Guiding Labelling Effort for Efficient Learning With Georeferenced Images

Takaki Yamada, Miquel Massot-Campos, Adam Prügel-Bennett, Oscar Pizarro, *Member, IEEE*
Stefan B. Williams, *Senior Member, IEEE,* and Blair Thornton, *Member, IEEE*

Abstract—We describe a novel semi-supervised learning method that reduces the labelling effort needed to train convolutional neural networks (CNNs) when processing georeferenced imagery. This allows deep learning CNNs to be trained on a per-dataset basis, which is useful in domains where there is limited learning transferability across datasets. The method identifies representative subsets of images from an unlabelled dataset based on the latent representation of a location guided autoencoder. We assess the method's sensitivities to design options using four different ground-truthed datasets of georeferenced environmental monitoring images, where these include various scenes in aerial and seafloor imagery. Efficiency gains are achieved for all the aerial and seafloor image datasets analysed in our experiments, demonstrating the benefit of the method across application domains. Compared to CNNs of the same architecture trained using conventional transfer and active learning, the method achieves equivalent accuracy with an order of magnitude fewer annotations, and 85 % of the accuracy of CNNs trained conventionally with approximately 10,000 human annotations using just 40 prioritised annotations. The biggest gains in efficiency are seen in datasets with unbalanced class distributions and rare classes that have a relatively small number of observations.

Index Terms—Semi-supervised learning, convolutional neural network, autoencoder, georeferenced imagery, pseudo-labelling

1 Introduction

Georeferenced visual images taken by aircraft, satellites and submersibles are widely used in environmental monitoring. Modern robotic surveys using aerial drones and Autonomous Underwater Vehicles (AUVs) can collect thousands to tens of thousands of georeferenced images in a single mission [1], [2], [3]. As the influx of images gathered by these platforms increases, the need for domain expertise to generate appropriate annotations becomes a bottleneck in our ability to efficiently interpret the data. Supervised machine learning techniques are potentially useful for automated interpretation. However, environmental studies have reported limited transferability of learning from generic training datasets [4], [5], citing the need for application-specific expert-annotated training examples. This is limiting since comprehensive training datasets do not yet exist for many environmental monitoring applications. The main reasons for this are the high sensitivity of image appearance to environmental conditions (e.g., lighting, atmosphere/water turbidity), observation variables (e.g., range to target, spatial resolution, observation footprint), the large variability in the appearance of unstructured scenes and the complexity of the annotation schemes used in environmental monitoring applications [6], [7]. These factors combined with the large number and different specification of the imaging platforms used (e.g., wavelength sensitivity, dynamic range, illumination source for underwater applications) limit crossover between datasets. Although unsupervised methods can efficiently process large volumes of imagery without relying on human annotations, their outputs typically do not align with the class boundaries of interest to experts, which limits their value for environmental monitoring and infrastructure inspection [8], [9].

This paper develops a novel semi-supervised method that improves learning efficiency when using georeferenced imagery, and reduces the human effort needed to train classifiers for environmental monitoring applications. The method is designed for whole image classification of natural scenes in downward looking imagery and consists of the following parts:

- Unsupervised learning - extracting latent representations of an unlabelled image dataset
- Prioritised labelling - identifying a subset of representative images for human annotation, and assigning predictive pseudo-labels to the remaining data.
- Supervised learning - use of prioritised annotations and pseudo-labels to train CNNs

For unsupervised learning, we investigate the impact on downstream accuracy when two different types of autoencoder are used to learn latent representations. The first uses only the information in images and the second is a location guided autoencoder (LGA) that also uses georeference information to regularise learning [9]. For prioritised labelling, we investigate the impact of using different methods to automatically identify a small subset of images for prioritised annotation and estimate class decision boundaries when assigning predictive pseudo-labels in unannotated images.

- *T. Yamada, M. Massot-Campos, A. Prügel-Bennett and B. Thornton are with the Faculty of Engineering and Physical Science, University of Southampton, Southampton, SO16 7QF, United Kingdom.*
  E-mail: {T.Yamada,miquel.massot-campos,B.Thornton}@soton.ac.uk, apb@ecs.soton.ac.uk
- *S. Williams and O. Pizarro are with Australian Centre for Field Robotics, The University of Sydney, NSW, Australia.*
  E-mail: {stefan.williams, o.pizarro}@sydney.edu.au
- *B. Thornton is also with Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba Meguro-ku, Tokyo 153-8505, Japan.*

*Manuscript received ** **, 2020; revised ** **, 2021.*

The prioritised annotations and pseudo-labels can be used to train different CNNs. We analyse the success and sensitivity of the proposed method using four different real-world datasets consisting of tens of thousands of georeferenced environmental monitoring image patches that have expert human labels for training and validation. The gains in learning efficiency are assessed based on the achieved accuracy and number of human annotations used in comparison to CNNs trained using well established transfer and active learning methods.

The advantages of the semi-supervised method for downstream classification tasks are:

- Unlike unsupervised methods, classifier outputs are aligned with class boundaries of interest to humans
- Accurate results can be achieved with significantly reduced human annotation effort compared to conventional supervised methods, and significantly reduced human and computational effort compared to iterative training approaches (e.g., active learning)

The reduction in human effort needed to achieve an equivalent accuracy to current state of the art approaches means that end-to-end training can be achieved on a per-dataset basis, making our approach suitable for use in domains where there is limited transferability of learning between datasets. The rest of this paper is structured as follows; section 2 reviews relevant machine learning literature and section 3 describes the semi-supervised training method. Experimental results for georeferenced seafloor and aerial image datasets are presented in section 4.

2 Background

2.1 Machine Learning for Environmental Monitoring

Determining the distribution of land cover, land use, habitats, substrates and infrastructures are tasks that lie at the core of environmental monitoring. One way of achieving these tasks is to interpret imagery using established classification schemes [10], [11], where often a small subset of images are selected for human annotation from which aggregate statistics can be derived. For more comprehensive analysis, many groups have reported automated interpretation of imagery using machine learning, with representative literature described in the following subsections.

2.1.1 Supervised Learning

A large proportion of automated classifiers have used a combination of hand-picked features chosen based on expert knowledge of the application domain or through a reward-based selection process [12], [13]. In [12] the authors apply a Support Vector Machine (SVM) to texture- and colour-based features designed to classify seafloor images into different substrates types for reef ecology surveys. In [14] hand-picked geometric features are combined with SVM for classification of satellite images. In [13] a similar approach is applied for seafloor mineral prospecting. Spatial invariant features such as Local Binary Patterns (LBP) [15] and Spatial Pyramid Matching (SPM) [16] have also been effectively applied to classification problems for land [17], [18] and seafloor imagery [19], [20]. However, these types of features require manual tuning of parameters, or feature engineering, to efficiently describe each independent dataset. Furthermore, a separate classification process is needed, which typically requires further parameter tuning. As such these methods often require expert knowledge of both the data and application domain, and have limited versatility when applied to multiple datasets.

A key advantage of deep learning techniques is that both the latent representation of data and classification can be simultaneously optimised in a single end-to-end training process. This avoids the need for costly and potentially subjective feature engineering and reduces the need for parameter tuning, making deep learning techniques a compelling choice for image classification tasks. Deep learning techniques are widely used for interpreting aerial and satellite imagery [21]. In [22] the ResNet [23] deep learning CNN is used to classify images of coral into nine separate classes, achieving higher classification resolution than prior studies and demonstrating the ability of deep learning to effectively model class boundaries used in scientific taxonomy. However, to work effectively, deep learning classification techniques typically require a large number of annotated examples of each class. Although labelling platforms tailored to aerial imagery [24] and seafloor imagery exist [2], [5], the sensitivity of images to environmental and acquisition conditions, complexity of annotation schemes and comparatively small size of each environmental monitoring community means that large-scale label repositories such as those in terrestrial imaging [25] and autonomous driving [26] do not yet exist. Several annotated datasets exist for satellite imagery [27]. However, most of these target built environments and artificial objects, and the annotations are not suitable for monitoring and conservation of the natural environment, where standardised but complex hierarchical annotation schemes that consist of hundreds to several thousands of terms are used [6], [7]. Furthermore, for sub-sea imaging, most groups gather images using custom built imaging hardware, where in [28] the authors reported that even small differences in sub-sea imaging hardware limits learning transferability and distorts deep learning classifier outputs. In [29] a pipeline to make training datasets transferable for inference on images from other datasets is proposed for segmentation of marine organism. The work proposes how to reduce scale variance across multiple datasets, which is highlighted as an important consideration for seafloor imagery. A detailed description of this and other domain specific distortions (e.g., blur, haziness, and colour distortion) that affect seafloor imagery can be found in Appendix A. For datasets where these disturbances are non-negligible, training on a per-dataset basis as is common in unsupervised learning can be considered a potentially effective solution.

Under these constraints, a reasonable approach for effective use of deep learning techniques is to train models on the target dataset itself. However, the implied requirement to annotate large numbers of images every time a new dataset is obtained is unlikely to be justified for most applications, forming a barrier to wide-spread adoption of deep learning for image interpretation in environmental monitoring applications. This motivates research into techniques for effort reduction.

2.1.2 Unsupervised Learning

Unsupervised learning techniques have great potential for image interpretation in environmental monitoring because they do not require annotations, and so can be efficiently trained and applied on a per-dataset basis. As with any automated image analysis, feature engineering is crucial for effective interpretation. In [8] LBP [15] features derived from greyscale images, 3D rugosity and colour are applied to seafloor image clustering. The authors later applied SPM [16] as a more generic approach to describe seafloor images [30]. These scale invariant features are also used for clustering of aerial and satellite imagery [31]. In [32], the non-parametric Bayesian clustering technique used in [8] and [30] is extended to incorporate annotations made during active learning [33] for seafloor imagery. In [34] the accumulated histogram of oriented gradients from keypoints are used to describe each image, and this is applied to clustering and anomaly detection. More recently, Shields et al. [35] used unsupervised clustering results generated from visual images as labels for supervised learning of terrain elevation datasets. To avoid the demanding trial and error process of feature engineering, we developed an unsupervised deep learning LGA in our previous work [9]. The proposed LGA learns latent representations without the need for feature engineering. The georeference information attached to each image is used to regularise learning, allowing CNN architectures to leverage this information and describe patterns that occur on spatial scales larger than a single image frame in a single end-to-end process. Since the LGA does not require any human annotations, it can be efficiently trained and applied on a per-dataset basis, and this has been shown to be effective for clustering and content-based query of seafloor images. Tile2Vec [36] is a method proposed for representation learning of aerial and satellite imagery, where a similar approach based on the physical distances between cropped image patches are leveraged during training.

However, a disadvantage of unsupervised approaches is that the resulting clusters do not attempt to align with the class boundaries of interest to humans, and when latent representations are optimised on a per-dataset basis, it is not possible to make direct comparisons between clusters or perform content-based queries across multiple processed datasets.

2.2 Methods to Reduce Annotation Effort

The shortage of annotations is a common problem when supervised learning is applied to real-world problems, and a number of concepts have emerged to address this issue.

2.2.1 Transfer Learning

Transfer learning allows supervised learning models to be trained using a relatively small number of annotations in the target dataset by making use of much larger annotated datasets from a different domain. Several frameworks have been proposed to implement this concept [37]. Network-based transfer learning has been applied in many application domains including medical [38], satellite [39], and seafloor imaging [40]. This approach works by reusing networks that have been pre-trained using large, generic datasets (e.g., ImageNet [41], COCO [42], Pascal VOC [43]) that consist of hundreds of thousands to more than ten million labels as an initial model. Though the number of dataset specific annotations needed depends on the domain, number of classes and data augmentation methods used, previous studies on satellite [44] and medical imagery [38] have required several hundreds of domain specific labels for effective use.

