Predicting Seafloor Visual Classes from Multimodal Remote Sensed Priors using Location-Guided Self-supervised Learning

Cailei Liang, Jose Cappelletto, Adrian Bodenmann
Stephen Turnock, Blair Thornton*
University of Southampton, United Kingdom
C.Liang@soton.ac.uk

Veerle A.I. Huvenne, Catherine Wardell
Ocean BioGeosciences,
National Oceanography Centre,
Southampton, UK

***Abstract*—Remote sensed mapping data and seafloor in-situ imagery are often gathered to infer benthic habitat distributions. However, leveraging multimodal data is challenging because of inherent inconsistencies between measurement modes (e.g., resolution, positional offsets, shape discrepancies). We investigate the impact of using location metadata in multimodal, self-supervised feature learning on habitat classification. Experiments were carried out on a multimodal dataset gathered using and Autonomous Underwater Vehicle (AUV) at the Darwin Mounds Marine Protected Area (MPA). Introducing location metadata improved F1 classification performance of a Bayesian classifier by an average of 27.7% over all conditions tested in this work, with a larger improvement of 32.9% achieved when multiple remote sensing data modes are combined for the analysis.**

***Index Terms*—multimodal feature learning, seafloor habitat classification, self-supervised learning, inference**

I. Introduction

Evaluating habitat distributions is an important part of understanding seafloor environments. One way to collect information about habitats is to use camera equipped Autonomous Underwater Vehicles (AUVs) to photograph the seafloor at sub-cm resolution. However, the strong attenuation of light in water limits seafloor imaging altitudes to 2-10 m, where the small image footprint (edge lengths of 2-10 m) is compounded by the slow speed and low endurance of AUVs (0.2-1 m/s with deployments lasting hours to days). Area cover ranges between 1,400-40,000 m^2/h, which is significantly smaller than the extent over which habitat distributions typically need to be understood, e.g., UK marine protected areas (MPAs) are 4-100,000 km^2. Acoustic remote sensing methods such as Side-Scan Sonar (SSS) and Multibeam Echo Sonar (MBES) [1] cover significantly larger areas than imaging, gathering lower resolution (tens of cm to metre order) information about different aspects of the environment. For example, calibrated SSS intensities are a proxy for seafloor hardness, while MBES reveals the depth and slope of the seafloor. Although these data modes do not directly identify habitat types, when combined with models or expert human judgement, they form priors that help infer likely habitat distributions over the spatial scales that are more relevant for statutory monitoring. The combined use of remote sensing with seafloor imaging for habitat class verification is becoming increasingly routine in seafloor monitoring.

Both feature engineering and feature learning have been used to extract information from remote sensed environmental priors [2]. Although feature engineering requires human expertise to extract and combine appropriate features, feature learning automatically extracts feature combinations from the data using machine-learning techniques. An examples of feature engineering is [1], where Grey Level Co-occurrence Matrices (GLCM) were deployed to extract 64 features to represent environmental priors over various spatial footprints, which were then combined for downstream classification tasks. The need for manual selection and combination in feature engineering limits how well a particular set of features generalises to capture information across different datasets. In contrast, feature learners automatically adapt the features they extract to best describe the data they are trained on. This automatic optimisation to the data is beneficial when multiple data modalities need to be considered and combined as it bypasses effort intensive optimisation studies needed for effective feature engineering.

Deep-learning convolutional neural networks (CNN) can learn features from geospatial data using supervised learning techniques. However, this requires large labelled training datasets, where the appearance of remote sensed seafloor dataset differs significantly from the terrestrial satellite imagery that constitutes most geospatial training repositories. Since preparation of reliable repositories for the diverse seafloor sensing modalities and classification targets is time-consuming, self-supervised learning approaches form an attractive alternative as as they can efficiently learn features that describe intrinsic patterns in unlabelled training datasets. These features can be subsequently used in downstream classification tasks where recent examples of semi-supervised learning have achieved state-of-the-art performance using far

Datasets used were collected during NERC CLASS (NE/R015953/1) and Oceanids BioCam cruises of the RRS Discovery (NE/P020887/1 and NE/P020739/1) with the support of the NOC MARS team. This research was supported in part by the Kakenhi Grant Aid B (22H01695) of the Japan Society for the Promotion of Science. *Blair Thornton is adjunct at the Institute of Industrial Science, The University of Tokyo.

fewer labelled class examples than conventional supervised learning [11]. Contrastive learning is a self-supervised approach that augments the same data in different ways and generates a feature space that embeds them nearby [3], [4]. In [5], a contrastive approach that ensures consistency of clustering outputs between different augmentations of the same data was developed. In [11], the authors developed a location-guided contrastive approach that embeds data augmentations from nearby locations to nearby regions of the feature space. Recently, transformer models based on advances in natural language processing have also been used for self-supervised feature learning [6], achieving comparable performance to CNNs.

