T. Yamada, M. Massot-Campos, A. Prugel-Bennett, S. B. Williams, O. Pizarro and B. Thornton, "Leveraging Metadata in Representation Learning with Georeferenced Seafloor Imagery,"
in IEEE Robotics and Automation Letters, doi: 10.1109/LRA.2021.3101881. https://ieeexplore.ieee.org/document/9507317

Leveraging Metadata in Representation Learning with Georeferenced Seafloor Imagery

Takaki Yamada[1], Miquel Massot-Campos[1], Adam Prügel-Bennett[1],
Stefan B. Williams[2], Oscar Pizarro[2] and Blair Thornton[1,3]

***Abstract*—Camera equipped Autonomous Underwater Vehicles (AUVs) are now routinely used in seafloor surveys. Obtaining effective representations from the images they collect can enable perception-aware robotic exploration such as information-gain-guided path planning and target-driven visual navigation. This paper develops a novel self-supervised representation learning method for seafloor images collected by AUVs. The method allows deep-learning convolutional autoencoders to leverage multiple sources of metadata to regularise their learning, prioritising features observed in images that can be correlated with patterns in their metadata. The impact of the proposed regularisation is examined on a dataset consisting of more than 30k colour seafloor images gathered by an AUV off the coast of Tasmania. The metadata used to regularise learning in this dataset consists of the horizontal location and depth of the observed seafloor. The results show that including metadata in self-supervised representation learning can increase image classification accuracy by up to 15% and never degrades learning performance. We show how effective representation learning can be applied to achieve class balanced representative image identification for summarised understanding of imbalanced class distributions in an unsupervised way.**

***Index Terms*—Marine Robotics, Representation Learning, Visual Learning, Computer Vision, Metadata**

I. Introduction

IMAGES gathered by camera equipped Autonomous Underwater Vehicles (AUVs) are now used in a wide range of seafloor survey applications. The captured images are used to characterise seafloor scenes where both manual and automatic methods are used for interpretation. Typical AUV missions will gather tens to hundreds of thousands of images during a single dive, where the high resolution and large redundancy of imagery pose a significant challenge for automated interpretation. In order to take full advantage of images for robotic applications, it is necessary to obtain compact representations that efficiently preserve the most valuable information in the original images. Once these are generated, algorithmic analysis can be performed with low latency using relatively limited computational resources. Examples of robotic applications that could be facilitated using image representation include information-gain-aware path planning for representative surveys [1], [2], target-aware seek and sample missions [3], curiosity-driven exploratory surveys [4], and real-time habitat inference [5], [6].

The aim of this paper is to develop a self-supervised learning method that can use metadata gathered with seafloor imagery to efficiently generate low-dimensional latent representation spaces that are useful for image interpretation. Effective low-dimensional representations form the basis of semantic interpretation, where classification, clustering, and content based retrieval are examples of tasks that can be readily applied to achieve efficient understanding of underwater scenes. Fig. 1 illustrates a typical AUV survey scenario. Data is often gathered over multiple dives, where ships transport AUVs between sites between their dives. These locations can be separated by distances far larger than that traversable by an individual AUV. Observations typically cover spatial extents several orders of magnitude larger than the footprint of a single image frame, which typically have edge lengths of a few metres, and span a wide range of seafloor depths. Habitats and substrates vary over spatial scales larger than each image and exhibit patterns with depth, especially in shallow water due to the influence of sunlight. Therefore, images taken close to each other, or separated but with similar depths, are more likely to share visual characteristics than would otherwise be the case. To leverage this information, we implement our metadata regularised learning method using horizontal location and depth information. A key advantage of this approach is that regularisation can be applied to data gathered in remote locations during different dives based on depth information. The novel contributions of this work are:

- Development of a regularisation method that leverages metadata when training deep-learning Convolutional Neural Network (CNN) based autoencoders for efficient latent representation of seafloor imagery.
- Implementation of the proposed method where an AlexNet [7] based autoencoder is trained on seafloor images, regularised by a loss function that introduces domain relevant assumptions on georeference (horizontal location and depth) information.
- Performance validation on a dataset gathered off the coast of Tasmania that consists of more than 30k seafloor images and 2.2k human annotations, taken over six AUV dives between depths of 28 and 96 m, and demonstration of an application to unsupervised representative image selection to generate semantic summaries of the observations.

Manuscript received: February, 23, 2021; Revised June, 3, 2021; Accepted July, 8, 2021. This paper was recommended for publication by Editor P. Pounds upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the UK Natural Environment Research Council's Oceanids Biocam project NE/P020887/1, Australian Research Council's Automated Benthic Understanding Discovery project DP190103914 and EU Horizon 2020 TechOceanS Grant ID 101000858. [1]Takaki Yamada, Miquel Massot-Campos, Adam Prügel-Bennett and Blair Thornton are with Centre for In Situ and Remote Intelligent Sensing, Faculty of Engineering and Physical Science, University of Southampton, U.K. T.Yamada@soton.ac.uk [2]Stefan B. Williams and Oscar Pizarro with Australian Centre for Field Robotics, The University of Sydney, Australia. [3]Blair Thornton is also with Institute of Industrial Science, The University of Tokyo, Japan. Digital Object Identifier (DOI): see top of this page.