2.2.2 Prioritised Labelling

Images in a dataset do not have equal value for CNN training. In [45] the authors demonstrate that training data selection can have a significant impact on learning, where CNNs trained on a well selected subset of annotations can outperform CNNs trained using a larger number of annotations. In [46] annotation efforts are prioritised using k means clustering to estimate the entropy of each sample, showing significant gains in performance compared to random selection.

In active learning [33], the learner interacts with human annotators by iteratively proposing data samples that it considers will most efficiently improve performance. Several strategies have been proposed to achieve this. Most approaches prioritise unlabelled samples that have the highest estimated uncertainty, or are predicted to have the biggest impact on the model. However, the heuristics used to suggest samples can only be calculated after the initial subset has been analysed by the algorithm. Although the initial subset can impact subsequent learning performance, its selection falls outside of the scope of most active learning techniques [33], [47].

In [40] an autoencoder is used to locate objects of interest in an unsupervised manner. The method highlights these regions to human experts in order to facilitate efficient use of time for manual segmentation. The approach leverages the assumption that interesting objects are relatively rare in the original image datasets they are applied to. Regions with a high autoencoder reconstruction loss value are considered likely to include targets of potential interest, and these regions are flagged for prioritised annotation by humans. Active learning is also applied for seafloor image interpretation in [32], [35], where the authors implemented this with SPM as the feature descriptor.

2.2.3 Group Labelling and Label Extrapolation

Group-based labelling [48], [49] is a technique that assigns annotations to subgroups of clustered data in order to reduce the human annotation effort. An advantage of this approach is that it can be applied to datasets with no labels by using unsupervised clustering methods to generate the groups. However, determining the annotation for a cluster of images can be more complex than per-sample based annotation, especially when unsupervised cluster decision boundaries are not aligned with the desired class boundaries, resulting in conflicted human annotations. In [50] the authors modified Gaussian mixture model based clustering to find clusters with high intra-cluster similarity since the samples in these clusters are considered to be more informative than others. Although these techniques have shown significant improvement in learning efficiency, the underlying assumption is that effective clustering can be achieved.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3140060, IEEE Transactions on Pattern Analysis and Machine Intelligence

Predictive pseudo-labelling [51] reduces human effort by first training a classifier on a small subset of data that requires fewer annotations than the target dataset. An advantage of this over group labelling is that annotators consider individual images. After initial training, the classifier predicts labels for the remaining data, and these pseudo-labels are used together with the original annotations to fine-tune a classifier. Li et al. [47] reports that SVM and Random Forest classifiers outperform CNNs when generating pseudo-labels from an initial annotated subset. Wu et al. [52] uses pseudo-labelling to improve the classification performance for a hyperspectral satellite image dataset, demonstrating effective application of this approach to unstructured environmental monitoring data, where random subsets were used for initial training. The use of prioritisation methods for subset selection in pseudo-labelling has not previously been investigated.

3 Efficient Learning in environmental monitoring Imagery

Our aim is to develop a method to efficiently learn class boundaries of interest to humans with fewer annotations than existing methods, and apply this to environmental monitoring image classification problems. Fig. 1 shows the proposed semi-supervised learning pipeline. It learns latent representations of images in a dataset using the LGA [9] (section 3.1). Next, a subset of image samples are selected based on hierarchical k means clustering (section 3.2) for prioritised annotation by humans. Pseudo-labels are assigned to all remaining images (section 3.4) based on the annotated subset. The human annotations and algorithm generated pseudo-labels are then used to fine-tune a CNN, which can be used to solve a downstream classification task. The method is designed to work offline on a per-dataset basis, once the complete dataset has been gathered. The initial latent representation learning and identification of prioritised images for labelling are unsupervised, where all images in the dataset are available for these steps without the need for any human input. Human input is only needed to annotate the subset of prioritised images, where the number of prioritised images can be matched and optimised according to the availability of human effort. As such, the method is compatible with post data acquisition workflows associated with environmental survey field work. The LGA driven Semi-Supervised (LGA-SS) method is versatile as it allows a CNN to be both trained and applied to classification on a per-dataset basis, making it effective in domains where the transferability of learning between datasets is limited.

3.1 Location Guided Autoencoder

Patterns of interest in environmental monitoring often occur on spatial scales larger than the image patch size considered by CNNs during their optimisation. The LGA overcomes this problem by introducing georeference regularisation in autoencoder training using a modified loss function [9]. This is designed to reflect the assumption that *two images captured within a close distance tend to look more similar than two that are far away* due to the presence of patterns beyond the footprint of a single image frame. The approach allows the LGA to recognise patterns that recur in images that are close to each other and prioritise these in its learning without introducing artefacts due to imperfect image stitching. The latent representations obtained using the LGA have been shown to perform better than those obtained using a standard convolutional autoencoder when used for clustering and content-based image retrieval [9].

3.2 Data Selection for Prioritised Labelling

The standard CNN learning process expects class-balanced distributions in training datasets. Skewed class distributions, such as those found in natural scenes on land and on the seafloor, can result in overfitting of classes with relatively large numbers of samples. If M images are randomly selected for annotation, training datasets approximate the skewed class distributions of the parent populations, resulting in non-ideal conditions for training and carrying a risk that smaller classes may not be represented in training for small M values.

In the proposed pipeline, k means clustering is applied to the LGA's latent representation to identify densely populated regions. The number of clusters should be large enough to avoid missing small classes. As long as this condition is satisfied, the outputs are not strongly sensitive to small differences in k as the clusters attempt to evenly represent the different regions of the latent space. In this work we define, $k = \lceil k_e/10 \rceil \times 10$, where k_e is a number of clusters estimated by the elbow method [53]. The value of k is k_e rounded up to the nearest ten. Next, a subset of images for prioritised annotation are selected by taking $\lfloor M/k \rfloor$ or $\lceil M/k \rceil$ images from each cluster so that the total number of images is M. This generates a training class distribution that follows the cluster distribution, which eases the class imbalance problem as long as effective clustering is achieved. The way samples are chosen from within each cluster can also affect learning. In [46] it is assumed that the samples close to the cluster boundaries are important as they have a greater effect on classification decision boundaries. This assumption is reasonable if the boundaries of clustering and classification are comparable, but in situations where class boundaries are ambiguous, like in many environmental monitoring application, it is possible that variability in the annotations will degrade learning performance.

In this study, we consider that the samples provided for training should represent the variability within each cluster in order to deal with situations where the clustering resolution is not sufficient to resolve class boundaries. We implement two approaches to achieve this. The first approach uses k means clustering and randomly samples data from within each cluster so that each cluster in the LGA latent representation is evenly represented in the training data. We also investigate a more structured form of latent space representation, which we implement using hierarchical k means clustering. This approach is originally proposed in [54] where a multi-stage clustering process is introduced. The first stage explores the dominant patterns in the whole dataset, and the following stages attempt to select a representative set of samples from within each cluster. This approach has also been applied to extract representative data in text clustering problems [55]. In this work, we

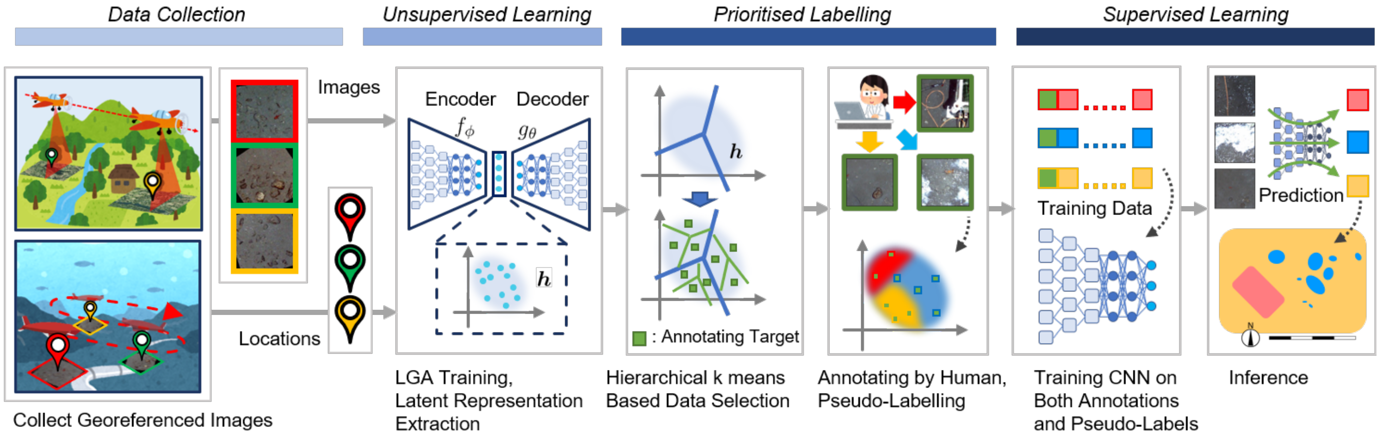

Data Collection
Unsupervised Learning
Prioritised Labelling
Supervised Learning
Images
Locations
Collect Georeferenced Images
Encoder f_ϕ
Decoder g_θ
h
LGA Training, Latent Representation Extraction
h
🟩 : Annotating Target
Hierarchical k means Based Data Selection
Annotating by Human, Pseudo-Labelling
Training Data
Training CNN on Both Annotations and Pseudo-Labels
Prediction
N
Inference

Fig. 1: A flow diagram of the proposed pipeline for LGA driven Semi-Supervised (LGA-SS) training of CNNs. Once a dataset is gathered, the latent representations of the images in the dataset are generated using the LGA [9] (section 3.1), after which hierarchical k means clustering (section 3.2) is used to identify a prioritised subset of images for human annotation. These annotations are used together with a set of algorithmically generated pseudo-labels for the remaining unannotated data to train a CNN that can be used for downstream classification tasks. The proposed LGA-SS method allows a CNN to be trained and applied to classification tasks on a per-dataset basis, making it effective in domains where there is limited transferability of learning between datasets.

consider that it is important to guarantee that samples are selected from dense regions of the latent representation, and so after the first k means clustering, we generate $\lfloor M/k \rfloor$ or $\lceil M/k \rceil$ sub-clusters within each cluster and select samples that are closest to each sub-cluster centroid so that the total number of samples is M.

3.3 Data Augmentation

Data augmentation [56] plays an important role in reducing the risk of overfitting during CNN training. Since most visual features in downward looking images of the seafloor and of land can be considered invariant to rotation and flipping [57], we apply these augmentations randomly during the training process, together with random shift operations to account for uncertainty in position. These transformations are applied with different parameters (i.e., rotation angle and offset) that are randomly assigned every time an image is fed into the model during training. Weighted sampling is also applied at each epoch to balance the number of samples in each class. Data augmentation is not applied to colour and scale distortions since it can be consistently corrected taking into account illumination and turbidity conditions and lens distortions [9].

3.4 Pseudo-Labelling

We predict pseudo-labels for each unseen image based on its location relative to annotated samples in the LGA latent space. Although the clustering results used to identify images for prioritised annotation can be used for this purpose, the decision boundaries of clusters and classes are not necessarily aligned. Therefore, we investigate different approaches to estimate class decision boundaries, comparing the performance of nearest neighbour (1-NN), Random Forest and SVM [58] with linear and Radial Basis Function (RBF) kernels as methods capable of expressing varying degrees of complexity of class boundaries in the latent space.