In this research, we investigate the effectiveness of location-guided self-supervised learning [11] for multimodal feature extraction. A factor that can impact performance is the stage at which multimodal information is fused. Three different stages for fusion are recognised in the literature [7]; early fusion, middle fusion and late fusion. Early fusion merges raw data before any feature extraction takes place (i.e., fusing raw or pre-processed data before any extraction of information). In our context, this ensures joint features can be captured from spatially overlapping regions. However, early fusion is susceptible to artifacts such as the positional offsets that can exist between data modes due to limitations in sub-sea localisation. Middle fusion extracts feature from individual data modes and fuses these in a hidden layer. The fused features are used to generate the final results. Late fusion employs completely separate models to extract features from different types of data, and these features are directly used for classification tasks. Compared to early fusion, middle fusion and late fusion are less sensitive to mismatches between raw data modes, but risk missing correlation between potentially useful patterns. Here, we investigate early fusion, and specifically whether fusion of multiple data modes improves performance, and if location-guiding improves robustness when identifying geospatial patterns in multimodal data.

II. Method

Our approach predicts visual class distributions over large, remotely sensed areas that have only partially been imaged. A location-guided contrastive learning approach [11] is used to learn features from early-fused multimodal remote sensing data. Gaussian Processes (GP) are trained to model the relationship between these features and the visual-class determined for overlapping images. An advantage of using GPs is that in addition to making class predictions, they also predict the class uncertainty, which is important as remote sensed features are not guaranteed to capture relevant information about an associated visual classes. The following factors also require consideration:

- Non-uniform data resolutions and extents between data modes
- Physical change in the mapped environment between acquisition of different data modes
- Positional offsets and deformation due to imperfect instrument calibration and localisation errors

We assess the performance of our approach on a multimodal dataset gathered at the Darwin Mounds MPA, where SSS, MBES and imaging data were acquired in overlapping regions on seperate AUV dives, where each data mode is pre-pocessed to compensate for any instrument calibration and attenuation effects.

A. Feature learning

Different remote sensed environmental priors are early fused before feature learning takes place. Each data mode is pre-processed using standard workflows to remove artifacts and reflect instrument calibrations. Since the patterns that describe substrates and habitats cover various spatial scales [1], [8], [9], it is important to consider how these can be captured in the learning process. CNNs consider a spatial footprint that is upper bounded by the size of the convolutional window in their first layer. Within this footprint, or patch, patterns of various scale are captured by layers in the CNN architecture. Although, patterns larger than the patch size typically do not influence learning, location-guided feature learners address this limitation by implementing a proximity assumption: Locations that are physically close are more likely to have similar seafloor characteristics than locations that are far-apart. This has been demonstrated through the use of modified autoencoder loss functions [14] and contrastive learning by assuming similarity of sample pairs taken within some distance constraint [11]. Although these methods allow features learning to capture information over various scales, the CNN patch size still determines the resolution of the final class maps and must and must satisfy some constraints.

To capture spatial patterns the patch must contain multiple pixels across all data modalities. A sensible range is between 32×32 pixels [15] and 256×256 pixels [16] based on the literature. The patch must also be sufficiently large to absorb the impact of positional offsets between different data modes that are being combined. Positional offsets can occur due to the inherent localisation uncertainty of AUVs, typically 1 % depth with acoustic localisation and 1 % of distance travelled with relying on Doppler aided inertial navigation. Other factors that can affect positional offsets include instrument calibration, and actual change in the environment for asynchronously acquired data. Conventional SSS does not consider terrain profile and so contains inherent projection errors that are not present in fully 3D measurement modes such as MBES. While it may be possible to match features between data modes, here we focus on understanding the impact of such geometric distortions on classification performance. For actual change in the environment, this depends on the habitats in the study area, where for targets that exhibit seasonal patterns and have rapid change (e.g., seagrass, hydrothermal vents) are inherently more sensitive to the interval between different remote sensing and imaging data acquisition than slower changing habitats (e.g., cold water coral, substrates, managanese deposites). Finally, the patch must be smaller than the scale of the habitats that are being characterised. This can be determined using methods such as autocolerration or based on prior knowledge of habitats at the site.