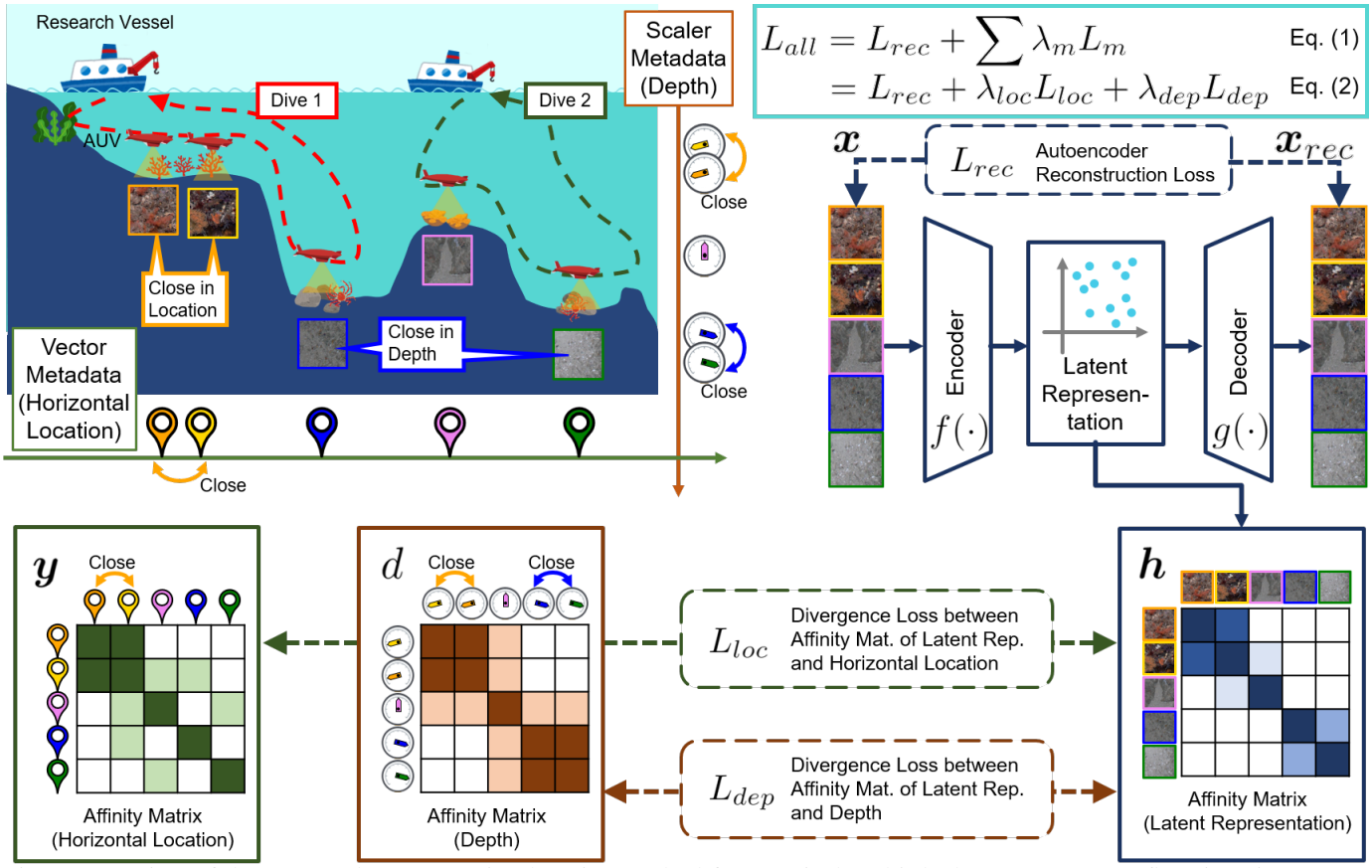


Fig. 1: Overview of the proposed representation learning method for a typical multi-deployment AUV seafloor imaging survey. The method regularises the representation space of an autoencoder by embedding images taken at similar horizontal locations, or separated but with similar depths, in nearby regions of the latent representation space. This is achieved by minimising the Kullback–Leibler divergence between the affinity matrix of the image latent representation with horizontal location metadata using the loss function L_{loc}, and with depth metadata using the loss function L_{dep}. These are optimised together with the autoencoder reconstruction loss L_{rec} to regularise the latent representation space according to these metadata.

II. Background

Seafloor habitats and substrates can be identified by unique patterns in their appearance, and various machine learning techniques have been applied to automate image interpretation. These can be broadly split into studies that use feature engineering, where descriptors are manually chosen or tuned by human experts, and representation learning, where descriptors are directly learnt from the data. In both cases, the reduced dimensions of the representations allow for more effective identification of patterns in the data.

Manually engineered feature descriptors have been investigated by several groups for efficient image representation [5], [6], [8], [9], [10], [11]. In [9], [11], colour-based descriptors were designed based on prior knowledge of targets that are of specific scientific interest. Generic descriptors such as Local Binary Patterns (LBP) [12] and Sparse Coding Spatial Pyramid Matching (ScSPM) [13] have also been applied to identify spatially invariant patterns that appear at different scales within images of the seafloor [5], [6]. In [10], accumulated histograms of oriented gradients from image keypoints were used to describe seafloor images for the purpose of clustering and anomaly detection. However, these types of descriptor often require manual tuning of parameters, or feature engineering, to effectively describe the datasets they are applied to.

CNNs avoid the need for feature engineering by learning the latent representations needed to best describe the datasets they are applied to. This is typically achieved by using labels generated by humans experts to supervise CNN training, which simultaneously optimises the latent representations and class boundaries to best describe the patterns of interest in a training dataset. In [14], the ResNet [15] deep-learning CNN was trained to distinguish between nine different classes of coral in a seafloor image dataset, demonstrating higher classification resolution than traditional feature engineering based methods. However, the need for large volumes of annotated images to supervise CNN learning limits wide scale use in marine applications since generic training datasets do not exist.