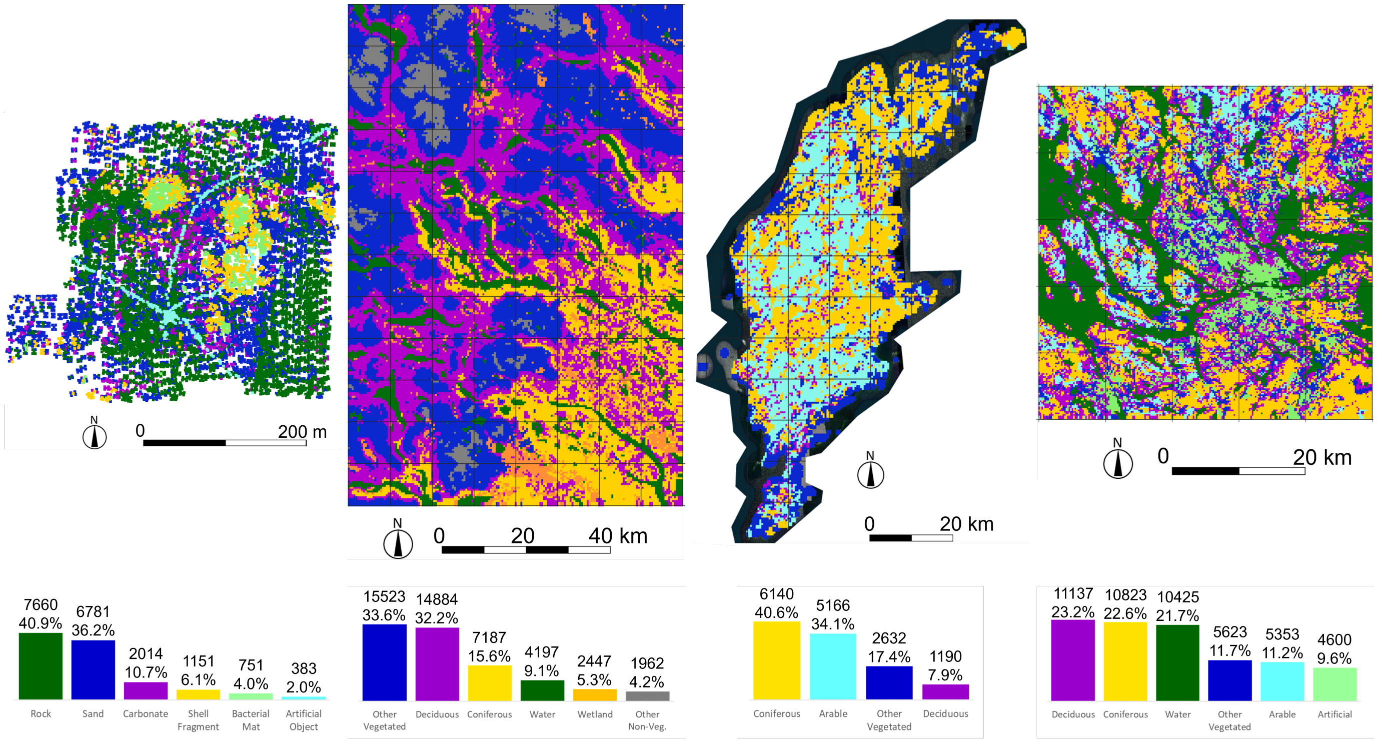
Although the original pseudo-labelling implementation for deep learning applies a single winner takes all class label to unseen data [51], recent research has demonstrated that taking the uncertainty of each pseudo-label into consideration can improve downstream classification accuracy [59], [60]. Class boundaries in environmental monitoring data are often ambiguous and so to address uncertainty near class decision boundaries, we implement probabilistic pseudo-labelling using a Gaussian Process classifier [61] to predict class conditional probability distributions for each sample in the latent space for comparison with the other methods.

Both the annotations and pseudo-labels assigned to the remaining images are used to train CNNs, where for probabilistic pseudo-labelling, the conditional probability distributions are applied to the softmax loss of CNN training in order to describe the pseudo-label uncertainty. The suitability of these classifiers for pseudo-labelling is determined through validation against human annotations.

4 Experiment

4.1 Dataset

The proposed method is applied to four different environmental monitoring image datasets. Fig. 2 shows the spatial and class distributions of the ground truth for each dataset. The Seafloor dataset (Fig. 2a, see Appendix A for further details) consists of seafloor visual images collected by an AUV, and the aerial image datasets (Fig. 2b - 2d, see Appendix B for further details) are of different types of scene (Mountain, Island and Urban). The class distributions in these spatially continuous datasets are highly skewed compared to the generic datasets that are often used in benchmarking studies. Our experiments consider each class to be of equal importance. The results are assessed based on the macro-averaged F_1-score, where we take the mean and standard deviation (SD) of 10 repeated sets of experiments under each test configuration.



(a) Seafloor (b) Mountain (c) Island (d) Urban

Fig. 2: Spatial patterns (top) and class distributions (bottom) of ground truth classes in four environmental monitoring datasets. Each natural or artificial object class shows a unique spatial pattern in each dataset. The class distributions are highly skewed since all the images in the corresponding areas are included in the datasets without any manual selection process. The Seafloor dataset (2a) consists of colour seafloor imagery collected by an AUV. The three aerial datasets (Mountain, Island and Urban) consist of aerial images cropped from ESRI World Imagery. Details of these datasets can be found in Appendices A and B, respectively.

4.2 Classification with Conventional Classifiers

We investigate the performance of conventional (non-CNN) classifiers in order to generate effective pseudo-labels from a small subset of annotated examples. Five well established classifiers; k-NN with $k = 1$ (1-NN), Random Forest (RF), SVM with linear (L-SVM) and RBF kernels (R-SVM) [58] and Gaussian Process (GP) [61] classifiers are applied to the latent space mapped by an LGA that has been trained on all available image patches. The results are compared with those of a standard convolutional autoencoder that uses the same architecture as the LGA except for the geo-reference regularisation. To evaluate the performance with a small number of annotations, an adjusted cross-validation is applied. First, half of the annotated image patches are randomly selected as a test subset, preserving the class distribution of the entire dataset in each dataset. Then M images are selected from the remaining patches based on random selection, k means based selection, and the proposed hierarchical k means based selection. Following the equation defined in section 2.2.2, $k = 20$ is used for both k means and hierarchical k means based selection for all the datasets. In k means based selection, $M/20$ images are selected randomly from each cluster. In hierarchical k means based selection, the second stage k means is applied to each cluster to find $M/20$ sub-cluster centroids, and the images closest to each centroid are selected for annotation. Training and testing are executed ten times for each configuration with M = 20, 40, 100, 200, 400, 1000 and 7500 (for the aerial datasets) or 9370 (for the Seafloor dataset).

TABLE 1 and TABLE 2 show the mean and SD of the F_1-scores for the ten-time cross-validation with each configuration (A1 - A20 in TABLE 1 and A'1 - A'20 in TABLE 2) on the seafloor and aerial datasets, respectively. The data selection strategy has a greater impact on performance than the choice of classifier, with all classifiers benefiting significantly from hierarchical k means prioritisation. The relative gains in accuracy compared to random selection are especially large for small values M (20, 40 and 100), confirming the importance of the data selection strategy when training with a small number of annotations. For the Seafloor dataset (TABLE 1), the combination of LGA based pre-training and hierarchical k means based data selection with a R-SVM (configuration A14) performs the best among the tested cases for all values of M. The L-SVM and GP generally perform better than 1-NN and RF, where the L-SVM tends to be better for small values of M and GP better for larger M. A similar trend is observed with the aerial datasets (TABLE 2). For small values of M, L-SVM outperforms R-SVM; however, the difference is marginal. The largest efficiency gains are achieved in the datasets that have rare classes with the smallest number of relative observations (i.e., Seafloor and Mountain).

The standard deep learning autoencoder (configuration

TABLE 1: F_1-Score (Macro-Average) Mean and SD (%) of the Classification Result with Conventional Classifiers on the Seafloor Dataset

Config. Label	Feature Learning	Data Selection	Classifier	Number of Annotations (M)						
				20	40	100	200	400	1000	9370
A1	LGA	random	1-NN	31.8±9.1	40.1±3.0	44.0±4.5	47.6±3.3	48.9±3.1	50.6±1.5	54.0±0.5
A2			RF	27.6±6.8	38.3±4.1	43.0±4.5	48.7±4.9	51.8±3.3	56.4±1.4	61.0±0.3
A3			L-SVM	33.8±9.6	43.5±3.9	48.2±4.4	52.6±3.4	54.3±2.5	56.3±1.9	60.0±0.5
A4			R-SVM	31.6±7.6	42.4±3.7	48.4±3.7	54.4±3.3	57.4±3.7	60.2±0.8	**63.3**±0.7
A5			GP	29.6±5.2	39.1±6.2	46.1±3.9	47.2±5.2	52.5±2.2	56.9±2.0	***63.2***±0.6
A6		k means	1-NN	41.6±5.1	46.2±5.2	47.2±4.3	50.8±1.7	51.0±1.9	52.5±1.1	54.0±0.5
A7			RF	33.7±5.7	43.1±5.2	49.2±4.0	54.7±1.2	56.9±1.6	59.1±0.7	61.0±0.3
A8			L-SVM	43.2±5.4	47.9±5.8	51.5±4.2	56.6±1.3	57.1±1.4	59.9±1.0	60.0±0.5
A9			R-SVM	42.0±6.0	50.7±5.3	***55.1***±4.4	***59.2***±1.5	***60.5***±1.3	***62.4***±0.8	**63.3**±0.7
A10			GP	42.1±6.8	45.8±6.8	51.8±2.5	55.1±1.5	57.2±1.6	60.0±0.8	***63.2***±0.6
A11		H-k means	1-NN	46.9±7.2	48.6±4.3	48.9±2.9	52.2±2.4	52.3±1.7	53.0±0.8	54.0±0.5
A12			RF	42.1±7.4	47.9±3.9	51.8±2.5	55.8±1.5	57.6±1.5	59.3±0.9	61.0±0.3
A13			L-SVM	***47.4***±8.1	50.9±4.7	53.6±3.0	56.8±1.6	58.3±1.2	60.8±0.9	60.0±0.5
A14			R-SVM	**48.0**±8.3	**54.8**±2.3	**56.9**±2.0	**60.1**±1.0	**61.0**±1.0	**62.7**±0.7	**63.3**±0.7
A15			GP	44.5±7.7	***51.4***±3.8	***55.1***±2.3	56.1±2.1	59.5±1.2	61.2±1.1	***63.2***±0.6
A16	Auto-encoder	H-k means	1-NN	25.5±1.3	30.5±1.5	33.2±1.0	33.8±1.2	35.6±1.4	36.6±0.8	38.3±0.5
A17			RF	24.4±1.7	29.0±3.0	32.0±1.6	33.6±2.2	35.6±1.1	39.1±0.8	41.1±0.4
A18			L-SVM	10.0±5.6	8.3±4.5	6.0±3.4	8.5±8.5	6.7±2.6	10.9±3.1	34.9±0.7
A19			R-SVM	21.7±3.4	28.2±2.6	29.6±4.0	35.0±1.8	38.3±1.5	42.0±0.9	44.9±0.6
A20			GP	9.7±0.0	9.7±0.0	9.7±0.0	10.3±1.4	14.9±1.3	18.9±0.8	21.5±0.3

A16 - A20 in TABLE 1 and A'16 - A'20 in TABLE 2) is significantly less effective than the LGA for all the datasets investigated in this work. This is an expected result since our previous work has already shown that the autoencoder achieves poor clustering performance without georeference regularisation [9], and the underlying assumption behind the data selection strategies investigated here is that effective clustering can be achieved. The results demonstrate that the proposed location guided latent representation learning and representative image selection are effective for environmental applications using georeferenced image datasets across application domains.