After considering these pointsm the survey region is split into patches of uniform geo-spatial dimensions and a convolutional kernel K_s of size $r_d \times r_d$ is applied to fuse the data from each mode s as follows [10]:

$$\mathbf{y}(N, E) = \sum_{s=1}^{S} \sum_{m,n=-r_d/2}^{r_d/2} I_s(N+m, E+n) \cdot K_s(m, n) \quad (1)$$

where N, E represents the north-south geographical location. I_s shows the intensity of the remote sensed data and the subscript s is the index of modes being combined. The Kernel location is determined by m, n. $\mathbf{y}$ forms the early fused multimodal input for feature learning. This early fusion strategy is applicable to most CNN models, and can also be applied to vision transformer models, where image patches are converted to embeddings via a learnable CNN block.

Our study uses GeoCLR [11], which is a location-guided contrastive feature learner that extends SimCLR [12] to deal with geospatial data. SimCLR [12] generates augmented views of the same patch by applying random distortions (e.g., crop, colour distortion) and ensures these are embedded nearby in the feature space. GeoCLR implements the proximity assumption by sampling patches within some fixed relative-distance constraint, and ensures these appear nearby in the feature space after applying random distortions (see Fig. 1). We investigate the impact of location-guiding by comparing the performance of these methods.

B. Predicting visual classes

Gaussian Process Regression (GPR) is used to model the relationship between the multimodal features extracted from remote sensed priors with visual classes. GPs are non-parametric models that can learn relationships using labelled training data. Important characteristics of GPs for this work are their robustness when trained over a range of small to medium sized datasets, and their ability to predict not just class probabilities, but also the uncertainty of their predictions. For the first point, our method guarantees visual class labels to train the GPs at any patch that has overlapping images. However, the number of patches with overlapping images depends on both the patch size and the survey design and so robust performance over a wide range of training data sizes is important. GPs are known to be slow when fitting large volumes of training data (>5000). This can be mitigated by randomly sub-sampling training data to a manageable size, while still maintaining robust performance when available training examples are limited (<500).

In general, the patches from which multimodal features are extracted are larger than the footprint of a single AUV image frame. Therefore, if overlapping images exist, patches can have several corresponding images that may represent different visual classes. We deal with this by assigning probabilities for each visual class based on their normalised frequency within each patch. The GPR models the probability of each visual class given each set of multimodal features wherever images overlap with the multimodal patches. Once trained, the models predict class probabilities for all patches in the dataset.

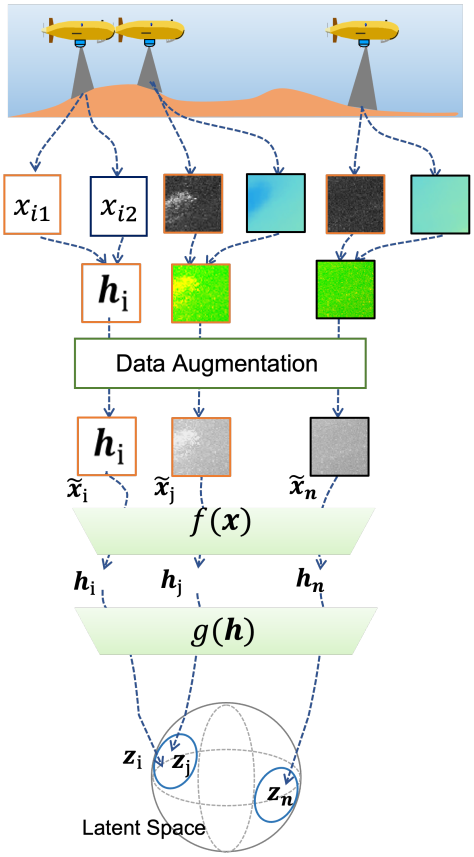


Fig. 1: GeoCLR takes nearby patches and applies random distortions to these in a similar way to SimCLR. It embeds these similar pairs (e.g. z_i and z_j) closer in the feature space than a dissimilar pair (e.g. z_i and z_n) sampled from a random location.

