Learning Features from Georeferenced Seafloor Imagery with Location Guided Autoencoders

Takaki Yamada
School of Engineering, Faculty of Engineering and Physical Science
University of Southampton
Southampton SO16 7QF, U.K.
T.Yamada@soton.ac.uk

Adam Prügel-Bennett
School of Electronics and Computer Science, Faculty of Engineering and Physical Science
University of Southampton
Southampton SO17 1BJ, U.K.
apb@ecs.soton.ac.uk

Blair Thornton
Centre for In Situ and Remote Intelligent Sensing
School of Engineering, Faculty of Engineering and Physical Science
University of Southampton
Southampton SO16 7QF, U.K.
Institute of Industrial Science
The University of Tokyo
4-6-1 Komaba Meguro-ku, Tokyo 153-8505, Japan
B.Thornton@soton.ac.uk

Abstract

Although modern machine learning has the potential to greatly speed up the interpretation of imagery, the varied nature of the seabed and limited availability of expert annotations form barriers to its widespread use in seafloor mapping applications. This motivates research into unsupervised methods that function without large databases of human annotations. This paper develops an unsupervised feature learning method for georeferenced seafloor visual imagery that considers patterns both within the footprint of a single image frame and broader scale spatial characteristics. Features within images are learnt using an autoencoder developed based on the AlexNet deep convolutional neural network. Features larger than each image frame are learnt using a novel loss function that regularises autoencoder training using the Kullback-Leibler divergence function to loosely assume that images captured within a close distance of each other look more similar than those that are far away. The method is used to semantically interpret images taken by an autonomous underwater vehicle at the Southern Hydrates Ridge, an active gas hydrate field and site of a seafloor cabled observatory at a depth of 780 m. The method's performance when applied to clustering and content-based image retrieval is assessed against a ground truth consisting of more than 18,000 human annotations. The study shows that the location based loss function increases the rate of information retrieval by a factor of two for seafloor mapping applications. The effects of physics based colour correction and image rescaling are also investigated, showing that the improved consistency of spatial information achieved by rescaling is beneficial for recognising artificial objects such as cables and infrastructures, but is less effective for natural objects that have greater dimensional variability.

1 Introduction

Autonomous Underwater Vehicles (AUVs) and Remotely Operated Vehicles (ROVs) routinely collect tens to hundreds of thousands of georeferenced seafloor images during their dives. Although modern machine learning has the potential to greatly speed up the interpretation of the images obtained, the absence of large annotated training datasets and sensitivity of data quality to imaging conditions has limited their application. Unlike the aerial and satellite images used in terrestrial mapping, the strong attenuation of light in water means that seafloor imaging typically requires underwater terrains to be followed at close altitudes of between two and ten metres and the use of strobed illumination. Under these conditions, even small fluctuations in altitude strongly affect the colour balance, spatial resolution and area covered, reducing the consistency between image frames. Furthermore, the footprint of each image is limited to an edge length of just a few metres, which is significantly smaller than many of the geological and ecological features that are of interest for scientific analysis and statutory monitoring. Although efforts to develop shared annotation schemes and datasets exist within the marine imaging community (Bewley et al., 2015a; Langenkämper et al., 2017), the variability of seafloor environments, imaging systems and the limited number of experts with domain specific knowledge mean that the development of comprehensive annotated training datasets similar to those on land (e.g. SpaceNet, ImageNet, COCO, Pascal VOC) are unlikely to be developed.

This paper investigates the use of unsupervised learning to extract features and perform semantic interpretation of seafloor imagery. A key advantage of unsupervised methods is that they do not require annotated datasets for training. Although the use of unsupervised learning for clustering seafloor acoustic imagery (Hasan et al., 2014) and visual imagery (Steinberg et al., 2011; Steinberg, 2013; Kaeli and Singh, 2015) have been reported, most previous work has used manually selected features that have been defined based on domain specific knowledge, limiting their ability to generalise across datasets. More recently, unsupervised frameworks that learn suitable features from a dataset have shown great promise as a general tool for semantic interpretation of seafloor imagery (Rao et al., 2017; Flaspohler et al., 2017). This work aims to further develop this concept and improve the extraction of information about geological and ecological features that exist on spatial scales larger than the footprint of a single image. This is achieved by developing an autoencoder framework that regularises features learning using georeferencing information. The main contributions of this paper are:

- Development of an autoencoder feature learning framework that can take into account georeferencing information using a novel loss function based on Kullback-Leibler divergence.
- Investigation of the effectiveness of georeference regularisation, physics based colour correction and spatial scale information on learning using an expert labelled ground truth.
- Demonstration of semantic mapping applications of learnt features through clustering and content-based image retrieval.

The autoencoder developed in this work learns features using a deep learning convolutional neural network based on AlexNet (Krizhevsky et al., 2012). A novel loss function that uses georeference information is used to regularise learning by minimising the Kullback-Leibler divergence between affinity in the latent feature space and geographic location. The proposed method is applied to semantically interpret images collected by the AUV ae2000f of the University of Tokyo, Japan. The dataset consists of more than 12,000 images collected from an altitude of 6 m off the seafloor at the Southern Hydrate Ridge, an active gas hydrate field at a depth of 780 m and site of the Ocean Observation Initiative's seafloor cabled observatory (Cowles et al., 2010). The effectiveness of the proposed method is assessed using more than 18,000 expert annotations.

2 Background

2.1 Semantic interpretation of seafloor imagery

Feature engineering is crucial to effectively interpret visual imagery. Seafloor images have unique characteristics compared to terrestrial datasets, and several studies have demonstrated semantic interpretation using manually selected features that are tailored to specific subsea applications (Pizarro et al., 2008; Maki et al., 2010; Thornton et al., 2012; Beijbom et al., 2012). These have been used for classification and segmentation within images and mosaiced reconstructions. More recently, an attempt to develop a generic feature extraction method by (Steinberg et al., 2011) used Local Binary Pattern (LBP) (Ojala et al., 2002) features derived from greyscale images together with 3D rugosity features and colour features for unsupervised clustering of seafloor stereo images. In (Steinberg, 2013), the author proposed Sparse Coding Spatial Pyramid Matching (ScSPM) (Yang et al., 2009) as a more generic approach. However, this required additional techniques to reduce the dimensionality of ScSPM outputs in order to perform classification. In (Kaeli and Singh, 2015), the accumulated histogram of oriented gradients from keypoints were used to describe each image, and this was applied to clustering and anomaly detection. A common characteristic of these methods is that they can preserve multi-scale features in the images, which is important as the size of seafloor targets can vary. Though these approaches are intrinsically robust to the scale variance, there exist many hyperparameters which require manual tuning to optimise performance for each dataset. Moreover, features larger than the footprint of a single frame cannot be captured.

For supervised learning applications, LBP and ScSPM features have been shown to be effective (Bewley et al., 2015b; Rao et al., 2017). Deep learning techniques can optimise feature learning and classification simultaneously within the same end-to-end training process. In (Mahmood et al., 2018), a convolutional neural network (ResNet (He et al., 2016)) was successfully applied to classify coral images. While prior works had aimed to classify broad seafloor categories such as 'Rock', 'Sand', 'Coral', the authors accurately classified nine classes of coral to prove the effectiveness of deep learning for detailed seafloor image interpretation. On the other hand, little has been reported on the application of deep learning techniques for unsupervised feature learning in seafloor visual imaging applications.

2.2 Autoencoders

The autoencoder is a variation of the artificial neural network that is useful for unsupervised feature learning. It consists of two parts; an encoder and a decoder. The encoder maps original data $\boldsymbol{x}$ into a latent representation $\boldsymbol{h}$ of lower dimensionality and can be expressed as $\boldsymbol{h} = f_\phi(\boldsymbol{x})$. The decoder is expressed as $\boldsymbol{x_r} = g_\theta(\boldsymbol{h})$, and reconstructs $\boldsymbol{x_r}$ to be as similar to the original sample $\boldsymbol{x}$ as possible for a given latent representation. When the values in $\boldsymbol{x}$ are continuous, the difference between $\boldsymbol{x}$ and $\boldsymbol{x_r}$ can be measured as the mean squared error. Given n samples in a dataset, the autoencoder's objective function can be formulated as follows,

$$\min_{\phi,\theta} L_{rec} = \min \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{x_r}_i\|^2 , \tag{1}$$

where ϕ and θ denote the parameters of the encoder and decoder, respectively. The biggest advantage of the autoencoder is that the networks used can be trained without the need for expert annotations. Since $\boldsymbol{x_r}$ is reconstructed from a latent representation $\boldsymbol{h}$ that preserves key information in $\boldsymbol{x}$ in a lower dimensional space, $\boldsymbol{h}$ can be thought of as the set of features of a given size that best represents the original data. In (Rao et al., 2017), autoencoders are applied partly to learn mid-level features in visual imagery after extracting low-level features with ScSPM. (Flaspohler et al., 2017) applies convolutional autoencoders for unsupervised feature learning from seafloor imagery and shows that they outperform hand-designed features in discovering characteristic patterns.