An alternative approach to train CNNs is to use self-supervised learning techniques. In domains where continuity exists between the samples in a dataset, this continuity can be used to help regularise representation learning without the need for direct human supervision. In natural language processing, Word2vec [16] and GloVe [17] leverage the assumption that words found in similar contexts are likely to have similar meanings. This continuity was used to generate continuous rep-

resentations of different word nuances. For image processing applications, Tile2Vec [18] extends the assumption to spatially distributed data, demonstrating its effectiveness for satellite image interpretation. In [19], we developed a Location Guided Autoencoder (LGA) that regularises autoencoder learning using horizontal geo-location information for efficient clustering and content-based retrieval of seafloor imagery. In [20], a similar assumption is introduced for CNN-based coral detection from seafloor imagery, where object tracking results in sequential frames are used for semi-supervised training.

The method developed in this paper advances the state of the art for seafloor image interpretation. First, the use of metadata is advanced by incorporating depth information in parallel to horizontal location information for learning regularisation. This is significant as even though our previous LGA method used horizontal geo-location to regularise learning [19], this method cannot regularise learning across large horizontal spatial discontinuities in observation, as is often seen between different AUV dives. Additionally, although the effectiveness of the method when applied to dense survey trajectories that fully cover a 2D region of the seafloor has been demonstrated, it is not clear how effective the method is for sparse trajectories. Dense survey trajectories guarantee that each image has many other images in its neighbourhood that it can be paired with to regularise learning. However, sparse trajectories are often used when surveying larger regions of the seafloor, and under these conditions only a small number of neighbourhood image pairs are available, which potentially limits the effectiveness of the horizontal location based regularisation. In contrast, depth information can provide a large number pairings for sparse surveys and regularise learning across different dives if depth related distribution patterns exist. Next, we apply contrastive learning methods to improve the regularisation effect of the metadata, where this is the first time contrastive learning has been applied to seafloor imagery. We demonstrate these concepts on a dataset that consists of seafloor imagery gathered over 6 AUV dives, with observations that are sparsely distributed over a 1.6 $\times$ 1.7 km region spanning a depth range of 28 to 96 m.

III. Metadata regularised autoencoder

A. General Concept

An autoencoder consists of an encoder $f(\cdot)$ and a decoder $g(\cdot)$. The encoder $f(\cdot)$ maps a set of seafloor images $\boldsymbol{x}$ to a lower-dimensional tensors $\boldsymbol{h}$ ($\boldsymbol{h}$=$f(\boldsymbol{x})$), and the decoder $g(\cdot)$ reconstructs the images $\boldsymbol{x}_{rec}$ from $\boldsymbol{h}$ ($\boldsymbol{x}_{rec}$=$g(\boldsymbol{h})$) so that the reconstructed images become as similar as possible to the original images. The optimisation minimises the mean squared error loss function L_{rec}=$\frac{1}{n}\sum^{n}\|\boldsymbol{x}_{rec}-\boldsymbol{x}\|^2$, where n is the total number of images.Here $\boldsymbol{h}$ can be regarded as reasonable latent representations of $\boldsymbol{x}$ since they preserve key information in $\boldsymbol{x}$ so that $\boldsymbol{x}_{rec}$ can be reconstructed properly. The key advantage of an autoencoder is that the encoder $f(\cdot)$ can be trained in a self-supervised manner, where only the input images are used and no additional human annotations are needed. To incorporate metadata into autoencoder training, we minimise a loss function of the following form:

$$L_{all} = L_{rec} + \sum \lambda_m L_m. \tag{1}$$

m is an index for each type of metadata used for learning regularisation, where these can be any number of continuous scalar or vector quantities that can be associated with the images. L_m is the loss function that regularise autoencoder training based on the values of metadata m. λ_m is a hyperparameter used to balance the loss contributions.

B. Implementation for Georeferenced Imagery

AUVs typically measure their horizontal location, depth and altitude for basic navigational functionality. This metadata can be leveraged to regularise autoencoder training by formulating eq. (1) as follows:

$$L_{all} = L_{rec} + \lambda_{loc} L_{loc} + \lambda_{dep} L_{dep}, \tag{2}$$

where L_{loc} is the loss function for the horizontal location based regularisation, L_{dep} is for the depth based regularisation, λ_{loc} and λ_{dep} are hyperparameters to balance their relative contributions. In our implementation, AlexNet [7] and its inverted architecture are used as the encoder and decoder, respectively, where any type of neural network can be used to construct autoencoder in a similar way. Our previous LGA method [19] can be regarded as a specific case of eq. (1), where only L_{loc} and λ_{loc} are used.