III. Results and Analysis

A. Dataset description

We investigate our approach using a multimodal dataset collected at the Darwin Mounds MPA in the UK in 2019 at a depth of 1000 m. The Darwin Mounds MPA is home to cold-water-coral that grow around the edges of small mounds that are approximately 70 m in length and <10 m high. The rest of the seafloor consists of sediments, with tails formed in the wake of the mounds, where large numbers of Xenophyophores are found.

Approximately 19,000 images, where Sediment, Tail, Mound Edge, and Mound Top represent 81%, 16%, 2%, and 1% respectively, were collected from 5 m altitude using the National Oceanography Centre's Autosub6000 equipped with the University of Southampton's BioCam 3D camera system [17]. Fig. 2 shows examples of visual classes at the

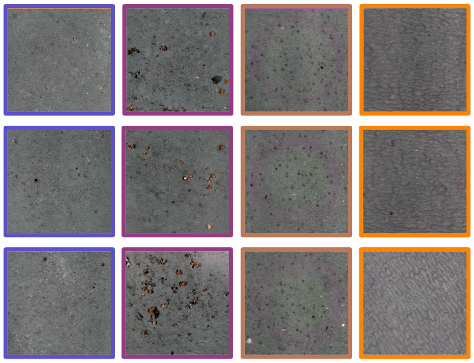


Fig. 2: Example seafloor images representing four visual classes, (left) Mound Top, Mound Edge, Tail and (right) Sediment. These were classified using the methods described in [18]

site. All images were classified following the method described in [18].

SSS and MBES data were also collected using Autosub6000 during the same cruise on a separate dive within days of the image survey. Since this is a slow changing environment ((cold-water-coral growth is $<$ 3 mm per year [13])) no detectable change is expected. The majority of positional offsets between the remote sensing data modes and imagery are expected to be due to AUV positional offsets (approximately 10 m at this depth) and instrument calibration and projection effects. Although the surveyed region is slightly sloped downward to the west, the dominant habitat characteristics are related to the mounds, and so our analysis considered the relative depth (RD) of each patch about its mean depth, and not the absolute depth of the seafloor. In addition, the resolution of the SSS and RD used in this study is 0.2m/pixel and 2m/pixel. They are divided into 50×50 patches with an overlap of 25×25.

B. Evaluation metrics

We evaluate the performance of single and multimodal models by utilizing features obtained through SimCLR and GeoCLR. To measure performance, the data set is divided into 80% for training and 20% for testing. The classification accuracy is then assessed using the F1 score, as defined in Eq. refeq:F1 score.

$$F_1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{2}$$

where TP is the number of true positives, FN is the number of false negatives, FP is the number of false positives. The F1 score is a more appropriate metric compared to accuracy due to the significant class imbalance present in the dataset.

C. Classification results

We perform the single-mode class inference using SSS and RD, respectively and multimodal class inference for the early-fused multi-modal input, i.e. SSS+RD. Tab.I gives F1 scores for single-mode and multimode class inference. In single-mode class inference, the F1 score for SSS is substantially higher than for RD. In addition, GeoCLR improves performance by 27.7 % over SimCLR on average, demonstrating more effective feature learning by making use of location metadata. For multi-modal class inference, GeoCLR improves over SimCLR by 32.9 %. Significantly, SimCLR where combining data modes reduces performance compared to the best performing single mode by -10.1 %, GeoCLR improves performance by 3.4 %. A possible explanation is that spatial inconsistencies between the data modes may cause confusion during the feature learning process in SimCLR if spatial offsets are large relative to the patch size. However, with location guiding, patterns larger than a single patch can influence feature learning, which may improve robustness to positional offsets between data modes.

TABLE I: F1 scores (%) of single-mode and multimodal class inference for SimCLR and the location-guided GeoCLR