To enhance the unsupervised feature learning performance of autoencoders, several studies have demonstrated training of autoencoders with additional loss functions designed to maximise clustering in the latent representation space (Aljalbout et al., 2018; Min et al., 2018). A typical loss function can be formulated as,

$$L_{all} = (1 - \lambda) L_{rec} + \lambda L_{clust}, \tag{2}$$

where L_{clust} is a clustering loss, and λ is a hyperparameter designed to balance L_{rec} and L_{clust}. In (Yang et al., 2017), the use of such a loss function for k-means clustering significantly improved clustering performance. In (Xie et al., 2016), L_{clust} is formulated as follows,

$$L_{clust} = \mathrm{KL}(P \| Q) = \sum_i \sum_k p_{ik} \log \frac{p_{ik}}{q_{ik}}, \quad q_{ik} = \frac{\left(1 + \|\boldsymbol{h}_i - \boldsymbol{\mu}_k\|^2\right)^{-1}}{\sum_{k'} \left(1 + \|\boldsymbol{h}_i - \boldsymbol{\mu}_{k'}\|^2\right)^{-1}}, \quad p_{ik} = \frac{q_{ik}^2 / f_k}{\sum_{k'} q_{ik'}^2 / f_{k'}}, \tag{3}$$

where $\boldsymbol{\mu}_k$ is the centroid of cluster k in the latent representation space, p_{ik} and q_{ik} are the $[i, k]$ th elements of the probabilistic distributions P and Q, and f_j is soft cluster frequency which is defined as $\sum_i q_{ij}$. $\sum_{k'}$ means that the values of $\left(1 + \|\boldsymbol{h}_i - \boldsymbol{\mu}_{k'}\|^2\right)^{-1}$ are calculated for all the clusters (k') and summed for use as a normalisation factor. The element q_{ik} can be interpreted as the probability of assigning $\boldsymbol{h}_i$ to cluster k, defined with the Student's t-distribution as a kernel following t-SNE algorithm (Maaten and Hinton, 2008). The element p_{ik} is the target value derived from q_{ik} to maximise the separation between cluster k and the other clusters. L_{clust} is trained after training L_{rec} by minimising the Kullback-Leibler (KL) divergence between P and Q. Since L_{clust} in Equation 3 is derived as a soft cluster assignment and is differentiable, it can be efficiently optimised using back-propagation. For public datasets, the use of a clustering loss was shown to improve clustering accuracy by up to 2.5 % for the MNIST dataset. However, both studies require the number of clusters to be manually set, which is not practical for seafloor images or other natural scenes where the appropriate number of clusters is not known.

Another important application of autoencoders is anomaly detection since anomalous data which are rarely observed in the dataset can not be reconstructed precisely and have a large value of L_{rec}. (Zurowietz et al., 2018) applies autoencoders to detecting anomalous regions in seafloor images as candidates for living organisms, since they are less frequently observed than backgrounds (i.e. rocks and sand).

3 Feature learning for seafloor imagery

3.1 Pre-processing

Images taken underwater are distorted by the water column. Colour and geometry corrections can be applied to improve the consistency of datasets prior to feature learning. In this work, colour correction parameters are estimated based on the altitude each image was taken at to compensate for wavelength dependent attenuation in the water column. Altitude is also used to rescale undistorted images and reduce the scale variance caused by differences in range to targets. Figure 1 shows sample images of raw (Figure 1A) and pre-processed (Figure 1B to 1D) strobed images taken at 5.1 m (top) and 7.0 m (bottom) altitudes together with their colour histograms. The methods used for correction are described in the following sections.

3.1.1 Colour correction

Light attenuation in water differs for each wavelength that constitutes the RGB channels. Since red attenuates more aggressively than green or blue wavelengths, uncorrected under-

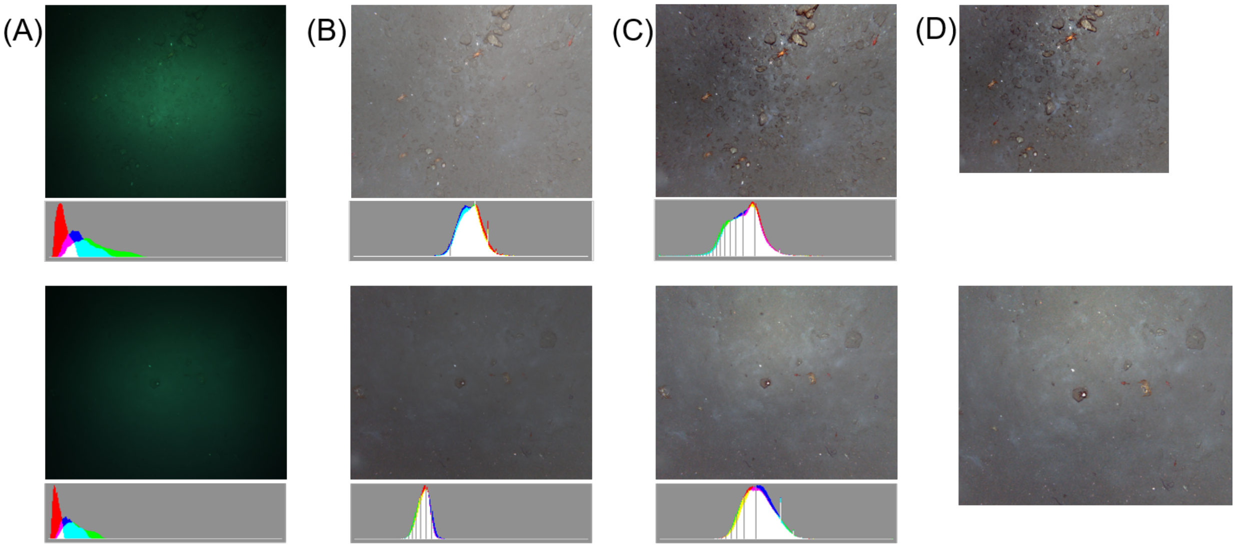


Figure 1: Seafloor images captured at 5.1 m (top, 5 mm/pixel) and 7.0 m (bottom, 7 mm/pixel) altitude for (A) raw, (B) pixel-wise normalisation, (C) attenuation correction and pixel-wise normalisation and (D) undistortion and rescaling to a constant spatial resolution (6 mm/pixel, equivalent to 6.0 m altitude).

water images appear blue and green (Jaffe, 1990) (Figure 1A). Seafloor images captured at low altitudes (Figure 1A top) are also brighter than the images captured at high altitudes (Figure 1A bottom). Often wide angle lenses are used to maximise the imaged area, and this can cause pixels at the centre of each image to be brighter than those at its edges. Pixel-wise colour correction normalises each pixel by the mean and standard deviation of the same pixel across an entire dataset based on the grey-world assumption (Buchsbaum, 1980). This can improve the imbalance between colour channels and uneven brightness within each image (Figure 1B). However, pixel-wise normalisation cannot correct colour variations caused by altitude differences within a dataset. To compensate for these variations, (Bryson et al., 2013) proposed a practical method that improves colour consistency by taking into account the attenuation of the different colour channels. This work applies a similar approach, where the attenuation is approximated as follows,

$$\mu_x(u,v,\nu,d) = a(u,v,\nu)\exp(-b(u,v,\nu)d) + c(u,v,\nu). \tag{4}$$

The indices $[u, v]$ specify each pixel's location in the image frame, ν is the colour channel, d is range from the centre of the camera to the seafloor when the image was taken, and $\mu_x(u, v, \nu, d)$ is the mean of all intensities in the dataset. Parameters $a(u, v, \nu)$, $b(u, v, \nu)$ and $c(u, v, \nu)$ model the effects of the water column for each pixel and colour channel. These parameters are identified through regression of the image dataset. In (Bryson et al., 2013), the range d is estimated by stereo image matching and the regression is calculated with a non-linear least squares fitting. Since stereo images are not always available, altitude values from range measurements made by a Doppler Velocity Log (DVL) are used for estimating d in this work. This method assumes the seafloor is flat, which is reasonable when the vertical profile in each image is small relative to the altitude. The pixel-wise normalisation also corrects for vignetting. Since outliers in the dataset disturb the normalisation, the 10 % extreme intensity values for each pixel location are trimmed before determining the model parameters. Figure 1C shows the result of the proposed colour correction. Compared to Figure 1B, the brightness between images taken at different altitudes is more uniform.

3.1.2 Geometry correction

The 3D information needed to fully compensate for scale effects within an image frame is not always available. Therefore, this work approximates scale effects from the imaging altitude and the lens field of view on a per image basis. Geometric distortions are also corrected using lens calibration data. Each image is downsampled to a consistent spatial resolution of 10 mm/pixel, which is considered appropriate for the imaging setup used for the experiments analysed in this paper (see Table 1). In this study, the roll and pitch of the images are not taken into account. This is reasonable for correctly trimmed underwater vehicles with downward-looking imaging systems.

3.2 Autoencoder based feature learning

3.2.1 Learning features within an image

This work uses AlexNet (Krizhevsky et al., 2012) as the basic architecture for feature learning. AlexNet was originally designed for supervised classification and is one of the first successful deeply stacked artificial neural networks with convolutional layers. It is effective at learning scale and rotation invariant features. (Cheng et al., 2016) successfully applies AlexNet for extracting rotation invariant features from satellite image datasets. As with satellite imagery, seafloor targets captured by downward-looking cameras do not have distinct rotations. For the encoder, AlexNet's original architecture is applied. For the decoder, an inverted version of AlexNet is developed where the convolutional layers are transposed to trans-convolutional layers, and the max pooling layers are transposed to max unpooling layers. An advantage of applying a common convolutional neural network architecture is that pre-trained parameters are available as initial values to accelerate training for each data set.

Though smaller dimensionality of the latent representation $\boldsymbol{h}$ is preferable for application to clustering and image search, excessive compression would lead to loss of information needed to represent the original data. In this work, the appropriate dimension was determined experimentally to minimise the size of the latent representation while making sure the reconstruction loss L_{recon} remained small.

Another difference from the original AlexNet is that the dropouts which appear in the original's fully-connected layers are not used. Instead, batch normalisation (Ioffe and Szegedy, 2015) is applied to each layer to accelerate training and stabilise convergence. The last layers of both the encoder and decoder have no activation function because their outputs, i.e. the latent representation $\boldsymbol{h}$ and the reconstructed data $\boldsymbol{x_r}$, are continuous values. To enhance the robustness of the autoencoder, data augmentation and noising (Vincent et al., 2010) is applied. The images in datasets are augmented by randomly rotating, flipping and shifting the bounding box of each image patch within ± 25 % of the patch size at each iteration during training. As a noise model, additive isotopic Gaussian noise is selected to model the noise in seafloor imagery as the main source of error is expected to be in the estimation of the parameters in Equation 4. The colour histograms of each image are first shifted randomly, then isotopic Gaussian noise is added to each pixel in the image. The reconstruction loss L_{rec} in Equation 1 is calculated as the difference between augmented images before noising and the reconstructed images.