4.3 Classification with CNN

This section evaluates the proposed LGA-SS learning pipeline's performance using CNNs. The M training images and test images are selected in the same way as in section 4.2. When M is smaller than the total number of available training data, data augmentation (section 3.3), pseudo-labelling (PL) or probabilistic pseudo-labelling (PPL) (section 3.4) are applied so that the number of training images at each epoch is the same as the total number of the training images to allow for fair comparison of the results. Here we evaluate the proposed pipeline on the dataset with the largest class imbalance (i.e., Seafloor), since efficient handling of skewed class distributions is an important consideration when interpreting natural environment datasets.

4.3.1 CNN Architecture Comparison (B1-B8)

The proposed LGA-SS training method can be applied to any CNN architecture. Here, we investigate the impact of using three well established CNN architectures on classification accuracy: AlexNet, ResNet18 and ResNet152 [23]. The accuracy of each configuration is evaluated based on the mean F_1-score (macro-average).

Each CNN is pre-trained using ImageNet, where experiments are performed with all layers and only last layer training on the dataset following the network-based transfer learning process described in [37]. Since AlexNet is used as the basic architecture of the LGA implemented in this work, the LGA's encoder can be regarded as an AlexNet classifier where the weight values have been optimised to describe all the available images in the target dataset through latent representation learning. The performance of the LGA pre-trained CNN is compared to traditional ImageNet pre-trained CNNs to assess the effectiveness of embedding georeference information using the LGA.

The following parameters are experimentally determined to achieve the best performance with each CNN architecture: Mini-batch sizes of 128 samples are used for AlexNet (all layer and final layer training) and ResNet18 (all layer training), 32 for ResNet18 (final layer training) and 16 for ResNet152, Adam [62] is used as the optimiser and the learning rate is set to 1e-5 except for ResNet18 (final layer training) where it is set to 1e-4, and the number of training epochs is 50 for all configurations.

TABLE 3 shows the results for configuration B1 to B8. As expected, the accuracy improves when a larger number of annotations are used to train each CNN architecture. Overall, B4, which corresponds to AlexNet pre-trained using the LGA where only the last layer is trained on the dataset, shows the best performance except for when $M = 40$ and 9370. The performance gap between B4 and B8, where all the layers are trained on the dataset, is potentially caused by overfitting due to high model flexibility of B8. Though B8 outperforms B4 for $M = 40$, the difference in performance here is marginal. B8 shows a similar level of accuracy to B5, where ImageNet is used for pre-training instead of the LGA, indicating that the advantage of LGA pre-training is lost when all the layers are trained. The fact that B4 generally outperforms these configurations demonstrates the

TABLE 2: F_1-Score (Macro-Average) Mean and SD (%) of the Classification Result with Conventional Classifiers on Aerial Datasets (Mountain/Island/Urban Dataset)

Config. Label	Feature Learning	Data Selection	Classifier	Number of Annotations (M)						
				20	40	100	200	400	1000	7500
A′1	LGA	random	1-NN	43.8±4.2	49.3±3.0	53.4±2.4	55.4±1.5	57.6±0.9	59.4±0.8	62.1±0.4
				42.0/45.3/44.0	49.6/47.9/50.4	54.5/50.9/54.8	56.8/52.5/57.0	60.3/53.7/58.8	63.0/55.0/60.3	66.9/56.4/63.0
A′2			RF	42.7±6.2	49.4±5.0	55.8±3.2	57.9±2.3	60.8±1.2	63.3±0.9	66.6±0.3
				39.0/45.4/43.8	48.5/49.0/50.7	55.8/53.2/58.4	59.0/53.2/61.5	63.2/55.1/64.2	66.6/56.8/66.4	71.2/58.8/69.8
A′3			L-SVM	46.3±4.5	52.1±3.6	58.6±2.3	61.1±1.4	63.3±1.0	65.3±0.5	66.9±0.4
				45.6/46.8/46.4	52.5/50.2/53.6	60.3/55.5/59.9	62.7/58.2/62.5	65.0/59.7/**65.4**	66.7/61.9/67.3	68.1/63.4/69.2
A′4			R-SVM	42.8±4.7	51.1±3.5	58.5±2.2	61.4±1.4	**63.8**±1.0	**65.7**±0.5	69.0±0.3
				39.0/46.7/42.8	50.4/51.3/51.6	59.8/**55.9**/59.9	62.9/**58.5**/62.8	65.6/**60.3**/65.3	67.5/**62.3**/67.4	70.7/65.3/**70.9**
A′5			GP	44.0±4.1	50.1±3.2	55.1±2.7	57.6±2.0	60.5±1.3	63.2±1.0	68.1±0.4
				42.9/45.1/44.0	51.4/48.0/50.9	56.5/51.9/57.0	59.0/52.7/60.9	63.5/54.4/63.7	67.2/56.1/66.3	72.9/60.5/70.8
A′6		k means	1-NN	47.1±4.3	51.1±2.4	54.0±2.1	55.9±1.4	57.1±1.1	59.0±0.7	61.7±0.4
				46.5/47.4/47.4	53.3/49.7/50.3	56.6/50.3/55.1	58.9/52.0/56.9	60.7/53.0/57.8	62.3/54.4/60.3	66.3/56.2/62.7
A′7			RF	45.3±5.7	51.4±3.0	56.0±1.9	58.4±1.7	60.6±1.2	63.1±0.9	66.4±0.4
				42.8/47.9/45.1	51.5/50.9/51.8	57.0/52.4/58.5	59.9/54.0/61.2	63.3/55.0/63.6	66.2/56.9/66.3	71.2/58.7/69.3
A′8			L-SVM	49.1±4.5	54.4±2.3	58.6±1.9	61.4±1.3	63.3±1.0	64.8±1.0	66.4±0.4
				47.3/50.0/50.0	56.1/52.1/54.9	60.9/54.7/60.1	63.7/57.3/63.2	65.4/59.6/65.1	66.5/60.6/67.4	68.0/62.3/69.0
A′9			R-SVM	45.6±4.6	53.4±3.5	58.0±2.5	61.3±1.4	63.3±1.1	65.2±0.9	68.3±0.5
				42.9/48.3/45.7	53.9/**53.3**/52.9	59.2/55.2/59.5	63.1/57.6/63.1	65.9/58.8/65.1	67.2/60.8/**67.6**	70.6/63.6/70.8
A′10			GP	47.6±4.4	51.9±3.0	56.1±1.8	58.5±1.7	60.4±1.2	63.0±0.8	67.7±0.4
				47.4/48.1/47.4	53.9/50.8/51.0	58.6/51.8/57.9	61.0/53.8/60.6	63.3/54.5/63.4	66.3/56.4/66.4	72.5/60.1/70.5
A′11		H-k means	1-NN	50.7±3.0	52.7±2.4	56.6±1.6	58.2±1.1	59.1±0.7	60.3±0.6	62.4±0.4
				50.8/49.9/51.4	55.8/49.2/52.9	60.9/51.3/57.5	61.6/53.9/59.1	63.5/53.8/60.0	64.5/54.8/61.4	67.4/56.4/63.3
A′12			RF	49.0±3.6	52.4±2.9	57.9±1.7	59.9±1.7	62.1±1.2	63.8±0.7	66.9±0.4
				47.1/49.3/50.7	51.5/52.1/53.6	59.3/54.3/60.0	60.4/56.6/62.8	64.8/56.7/64.6	67.7/56.9/66.7	72.3/58.9/69.6
A′13			L-SVM	**51.8**±3.1	**55.0**±2.2	**59.8**±1.6	61.8±1.4	63.2±0.8	65.0±0.7	67.1±0.3
				52.5/50.2/**52.7**	**58.0**/50.5/**56.4**	**63.0**/54.1/**62.3**	**65.0**/57.2/63.1	65.9/58.7/65.0	66.9/61.0/67.0	69.0/63.5/68.9
A′14			R-SVM	50.8±3.3	53.8±2.4	59.3±1.9	**61.9**±1.3	63.5±0.7	65.5±0.5	**69.3**±0.3
				51.4/**52.1**/48.9	55.5/52.5/53.5	62.0/54.4/61.5	64.8/57.5/**63.3**	**66.8**/58.4/**65.4**	68.1/61.0/67.4	71.5/**65.4**/**70.9**
A′15			GP	50.9±3.1	53.6±2.6	58.2±2.0	60.5±1.8	62.5±1.2	64.1±0.7	68.2±0.4
				51.0/50.1/51.7	56.1/50.7/54.0	61.7/53.1/60.0	63.0/55.8/62.8	66.0/56.3/65.0	**68.3**/57.0/67.0	**73.4**/60.6/70.8
A′16	Auto-	H-k means	1-NN	45.5±3.2	48.0±2.3	51.8±1.8	53.0±1.2	53.6±0.8	54.8±0.7	57.2±0.3
	encoder			47.6/44.3/44.5	54.6/44.9/44.5	58.2/47.5/49.7	57.8/49.1/52.2	58.7/49.3/52.8	59.5/50.8/54.2	62.2/52.5/56.7
A′17			RF	44.8±3.3	48.3±2.8	51.7±2.1	54.6±1.1	55.9±0.9	58.1±0.6	61.5±0.4
				44.3/47.3/42.7	54.0/46.6/44.3	55.8/48.7/50.6	58.3/51.1/54.4	58.9/52.1/56.8	61.4/53.3/59.7	65.7/54.6/64.3
A′18			L-SVM	47.5±2.7	49.4±3.2	53.6±2.1	54.7±1.4	56.4±1.3	58.5±1.5	62.7±0.4
				49.5/46.2/46.7	52.7/48.1/47.4	57.9/51.0/52.0	57.0/51.3/55.6	58.2/53.5/57.4	58.9/56.2/60.5	66.1/59.1/62.9
A′19			R-SVM	46.8±3.5	49.6±3.3	53.6±2.0	55.3±1.6	57.8±1.1	60.2±1.6	65.2±0.3
				50.0/45.9/44.6	52.7/49.2/47.0	57.0/51.4/52.6	57.6/53.5/54.8	60.9/55.1/57.5	63.2/57.4/60.0	68.8/61.1/65.6
A′20			GP	46.4±3.4	49.1±2.6	52.6±1.8	55.0±1.3	56.3±1.0	58.4±0.5	62.7±0.5
				48.1/46.0/45.1	55.9/46.5/44.9	56.9/49.5/51.4	58.4/51.5/55.1	59.3/52.5/57.3	61.3/53.5/60.4	67.2/55.6/65.4

The standard deviation values shown are the mean values of the standard deviations calculated for the three datasets.

advantage of embedding georeference information through LGA pre-training using the target dataset. When $M = 9370$, B6, corresponding to the case where all the layers of an ImageNet pre-trained ResNet18 are trained on the dataset, shows the best accuracy. This suggests that ResNet18's deeper architecture and use of residual blocks allows for better performance than AlexNet when a sufficient number of training examples is available. However, B4 is the best option overall for $M \leq 1000$, which is significant for this study since we are interested in efficient training with a small number of annotated examples.

The comparison between B1 to B3 (last layer only) and B5 to B7 (all layer) indicates that training only the last layer limits the performance of each architecture for large values of M, indicating that there is a significant difference between the low-level and mid-level features of ImageNet and the environmental monitoring dataset. In the proposed pipeline, the number of training examples can be considered large due to the use of pseudo-labels. Therefore, we choose to investigate B6 as it demonstrate the best capacity for learning among B1 to B8, and we also examine B4 since it is the most efficient learner for $M \leq 1000$.