1) Vector Based Regularisation: The horizontal location loss L_{loc} is introduced to regularise autoencoder training following the assumption that two images captured within a close distance tend to look more similar than two that are far away. In representation learning, if two images look similar and potentially belong to the same class, their latent representations should be located within a close distance in the latent space. In order to make the distribution of latent representations $\boldsymbol{h}$ reflect the 2D horizontal location vector $\boldsymbol{y}$ where the images $\boldsymbol{x}$ are taken, we introduce a loss function that has a similar structure to the loss function of t-SNE [21]. In t-SNE, original high-dimensional data $\boldsymbol{x}_{org}$ is embedded into a 2D or 3D space $\boldsymbol{x}_{emb}$ so that data with close relative distances in the original space are represented with high probability in the embedded space. In our problem, $\boldsymbol{y}$, which controls the distribution in the latent space corresponds to $\boldsymbol{x}_{org}$, and the latent representations $\boldsymbol{h}$ corresponds to $\boldsymbol{x}_{emb}$. Following the t-SNE loss function, the probability p_{ij}, which is proportional to the distance between $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$, is defined for i≠j as:

$$p_{j|i} = \frac{\exp\left(-\left\|\boldsymbol{y}_i - \boldsymbol{y}_j\right\|^2 / 2\sigma_{loc}^2\right)}{\sum_{k\neq i}\exp\left(-\left\|\boldsymbol{y}_i - \boldsymbol{y}_k\right\|^2 / 2\sigma_{loc}^2\right)}, \tag{3}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}, \tag{4}$$

where p_{ij}=0 when i=j, σ_{loc} is a normalising factor for $\boldsymbol{y}$.The probability q_{ij} is derived from $\boldsymbol{h}$, and is optimised based

on p_{ij}. For q_{ij} when $i{\neq}j$, it is defined by the Student's t-distribution as:

$$q_{ij} = \frac{\left(1 + \|\boldsymbol{h}_i - \boldsymbol{h}_j\|^2\right)^{-1}}{\sum_{k\neq l}\left(1 + \|\boldsymbol{h}_k - \boldsymbol{h}_l\|^2\right)^{-1}}, \tag{5}$$

where $q_{ij}{=}0$ for $i{=}j$.

By defining the affinity matrices P and Q with p_{ij} and q_{ij} as their elements, the horizontal location loss L_{loc} is defined as the Kullback–Leibler (KL) divergence of P from Q:

$$L_{loc} = \mathrm{KL}(P\|Q) = \sum_{i\neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{6}$$

Minimising L_{loc} forces Q to approach P, which embeds the correlation between the image representations and the horizontal location metadata into the latent representation. Eq. (3) - (6) are implemented in a similar way to the loss function of t-SNE, where $\boldsymbol{y}$ is used to derive the target probabilistic distribution instead of $\boldsymbol{x}_{org}$, and $\boldsymbol{h}$ is optimised instead of $\boldsymbol{x}_{emb}$.

2) Scalar Based Regularisation: The depth loss L_{dep} can be formulated in a similar way to the horizontal location loss L_{loc} defined earlier. Given that the seafloor depth where an image $\boldsymbol{x}_i$ is captured is a scalar value d_i, the probability r_{ij} is defined to be proportional to the difference between d_i and d_j where the observations are made:

$$r_{j|i} = \frac{\exp\left(-(d_i - d_j)^2/2\sigma_{dep}^2\right)}{\sum_{k\neq i}\exp\left(-(d_i - d_k)^2/2\sigma_{dep}^2\right)}, \tag{7}$$

$$r_{ij} = \frac{r_{j|i} + r_{i|j}}{2n}, \tag{8}$$

where $r_{ij}{=}0$ when $i{=}j$. σ_{dep} is a normalising factor. The depth loss is formulated as the KL divergence $L_{dep}{=}\mathrm{KL}(R\|Q)$, where R is the affinity matrix with elements r_{ij}.

3) Generalised Regularisation Behaviour: An important characteristic of the proposed method is that multiple regularisation methods can be applied without risk of significantly degrading performance. As elements in the affinity matrices (e.g. P and R), become further apart in the metadata space (i.e. the distance between $\boldsymbol{y_i}$ and $\boldsymbol{y_j}$ or d_i and d_j increases), the values of p_{ij} or r_{ij} become less sensitive to the separating distance. Furthermore, since the t-distribution used in this work is heavy-tailed compared to Gaussian distributions, it avoids the "crowding problem" that can occur when high-dimensional data is embedded into a lower-dimensional space when generating a t-SNE. This is preferable to avoid over-regularisation by the metadata, since pairs of images that are far apart are less strongly constrained by the regularisation and can be flexibly embedded in the latent space. Since the loss function only loosely constrains autoencoder training based on probabilistic distributions, it is inherently robust to over-fitting metadata. Furthermore, if the training process finds a particular type of metadata to have little correlation with the appearance of images, it gets automatically ignored, and where a particular type of metadata is found to have a strong correlation with image appearance it gets increasingly prioritised. This self-regulating characteristic is important in situations where many different types of metadata can be applied as the method can automatically prioritise the most significant metadata and mitigate any negative impact without additional human input or tuning.

Here P and R are formulated for $\boldsymbol{y}$ and d, which are 2D (latitude-longitude) vectors and scalar values, respectively. However, the proposed loss function can be implemented for any combination of vector or scalar metadata where the similarity between its values can be defined. This is important as it allows the proposed concept of metadata based regularisation to be readily applied to different types of samples (e.g. seafloor imagery, water column microscopy) and available metadata (e.g. acoustic back-scatter intensity, terrain rugosity, seawater temperature, pH) depending on the configuration of the data gathering platforms.