3.2.2 Georeference regularised learning

Geological and ecological features of the seafloor such as sediments, bacterial mats and seafloor infrastructures and background substrates such as sands and rocks exist over spatial scale larger than the footprint of a single image frame. To capture this property, the following assumption is made:

Assumption. *Two images captured within a close distance tend to look more similar than two that are far away.*

In general, a favourable feature learner should embed $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ closer in the latent representation space, if the original data $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are similar. Based on the assumption, the affinity between $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ in the latent representation space should be modified to account for the affinity between the geographical locations $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ at which $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are measured. For seafloor imagery, it is reasonable to assume that $\boldsymbol{y}$ is known since they are captured by AUVs or other platforms with navigational sensors and methods to determine position are well documented (Paull et al., 2013). To implement this idea, the Student's t-distribution is used as a kernel to measure affinity (Maaten and Hinton, 2008; Xie et al., 2016) in both the latent representation ($\boldsymbol{h}$) space and the geographical ($\boldsymbol{y}$) space. Thus q'_{ij}, which is the value of the affinity matrix at index (i, j) in the latent representation space Q' can be defined as,

$$q'_{ij} = \frac{\left(1 + \|\boldsymbol{h}_i - \boldsymbol{h}_j\|^2\right)^{-1}}{\sum_{i'}\sum_{j'}\left(1 + \|\boldsymbol{h}_{i'} - \boldsymbol{h}_{j'}\|^2\right)^{-1}}. \tag{5}$$

Likewise, p_{ij} which is the element of the affinity matrix P' in physical space for the georeferenced data can be defined as,

$$p'_{ij} = \frac{\left(1 + \mathrm{d}(\boldsymbol{y}_i, \boldsymbol{y}_j)\right)^{-1}}{\sum_{i'}\sum_{j'}\left(1 + \mathrm{d}(\boldsymbol{y}_{i'}, \boldsymbol{y}_{j'})\right)^{-1}}, \tag{6}$$

where $\mathrm{d}(\boldsymbol{y}_i, \boldsymbol{y}_j)$ is defined as $\mathrm{d}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \min(\left\|\boldsymbol{y}_i - \boldsymbol{y}_j\right\|^2, d_{max}^2)$ and d_{max} is the user-defined upper limit of the distance between the two locations that are correlated. This limit prevents P' from overfitting images that are at a large distance apart. To capture features larger than a single image, d_{max} should be enough larger than the physical footprint sizes of images. An appropriate value d_{max} works for most targets and backgrounds, though the sizes of them vary highly. This is because even if two images belonging to a continuous pattern are separated farther away than d_{max}, some images belonging to the same pattern would exist continuously between them, and all of these images would be embedded close together in the latent representation space. The autoencoder is trained so that the KL divergence between the two affinity matrices P' and Q' is minimised. The proposed loss function becomes,

$$L_{all} = L_{rec} + \lambda L_{geo} = L_{rec} + \lambda \mathrm{KL}(P' \| Q'). \tag{7}$$

λ is a hyperparameter for balancing L_{geo} and L_{rec}. This can be optimised iteratively using a mini-batch. However, if the many of images in a mini-batch are sampled from the locations separated farther than d_{max}, the elements of P' become similar thus the training is not regularised as intended. To avoid this issue, at each mini-batch sampling, the first image is randomly chosen from the whole dataset and the other images to populate the batch are chosen according to their physical proximity to the first image.

Since t-distribution is a heavy tailed distribution, L_{geo} loosely regularises the latent representation space to follow the assumption. The appropriate setting of λ is also necessary because if the regularisation is too forceful, L_{rec} is ignored and the latent representations become meaningless as features for semantic interpretation. In addition, to deal with the uncertainty of georeference information, images are randomly shifted within 25 % of each image patch size at every sampling step. The similarity assumption has no hypothesis on the rotation of images, and so random rotations are applied to the images to avoid fitting rotation variances in the dataset.

Figure 2 gives an overview of the proposed feature learner. The proposed autoencoder learns local features within an image using a convolutional neural network. Learning is also regularised by the georeference loss function to account for patterns larger than each image footprint. Once the autoencoder has been trained to minimise Equation 7, its encoder can be used as a feature extractor. Denoising, random rotation and shifting are not applied to extract features by the encoder. It should be noted that georeference information $\boldsymbol{y}$ is also not used in the feature extraction phase. This is because the aim of embedding georeference information is not to map the absolute coordinates of the images to the latent representation space, but to control the feature mapping by embedding the assumption into the trained autoencoder. This allows the encoder to extract features from datasets unrelated to the training dataset, and datasets without georeferencing information.

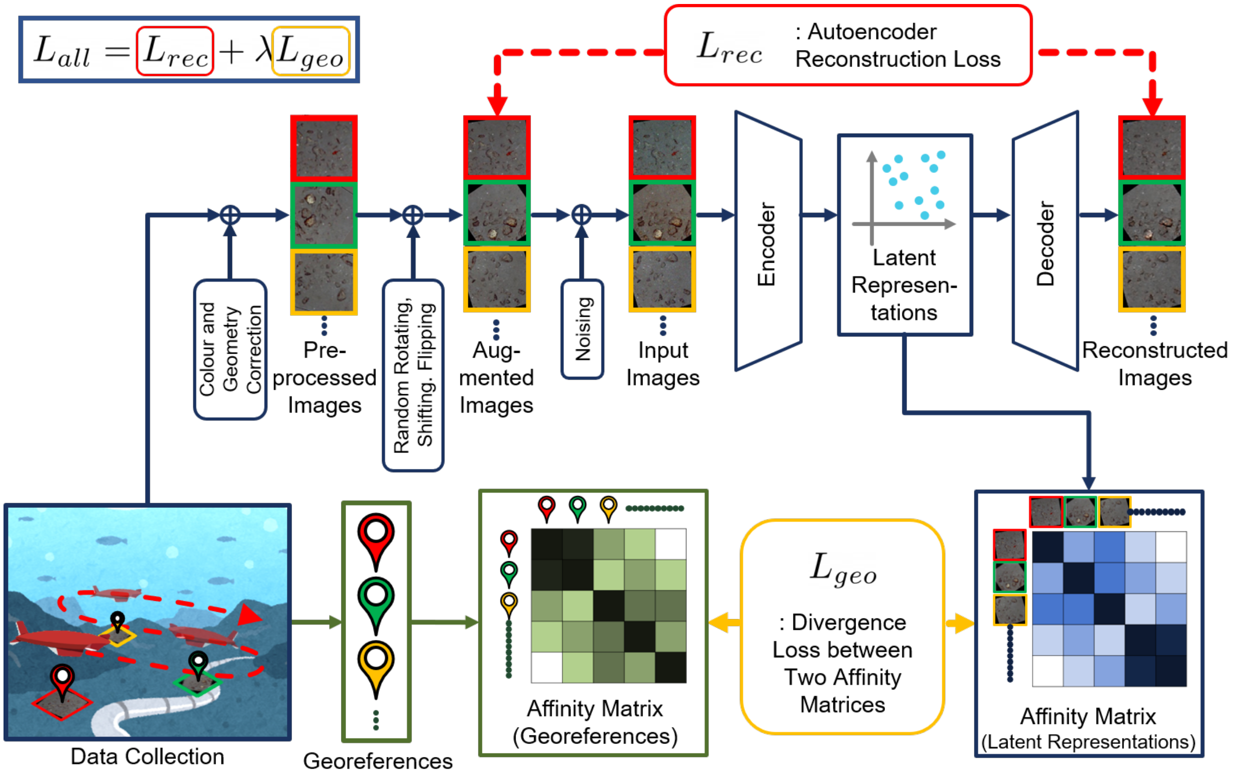

$L_{all} = L_{rec} + \lambda L_{geo}$
L_{rec} : Autoencoder Reconstruction Loss
Colour and Geometry Correction
Pre-processed Images
Random Rotating, Shifting, Flipping
Aug-mented Images
Noising
Input Images
Encoder
Latent Represen-tations
Decoder
Reconstructed Images
Data Collection
Georeferences
Affinity Matrix (Georeferences)
L_{geo} : Divergence Loss between Two Affinity Matrices
Affinity Matrix (Latent Representations)

Figure 2: Flow diagram for calculating the proposed loss function L_{all} (Equation 7). L_{rec} is the reconstruction loss of the autoencoder (Equation 1). L_{geo} is the divergence loss between the two affinity matrices in the latent representation space (Equation 5) and in the physical space (Equation 6).

4 Applications of feature learning to semantic interpretation

4.1 Clustering

Clustering is a useful technique for semantic interpretation of the features obtained with the proposed autoencoder since it does not require ground truth and interprets the data in a completely unsupervised manner. If the dimensionality of the latent representation $\boldsymbol{h}$ is small enough, clustering techniques can be applied directly without any further dimensional reduction. In order to automatically determine the appropriate number of clusters for the latent representation, the non-parametric Bayesian method described in (Blei et al., 2006) is applied.

4.2 Image search

Once an interesting target is found in a dataset, images that are similar in appearance and their geographic distribution are also likely to be of interest. This information can be automatically retrieved from large volumes of imagery by calculating the similarity between the query image and other remaining images in the latent representation space. This is useful as clustering techniques typically do not assign an independent cluster to categories with a small number of samples, and have difficulty with ambiguous categories that have continuously varying characteristics.

The similarity between a pair of images $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ can be derived from the latent representation $\boldsymbol{h}$ of each image, where established similarity metrics such as the Euclidean distance and cosine distance can be used (Wu et al., 2013). Since the similarities are defined in the latent representation space, georeference information is unnecessary for this application once the autoencoder has been trained. However, predicting the performance of the two metrics is difficult for features learnt by an autoencoder since the interpretation of their meaning is non-trivial. Therefore this work compares their performance experimentally.