4.3.2 Active Learning Comparison (C1-C12)

Active learning methods attempt to improve learning efficiency by training classifiers on a subset of annotated samples, and proposing which samples should be annotated next based on their prediction uncertainty [33]. CNNs are well suited to this iterative process of prediction and prioritised annotation as their outputs are already conditional probabilities against labels and so uncertainty metrics can be easily derived. Common strategies for uncertainty based prioritisation include Least Confidence (LC) sampling, margin sampling and entropy based sampling, all of which have previously been demonstrated to be effective for environmental monitoring applications [32].

Conventional active learning starts the iterative training process with a randomly selected subset of samples. However, its performance is sensitive to this initial selection

TABLE 3: F_1-Score (Macro-Average) Mean and SD (%) of the Classification Result with CNN Trained by Standard Supervised Learning (Section 4.3.1), Active Learning (Section 4.3.2) and the Proposed Pipeline (Section 4.3.3).

Config. Label	CNN	Pre-training	Trained Layer	Data Selection	Number of Annotations (M)						
					20	40	100	200	400	1000	9370
B1	AN	IN	last	random	36.6±5.1	38.0±7.1	50.2±5.8	57.7±1.5	59.4±1.7	59.7±1.1	60.5±0.9
B2	RN18	IN	last	random	36.3±6.4	42.3±3.8	48.3±5.2	53.4±5.5	57.7±2.6	60.3±2.7	62.8±0.7
B3	RN152	IN	last	random	34.9±7.0	43.4±5.0	49.7±6.6	54.2±3.8	58.4±2.4	58.8±2.6	61.4±1.1
B4	AN	LGA	last	random	39.2±7.4	43.2±6.3	51.2±5.3	58.3±1.9	62.0±2.5	65.8±0.9	67.7±0.7
B5	AN	IN	all	random	31.1±7.5	39.2±6.7	48.1±5.9	53.8±3.8	57.3±2.2	60.5±2.0	***68.6***±0.7
B6	RN18	IN	all	random	34.1±7.0	38.5±9.9	50.7±6.4	54.9±5.1	58.5±3.1	61.9±1.1	**69.4**±0.6
B7	RN152	IN	all	random	35.3±6.4	38.2±8.2	50.3±5.8	51.7±3.3	57.5±2.0	59.1±1.8	64.9±1.1
B8	AN	LGA	all	random	32.9±7.0	44.6±4.0	44.9±5.8	54.7±3.5	57.7±2.7	60.1±1.4	66.3±0.9
C1	AN	LGA	last	random+LC	30.5±6.2	34.9±6.7	47.2±6.6	57.0±3.8	62.0±1.8	63.7±0.8	65.5±1.1
C2	AN	LGA	last	random+margin	32.9±5.6	40.3±4.4	49.6±9.1	55.8±7.4	60.5±2.3	61.8±1.4	64.8±1.5
C3	AN	LGA	last	random+entropy	36.9±7.9	41.3±8.7	53.4±4.9	58.5±3.5	62.0±1.5	63.7±1.3	66.2±0.5
C4	AN	LGA	last	k means+LC	49.6±4.7	53.7±5.4	56.5±4.3	59.6±2.0	62.2±1.8	62.8±1.5	65.6±1.0
C5	AN	LGA	last	k means+margin	48.9±4.2	52.5±2.7	56.3±2.7	57.8±1.9	60.5±1.2	61.7±1.4	64.2±1.4
C6	AN	LGA	last	k means+entropy	46.6±5.2	49.8±5.4	55.7±3.4	58.3±3.4	62.5±1.3	63.3±1.2	65.6±0.6
C7	RN18	IN	all	random+LC	33.4±7.8	43.4±6.7	53.1±4.5	56.8±2.2	58.6±1.2	59.4±0.9	63.5±0.7
C8	RN18	IN	all	random+margin	38.2±4.2	42.9±6.4	52.8±5.8	54.3±2.9	57.3±2.0	58.2±1.8	63.9±0.5
C9	RN18	IN	all	random+entropy	35.9±6.5	47.7±6.2	55.2±2.1	56.2±3.7	57.6±1.8	59.1±1.3	63.6±0.8
C10	RN18	IN	all	k means+LC	50.3±5.7	53.5±4.8	56.3±2.0	56.4±2.3	59.1±1.3	59.5±1.4	64.0±0.9
C11	RN18	IN	all	k means+margin	49.1±6.2	50.8±6.1	53.4±4.9	54.5±2.9	57.3±1.1	58.5±1.0	63.7±0.5
C12	RN18	IN	all	k means+entropy	49.2±7.0	52.1±5.7	55.4±2.9	57.5±2.9	59.0±2.1	60.3±1.3	63.6±0.7
D1	AN	LGA	last	k means	43.6±4.0	51.4±4.8	56.7±2.9	60.9±2.0	**64.5**±1.0	66.0±0.9	67.7±0.7
D2	AN	LGA	last	H-k means	44.9±6.4	53.2±4.2	58.1±2.2	61.5±1.8	***64.4***±1.1	**66.9**±0.8	67.7±0.7
D3	AN	LGA	last	H-k means+PL	***50.4***±8.3	***57.8***±3.0	**60.4**±2.6	***62.8***±1.0	62.7±1.2	64.7±0.8	67.7±0.7
D4	AN	LGA	last	H-k means+PPL	31.3±2.7	40.1±2.8	52.0±2.6	57.2±1.6	62.3±0.9	65.4±0.8	67.7±0.7
D5	RN18	IN	all	k means	45.5±8.0	49.6±7.2	55.4±3.9	57.2±2.3	59.5±1.6	62.0±1.1	**69.4**±0.6
D6	RN18	IN	all	H-k means	44.7±8.1	53.0±5.3	57.9±1.9	59.3±1.7	59.2±2.7	62.1±1.5	**69.4**±0.6
D7	RN18	IN	all	H-k means+PL	**51.9**±7.6	**59.1**±2.7	**60.4**±2.4	**62.9**±0.7	64.2±1.0	64.8±0.8	**69.4**±0.6
D8	RN18	IN	all	H-k means+PPL	46.2±3.2	51.2±2.5	55.1±2.4	58.9±1.3	52.3±1.5	***66.4***±1.1	**69.4**±0.6

The proposed method (D3 and D7) outperforms other configurations when $M \leq 200$. When $M = 9370$, all available training images are used making the selection strategy irrelevant. Bold and bold italics indicate the best and next best performer for each value of M.

and so we investigate whether an initial selection of samples nearest to the centroids of the k means clusters in the LGA latent space improves their performance. Subsequent batches of samples (20 when $M \leq 1000$ or 1000 when $M > 1000$) are selected based on the active learning query strategies and iteratively added to the subset of annotated samples for training. A training epoch of 10 was chosen so that the total number of epochs is comparable to the standard supervised learning results (B1-8) and proposed methods (D1-D8).

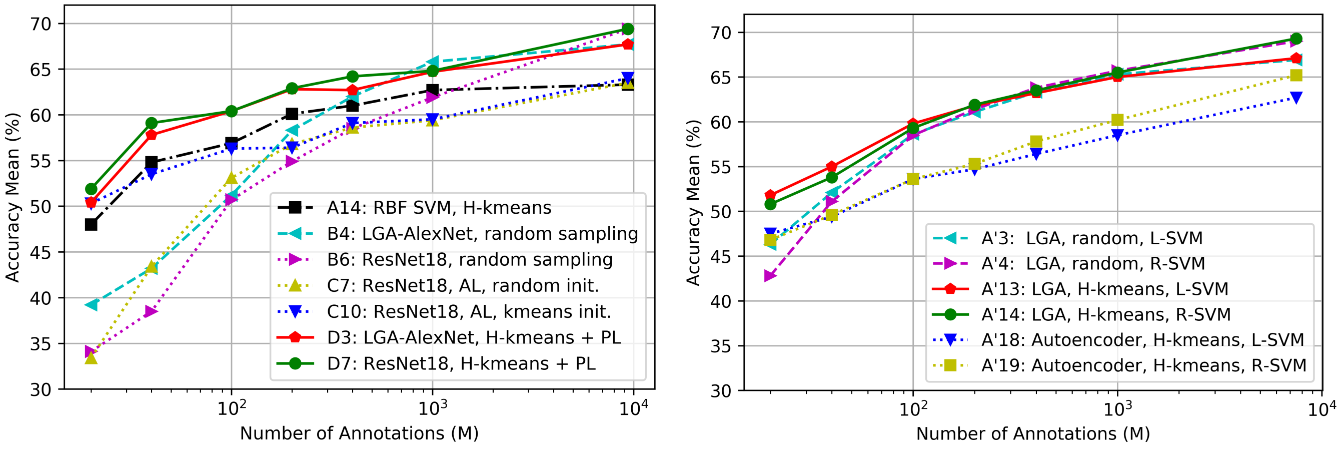
In our experiment, we assess two different CNN architectures (AlexNet and ResNet18), and compare the performance of three well established active learning iterative sampling techniques (LC sampling, margin sampling and entropy based sampling). The active learning process is initialised using two different initial subset selection methods; first where the initial subset is randomly sampled (corresponding to traditional active learning workflows), and second where active learning initialised by a k means centroid based sample initialisation (taking advantage of the georeference embedded latent representations learnt during LGA pre-training).

Configuration C1 to C12 in TABLE 3 show the accuracy scores for CNNs trained using the different configurations for active learning. Comparing the LGA pre-trained AlexNet configurations (C1 to C3) with their transfer learning counterpart (B4) shows that the active learning reduces accuracy. However, for ResNet18, the accuracy increases when active learning is applied (B6 and C7 to C9) for small values of $M < 1000$. It is noticeable that for larger M (particularly $M = 9370$), active learning degrades performance, possibly due to overfitting of CNN weights at an early phase of the iterative learning process trapping them in local minima. This is because the CNN is trained sequentially on discrete subsets of data, where the stored weights are used to initialise the optimisation of the next subset to limit the total number of training epochs required [63]. Although overfitting is potentially mitigated by resetting the CNN weights between each training subset [64], this requires a large number of training epochs, making it impractical for use in domains that require per-dataset training.

The use of the LGA k means centroids for initial sample selection significantly improves performance (C4 to C6 and C10 to C12), where the gains are largest for small numbers of training examples, i.e., $M \leq 100$. Although this advantage is lost as M increases, it does not cause any significant degradation in performance compared with the random initial subset selection. The difference between the active learning strategies is marginal for both the random and k means initial selection. Although different hyperparameters (e.g., number of epochs for each iteration) may improve active learning performance, optimisation of these is outside the scope of this work since there are no systematic methods available to determine them.

4.3.3 Data Selection Strategy Comparison (D1-D8)

Four data selection strategies; k means, hierarchical k means, and hierarchical k means with pseudo-labelling or probabilistic pseudo-labelling, are validated in this section.



(a) Annotations (M) vs F_1 (macro-average) mean for the Seafloor Dataset

(b) Annotations (M) vs F_1 (macro-average) means averaged over the Mountain, Island and Urban Aerial Datasets

Fig. 3: Comparison of classification performance investigated in section 4. Mean of F_1 (macro-average) values against each M are shown. Representative configurations are chosen from TABLE 1 and 3 for Fig. 3a and TABLE 2 for Fig.3b. The proposed georeference embedded sample selection method improves performance for all the datasets analysed in our experiments. Larger gains in learning efficiency are achieved in datasets that have a more heavily skewed class distribution.

