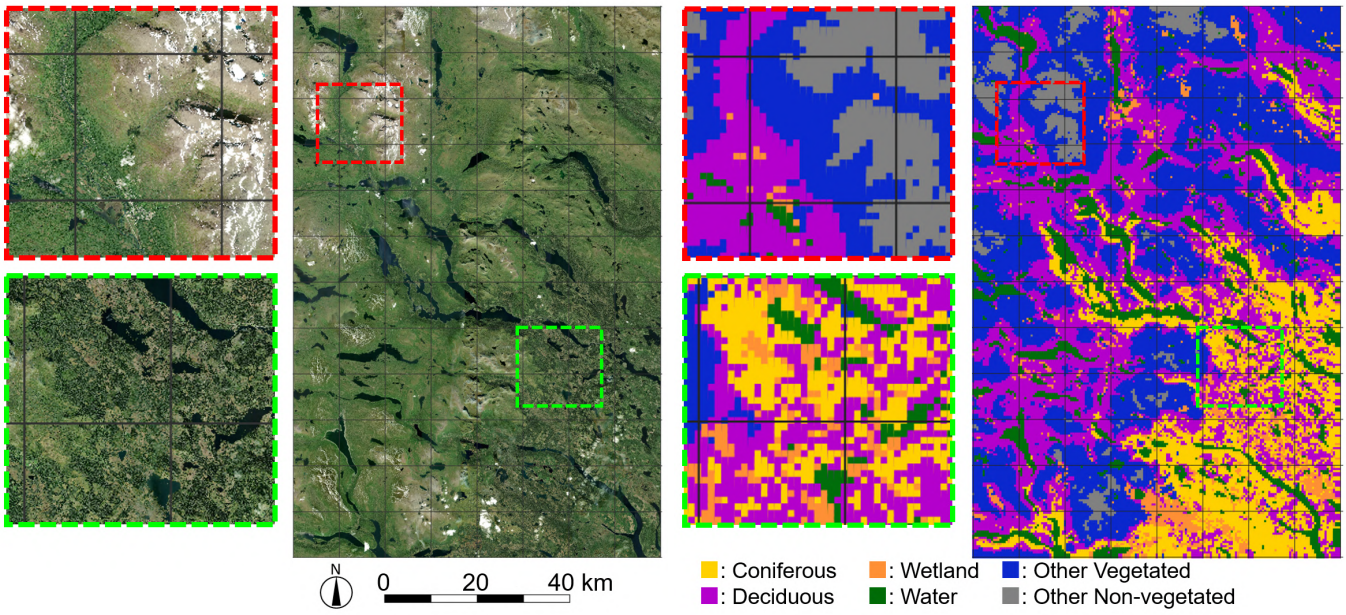
[7] A. Mahmood, M. Bennamoun, S. An, F. A. Sohel, F. Boussaid, R. Hovey, G. A. Kendrick, and R. B. Fisher, "Deep image representations for coral image classification," *IEEE Journal of Oceanic Engineering*, vol. 44, no. 1, pp. 121–131, 2018.

[8] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[9] J. Walker, T. Yamada, A. Prugel-Bennett, and B. Thornton, "The effect of physics-based corrections and data augmentation on transfer learning for segmentation of benthic imagery," in *2019 IEEE Underwater Technology (UT)*. IEEE, 2019, pp. 1–8.

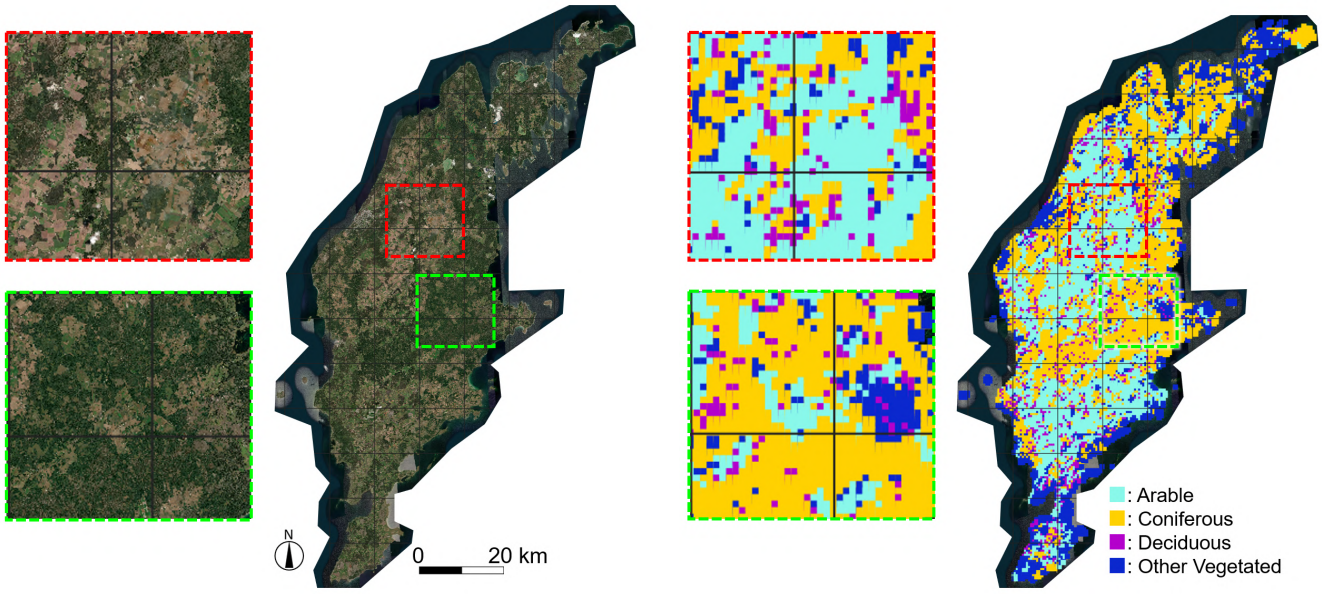
[10] I. Mahon, S. B. Williams, O. Pizarro, and M. Johnson-Roberson, "Efficient view-based SLAM using visual loop closures," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1002–1014, 2008.