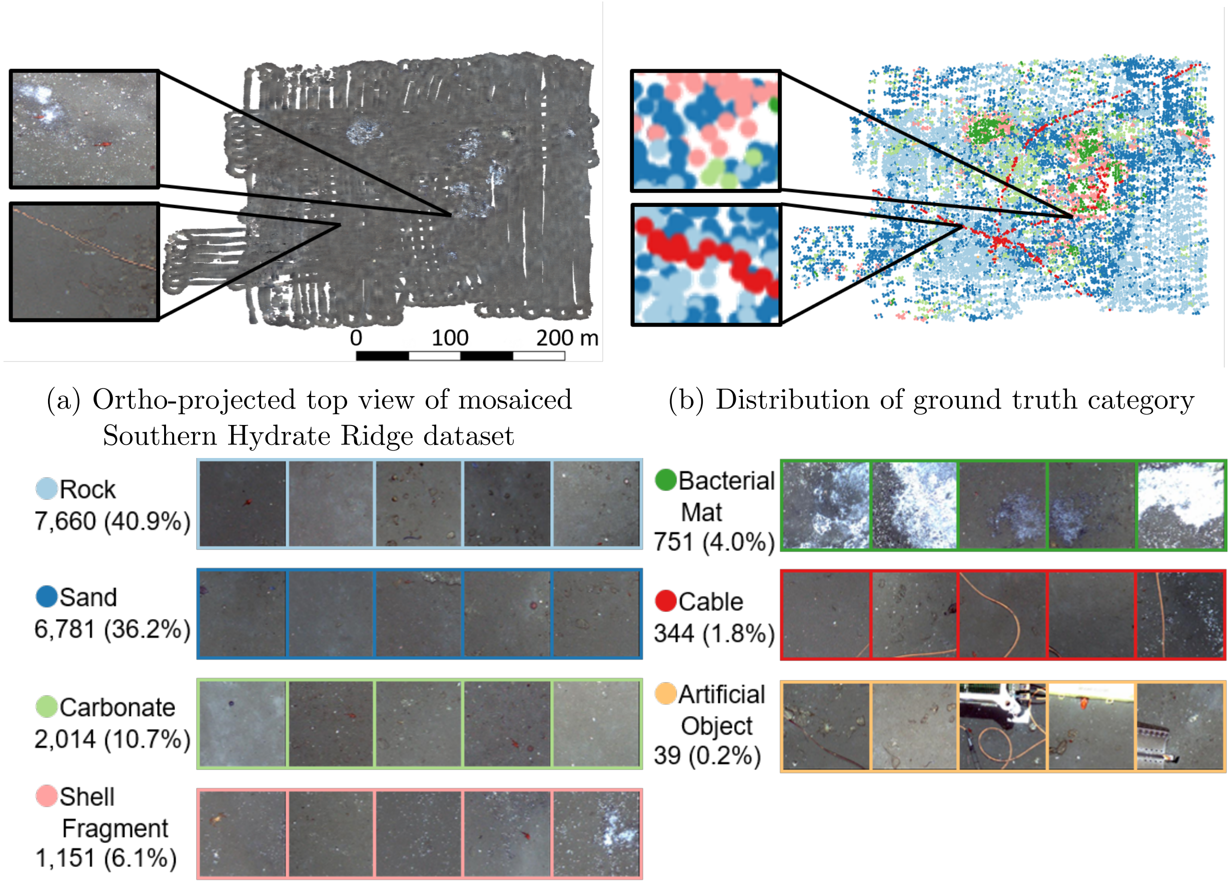
5 Experiment

The methods developed in this work are applied to seafloor imagery obtained at the Southern Hydrate Ridge, a gas hydrate field that is home to a seafloor cabled observatory (Cowles et al., 2010) located 100 km off Oregon, USA (Table 1). Over 12,000 images of the site were collected using the AUV ae2000f of the Institute of Industrial Science, University of Tokyo, Japan, during the Schmidt Ocean Institute's FK180731 #Adaptive Robotics campaign in August 2018. Table 1 gives an overview of the dataset, and Figure 3a shows an ortho-projected mosaic created from the images in the dataset using a stereo SLAM pipeline developed by the Australian Centre for Field Robotics, University of Sydney, Australia (Mahon et al., 2008; Johnson-Roberson et al., 2010).

Five small patches are cropped from each image at this size from the four corners and the centre to obtain the proper size of images for the proposed AlexNet based autoencoder (227 × 227, overlapping partially). When rescaling is applied, the original images are first scaled so that they have a constant spatial resolution of 10 mm/pixel. The total number of patches for autoencoder training is 62,875. The georeference information (position where each image was captured) is obtained through the visual SLAM pipeline developed in (Mahon et al., 2008). This has been applied to data collected by the AUV's navigational sensors, consisting of an iXblue Quadrans IMU, RDI 300 kHz DVL, Paroscientific depth sensor, iXblue Gaps USBL and stereo imagery collected by the SeaXerocks mapping system of the University of Tokyo, Japan (Thornton et al., 2016). The relative position accuracy using this combination is estimated to be <1 m across the dataset. This is of a similar order to the randomly allocated shifting of images applied for data augmentation when training the autoencoder (25 % of 2.27 m). This allows the autoencoder to take localisation uncertainty into consideration and avoids overfitting to the georeference information.

5.1 Ground truth for evaluation

Ground truth annotations were generated using SQUIDLE+ (Bewley et al., 2015a) by experts for 18,740 (approx. 30 %) image patches randomly selected from the original 62,875 image patches. Figure 3 shows the spatial distributions, the numbers and the examples of each category. Boundaries between some categories are ambiguous, especially for natural features, e.g. 'Rock', 'Sand' and 'Carbonate', where the density of the relevant targets vary on a continuum. From the appearances of the ground truth categories shown in Figure 3c, it is noticeable that these categories form the larger patterns than the footprint of images, thus the proposed georeference regularisation is assumed to be effective. In this experiment, only the dominant label is given to each image patch based on individual annotator's judgement. Although this complicates the quantitative evaluation of performance, the relative performance between different conditions of the proposed feature learning can be used to verify how effective the methods developed in this paper are for semantic interpretation.



(a) Ortho-projected top view of mosaiced Southern Hydrate Ridge dataset

(b) Distribution of ground truth category

(c) Samples from each ground truth category

Figure 3: Overview of Southern Hydrate Ridge dataset. (a) and (b) show that each ground truth category has a characteristic spatial distribution. Many of samples in the dataset are categorised as either 'Rock' or 'Sand'. 'Shell Fragment' is often observed around 'Bacterial Mat'. 'Cable' has an characteristic distribution.

5.2 Autoencoder training

To evaluate the effectiveness of the novel aspects of the proposed method, the autoencoder is trained to learn features in the dataset with/without colour attenuation correction (Section

3.1.1), rescaling (Section 3.1.2) and the georeference regularisation (Section 3.2.2). The dimensionality of $\boldsymbol{h}$ is set as 16 since the L_{rec} does not vary significantly even if larger values are used. The weights in the autoencoder are initialised with the original AlexNet trained with the ImageNet dataset. The mini-batch size is fixed as 256, and an Adam optimiser (Kingma and Ba, 2014) is used. The value for d_{max}, which limits $\mathrm{d}(\boldsymbol{y}_i, \boldsymbol{y}_j)$ in Equation 6, is set as 8.0 m. This is approximately 3.5 times the edge length of each image patch, where this value is appropriate for describing even large scale features as the images in this dataset constitute a dense grid with continuous cover between adjacent image pairs. A value of $\lambda = 1e5$ is used in Equation 7 for the georeference regularisation, and the number of epochs is set as 2,000. These parameters are empirically determined as values where both L_{rec} and L_{geo} in Equation 2 decrease monotonically during the training. The values of L_{rec} are not considerably different at the end of training regardless of whether the georeference regularisation is applied or not, demonstrating that the value used for λ does not over utilise the georeference information. When training without the georeference regularisation, each epoch contains all of the image patches in the data set. With the georeference regularisation, this is not guaranteed because of the unique sampling strategy described in Section 3.2.2. However, the large number of epochs ensures that the data is evenly sampled for autoencoder training. After autoencoder training, the latent representation $\boldsymbol{h}$ of each image $\boldsymbol{x}$ is obtained by processing $\boldsymbol{x}$ with the trained encoder without any rotating, shifting or addition of noise.

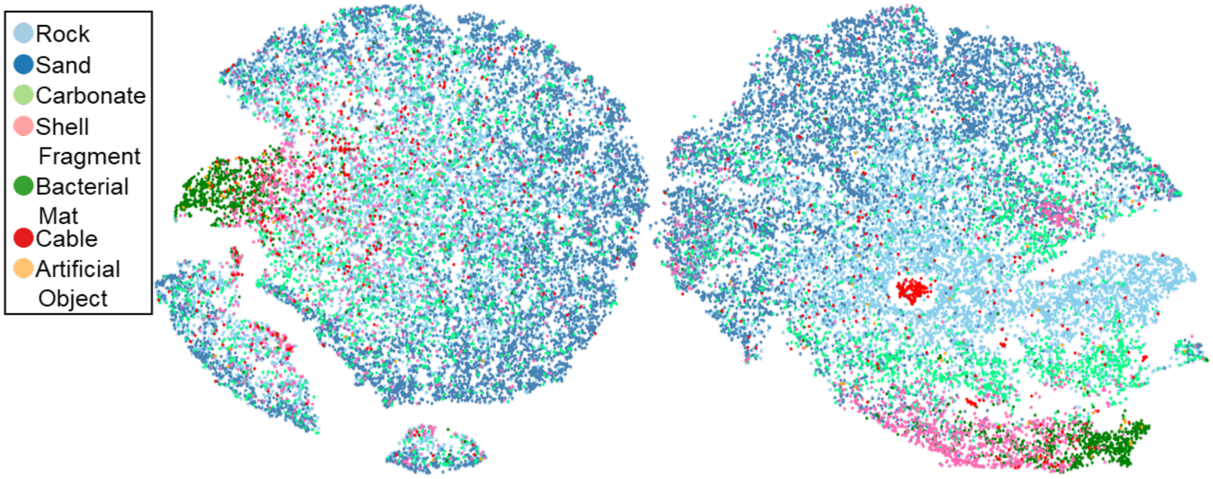
It can be said that a better feature extractor outputs smaller distances between samples for the same category and larger distances for the different categories in the latent representation space. Since this viewpoint is the same as internal evaluation metrics for clustering performance, the proposed feature learning can be evaluated through the metrics by inputting ground truth instead of clustering results. Silhouette score (Rousseeuw, 1987), Calinski and Harabasz score (CH) (Caliński and Harabasz, 1974) and Davies-Bouldin score (DB) (Davies and Bouldin, 1979) are used for the evaluation in this experiment. However, it should be noted that while these are the most widely used metrics to assess clustering performance, it has been reported that these existing metrics cannot completely take into account imbalances in datasets (Krawczyk, 2016). Although the dataset analysed in this work is highly skewed (see Figure 3c) these metrics are used since no standard methods are available that can overcome these limitations.

