UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering and Physical Sciences
School of Engineering

Leveraging Domain Knowledge in Machine Learning for Seafloor Image Interpretation

by

Takaki Yamada

BEng, MSc

ORCiD: 0000-0002-5090-7239

A thesis for the degree of
Doctor of Philosophy

October 2021

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences
School of Engineering

Doctor of Philosophy

Leveraging Domain Knowledge in Machine Learning for Seafloor Image Interpretation

by Takaki Yamada

This thesis develops a method to incorporate domain knowledge into modern machine learning techniques when interpreting large volumes of robotically obtained seafloor imagery. Deep learning has the potential to automate tasks such as habitat and animal recognition in marine monitoring. However, the large input of human effort needed to train the models is a bottleneck, and this motivates research into methods that reduce the human input requirements. This research investigates how metadata gathered during robotic imaging surveys, such as the location and depth information, can be used to constrain learning based on expected metadata patterns. Two self-supervised representation learning methods are developed. The first uses deep learning convolutional autoencoders that leverage location and depth information to impose soft constraints based on the assumption that images taken in physically nearby locations or similar depths are more likely to share important features than images that are taken far apart or at different depths. The second method uses contrastive learning techniques where three-dimensional position information acts as a hard constraint on representation learning. Self-supervision allows both methods to be implemented on a per-dataset basis with no human input. The representations learned can be used for different downstream interpretation tasks, where applications to unsupervised clustering and representative image identification (i.e. as tasks that do not require any human input) are demonstrated alongside content based retrieval and semi-supervised learning based classification (i.e. tasks that require a relatively small amount of human input). Three real-world seafloor image datasets are analysed. These consist of ~150k seafloor images taken over 16 dives by two different Autonomous Underwater Vehicles (AUVs) along sparse and dense survey trajectories spanning a seafloor depth range of 20 to 780 metres. The results show relative accuracy gains of 7 to 15 % compared to other state of the art self-supervised representation learning and supervised learning techniques, and achieves equivalent accuracy for an order of magnitude less human input. This offers a practical solution to the problem of training deep-learning neural networks in application domains where there is limited transfer of learning across datasets.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:
   Yamada, T., Prügel-Bennett, A., and Thornton, B. (2021b). Learning features from georeferenced seafloor imagery with location guided autoencoders. *Journal of Field Robotics*, 38(1):52–67
   Yamada, T., Massot-Campos, M., Prügel-Bennett, A., Williams, S. B., Pizarro, O., and Thornton, B. (2021a). Leveraging metadata in representation learning with georeferenced seafloor imagery. *IEEE Robotics and Automation Letters*, 6(4):7815–7822

Signed:..... Date:4 October 2021.....

Acknowledgements

This PhD thesis would not have been possible without the support of a huge number of people. First and foremost, I am eternally grateful to my supervisor, Prof. Blair Thornton, for invaluable advice and continuous support throughout this PhD programme. The discussions were always vital and impressive for me to think about what is really necessary for real-world applications in our research domain and what we can reasonably achieve with current technologies. I would also like to thank my co-supervisor, Prof. Adam Prügel-Bennett, for giving me helpful advice for applying machine learning techniques to new domains.

I would like to thank all the members of our research group: Adrian Bodenmann, Dr Miquel Massot-Campos, Dr Subhra Kanti Das, Jonathan Boschen-Rose, Jenny Walker, Jin Wei Lim, David Stanley, Jose Cappelletto and Emma Curtis, for sharing their expertise and giving me invaluable advice. Without their support, it would not have been possible to accomplish this work in marine robotics, which was a completely new research topic for me.

My deep appreciation goes out to the members of the Maritime Engineering Group. I was always stimulated by the discussions and talks with them on various research topics related to the ocean. I also appreciate Dr Veerle A.I. Huvenne and other National Oceanography Centre members who study seafloor and habitat mapping. The survey cruise DY108/109 we went on together was one of the most impressive experiences in my PhD programme. In addition, I would like to thank Prof. Stefan B. Williams and Dr Oscar Pizarro in Australian Centre for Field Robotics of the University of Sydney for their valuable guidance during my work.

And finally, I owe my greatest thanks to my wife Akane and my son Sotaro. Without their tremendous understanding and encouragement, it would have been impossible for me to complete my programme.