C. Mini-batch Sampling and Contrastive Learning

Ideally, L_{loc} and L_{dep} would be derived from all the samples in a dataset (i.e. n samples) so that they are globally optimised. However, due to computational limitations, mini-batch gradient descent is used for the simultaneous optimisation of L_{rec}, L_{loc} and L_{dep}. The number of images considered at each iteration is limited to a mini-batch size n^*, where a strategy is needed to avoid over-fitting to local minima in L_{loc} or L_{dep} when sampling n^* images. Since the regularisation effect is diminished as the number of horizontal location and depth neighbourhood pairs reduces, we introduce a sampling method that balances the number of images that are nearby and far away in each metadata space. First, two images are randomly selected at each iteration. Next $n^*/3$ images are selected from the first image's horizontal location neighbourhood, and another $n^*/3$ images are selected from the second image's depth neighbourhood, and the final $n^*/3$ images are randomly selected from the whole dataset in accordance with the principles of triplet loss contrastive learning demonstrated in [22]. This ensures a large variety is maintained in the values of the affinity matrices P and R, which prevents over-regularisation and allows similar images and dissimilar images to be evenly considered at each batch iteration.

IV. Experiment

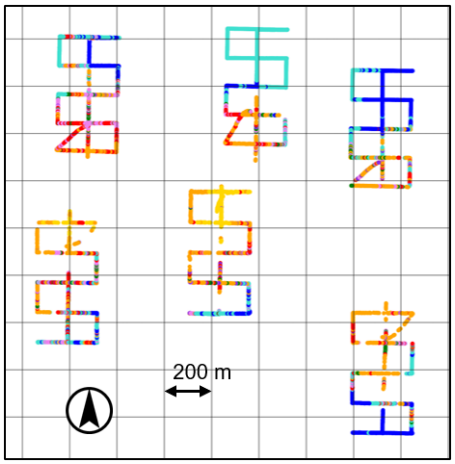
A. Dataset

The proposed method is applied to seafloor imagery obtained off the east coast of Tasmania [23]. Analysis is performed on 32,097 seafloor images taken by the Australian Centre for Field Robotic's Sirius AUV from an altitude of ~2 m. The data analysed here was gathered over six dives sparsely covering a 1.6×1.7 km region of the seafloor between 28 and 96 m depth. Details of the survey are given in TABLE I.

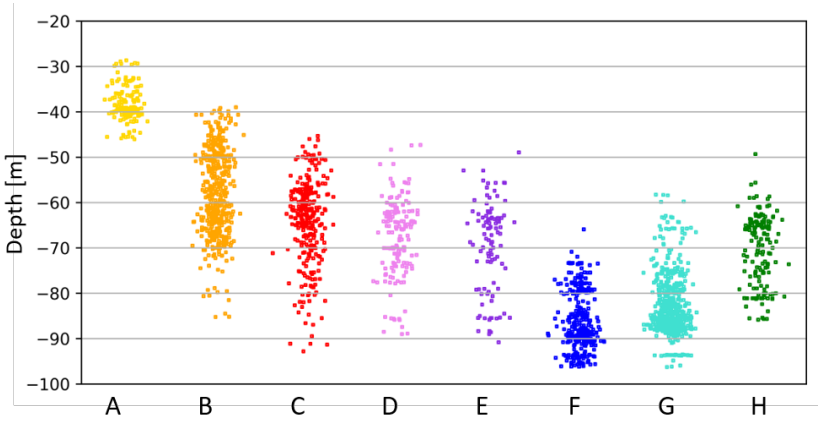
The images show various habitat and substrate distributions, including kelp (A), a registered essential ocean variable, and rocky reefs (B) - (E), which can form habitats for various conservation targets such as coral and sponges [24]. The original resolution of the images is 1,360 × 1,024. Each image in the

Label	A	B	C	D	E	F	G	H
Count	152 (6.8%)	449 (20.2%)	285 (12.8 %)	157 (7.1 %)	101 (4.6 %)	315 (14.2%)	605 (27.2%)	157 (7.1%)

(a) Class example images together with the number expert human annotations in each class.



(b) Horizontal distribution.



(c) Class vs depth.

Fig. 2: Overview of the Tasmania dataset. The number of expert human annotations in each image class are shown together with example images in a), with class labels: A - Kelp, B - High Relief Reef, C - Low Relief Reef, D - Patch Reef, E - Reef & Sand, F - Screw Shell Rubble, G - Screw Shell Rubble & Sand, and H - Coarse Sand. The horizontal spatial distribution of the human annotated classes are shown in b) and the depth distribution of each class is shown in c), where the same colour scheme has been used throughout the figure. The horizontal location loss L_{loc} regularises learning based on the horizontal distribution of the images, and the proposed depth loss L_{dep} regularises learning based on their depth distribution.

dataset is re-scaled to a resolution of 2 mm/pixel based on the camera field of view (FoV) and imaging altitude. The centre 227 × 227 of each image is used in the analysis. The average distance between adjacent images is approximately 0.5 m and so the overlap between cropped images is negligible. 2,221 randomly selected images are annotated by human experts into 8 classes, as shown in Fig. 2a, where these are used to validate the performance of the proposed method. Fig. 2b shows the horizontal distribution of each ground truth class in the dataset. The figure shows that the classes form continuous spatial patterns along the sparse survey trajectories. Fig. 2c shows the depth distribution of annotated images in each class together the class labels. The figure shows that Kelp (A) is found at shallow depth ranges where energy from the sun can reach. High Relief Coral (B) and Low Relief Reef (C) start to appear at the depth of 40 m and 45 m, respectively. Other classes (D) - (H) also exhibit unique depth distributions, though there is considerable overlap beyond 50 m depth.