5.2.1 Results

The internal evaluation metrics corresponding to each training condition, labelled C_1 to C_9, are shown in Table 2. The latent representations $\boldsymbol{h}$ are normalised in each dimension as standard scores. Table 2 shows that the proposed georeference regularisation improves performance significantly for all metrics. The attenuation correction also increases performance, but the effectiveness of rescaling is less clear from these results alone. Figure 4 illustrates the distribution of expert annotations in the latent representation space $\boldsymbol{h}$ using t-SNE visualisation (Maaten and Hinton, 2008). Figure 4a and 4b are for autoencoders trained without/with the georeference regularisation, respectively (corresponding to C_4 and C_8 in Table 2). The most distinguishing characteristic of the resulting representation is that the distribution corresponding to 'Cable' forms an obvious cluster at the centre of Figure 4b with clear separation from other categories, while it is widely distributed in 4a without the georeference regularisation. This illustrates how the georeference regularisation allows the autoencoder to prioritise features that are common between images taken in close proximity to each other over features that would be learnt without this regularisation. The other ground truth categories also gather more closely in Figure 4b than in Figure 4a, as reflected by the improved evaluation metrics in Table 2.

5.3 Clustering

In this experiment, Normalised Mutual Information (NMI) (Estévez et al., 2009) is used for evaluation, following the previous works which also use imbalanced seafloor datasets for



(a) Without georeference regularisation (C_4 in Table 2)

(b) With georeference regularisation (C_8 in Table 2)

Figure 4: A t-SNE visualisation of the latent representation $\boldsymbol{h}$ for the expert annotations.

their experiments (Steinberg et al., 2011; Beijbom et al., 2012; Steinberg, 2013; Kaeli and Singh, 2015; Rao et al., 2017; Flaspohler et al., 2017). A NMI score is bounded between 0 (no mutual information) and 1 (perfectly correlated) and it is theoretically equivalent to the V-measure. A large NMI score means that the clustering result has a large amount of mutual information with the ground truth and corresponds to superior clustering performance. The numbers of clusters found using the non-parametric Bayesian method are not guaranteed to be the same as the number of categories used in human annotation. NMI is a favourable metric for this experiment because it does not require the targets to have the same number of clusters or categories. However, the result should be carefully investigated since it does not completely manage imbalanced datasets (Krawczyk, 2016).

5.3.1 Results

Table 2 shows the number of clusters and the NMI scores for each autoencoder. The proposed georeference regularisation improves the NMI scores by a factor of 1.6 (C_2 to C_6) to 2.2 (C_4 to C_8) compared to equivalent analysis without this regularisation. The modification of the loss function in Equation 7 is effective at controlling the training process so that it obtains solutions closer to human interpretation. This can be expected as it leverages an assumption about the scale of seafloor habitats and features, compensating for the limited image footprints that can be achieved underwater. When georeference regularisation is used, the proposed light attenuation correction improves the NMI score by 23 % (C_7 to C_9) and 38 % (C_6 to C_8) compared to a simple grey-world assumption. In contrast, no increase in performance is observed when the georeference regularisation was not used. A possible explanation is that when autoencoder training is regularised to the local neighbourhood, colour information is used in the latent space since adjacent images will tend to show a similar colour of seafloor. Under this assumption, any colour artefacts will degrade clustering performance. With no georeference regularisation, the autoencoder can easily end up being trained using images that are far apart, where the actual seafloor colour would tend to be more varied. In this scenario, the autoencoder would not prioritise colour information in the latent representation space, and so be less sensitive to differences in the colour correction method used. The results for rescaling are inconclusive with no significant difference observed in the NMI scores compared to equivalent experiments without rescaling. Although it is thought that rescaling would be effective for images of objects with consistent physical sizes, objects in the natural scenes that dominate the dataset vary widely in size, and so no significant gains in NMI performance could be achieved. The maximum NMI score achieved is not high (0.227), which is in part due to the impact of imbalanced categories as reported by (Krawczyk, 2016), and therefore a category based evaluation is also necessary.

Representative images from each cluster in the result with the highest NMI score (C_8 in Table 2) are shown in Figure 5. The relationship between the ground truth and this clustering result are shown in Table 3. To obtain a better understanding of each identified cluster, a treemap (Bruls et al., 2000) is shown in Figure 6, which allows the relative sizes of each cluster and their representative samples to be visualised simultaneously. To discuss the performance of the clustering result quantitatively, the confusion matrix is shown in Figure 7. Since the non-parametric Bayesian method optimises the number of clusters automatically, some clusters are manually merged based on the appearance of their representative samples so that the number of merged clusters corresponds to the number of ground truth categories. For example, cluster 'A', 'B' and 'F' are merged and regarded as 'Rock', and they appear at the first column of the confusion matrix as a single merged cluster. Since the number of 'Artificial Object' in ground truth is extremely small compared to other categories, the category is merged with 'Cable' and a 6×6 confusion matrix is shown. Table 4 shows the precision, recall and F_1-score for each ground truth category, based on the confusion matrix. The table shows that the proposed method can separate 'Rock', 'Sand', and 'Bacterial Mat' with F_1-scores greater than 0.6. The F_1-score scores for other categories are lower since they are subjective classes where there is ambiguity in human judgement. For example, 'Carbonate', which shows the lowest F_1-score (0.25), is often confused with 'Rock' and 'Sand' as shown in the confusion matrix. This result is reasonable because the density of both rock and carbonate distributions on sandy backgrounds vary on a continuum. Further verification to distinguish carbonates and rocks would require physical sampling, and it can be said that the clustering provides a meaningful result, in line with human interpretation, considering the inherent limitations of visual observation.

5.3.2 Habitat map

Habitat maps are useful as they summarise the geological and ecological patterns observed in a seafloor region. Figure 8 shows the habitat map obtained by plotting the semantic clusters generated by the proposed method. Figure 8b shows the result with the highest NMI score (C_8 in Table 2), and Figure 8a is the clustering result for the same pre-processing steps but without the georeference regularisation (C_4 in Table 2). Comparison with the distribution of ground truth in Figure 3b illustrates that the habitat map in Figure 8b can identify areas corresponding to categories such as 'Bacterial Mat', 'Shell Fragment', and 'Cable' more effectively than the habitat map in Figure 8a. Since these categories have geographic distribution patterns larger than the footprint of an image, the proposed

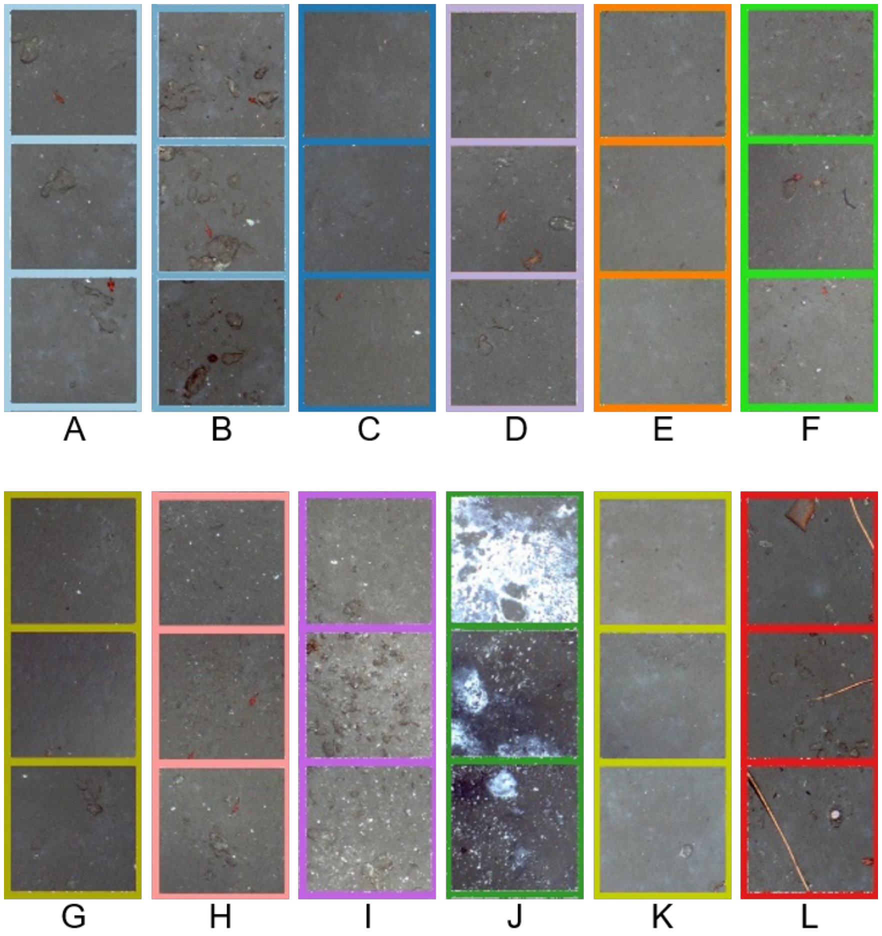


Figure 5: Representative samples of each cluster (C_8 in Table 2).

georeference regularisation is effective at extracting the features that are representative of these categories.