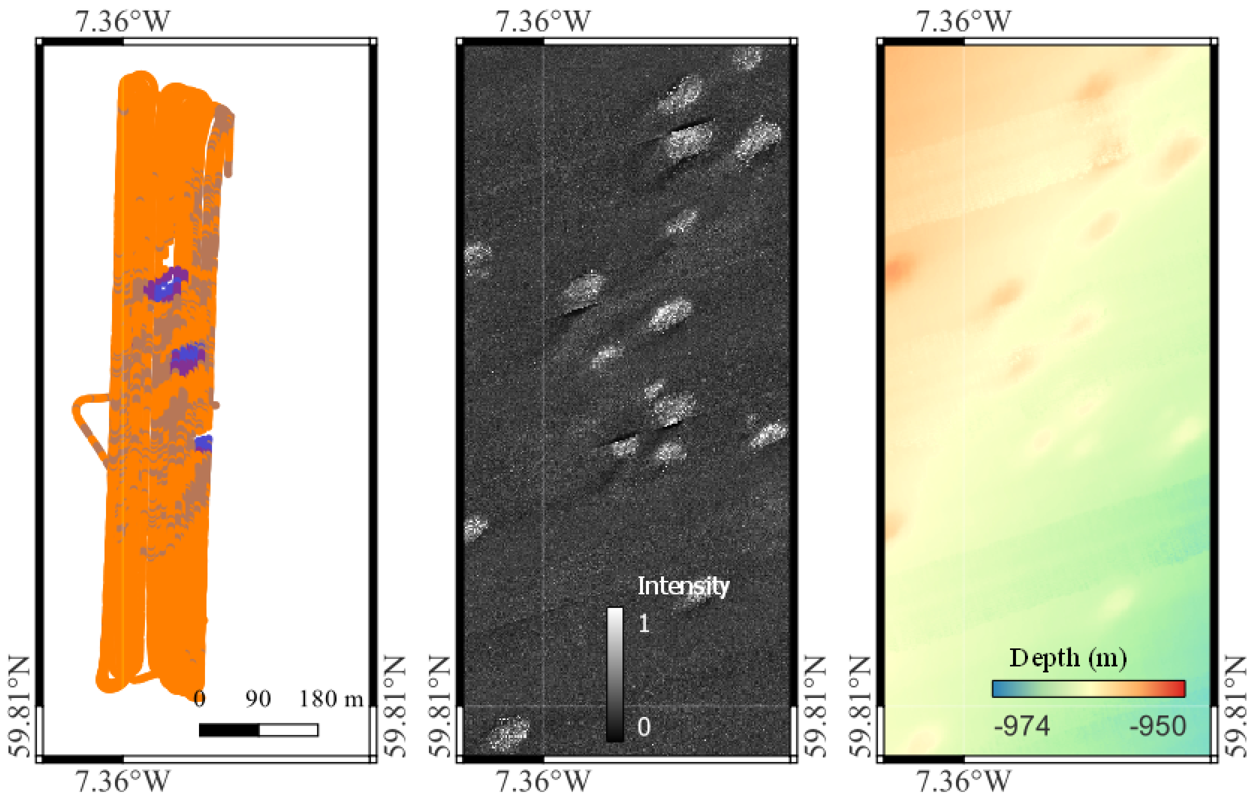
Environmental priors	SimCLR	GeoCLR
SSS	56.2 ± 1.2	75.6 ± 0.3
RD	31.3 ± 3.7	62.0 ± 2.0
SSS+RD	46.1 ± 1.2	79.0 ± 1.7

Fig. 4 shows the TSNE distribution of single-mode (SSS) and multimodal (SSS+RD) features for SimCLR and GeoCLR. The colours represent the visual classes. For SimCLR, the Mound Edge, Mound Top and Tail classes become significantly more scattered in the multimodal feature space (Fig. 4 (c)), compared to the single-mode feature space (Fig. 4 (a)), which makes it inherently more difficult for the GPRs to model the feature to visual class relationship. With GeoCLR however, the same three classes show a similar level of grouping in the multi-modal and single-mode feature spaces (Fig. 4 (d,a)), respectively.

Fig. 5(a)-(d) shows the probabilities of each class when using the multimodal feature space. The background classes Tails and Sediment both occur in flat areas with some distinct texture in the SSS data (see Fig. 3(b,c)). Mound Top and Mound Edge are well distinguished as seen in Fig. 5(a,b), with clear boundaries as seen in the reference visual classes extracted from the imagery (Fig. 3(a)). Fig. 5(e) shows the prediction uncertainties derived from the relative entropy. It demonstrates the regions where the classifier lacks confidence in its predictions, indicating the necessity for a more thorough survey in those areas.