Horizontal location and depth estimates for each image are generated based on the Simultaneous Localisation and Mapping (SLAM) pipeline described in [25]. Georeference errors smaller than σ_{loc} in eq. (3) or σ_{dep} in eq. (7) do not affect the optimisation. Where SLAM or other global localisation methods such as ultra-short baseline or long-baseline acoustic positioning are not used, horizontal position errors accumulate at a rate of approximately 1 % distance travelled using typical AUV navigational sensor suites [26]. In practical terms, this means that the position uncertainty between sequentially taken images will be negligible. For images taken nearby but with a longer period of separation, the position uncertainty should be estimated using established methods (e.g. an extended Kalman filter) and where the uncertainty exceeds σ_{loc}, the pair should be rejected. Error accumulation does not occur when using commercial grade pressure and altitude sensors to determine seafloor image depth and so depth regularisation can be performed as long as these sensors are properly calibrated.

B. Autoencoder Training

To investigate the effectiveness of the proposed regularisation, the autoencoder is trained (i) without regularisation,

TABLE I: Tasmania Dataset Description

Vehicle	Sirius AUV
Camera Resolution	1,360 × 1,024
Camera FoV	42 × 34 deg
Year	2008
Location	East Coast of Tasmania, Australia
Coordinate	43.08°S, 147.97°E
Extend	1.6 × 1.7 km
Depth	28 - 96 m
Altitude	1.0 - 3.0 m
No. of Images	32,097
No. of Annotations	2,221
No. of Classes	8 (See Fig.2a)
No. of Dives	6

(ii) with L_{loc}, (iii) with L_{dep}, (iv) with both L_{loc} and L_{dep} on all 32,097 images in the dataset. AlexNet [7] with batch normalisation is used as the encoder architecture, and its inverse is used as the decoder where the number of dimensions of the encoder output (equal to the number of dimensions of the decoder input) is set to 16 in accordance with our previous work [19]. The autoencoder weights are initialised with the values of AlexNet pre-trained on ImageNet. A mini-batch size of $n^* = 256$ is applied and random rotation, shifting, flipping and colour distortions are applied for data augmentation. In the experiments where either L_{loc} or L_{dep} are applied, $n^*/2$ images are selected from the metadata space neighbourhoods of each randomly selected sample, and remaining $n^*/2$ images are selected randomly from the entire dataset. σ_{loc} in eq. (3) is set to 10.0 m, and σ_{dep} in eq. (7) is set to 1.0 m since image appearance is expected to show some degree of correlation with horizontal location and depth within these ranges. Preliminary experiments indicated that the method is not highly sensitive to these parameters, where σ_{loc} values ranging from 3.0 to 20 m only had a marginal impact on performance. This is favourable for practical application since extensive parameter tuning via trial and error is not necessary. Both λ_{loc} and λ_{dep} in eq. (2) are set to 1×10^5, and a learning rate of lr=1×10^{-5} is used for the Adam optimiser. These hyperparameters are experimentally determined so that all loss terms that are applied decrease during training. This is also favourable in practical terms since decrease of the loss function is a necessary condition for successful training, where most workflows already confirm this happens before proceeding with further analysis. The number of epochs is set to 100 and each experiment configuration is executed three times.

C. Evaluation Metrics

The representation learning performance is evaluated based on the classification accuracy achieved using the acquired representations. The classifiers used to assess performance consist of a k-Nearest Neighbour with k=1 (1-NN), a Gaussian Process classifier (GP), Random Forest (RF), Support Vector Machine with Linear kernel (L-SVM) and with Radial basis function kernel (R-SVM). A 10-fold cross validation is performed to examine each autoencoder, where three autoencoders are used in each training configuration. To reduce the effect of class imbalance, the cost functions of RF, L-SVM and R-SVM are balanced considering the class counts. The F_1 score (macro average) is used for performance evaluation, where we consider all class to be of equal importance. Though this experiment considers classification to evaluate accuracy, the higher score indicates that the obtained representations are effective at describing the images, and so form a favourable basis for other applications such as clustering, contents retrieval, and use in observation-aware path planning methods.

TABLE II: F_1 Macro Average Scores for Each Regularisation Configuration and Classifier.

Regularisation	Classifier				
	1-NN	RF	GP	L-SVM	R-SVM
(i)	46.0±3.2	50.1±2.5	48.3±2.8	51.4±3.4	50.3±3.4
(ii)	49.4±3.8	**53.6**±3.3	53.3±3.7	56.3±3.6	56.6±4.0
(iii)	48.2±3.7	51.1±3.2	52.4±3.5	56.6±4.1	54.7±3.6
(iv)	**49.7**±2.8	53.4±3.7	**54.3**±2.7	**57.5**±3.8	**57.9**±4.1

The convolutional autoencoder is trained (i) without regularisation, (ii) with L_{loc}, (iii) with L_{dep}, (iv) with L_{loc} and L_{dep}. Five different classifiers are trained on the autoencoder embedded representations (1-Nearest Neighbour, Random Forest, Gaussian Process classifier, Linear kernel Support Vector Machine (SVM) and Radial basis function SVM. The F_1 Macro Average is computed based on human labels.