5.4 Image search

The performance of content based image search using Euclidean distance and cosine similarity are quantitatively evaluated by taking the average values of top-10 accuracy, defined as the rate of images retrieved with the same ground truth category as the query image for each ground truth category (Wu et al., 2013). Feature learning is achieved for the proposed autoencoder trained with/without the georeference regularisation and with/without rescal-

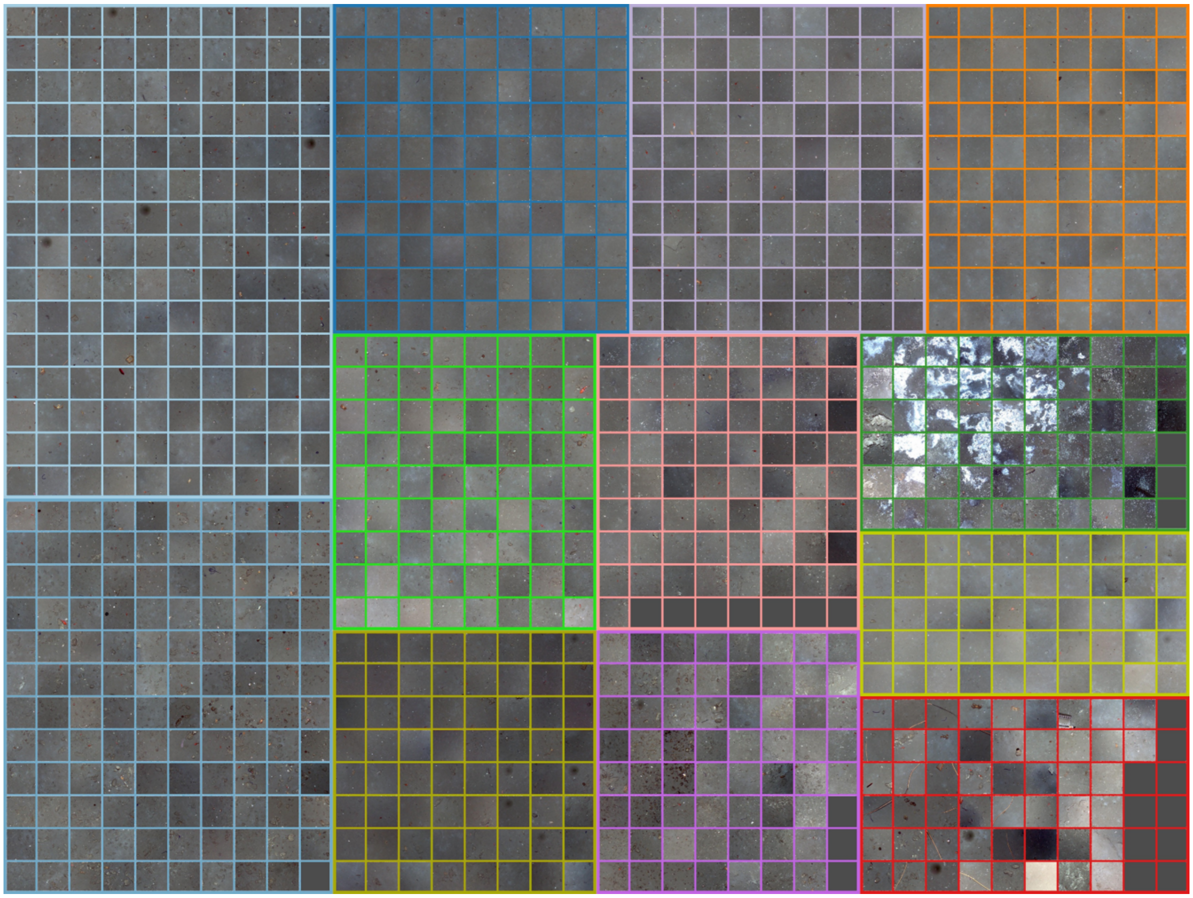
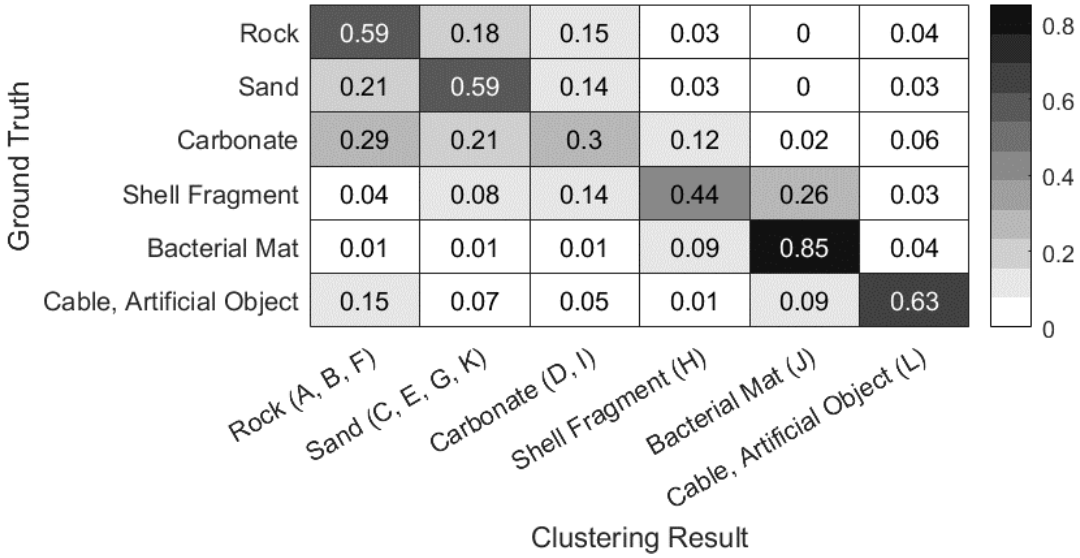


Figure 6: Visualisation of the size of each cluster (C_8 in Table 2) using a tree-map representation. The same colours as Figure 5 are assigned for each cluster and the areas are proportional to the number of image patches in each cluster.

ing. These correspond to autoencoders labelled C_4, C_5, C_8 and C_9, respectively in Table 2.

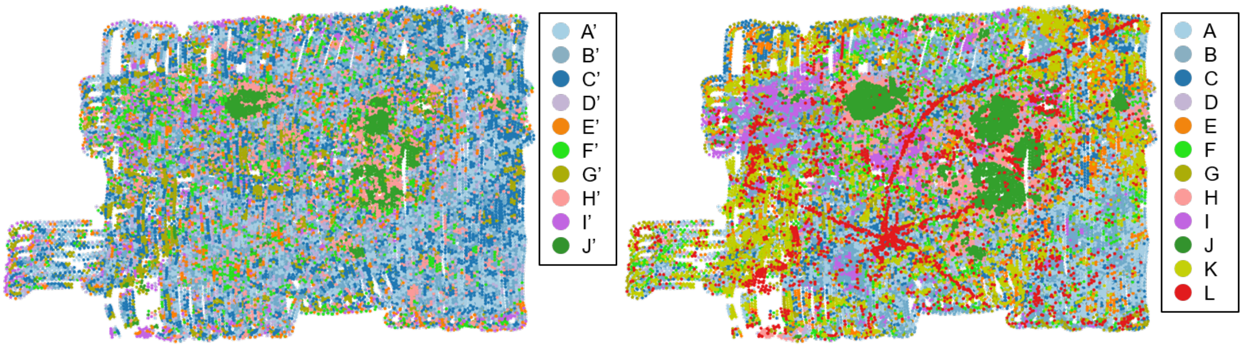
5.4.1 Results

The results in Table 5 show that the proposed georeference regularisation improves the performance in every category, with an overall increase in accuracy across all categories from 47 % to 59 %. The largest improvement is for 'Cable', from 10 % to 15 % accuracy without the georeference regularisation to a maximum value of 53.7 % with the regularisation. Although rescaling does not influence the accuracy of most categories, the accuracy of 'Cable' improves noticeably from 40 % to 52 %. The results indicate that rescaling is only effective when



	Rock (A, B, F)	Sand (C, E, G, K)	Carbonate (D, I)	Shell Fragment (H)	Bacterial Mat (J)	Cable, Artificial Object (L)
Rock	0.59	0.18	0.15	0.03	0	0.04
Sand	0.21	0.59	0.14	0.03	0	0.03
Carbonate	0.29	0.21	0.3	0.12	0.02	0.06
Shell Fragment	0.04	0.08	0.14	0.44	0.26	0.03
Bacterial Mat	0.01	0.01	0.01	0.09	0.85	0.04
Cable, Artificial Object	0.15	0.07	0.05	0.01	0.09	0.63

Figure 7: Confusion matrix between ground truth categories and the unsupervised clustering result using C_8 in Table 2. Some clusters and ground truth categories are manually merged based on the appearance of representative images. The values in the matrix are normalised, and diagonal elements correspond to the recall values in Table 4.



(a) Without georeference regularisation (C_4 in Table 2)

(b) With georeference regularisation (C_8 in Table 2)

Figure 8: Habitat maps based on unsupervised clustering result. The clusters corresponding to 'Bacterial Mat' ('J'), 'Shell Fragment' ('H') and 'Cable' ('L') appear clearly in Figure 8b. The results demonstrate that the proposed georeference regularisation enhances clustering performance over wide spatial distributions.

learning the features of objects with a fixed size, which are in this case 'Cable' and 'Artificial Object', improving their accuracy scores by a factor of 1.39 and 1.16 when rescaling is applied, while the other categories that have a large amount of natural variability show no improvement. This fact shows an important characteristic of the autoencoder, which prioritises meaningful features automatically. If scale variant features are found to be better descriptors, then the autoencoder will prioritise this property. The results show that for some categories such as 'Cable' and 'Artificial object', physical scaling can pose benefits. For many categories, the difference is negligible, illustrating that the autoencoder does no need to make such explicit assumptions.

Regarding the similarity metrics, Equation 7 for the proposed georeference regularisation assumes that the similarities of $\boldsymbol{h}$ are related to a t-distribution, which is derived from Euclidean distance ($\|\boldsymbol{h}_i - \boldsymbol{h}_j\|$). However, interpretation of the autoencoder learnt feature space is challenging, and the results indicate that Euclidean distance and cosine similarity are almost equivalent for the dataset used in this study.