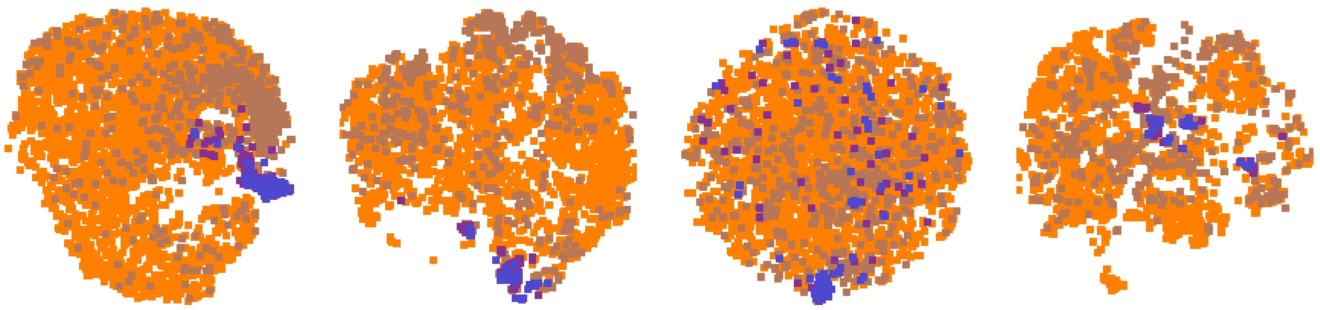
IV. Conclusion and future directions

We propose a method to predict seafloor visual classes onto wide-area multimodal remote sensing data. Our results show that location-guiding can improve the habitat class prediction of self-supervised feature learners for both single- and multi-modal remote sensing data. Furthermore, location-guiding allows feature learners to take advantage of early fused multi-modal inputs in scenarios where conventional feature learning cannot. It is thought that this relates to improved robustness to positional offsets between data modes when



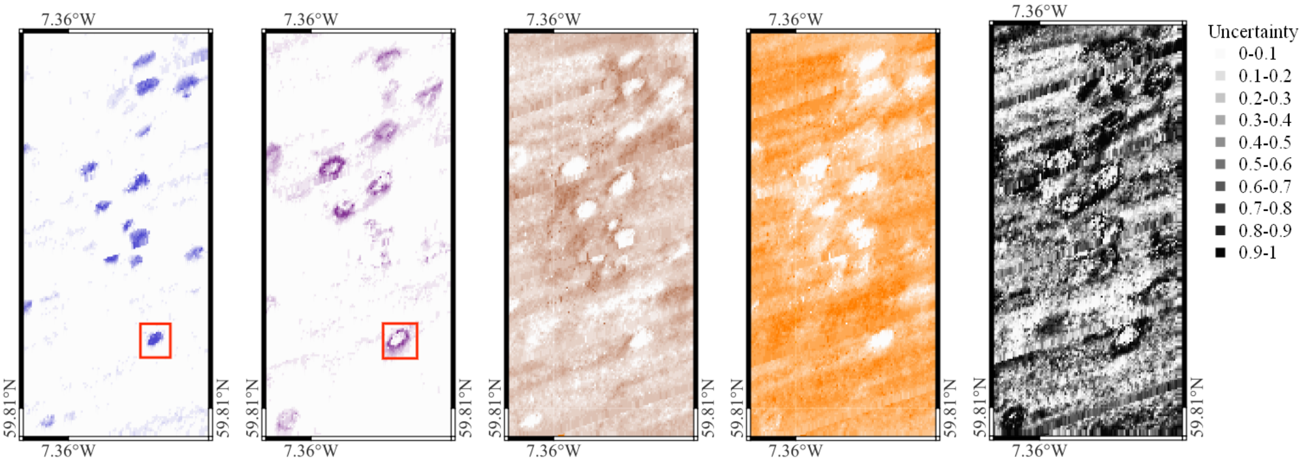
(a) (b) (c)

Fig. 3: Data from the Darwin Mounds MPA. (a) Visual classes determined for the seafloor images using the approach described in [18]. The colors correspond to Fig. 2; (b) and (c) show SSS intensity and seafloor depth derived from MBES. In our study, the relative depth (RD) is used to capture habitat relevant terrain features. This is combined with the SSS intensities to form multimodal inputs for the feature learner.



(a) (b) (c) (d)

Fig. 4: TSNE distribution of features in single-mode and multimodal class inference. Panels (a) and (b) depict single-mode class inference using SimCLR and GeoCLR, respectively, while panels (c) and (d) present multimodal class inference utilizing both SimCLR and GeoCLR.