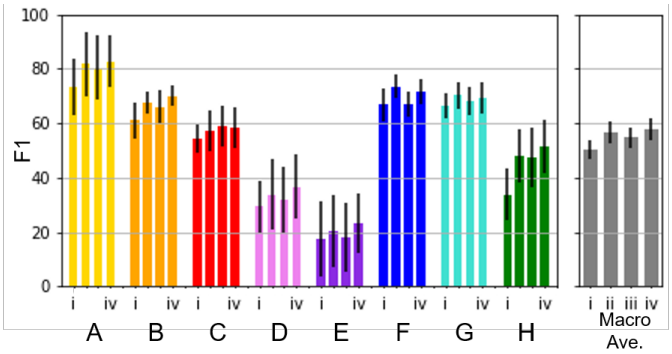
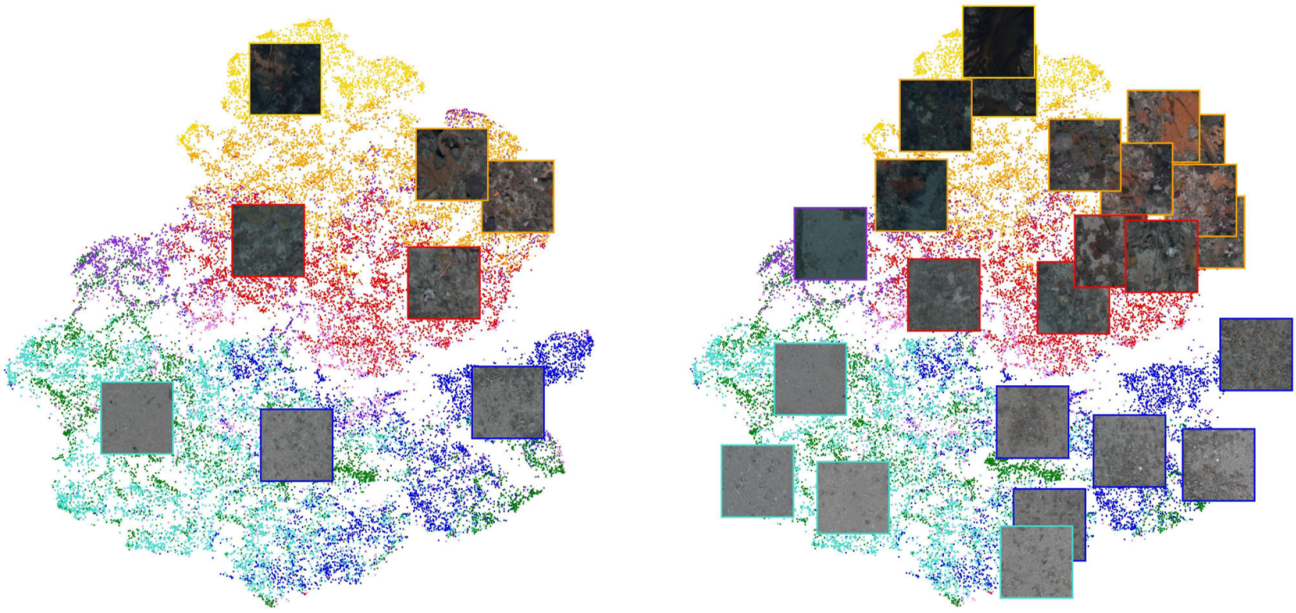


Fig. 3: Per-class F_1-scores and their macro average for (i) no regularisation, (ii) horizontal location regularisation, (iii) depth regularisation and (iv) horizontal location and depth regularisation. R-SVM is used as the classifier in this plot.

D. Result

TABLE II shows the mean and standard deviation of the F_1 scores for each autoencoder training configuration and classifier. For four of five classifiers; 1-NN, GP, L-SVM and R-SVM, the autoencoders trained with both L_{loc} and L_{dep} (configuration (iv)) show the best performance among the four configurations. For RF, configuration (ii), where only L_{loc} is applied, has the best score. However, the difference between (ii) and (iv) is marginal. Configurations (ii) - (iv) perform better than configuration (i), where no regularisation is applied, for all classifiers, achieving an average performance gain of (ii) 9.4 %, (iii) 6.9 % and (iv) 10.9 %, respectively. The results show that horizontal location metadata is more effective for learning latent representations than depth for this dataset. However, using both of horizontal location and depth information generally improves performances, and never causes any significant degradation. The biggest gains in performance are seen for the R-SVM classifier, where an improvement of (ii) 12.5%, (iii) 8.7% and (iv) 15.1%, are seen respectively compared to no regularisation (i). Another noticeable point is that for L-SVM, configuration (iii) shows a better score better than (ii). Among the five classifiers used in the experiment, L-SVM is the only linear classifier, which makes it relatively robust



(a) k-means clustering (k=8). (b) Hierarchical k-means clustering (k=8 and k'=3).

Fig. 4: The t-SNE latent representation learnt by the proposed method for both horizontal location and depth regularisation, i.e. configuration (iv). Representative images are selected based on k-means clustering for (a) and hierarchical k-means clustering for (b). The colours represent the classes determined by R-SVM and are used for illustrative purposes only.

against over-fitting. A different trend is observed compared to the other classifiers with depth only regularisation performing favourably. A possible explanation for this is that some over-fitting may be taking place with the non-linear classifiers when only depth regularisation is used.

Fig. 3 shows the per-class F_1 scores of the best performing classifier (R-SVM) for regularisation configurations (i)-(iv). Configurations (ii)-(iv) are superior to configuration (i) for all classes. Horizontal location regularisation (ii) performs better than depth regularisation (iii) for all classes except for C. The relative performance improvement with metadata regularisation is most significant for classes D, E, and H (24.9%, 33.5%, and 52.6% between (i) and (iv), respectively), which have relatively small populations in the dataset. This can be explained as optimising only the autoencoder reconstruction loss L_{rec} potentially leads to focusing on the appearances of majority classes, where the proposed regularisation avoids this form of over-fitting by effectively prioritising patterns in classes with smaller populations.