The previous section already confirmed that hierarchical k means based data selection is effective for small values of M when combined with conventional non-CNN classifiers. In order to allow for fair comparison, the number of training samples used by the CNN at each training epoch is fixed to the total number of available labelled training image patches (i.e., 9370 in this experiment). For configurations where all available labelled image patches are used in the training (i.e., all pseudo-label and probabilistic pseudo-label configurations and where $M = 9370$ without pseudo or probabilistic pseudo-labelling), each original labelled training image patch is used once, and these samples are individually subjected to data augmentations that randomise orientations, flipping and position offsets at each training epoch before being used by the CNN. For configurations where the number of labelled image patches used in the training is less than available labelled training image patches (i.e., $M < 9370$ with no pseudo or probabilistic pseudo labels), the selected original images are sampled multiple times (i.e., approximately $9370/M$ times) so that a fixed number of labelled training samples are provided to the CNN, where each sample is subjected to random data augmentation before being used by the CNN at each training epoch. In [51], pseudo-labels are determined by the k means clustering results, corresponding to 1-NN in TABLE 1. However, TABLE 1 shows that R-SVM consistently estimates better class decision boundaries, and so will be used to assign predictive pseudo-labels in this work. Although the GP classifier described in section 3.4 did not perform as well as the R-SVM, the prediction uncertainty may be useful for CNN training and so experiments are also performed using these outputs as probabilistic pseudo-labels.

Configuration D1 to D4 in TABLE 3 shows the performance metrics for each data selection strategy with the LGA pre-trained AlexNet CNN with last layer supervised training. D5 to D8 show the same comparison for ImageNet pre-trained ResNet18 CNN with all layer supervised training, where these base configurations where chosen since they performed best in our CNN architecture comparison (B4 and B6 in section 4.3.1). For both AlexNet and ResNet18, the combination of hierarchical k means and pseudo-labelling achieves the best performance for $M \leq 200$. Comparing the cases with pseudo-labelling (D3 and D7) to the cases without (D2 and D6) shows that pseudo-labelling consistently improves classification performance. D7, which applies hierarchical k means and pseudo-labelling to ResNet18, performs the best for $M \leq 200$ among all the configurations in TABLE 3. The accuracies achieved by D7 with $M = 20$, 40, 100 are similar to the metrics achieved for B1 to B4 with $M = 200$, 400, 1000, which have an order of magnitude more annotations. In particular, B6 and D7 use the same CNN architecture, showing that gains in learning efficiency can be attributed to the LGA-SS training method, resulting in a significant reduction in human effort to achieve a similar level of classification accuracy. Although the efficiency gains diminish as the number of human annotations available for training increases, the LGA-SS method never degrades the CNN's performance for an equal number of annotations. Another way to look at this is that the largest gains in learning efficiency are achieved when there is only a small amount of human effort available for annotation tasks, where D7 with 40 prioritised annotations reaches 85 % of the accuracy achieved by the best performing supervised CNN, B6, trained using 9370 human annotations, which represents just 0.4 % of the human effort. The data also shows that the combination of hierarchical k means and pseudo-labelling improves the repeatability between experiments under the same conditions, which is an important attribute for practical application of automated data interpretation.

Probabilistic pseudo-labelling outperforms pseudo-labelling only when $M = 1000$. This indicates that meaningful probabilistic expression of pseudo-labels can only be taken advantage of when a relatively large number of annotations are available. On the other hand D2, where pseudo-labelling is not applied, shows the best accuracy for $M = 1000$, and similarly D1 shows the best performance for $M = 400$ with D2 following it. This trend suggests that LGA pre-trained AlexNet is effective at describing the class boundaries when a sufficient number of annotated examples can be provided for fine-tuning. The equivalent training approach for D5 and D6 does not show this behaviour, indicating that this is a particular feature of using the LGA pre-trained network. The advantages of the proposed method with hierarchical k means for prioritised sample annotation and pseudo-labelling using R-SVM is significant for $M \leq 200$ for both CNN architectures (i.e., D3 and D7).

4.4 CNN and Conventional Classifier Comparison

Fig. 3 compares the performance metrics of several representative configurations in TABLE 1, 2 and 3. The result under configuration A14 are shown as this is the best performing conventional (i.e., non-CNN) classifier. For the CNN classifiers, configurations B4, B6, C7, C10, D3 and D7 are shown to demonstrate the effectiveness of the proposed pipeline compared to other data selection strategies (random selection and active learning) in Fig. 3a. Fig. 3b shows that the proposed LGA and H-k means based training data selection is also effective on the aerial datasets. The performance gains achieved here are is less than for the Seafloor dataset, where this is thought to be due to the more skewed class distribution in the Seafloor dataset (see Fig. 2), since the H-k means selection strategy is most effective when dealing with imbalanced classes.

Overall, the CNNs trained with proposed pipeline (D3 and D7) outperform the conventional classifier (A14) and the best performing CNN trained using active learning (C7 and C10), except for $M = 1000$. The outputs of the A14 form the inputs to train D3, where the same LGA is used for pre-training the AlexNet CNN. The improvement in performance shows that the CNN does not merely replicate the class boundaries found in the annotations and the pseudo-labels, but learns new boundaries that discriminate the classes more accurately. ResNet18 (D7) shows better performance than AlexNet (D3) when trained using the same outputs of A14, indicating an ability to more accurately model complex class boundaries. This was generally the case for all random selected training data and the proposed pipeline. Comparing $M = 1000$ and $M = 9370$, the conventional classifier's accuracy is not significantly improved even though almost 10 times the number of annotations are used for training. On the other hand, the CNNs achieves statistically significant increases from $M = 1000$ to $M = 9370$ in all cases. This supports the common understanding that deep learning CNNs are a better option than conventional classifiers when large training datasets are available, and that conventional classifiers are a reasonable option when only a small number of annotations are available for training.

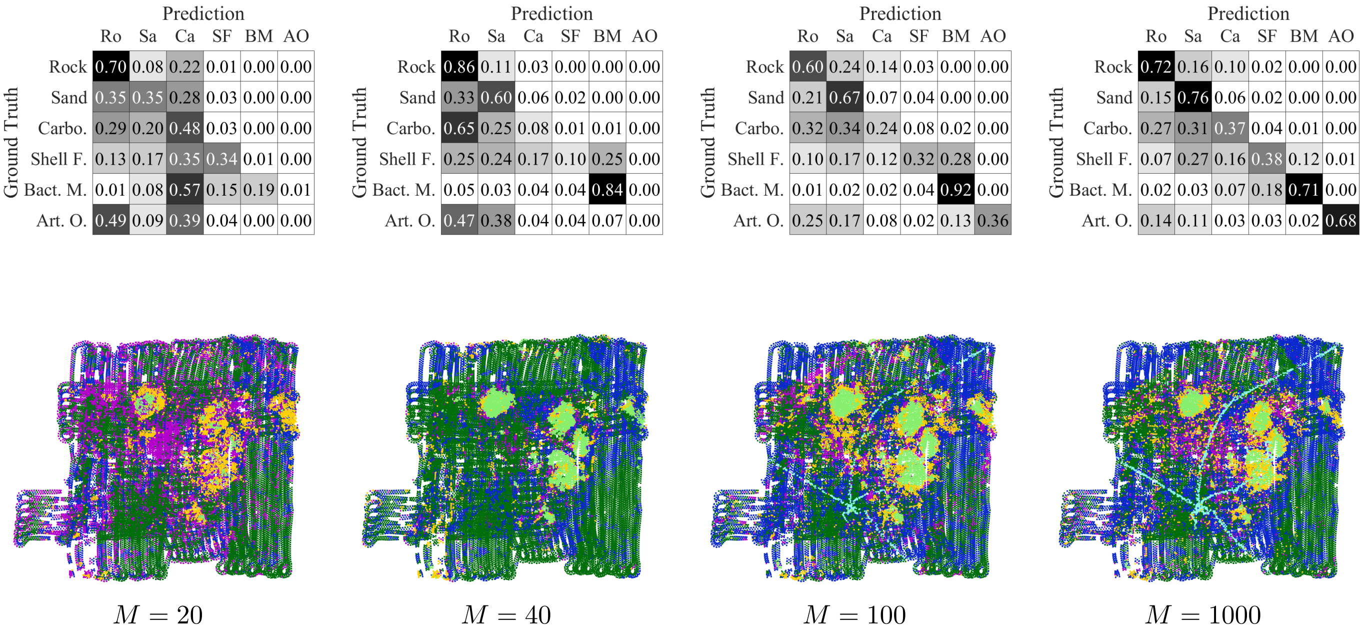
Active learning (C7 and C10) benefits from LGA based k means initialisation (C10), and shows better accuracy than standard training (B4 and B6) for small M, but the performance degrades when M is large due to overfitting as discussed previously. The proposed pipeline with prioritised annotation and pseudo-labelling significantly outperforms active learning for all M and both CNN architectures (D3, D7). Pseudo-labelling is more robust to overfitting than active learning since variability within the dataset is fully represented as all the available images are used for training.

Other factors that are important for practical application include the computational cost and the requirements for human input. Compared with CNNs, conventional classifiers require less time for training once LGA latent representations are generated and annotations have been made. In active learning, the three main steps i.e., training with annotated samples, inference for prioritising samples without annotations and annotating by humans, need to be repeated in sequence. This results in a large computational cost and also leads to inefficiencies as human annotators are forced to work around classifier retraining at each iteration. On the other hand, the time investment needed for the proposed pipeline is similar to conventional CNN training, since the unsupervised training and LGA based sample prioritisation do not require any human input, and the computation time for predicting pseudo-labels is negligible.

4.5 Per-class Performance

So far the macro-averaged F_1 score has been used as a metric to compare the overall performance of different classifiers. This is appropriate when we assume all classes in a dataset are of equal importance. However, there are applications where this is not the case, and in these scenarios it is more valuable to consider performance on a per-class basis. Figs. 4 and 5 compare the per-class confusion matrices for M values of 20, 40, 100 and 1000 for configurations B2 and D7. These represent the outputs of the best performing network, ResNet18, trained using standard transfer learning and the proposed LGA-SS pipeline, respectively. The values in each confusion matrix are normalised by the number of ground truth annotations so that the diagonal elements correspond to the recall value of each class. The confusion matrices corresponding to the trials with the closest F_1 score (macro-average) to the mean of ten repetitions (TABLE 3) are chosen for each value of M. The values of zeros for $M = 20$ and 40 in Fig. 4 suggest no images corresponding to 'Artificial Object' were selected in the random selection used for training and so predictions could not be made effectively for this class. On the other hand, Fig. 5 shows that all 6 classes in the dataset are predicted for all M, illustrating the advantage of using hierarchical k means based data selection to avoid minor classes from being overlooked even when the total number of annotated images is small.