Identifying the location of similar images is important to interpret spatial patterns of interesting targets. In comparison to clustering, which interprets the representative patterns in the dataset, content based image search can generate target specific distribution maps using the same unsupervised feature space. This can be useful when specific targets within a cluster are of interest, or where the target is rare and so does not form an independent cluster. Since the query target is known, the autoencoder and similarity metric used can be tailored to the type of object, where for human-made objects such as 'Cable' and 'Artificial Object', the georeference regularisation with rescaling and cosine similarity provided the best performance.

5.4.2 Utility map

The utility maps in Figure 9 show some results of image search and the locations of images that have a similar appearance. Figure 9a shows the result of a bacterial mat image search. On the whole, the areas with high similarity in the utility map show similar distributions 'Bacterial Mat' in the ground truth (Figure 3b). Since the similarity scores vary continuously, the result is useful for analysing small differences between images which are categorised as ‘Bacterial Mat’. Figure 9b shows the result when a typical image of a cable is chosen as a query. The image search successfully extracts cables deployed in this area, and the utility map shows the distribution of cables more clearly than the clustering result (Figure 8b). Features that are small, more sparsely distributed and few in numbers such as seafloor infrastructures and crabs are less likely to form independent clusters using the non-parametric Bayesian method (Figure 5). However, relatively minor categories such as these can be effectively found using content based image search, where Figure 9c and 9d show the different distributions in this area. These utility maps can form a useful tool for rapidly understanding complex, multi-parameter spatial patterns in georeferenced imagery. An important point is that the distributions in Figure 9 are spread widely and are not limited within the neighbouring area of the query images. This fact confirms that the proposed georeference loss function in Equation 7 allows meaningful features to be extracted from the images themselves without over-regularising the results of the image search. Looking more closely at Figure 9d shows that some of the results of the search do not include crabs, but instead contain other types of benthic organisms. To obtain a more precise result for these categories, supervised learning based approaches are more appropriate (Walker et al., 2019). The proposed content based image search may be useful to reduce the effort requires for manual annotation by filtering out candidate images that are more likely to contain the targets of interests.

6 Conclusion

This paper has described a novel, unsupervised feature learning method for semantic interpretation of seafloor visual imagery and applied it to a seafloor dataset consisting of more than 12,000 images. Although this work has focused on subsea visual mapping applications, the methods are equally applicable to terrestrial georeferenced imaging applications such as drone and satellite imaging. The study has demonstrated that:

- Autoencoders implemented using deep convolutional neural networks form an effective and generic method to learn features in seafloor visual imagery.

- The use of georeference regularisation implemented using the Kullback-Leibler divergence criteria leads to a factor of two improvement in the retrieval of information from the seafloor images analysed in this work. This includes geomorphological and ecological patterns that occur on spatial scales larger than a single image frame.

- Correction of colour information in seafloor imagery using physics based techniques improves information retrieval rates by more than 20% when the georeference regularisation is used.

- Correction for spatial scale and distortion of images prior to feature learning improves the recognition of artificial structures on the seafloor. However, for natural objects that exhibit significant variability in size and shape, the gains in performance achieved through scale correction are minimal.

- Non-parametric Bayesian unsupervised clustering and content-based image search can be implemented directly on features learnt by the proposed autoencoder for effective semantic interpretation and visualisation of spatial patterns in seafloor visual mapping data.

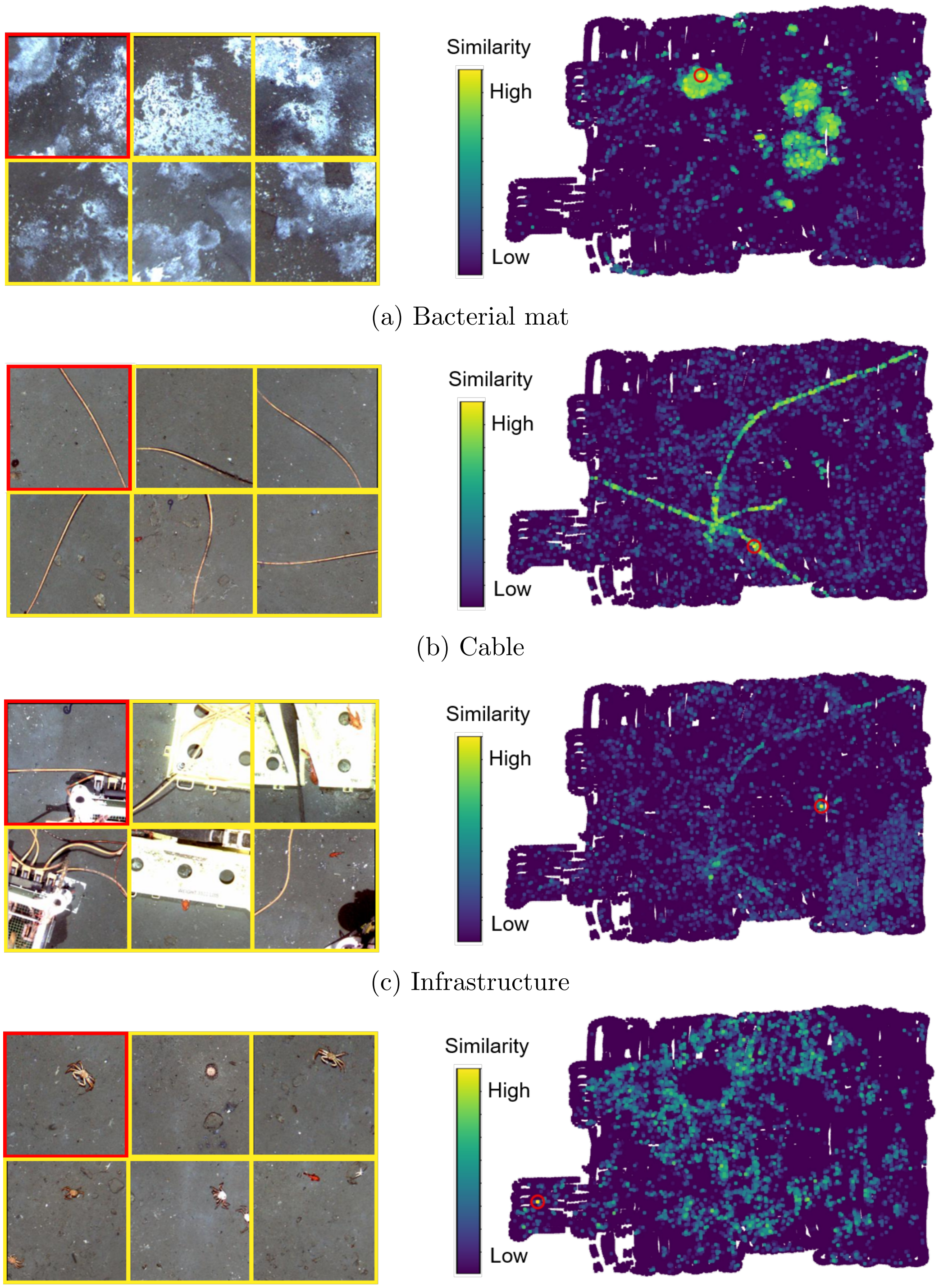
- No significant difference was found between the performance of content-based retrieval of images when using Euclidean distance and cosine similarity metrics in the latent feature space.

The images used in this study can be accessed via SQUIDLE+ (`http://soi.squidle.org`) as (Campaign: `fk180731[ID:53]`, deployment: `20180804_093404_20180804_143258_20180805_123456_20180809_083837_ae2000f_sx3[ID:711]`). The expert annotations for the images can be accessed at `SHR_AE2000_3000samples[ID:80]` and `SHR_AE2000_`

1000samples[ID:74] in uos-oplab-fk180731[ID:9] datasets. The colour correction and undistortion methods used to pre-process images in this work can be found on https://github.com/ocean-perception/oplab_pipeline/tree/master/correct_images.

Acknowledgements

This work was carried out under the UK Natural Environment Research Council's Biocam project NE/P020887/1, part of the Oceanids Marine Sensor Capital program. The data used in this work was collected during the Schmidt Ocean Institute's FK180731 #Adaptive Robotics campaign, with support from the Japanese Government's Zipangu in the Ocean Strategic Innovation Program. The authors thank Kazunori Nagano, Tetsuo Koike and Harumi Sugimatsu of the Institute of Industrial Science at the University of Tokyo, Japan, Adrian Bodenmann, Jennifer Walker, and Jim Wei Lim of the University of Southampton, UK, Gabriel Oliver and Miquel Massot-Campos of the SRV group at the University of the Balearic Islands, Spain, and the crew of the RV Falkor for their operational support deploying the AUV ae2000f to collect the imagery used in this work. The authors thank the Marine Systems group at the Australian Centre for Field Robotics, University of Sydney, for providing access to their SLAM and 3D reconstruction pipelines used to generate the navigational solution and mosaics used in this work, Jin Wei Lim and Jose Cappelletto for their assistance in annotating the dataset. The IRIDIS High Performance Computing Facility and associated support services at the University of Southampton were used in the completion of this work.



(a) Bacterial mat

(b) Cable

(c) Infrastructure

(d) Crab

Figure 9: Image search result. Red frame images are query and yellow frame images are Top 5 similarity images. The maps on the right show the similarity distributions between the queries and all the other images in the dataset. The red circles in the utility maps show the locations of the query images.

References

Aljalbout, E., Golkov, V., Siddiqui, Y., Strobel, M., and Cremers, D. (2018). Clustering with deep learning: Taxonomy and new methods. *arXiv preprint arXiv:1801.07648*.

Beijbom, O., Edmunds, P. J., Kline, D. I., Mitchell, B. G., and Kriegman, D. (2012). Automated annotation of coral reef survey images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1170–1177. IEEE.

Bewley, M., Friedman, A., Ferrari, R., Hill, N., Hovey, R., Barrett, N., Marzinelli, E. M., Pizarro, O., Figueira, W., Meyer, L., et al. (2015a). Australian sea-floor survey data, with images and expert annotations. *Scientific data*, 2:150057.

Bewley, M., Nourani-Vatani, N., Rao, D., Douillard, B., Pizarro, O., and Williams, S. B. (2015b). Hierarchical classification in auv imagery. In *Field and service robotics*, pages 3–16. Springer.