(a) (b) (c) (d) (e)

Fig. 5: Probabilities of class inference. (a), (b), (c) and (d) reveal the probability of Mounds Top, Mounds Edge, Tails and Sediment. Illustrated colours correspond to those in Fig. 3a. (e) shows the prediction uncertainties

larger-scale spatial patterns are taken into account during the feature learning process.

The use of GPRs to model the relationship between features extracted from multimodal remote sensing data and overlapping visual classes extracted from imagery allows both the visual class probabilities and the prediction uncertainties to be determined. The final aspect can potentially be used to develop novel survey strategies, where efforts in data acquisition are focused on minimising predictive uncertainty to improve confidence in our understanding of seafloor habitat distributions.

References

[1] A. Zelada Leon et al., “Assessing the repeatability of automated seafloor classification algorithms, with application in Marine Protected Area Monitoring,” Remote Sensing, vol. 12, no. 10, p. 1572, May 2020.

[2] D. Rao, M. De Deuge, N. Nourani–Vatani, S. B. Williams, and O. Pizarro, “Multimodal learning and inference from visual and remotely sensed data,” The International Journal of Robotics Research, vol. 36, no. 1, pp. 24–43, Dec. 2016.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, ”A simple framework for contrastive learning of visual representations,” in Proc. Int. Conf. Mach. Learn. (ICML), PMLR, pp. 1597-1607, Nov. 2020.

[4] S. Laine and T. Aila, ”Temporal ensembling for semi-supervised learning,” arXiv preprint arXiv:1610.02242, 2016.

[5] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, ”Unsupervised learning of visual features by contrasting cluster assignments,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 9912-9924.

[6] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, and P. Bojanowski, ”Dinov2: Learning robust visual features without supervision,” arXiv preprint arXiv:2304.07193, 2023.

[7] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, ”More diverse means better: Multimodal deep learning meets remote-sensing imagery classification,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 5, pp. 4340-4354, 2020.

[8] W. R. Tobler, ”A computer movie simulating urban growth in the Detroit region,” Econ. Geogr., vol. 46, no. sup1, pp. 234-240, 1970.

[9] W. D. Koenig, ”Spatial autocorrelation of ecological phenomena,” Trends Ecol. Evol., vol. 14, no. 1, pp. 22-26, 1999.

[10] H. Gunes and M. Piccardi, ”Affect recognition from face and body: early fusion vs. late fusion,” in Proc. 2005 IEEE Int. Conf. Syst., Man, Cybern., vol. 4, Oct. 2005, pp. 3437-3443.

[11] T. Yamada, A. Prügel-Bennett, S. B. Williams, O. Pizarro, and B. Thornton, ”Geoclr: Georeference contrastive learning for efficient seafloor image interpretation,” Field Robotics, vol. 2, pp. 1134-1155, 2022.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, ”A simple framework for contrastive learning of visual representations,” in Proc. Int. Conf. Mach. Learn., PMLR, 2020, pp. 1597-1607.

[13] L. Victorero, D. Blamart, E. Pons-Branchu, M. N. Mavrogordato, and V. A. Huvenne, ”Reconstruction of the formation history of the Darwin Mounds, N Rockall Trough: how the dynamics of a sandy contourite affected cold-water coral growth,” Mar. Geol., vol. 378, pp. 186-195, 2016.

[14] T. Yamada, A. Prügel-Bennett, and B. Thornton, ”Learning features from georeferenced seafloor imagery with location guided autoencoders,” J. Field Robotics, vol. 38, pp. 52-67, 2021.

[15] A. Krizhevsky, V. Nair, and G. Hinton, ”CIFAR-10 (Canadian Institute for Advanced Research),” [Online]. Available: http://www.cs.toronto.edu/ kriz/cifar.html

[16] J. Deng, W. Dong, R. Socher, Li-Jia Li, Li Kai, and F.-F. Li, ”ImageNet: A large-scale hierarchical image database,” in IEEE CVPR, 2009, pp. 248-255.

[17] B. Thornton et al., ”Visualizing multi-hectare seafloor habitats with BioCam,” Front. Ocean Observ., suppl. to Oceanography, vol. 34, pp. 92-93, 2021.

[18] T. Yamada, M. Massot-Campos, A. Prügel-Bennett, O. Pizarro, S. Williams, and B. Thornton, ”Guiding Labelling Effort for Efficient Learning With Georeferenced Images,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, pp. 593-607, 2023.