Comparing the habitat maps generated using the classification results to the ground truth annotations (Fig. 2a) shows that the random data selection (Fig. 4) requires a larger number of training samples M to capture the different spatial distribution patterns of each class. Using the proposed LGA-SS training method (Fig. 5) results in more consistent per-class performance, providing a better approximation of the ground truth class distribution patterns even for small values of M. The consistent performance for different numbers of input training data is an important attribute for practical application since the annotation resource available for different datasets is likely to vary. These points favour the proposed method over random sampling approaches



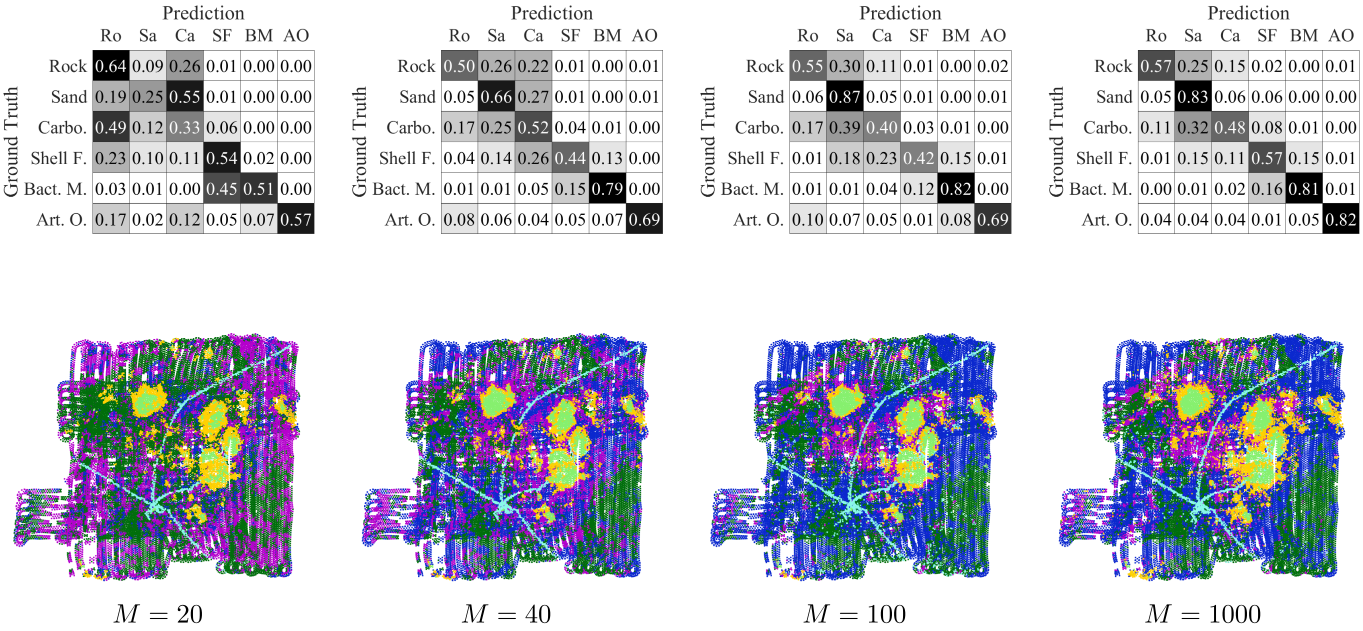
	Ro	Sa	Ca	SF	BM	AO
Rock	0.70	0.08	0.22	0.01	0.00	0.00
Sand	0.35	0.35	0.28	0.03	0.00	0.00
Carbo.	0.29	0.20	0.48	0.03	0.00	0.00
Shell F.	0.13	0.17	0.35	0.34	0.01	0.00
Bact. M.	0.01	0.08	0.57	0.15	0.19	0.01
Art. O.	0.49	0.09	0.39	0.04	0.00	0.00

	Ro	Sa	Ca	SF	BM	AO
Rock	0.86	0.11	0.03	0.00	0.00	0.00
Sand	0.33	0.60	0.06	0.02	0.00	0.00
Carbo.	0.65	0.25	0.08	0.01	0.01	0.00
Shell F.	0.25	0.24	0.17	0.10	0.25	0.00
Bact. M.	0.05	0.03	0.04	0.04	0.84	0.00
Art. O.	0.47	0.38	0.04	0.04	0.07	0.00

	Ro	Sa	Ca	SF	BM	AO
Rock	0.60	0.24	0.14	0.03	0.00	0.00
Sand	0.21	0.67	0.07	0.04	0.00	0.00
Carbo.	0.32	0.34	0.24	0.08	0.02	0.00
Shell F.	0.10	0.17	0.12	0.32	0.28	0.00
Bact. M.	0.01	0.02	0.02	0.04	0.92	0.00
Art. O.	0.25	0.17	0.08	0.02	0.13	0.36

	Ro	Sa	Ca	SF	BM	AO
Rock	0.72	0.16	0.10	0.02	0.00	0.00
Sand	0.15	0.76	0.06	0.02	0.00	0.00
Carbo.	0.27	0.31	0.37	0.04	0.01	0.00
Shell F.	0.07	0.27	0.16	0.38	0.12	0.01
Bact. M.	0.02	0.03	0.07	0.18	0.71	0.00
Art. O.	0.14	0.11	0.03	0.03	0.02	0.68

Fig. 4: Confusion matrices and habitat maps predicted by ResNet18 trained using the random data selection (configuration B6 in TABLE 3). This corresponds to conventional good practise, using a CNN pre-trained on the ImageNet annotation dataset and fine-tuning all the layers using randomly sampled annotated images with data augmentation. The results show that for a values of $M = 20$ the 'Artificial Object' and 'Bacterial Mat' class that contain the fewest samples are not efficiently learned, and even for $M = 40$, 'Artificial Object' is not recognised. The confusion matrix shows that even with $M = 1000$, there is still significant confusion when classifying 'Carbonate' and 'Shell Fragment'.



	Ro	Sa	Ca	SF	BM	AO
Rock	0.64	0.09	0.26	0.01	0.00	0.00
Sand	0.19	0.25	0.55	0.01	0.00	0.00
Carbo.	0.49	0.12	0.33	0.06	0.00	0.00
Shell F.	0.23	0.10	0.11	0.54	0.02	0.00
Bact. M.	0.03	0.01	0.00	0.45	0.51	0.00
Art. O.	0.17	0.02	0.12	0.05	0.07	0.57

	Ro	Sa	Ca	SF	BM	AO
Rock	0.50	0.26	0.22	0.01	0.00	0.01
Sand	0.05	0.66	0.27	0.01	0.00	0.01
Carbo.	0.17	0.25	0.52	0.04	0.01	0.00
Shell F.	0.04	0.14	0.26	0.44	0.13	0.00
Bact. M.	0.01	0.01	0.05	0.15	0.79	0.00
Art. O.	0.08	0.06	0.04	0.05	0.07	0.69

	Ro	Sa	Ca	SF	BM	AO
Rock	0.55	0.30	0.11	0.01	0.00	0.02
Sand	0.06	0.87	0.05	0.01	0.00	0.01
Carbo.	0.17	0.39	0.40	0.03	0.01	0.00
Shell F.	0.01	0.18	0.23	0.42	0.15	0.01
Bact. M.	0.01	0.01	0.04	0.12	0.82	0.00
Art. O.	0.10	0.07	0.05	0.01	0.08	0.69

	Ro	Sa	Ca	SF	BM	AO
Rock	0.57	0.25	0.15	0.02	0.00	0.01
Sand	0.05	0.83	0.06	0.06	0.00	0.00
Carbo.	0.11	0.32	0.48	0.08	0.01	0.00
Shell F.	0.01	0.15	0.11	0.57	0.15	0.01
Bact. M.	0.00	0.01	0.02	0.16	0.81	0.01
Art. O.	0.04	0.04	0.04	0.01	0.05	0.82

Fig. 5: Confusion matrices and habitat maps predicted by ResNet18 trained using the proposed LGA-SS method with hierarchical k means based data selection and pseudo-labelling (configuration D7 in TABLE 3). Compared to Fig. 4, the results show improved learning efficiency, especially for small values of M, where both the 'Artificial Object' and 'Bacterial Mat' classes are efficiently learned using just 20 human annotations, despite these being rare classes with a small number of data samples. The performance with $M = 100$ shows similar performance to when the same CNN architecture is trained using an order of magnitude more annotations from randomly selected data (i.e Fig. 4).

that are more sensitive to the number of available annotations, and require larger amounts of training data to achieve similar performance.

5 Conclusion

This paper proposes a novel semi-supervised learning pipeline to classify georeferenced imagery using deep learning CNNs. The main advantage of the proposed LGA-SS method is that it can interpret images according to class boundaries of interest for environmental monitoring more efficiently than the alternative methods tested in this work, requiring less human effort and achieving better accuracy. The method is designed for per-dataset training in order to achieve high performance with a realistic investment of human effort for practical application. Experiments on four georeferenced image datasets spanning aerial and seafloor environments show that the proposed georeference embedding and sample selection methods are effective across application domains, achieving the largest gains in efficiency are achieved on datasets that have highly skewed class distributions, which are a common feature in environmental monitoring applications. Other relevant advantages include reduced variability between multiple end-to-end training and classification runs under the same configurations, and more consistent performance with different sizes of input training data compared to traditional naive (i.e., random sampling) based transfer learning methods. These properties make the LGA-SS method suitable for use in domains where there is limited transfer of learning between datasets. Our results demonstrate that:

- The proposed LGA-SS can achieve classification accuracy equivalent to naively trained CNNs with an order of magnitude fewer human annotations (i.e., tens to hundreds, as opposed to thousands). The results demonstrate improvements in accuracy by a factor of 1.2 to 1.5 when a hundred or less annotations are used, where the largest gains in learning efficiency are achieved with small numbers of annotations. The method also reduces the statistical variability between independent trials under the same learning configurations to approximately 0.6 of that when random sampling is used. The proposed method reaches 85 % of the accuracy achieved by the best performing naively trained CNN (trained using 9370 human annotations) with just 40 prioritised annotations, which represents 0.4 % of the human effort.
- The strategy to select data for human annotation affects final classification performance. On the four datasets, introducing structure to prioritise annotation effort using hierarchical k means in the latent representation shows an average of 1.12 times improvement, and leveraging LGA instead of an autoencoder with the same CNN architecture achieved 1.23 times higher accuracy in terms of R-SVM classification results when the number of annotations is less than 100. A similar gain in performance is seen when the LGA based k means selection is used to initialise active learning, with a 1.25 factor improvement compared to equivalent randomly initialised active learning setups.
- The proposed method makes more efficient use of human effort than traditional active learning based techniques tested in this work, and is less prone to overfitting, achieving a factor 1.12 and 1.22 improvement in performance for AlexNet and ResNet18 respectively when compared to randomly initialised active learning across all values of M.
- CNN architectures are able to generalise class boundaries of interest to humans even when pseudo-labels are assigned to all data in a training set. The resulting CNN is able to improve the relative classification accuracy by an average of 6.4 % compared to the classification accuracy of the pseudo-labels themselves.
- The performance of conventional classifiers for pseudo-label generation is significantly improved using k means based selection compared to random selection when generating subsets of data for annotation. A factor of 1.30 improvement in classification accuracy is achieved for prioritised subsets with a hundred samples or less.
- Implementation of annotation effort prioritisation strategies relies on effective unsupervised clustering performance for seafloor images, where the use of georeferencing information by the LGA compared to an equivalent autoencoder that only uses information in images resulted in an improvement in classification accuracy by a factor of 1.4 to 8.9 (average 3.1) for the configurations tested in this work.

Acknowledgments

This work was carried out under the UK Natural Environment Research Council's Oceanids Biocam project NE/P020887/1 and Australian Research Council's Automated Benthic Understanding DP190103914 Discovery project. Data was collected during the Schmidt Ocean Institute's #Adaptive Robotics campaign. The University of Southampton IRIDIS High Performance Computing Facility was used in this work.

References

[1] I. Colomina and P. Molina, "Unmanned aerial systems for photogrammetry and remote sensing: A review," *ISPRS Journal of photogrammetry and remote sensing*, vol. 92, pp. 79–97, 2014.

[2] 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.

[3] 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.