Blei, D. M., Jordan, M. I., et al. (2006). Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143.

Bruls, M., Huizing, K., and Van Wijk, J. J. (2000). Squarified treemaps. In *Data visualization 2000*, pages 33–42. Springer.

Bryson, M., Johnson-Roberson, M., Pizarro, O., and Williams, S. B. (2013). Colour-consistent structure-from-motion models using underwater imagery. *Robotics: Science and Systems VIII*, page 33.

Buchsbaum, G. (1980). A spatial processor model for object colour perception. *Journal of the Franklin institute*, 310(1):1–26.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Cheng, G., Zhou, P., and Han, J. (2016). Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415.

Cowles, T., Delaney, J., Orcutt, J., and Weller, R. (2010). The ocean observatories initiative: Sustained ocean observing across a range of spatial scales. *Marine Technology Society Journal*, 44(6):54–64.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

Estévez, P. A., Tesmer, M., Perez, C. A., and Zurada, J. M. (2009). Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 20(2):189–201.

Flaspohler, G., Roy, N., and Girdhar, Y. (2017). Feature discovery and visualization of robot mission data using convolutional autoencoders and bayesian nonparametric topic models. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE.

Hasan, R. C., Ierodiaconou, D., Laurenson, L., and Schimel, A. (2014). Integrating multi-beam backscatter angular response, mosaic and bathymetry data for benthic habitat mapping. *Plos one*, 9(5):e97339.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Jaffe, J. S. (1990). Computer modeling and the design of optimal underwater imaging systems. *IEEE Journal of Oceanic Engineering*, 15(2):101–111.

Johnson-Roberson, M., Pizarro, O., Williams, S. B., and Mahon, I. (2010). Generation and visualization of large-scale three-dimensional reconstructions from underwater robotic surveys. *Journal of Field Robotics*, 27(1):21–51.

Kaeli, J. W. and Singh, H. (2015). Online data summaries for semantic mapping and anomaly detection with autonomous underwater vehicles. In *OCEANS 2015-Genova*, pages 1–7. IEEE.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Langenkämper, D., Zurowietz, M., Schoening, T., and Nattkemper, T. W. (2017). Biigle 2.0-browsing and annotating large marine image collections. *Frontiers in Marine Science*, 4:83.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Mahmood, A., Bennamoun, M., An, S., Sohel, F. A., Boussaid, F., Hovey, R., Kendrick, G. A., and Fisher, R. B. (2018). Deep image representations for coral image classification. *IEEE Journal of Oceanic Engineering*, 44(1):121–131.

Mahon, I., Williams, S. B., Pizarro, O., and Johnson-Roberson, M. (2008). Efficient view-based slam using visual loop closures. *IEEE Transactions on Robotics*, 24(5):1002–1014.

Maki, T., Kume, A., Ura, T., Sakamaki, T., and Suzuki, H. (2010). Autonomous detection and volume determination of tubeworm colonies from underwater robotic surveys. In *OCEANS'10 IEEE SYDNEY*, pages 1–8. IEEE.

Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., and Long, J. (2018). A survey of clustering with deep learning: From the perspective of network architecture. *IEEE Access*, 6:39501–39514.

Ojala, T., Pietikäinen, M., and Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):971–987.

Paull, L., Saeedi, S., Seto, M., and Li, H. (2013). Auv navigation and localization: A review. *IEEE Journal of Oceanic Engineering*, 39(1):131–149.

Pizarro, O., Rigby, P., Johnson-Roberson, M., Williams, S. B., and Colquhoun, J. (2008). Towards image-based marine habitat classification. In *OCEANS 2008*, pages 1–7. IEEE.

Rao, D., De Deuge, M., Nourani-Vatani, N., Williams, S. B., and Pizarro, O. (2017). Multimodal learning and inference from visual and remotely sensed data. *The International Journal of Robotics Research*, 36(1):24–43.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Steinberg, D. (2013). An unsupervised approach to modelling visual data. *PhD Thesis, University of Sydney, Australia.*

Steinberg, D., Friedman, A., Pizarro, O., and Williams, S. B. (2011). A bayesian nonparametric approach to clustering data from underwater robotic surveys. In *International Symposium on Robotics Research*, volume 28, pages 1–16.

Thornton, B., Asada, A., Bodenmann, A., Sangekar, M., and Ura, T. (2012). Instruments and methods for acoustic and visual survey of manganese crusts. *IEEE Journal of Oceanic Engineering*, 38(1):186–203.

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

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408.

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

Wu, P., Hoi, S. C., Xia, H., Zhao, P., Wang, D., and Miao, C. (2013). Online multimodal deep similarity learning with application to image retrieval. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 153–162. ACM.

Xie, J., Girshick, R., and Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487.

Yang, B., Fu, X., Sidiropoulos, N. D., and Hong, M. (2017). Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3861–3870. JMLR. org.

Yang, J., Yu, K., Gong, Y., and Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *2009 IEEE Conference on computer vision and pattern recognition*, pages 1794–1801. IEEE.

Zurowietz, M., Langenkämper, D., Hosking, B., Ruhl, H. A., and Nattkemper, T. W. (2018). Maia—a machine learning assisted image annotation method for environmental monitoring and exploration. *PloS one*, 13(11).

Tables

Table 1: Southern Hydrate Ridge dataset description. The dataset was collected using the AUV ae2000f during the Schmidt Ocean Institute's FK180731 #Adaptive Robotics campaign. A total of 62,875 image patches are cropped from the original images and used for autoencoder training.

Date	4/8/2018 - 9/8/2018 (4 dives)
Location	Southern Hydrate Ridge (N44.6°, W125.1°)
Seafloor Depth [m]	765 - 785
Altitude [m]	5.0 - 7.0
AUV	ae2000f
Camera System	SeaXerocks (Thornton et al., 2016)
Number of Images	12,575
Original Image Resolution	1280 × 1024
Original Space Resolution [mm/pixel]	approx. 5 - 7
FoV (Underwater)	68° × 57°
Total Area Covered [m^2]	118,000
Mapping Method	Dense grid with 30 % overlap between images
Ground Truth Categories	7 categories as shown in Figure 3c
Annotation Platform	SQUIDLE+

Table 2: Evaluation results of the proposed feature learning and clustering. The check (✓) and dash (-) marks illustrate whether each preprocessing or regularisation is applied or not, respectively. Each condition is labelled from C_1 to C_9 and these labels are referred to in the later sections. The best scores (the lowest for DB and the highest for Silhouette, CH and NMI) are shown in bold.

Condition Label	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
Pixel-wise Normalisation	-	✓	✓	✓	✓	✓	✓	✓	✓
Attenuation Correction	-	-	-	✓	✓	-	-	✓	✓
Rescaling	-	-	✓	-	✓	-	✓	-	✓
Georeference Regularisation	-	-	-	-	-	✓	✓	✓	✓
Silhouette	-0.020	-0.026	-0.025	-0.004	-0.019	0.003	0.010	0.032	**0.035**
CH	253	160	476	290	272	622	696	**1078**	772
DB	18.7	14.6	10.2	8.0	7.3	5.1	4.7	**3.4**	3.5
Num. of Clusters	15	10	9	10	8	15	13	12	11
NMI	0.078	0.101	0.111	0.103	0.104	0.165	0.176	**0.227**	0.216

Table 3: Contingency table of the clustering result. Rows and columns correspond to the ground truth and the clustering result using C_8 in Table 2, respectively.

	A	B	C	D	E	F	G	H	I	J	K	L	Total
Rock	1,741	2,039	436	425	229	776	489	206	755	22	207	335	7,660
Sand	1,096	62	1,448	937	1,237	298	660	207	3	3	646	184	6,781
Carbonate	161	116	99	305	74	317	147	233	299	43	103	117	2,014
Shell Fragment	15	3	20	142	13	30	35	504	17	305	27	40	1,151
Bacterial Mat	3	1	2	6	0	2	1	64	1	639	1	31	751
Cable	25	17	8	6	3	8	8	4	9	28	2	226	344
Artificial Object	2	2	0	2	0	5	3	1	3	6	1	14	39
Total	3,043	2,240	2,013	1,823	1,556	1,436	1,343	1,219	1,087	1,046	987	947	18,740

Table 4: Precision, recall and F_1-score for the clustering result using C_8 in Table 2. The same cluster merging as in Figure 7 is applied. The total accuracy across all categories is 0.56.

Category	Precision	Recall	F_1-score
Rock	0.68	0.59	0.63
Sand	0.68	0.59	0.63
Carbonate	0.21	0.30	0.25
Shell Fragment	0.41	0.44	0.43
Bacterial Mat	0.61	0.85	0.71
Cable, Artificial Object	0.25	0.63	0.36

Table 5: Mean top 10 accuracy of search in each category (%). 'l2' and 'cos' in the similarity metric correspond to the euclidean distance and cosine similarity, respectively. As with the clustering result in Table 2, the proposed georeference regularisation significantly improves the accuracy scores, especially for 'Cable' which has a characteristic spatial distribution.

Condition in Table 2	C_4		C_5		C_8		C_9	
Georeference Regularisation	-	-	-	-	✓	✓	✓	✓
Rescaling	-	-	✓	✓	-	-	✓	✓
Similarity Metric	l2	cos	l2	cos	l2	cos	l2	cos
Rock	53.1	51.5	52.8	52.6	**66.6**	**66.6**	63.2	65.5
Sand	56.5	57.0	55.2	55.3	63.9	**64.9**	63.9	62.0
Carbonate	16.7	16.1	15.1	15.0	**27.8**	27.3	25.6	24.1
Shell Fragment	19.0	18.8	16.0	15.7	**43.2**	41.2	39.3	38.7
Bacterial Mat	61.1	62.9	58.3	58.2	71.0	**72.1**	69.8	70.8
Cable	11.0	11.8	15.6	13.5	40.2	39.7	51.3	**53.7**
Artificial Object	3.8	4.4	4.1	4.4	4.9	4.1	6.4	**6.7**
NMI in Table 2 (for Reference)	0.10		0.10		**0.23**		0.22	