An important characteristic of the proposed method is that both regularisation methods can be applied without risk of significant performance degradation. This is due to the use of t-distributions and the loose regularisation constraints imposed during the loss function optimisation based on probabilistic distributions. We see this characteristic where configuration (iv) leads to and overall improvement in performance, and better class scores than configurations (ii) and (iii) for most classes. Where the scores for classes C, F and G are slightly degraded, the difference is negligible. Although horizontal location regularisation is generally more effective than depth regularisation for this dataset, the ability to improve performance using only depth information is valuable as accurate horizontal localisation in GPS denied subsea environments requires expensive navigational sensors that may not be available on some low cost AUVs and Remotely Operated Vehicles (ROVs). On the other hand, depth sensors are relatively cheap and so are available on almost all underwater platforms.

E. Application to Seafloor Survey

Latent representations can be applied to efficiently understand the characteristics of a dataset. One way to do this is by automatically identifying images that are most representative of the variety of scenes that exist in the data. Fig. 4 shows the automatically selected representative images, overlaid on the representations of Tasmania dataset using a t-SNE visualisation [21]. In Fig. 4a, k-means clustering is applied to the acquired latent representations, and the images closest to each of the k centroids are selected as representative images. Here, we use k=8 which is automatically determined based on the elbow-method [27]. In Fig. 4b, Hierarchical k-means clustering [28] is applied to identify a further k'=3 within each original cluster. This allows for representation of the range and sequential transitions of seafloor scenes. The results show that a relatively small number of representative images automatically identified by the system can efficiently describe the variety of scenes found in a dataset consisting of more than 30k images, including representative examples of classes with a small population. This is valuable for remote transmission of exemplary data over the limited bandwidths available using long-range underwater acoustics communications, or global communication satellites when platforms are at the water surface. Representative images may also benefit low-shot training of supervised and semi-supervised classifiers.

V. Conclusion

We have proposed a novel autoencoder regularisation method that can leverage any number and combination of vector or scalar metadata for seafloor image representation learning. The regularisation is effective when two images that are close in their metadata space tend to be more similar in appearance. By optimising loss functions using the KL divergence and t-distributions, it is possible to mitigate over-regularisation by metadata and avoid significant performance degradation when multiple sources of metadata are applied to regularise learning. The self regulating latent representation learning method was applied to a dataset consisting of more than 30k images taken during 6 AUV dives. Validation against 2.2k expert human annotations shows that:

- Combining multiple sources of metadata regularisation can outperform single metadata regularisation using the proposed method. Regularising learning using depth and horizontal location metadata improves the performance of five classifiers operating on the latent representations by an average of 10.9% compared to a standard convolutional autoencoder, with the R-SVM classifier showing the largest gain in performance at 15.1%.
- Horizontal location regularisation is more effective than depth regularisation for the sparse transect dataset analysed in this work, achieving an average improvement of 9.4% (as opposed to 6.9%) across five classifiers, and 12.5% (as opposed to 8.7%) for the best performing classifier. However, combining both in metadata regularisation reliably outperforms individual regularisation and never significantly degrades performance.
- The acquired latent representations allow representative images of large datasets with imbalanced class distributions to be automatically identified in a fully unsupervised way, which can help achieve an efficient understanding of underwater scenes and be applied to adaptive path planning using visual information.

References

[1] A. Bender, S. B. Williams, and O. Pizarro, "Autonomous exploration of large-scale benthic environments," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 390–396.

[2] J. Shields, O. Pizarro, and S. B. Williams, "Feature space exploration for planning initial benthic auv surveys," *arXiv preprint arXiv:2105.11598*, 2021.

[3] G. Flaspohler, V. Preston, A. P. Michel, Y. Girdhar, and N. Roy, "Information-guided robotic maximum seek-and-sample in partially observable continuous environments," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3782–3789, 2019.

[4] Y. Girdhar and G. Dudek, "Exploring underwater environments with curiosity," in *2014 Canadian Conference on Computer and Robot Vision*. IEEE, 2014, pp. 104–110.

[5] 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.

[6] 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.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[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] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[17] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[18] 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.

[19] T. Yamada, A. Prügel-Bennett, and B. Thornton, "Learning features from georeferenced seafloor imagery with location guided autoencoders," *Journal of Field Robotics*, vol. 38, no. 1, pp. 52–67, 2021.

[20] M. Modasshir and I. Rekleitis, "Enhancing coral reef monitoring utilizing a deep semi-supervised learning approach," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 1874–1880.

[21] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[22] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[23] S. B. Williams, O. R. Pizarro, M. V. Jakuba, C. R. Johnson, N. S. Barrett, R. C. Babcock, G. A. Kendrick, P. D. Steinberg, A. J. Heyward, P. J. Doherty, I. Mahon, M. Johnson-Roberson, D. Steinberg, and A. Friedman, "Monitoring of benthic reference sites: Using an autonomous underwater vehicle," *IEEE Robotics Automation Magazine*, vol. 19, no. 1, pp. 73–84, 2012.

[24] T. Moltmann, J. Turton, H.-M. Zhang, G. Nolan, C. Gouldman, L. Griesbauer, Z. Willis, A. M. Piniella, S. Barrell, E. Andersson *et al.*, "A global ocean observing system (goos), delivered through enhanced collaboration across regions, communities, and new technologies," *Frontiers in Marine Science*, vol. 6, p. 291, 2019.

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

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

[27] 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.

[28] 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.