[11] T. Cowles, J. Delaney, J. Orcutt, and R. Weller, "The ocean observatories initiative: Sustained ocean observing across a range of spatial scales," *Marine Technology Society Journal*, vol. 44, no. 6, pp. 54–64, 2010.



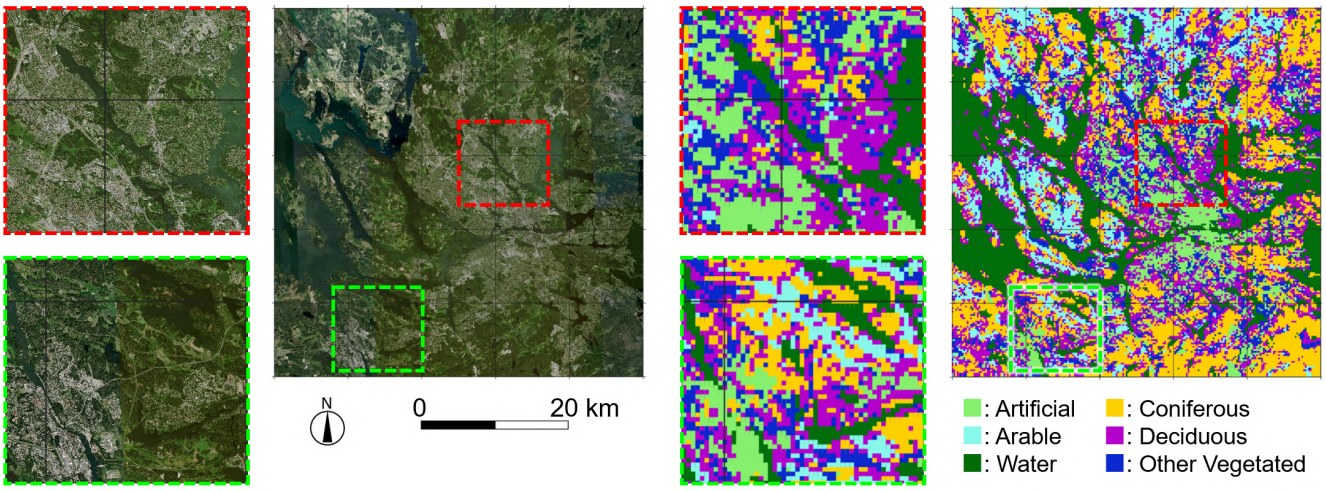
(a) Mosaiced aerial imagery of the Mountain dataset (b) Ground truth classes for the Mountain dataset

Fig. B1: Mountain dataset showing the area surrounding the Vindelfjällen Nature Reserve in Sweden. Six classes are observed in this area, where 'Wetland' and 'Other Non-vegetated' (corresponding to alpine peaks) are unique to this dataset in our experiments. The dataset also has 'Water', 'Coniferous', 'Deciduous' and 'Other Vegetated' regions, where these classes are shared across the different datasets studied in this work. The figure shows that the spatial distributions of the shared classes are different to their distributions in the Island (Fig. B2) and Urban (see Fig. B3) datasets.



(a) Mosaiced aerial imagery of the Island dataset (b) Ground truth classes of the Island dataset

Fig. B2: Island dataset showing Gotland island in Sweden, which consists of four classes, including large regions of farmland ('Arable' class) that dominate the open areas. The dataset also has 'Coniferous', 'Deciduous' and 'Other Vegetated' regions, where these classes are shared across the different datasets studied in this work. The figure shows that the spatial distributions of the shared classes are different to their distributions in the Mountain (Fig. B1) and Urban (see Fig. B3) datasets.



(a) Mosaiced aerial imagery of the Urban dataset

(b) Ground truth classes of the Urban dataset

Fig. B3: Urban dataset showing the area surrounding Stockholm in Sweden. The 'Artificial' class is used to describe the city and other built up areas, where this class is unique to this dataset in our experiments. The dataset also has 'Arable', 'Water', 'Coniferous', 'Deciduous' and 'Other Vegetated' regions, where these classes are shared across the different datasets studied in this work. The figure shows that the spatial distributions of shared classes are different to their distributions in the Mountain (Fig. B1) and Island (see Fig. B2) datasets.