[4] S. Kentsch, M. L. Lopez Caceres, D. Serrano, F. Roure, and Y. Diez, "Computer vision and deep learning techniques for the analysis of drone-acquired forest images, a transfer learning study," *Remote Sensing*, vol. 12, no. 8, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/8/1287

[5] D. Langenkämper, M. Zurowietz, T. Schoening, and T. W. Nattkemper, "BIIGLE 2.0 - browsing and annotating large marine image collections," *Frontiers in Marine Science*, vol. 4, p. 83, 2017. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmars.2017.00083

[6] F. Althaus, N. Hill, R. Ferrari, L. Edwards, R. Przeslawski, C. H. L. Schönberg, R. Stuart-Smith, N. Barrett, G. Edgar, J. Colquhoun, M. Tran, A. Jordan, T. Rees, and K. Gowlett-Holmes, "A standardised vocabulary for identifying benthic biota and substrata from underwater imagery: The catami classification scheme," *PLOS ONE*, vol. 10, no. 10, pp. 1–18, 10 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0141039

[7] J. N. Gomes-Pereira, V. Auger, K. Beisiegel, R. Benjamin, M. Bergmann, D. Bowden, P. Buhl-Mortensen, F. C. De Leo, G. Dionísio, J. M. Durden, L. Edwards, A. Friedman, J. Greinert, N. Jacobsen-Stout, S. Lerner, M. Leslie, T. W. Nattkemper, J. A. Sameoto, T. Schoening, R. Schouten, J. Seager, H. Singh, O. Soubigou, I. Tojeira, I. van den Beld, F. Dias, F. Tempera, and R. S. Santos, "Current and future trends in marine image annotation software," *Progress in Oceanography*, vol. 149, pp. 106 – 120, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0079661116301240

[8] D. Steinberg, A. Friedman, O. Pizarro, and S. B. Williams, "A bayesian nonparametric approach to clustering data from underwater robotic surveys," in *International Symposium on Robotics Research*, vol. 28, 2011, pp. 1–16.

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

[10] K. J. Morris, B. J. Bett, J. M. Durden, V. A. I. Huvenne, R. Milligan, D. O. B. Jones, S. McPhail, K. Robert, D. M. Bailey, and H. A. Ruhl, "A new method for ecological surveying of the abyss using autonomous underwater vehicle photography," *Limnology and Oceanography: Methods*, vol. 12, no. 11, pp. 795–809, 2014.

[11] J. Escartín, R. García, O. Delaunoy, J. Ferrer, N. Gracias, A. Elibol, X. Cufi, L. Neumann, D. J. Fornari, S. E. Humphris, and J. Renard, "Globally aligned photomosaic of the lucky strike hydrothermal vent field (mid-atlantic ridge, $37°18.5'$ N): Release of georeferenced data, mosaic construction, and viewing software," *Geochemistry, Geophysics, Geosystems*, vol. 9, no. 12, 2008.

[12] 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.

[13] U. Neettiyath, B. Thornton, M. Sangekar, Y. Nishida, K. Ishii, A. Bodenmann, T. Sato, T. Ura, and A. Asada, "Deep-sea robotic survey and data processing methods for regional-scale estimation of manganese crust distribution," *IEEE Journal of Oceanic Engineering*, pp. 1–13, 2020.

[14] J. Inglada, "Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features," *ISPRS journal of photogrammetry and remote sensing*, vol. 62, no. 3, pp. 236–248, 2007.

[15] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 7, pp. 971–987, 2002.

[16] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *2009 IEEE Conference on computer vision and pattern recognition*. IEEE, 2009, pp. 1794–1801.

[17] C. Chen, B. Zhang, H. Su, W. Li, and L. Wang, "Land-use scene classification using multi-scale completed local binary patterns," *Signal, image and video processing*, vol. 10, no. 4, pp. 745–752, 2016.

[18] S. Chen and Y. Tian, "Pyramid of spatial relatons for scene-level land use classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 1947–1957, 2014.

[19] M. Bewley, N. Nourani-Vatani, D. Rao, B. Douillard, O. Pizarro, and S. B. Williams, "Hierarchical classification in AUV imagery," in *Field and service robotics*. Springer, 2015, pp. 3–16.

[20] 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, 2017.

[21] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS journal of photogrammetry and remote sensing*, vol. 152, pp. 166–177, 2019.

[22] 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.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] D. Hou, Z. Miao, H. Xing, and H. Wu, "V-rsir: an open access web-based image annotation tool for remote sensing image retrieval," *IEEE Access*, vol. 7, pp. 83 852–83 862, 2019.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[26] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[27] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, "Spacenet: A remote sensing dataset and challenge series," *arXiv preprint arXiv:1807.01232*, 2018.

[28] D. Langenkämper, R. van Kevelaer, A. Purser, and T. W. Nattkemper, "Gear-induced concept drift in marine images and its effect on deep learning classification," *Frontiers in Marine Science*, vol. 7, p. 506, 2020.

[29] M. Zurowietz and T. W. Nattkemper, "Unsupervised knowledge transfer for object detection in marine environmental monitoring and exploration," *IEEE Access*, vol. 8, pp. 143 558–143 568, 2020.

[30] D. Steinberg, "An unsupervised approach to modelling visual data," *PhD Thesis, University of Sydney, Australia*, 2013.

[31] T. Vigneshl and K. Thyagharajan, "Local binary pattern texture feature for satellite imagery classification," in *2014 International Conference on Science Engineering and Management Research (ICSEMR)*. IEEE, 2014, pp. 1–6.

[32] A. Friedman, D. Steinberg, O. Pizarro, and S. B. Williams, "Active learning using a variational dirichlet process model for pre-clustering and classification of underwater stereo imagery," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2011, pp. 1533–1539.

[33] B. Settles, "Active learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2009.

[34] J. W. Kaeli and H. Singh, "Online data summaries for semantic mapping and anomaly detection with autonomous underwater vehicles," in *OCEANS 2015-Genova*. IEEE, 2015, pp. 1–7.

[35] J. Shields, O. Pizarro, and S. B. Williams, "Towards adaptive benthic habitat mapping," *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[36] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon, "Tile2vec: Unsupervised representation learning for spatially distributed data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3967–3974.

[37] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *International conference on artificial neural networks*. Springer, 2018, pp. 270–279.

[38] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[39] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 6, pp. 3660–3671, 2016.

[40] M. Zurowietz, D. Langenkämper, B. Hosking, H. A. Ruhl, and T. W. Nattkemper, "MAIA—A machine learning assisted image annotation method for environmental monitoring and exploration," *PloS one*, vol. 13, no. 11, 2018.

[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[42] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[43] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[44] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using imagenet pretrained net-

works,” *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 1, pp. 105–109, 2015.

[45] A. Lapedriza, H. Pirsiavash, Z. Bylinskii, and A. Torralba, “Are all training examples equally valuable?” *arXiv preprint arXiv:1311.6510*, 2013.

[46] S. Paul, J. H. Bappy, and A. K. Roy-Chowdhury, “Efficient selection of informative and diverse training samples with applications in scene classification,” in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 494–498.

[47] Z. Li, B. Ko, and H.-J. Choi, “Naive semi-supervised deep learning using pseudo-label,” *Peer-to-Peer Networking and Applications*, vol. 12, no. 5, pp. 1358–1368, 2019.

[48] D. Dai, M. Prasad, C. Leistner, and L. Van Gool, “Ensemble partitioning for unsupervised image categorization,” in *Proceedings of the 12th European conference on Computer Vision-Volume Part III*, 2012, pp. 483–496.

[49] M. Wigness, B. A. Draper, and J. R. Beveridge, “Efficient label collection for unlabeled image datasets,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 4594–4602.

[50] Y. Tian, W. Liu, R. Xiao, F. Wen, and X. Tang, “A face annotation framework with partial clustering and interactive labeling,” in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[51] D.-H. Lee, “Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks,” in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013.

[52] H. Wu and S. Prasad, “Semi-supervised deep learning using pseudo labels for hyperspectral image classification,” *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1259–1270, 2017.

[53] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, “Finding a” kneedle” in a haystack: Detecting knee points in system behavior,” in *2011 31st international conference on distributed computing systems workshops*. IEEE, 2011, pp. 166–171.

[54] D. Nister and H. Stewenius, “Scalable recognition with a vocabulary tree,” in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, vol. 2. Ieee, 2006, pp. 2161–2168.

[55] H. S. Gowda, M. Suhil, D. Guru, and L. N. Raju, “Semi-supervised text categorization using recursive k-means clustering,” in *International Conference on Recent Trends in Image Processing and Pattern Recognition*. Springer, 2016, pp. 217–227.

[56] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

[57] G. Cheng, P. Zhou, and J. Han, “Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7405–7415, 2016.

[58] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[59] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, “Label propagation for deep semi-supervised learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5070–5079.

[60] E. Arazo, D. Ortego, P. Albert, N. E. O’Connor, and K. McGuinness, “Pseudo-labeling and confirmation bias in deep semi-supervised learning,” in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[61] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006, vol. 2, no. 3.

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

[63] Z. Zhou, J. Shin, L. Zhang, S. Gurudu, M. Gotway, and J. Liang, “Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7340–7351.

[64] Y. Gal, R. Islam, and Z. Ghahramani, “Deep bayesian active learning with image data,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 1183–1192.

Takaki Yamada received a BEng and MSc from the University of Tokyo, Japan in 2009 and 2011, respectively, and a PhD from the University of Southampton, United Kingdom in 2021. He is a member of Centre for In Situ and Remote Intelligent Sensing and is interested in sensing and perception for autonomous underwater vehicles.

Miquel Massot-Campos received a MEng from BarcelonaTech, Spain in 2011, MSc and PhD from the University of the Balearic Islands, Spain in 2013 and 2019, respectively. He is currently with the University of Southampton, United Kingdom. He is a member of Centre for In Situ and Remote Intelligent Sensing and is interested in the scalability of autonomous underwater vehicle’s missions.

Adam Prügel-Bennett obtained a BSc in Physics at Southampton University in 1984 and a PhD in theoretical Physics at Edinburgh University in 1989. He worked in research jobs in Oxford, Paris, Manchester, Copenhagen and Dresden before finally returning to Southampton. He is currently Professor of Electronics and Computer Science where his main research interests are in machine learning and particularly deep learning.

Stefan B. Williams (S’99–A’01–M’02) received a BSc in systems engineering design from the University of Waterloo in 1997 and a PhD in field robotics from the University of Sydney in 2002. He is Professor of Marine Robotics at the University of Sydney’s Australian Centre for Field Robotics, and heads Australia’s Integrated Marine Observing System AUV Facility. His research interests include simultaneous localization and mapping in underwater environments, autonomous navigation and data interpretation.

Oscar Pizarro (S’93–M’04) received a BSc in electronic engineering from the Universidad de Concepcion in 1997, a dual MSc in ocean engineering and electrical engineering and computer science and PhD in oceanographic engineering from the MIT-WHOI Joint Program,in 2003 and 2004, respectively. He joined the University of Sydney’s Australian Centre for Field Robotics in 2005, where he is Principle Research Fellow. His research interests include scalable approaches to seafloor imaging and habitat characterisation.

Blair Thornton (M'07) obtained a BEng in Naval Architecture and PhD in Underwater Robotics from Southampton University in 2002 and 2006, respectively. In 2003 he joined the Underwater Robotics and Application Lab., Institute of Industrial Science, UTokyo, before rejoining Southampton in 2016 where he is Professor of Marine Autonomy. He has spent 450+ days at sea deploying robotic systems and is dedicated to generating data and insight in marine science through improved sensing and autonomy.