Definitions and Abbreviations

Definition

$f(\cdot)$	Encoder or feature descriptor
$f_{cls}(\cdot)$	Conventional (non-deep learning) classifier
$f_{dlc}(\cdot)$	Deep learning classifier
$g(\cdot)$	Decoder
g_{east}	Easting georeference
g_{north}	North georeference
g_{depth}	Depth georeference
$\boldsymbol{h}$	Latent representation or feature vector
$\mathrm{KL}(\cdot\|\|\cdot)$	Kullback–Leibler divergence
L_{rec}	Autoencoder reconstruction loss
M	Number of annotations used for training
n	Number of images in a dataset
n^*	Number of images loaded per mini-batch
$\boldsymbol{x}$	Image tensor
$\boldsymbol{x_r}$	Autoencoder reconstructed image tensor
y	Ground truth
$\hat{y}$	Class prediction
$\boldsymbol{z}$	Latent representation vector for SimCLR compressed space

Abbreviations

AUV	Autonomous Underwater Vehicle
CNN	Convolutional Neural Network
k-NN	k-Nearest Neighbour
LBP	Local Binary Pattern
NMI	Normalised Mutual Information
PCA	Principal Component Analysis
RF	Random Forest
ScSPM	Sparse coding Spatial Pyramid Matching
SVM	Support Vector Machine
t-SNE	t-distributed Stochastic Neighbour Embedding

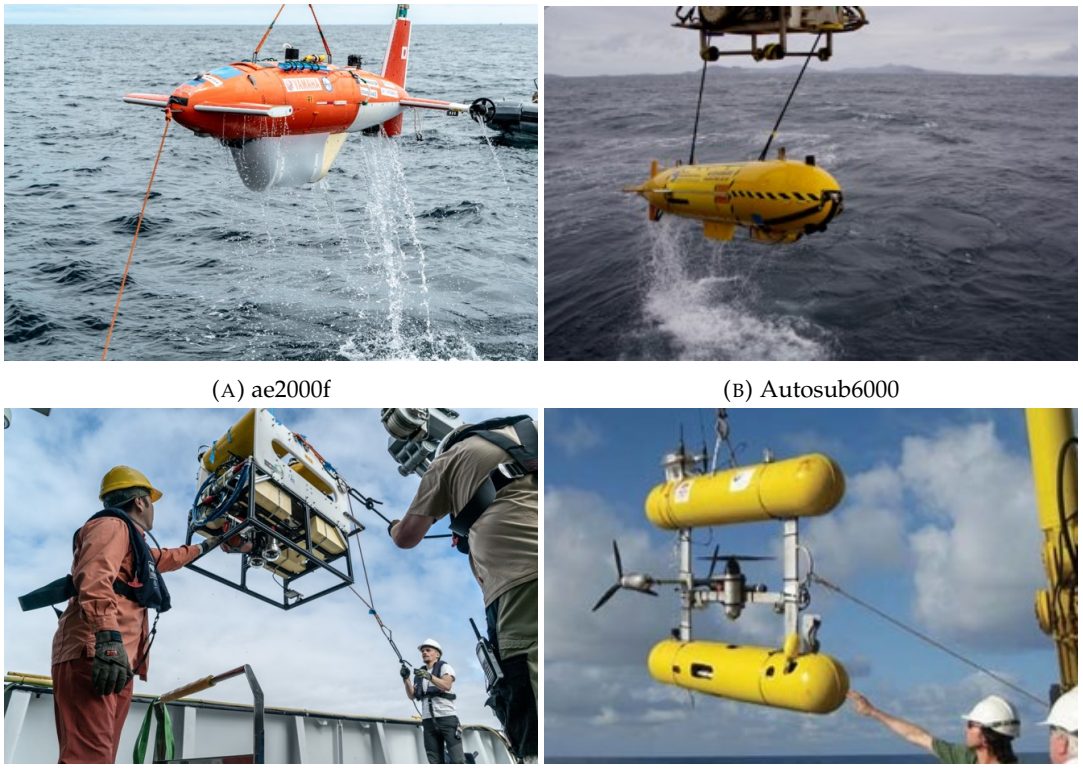
Chapter 1

Introduction

1.1 Motivation

Autonomous Underwater Vehicles (AUVs) equipped with acoustic sensors and camera systems can map large areas of the seafloor at high-resolution. Data gathered by these systems is used for a wide range of scientific and commercial purposes. However, our ability to gather data is not currently matched by our ability to interpret it and efficiently generate the information needed for marine monitoring, conservation and subsea inspection. This forms a bottleneck that limits the usefulness of mobile robotic platforms.

In particular, the large scale colour image datasets that are now routinely gathered by camera equipped AUVs present a major challenge for analysis. During a typical seafloor imaging survey, AUVs collect tens to hundreds of thousands of seafloor images during their dives (Figure 1.1). The high-resolution colour images they gather are informative for estimating the distribution of seafloor habitats and substrates. However, interpreting these images often requires domain specific expertise, and so only a small number of human experts are able to perform this task. As a result, our ability to interpret the data cannot keep up with the influx of data complexity and availability, limiting our capacity to build knowledge and insight. Modern machine learning techniques have the potential to significantly speed up the interpretion of these images; however, the domain-specific characteristics of seafloor imaging pose significant challenges for direct applications of these techniques. In particular, the absence of large annotated image datasets, which are available in other domains, i.e. general objects (ImageNet (Deng et al., 2009a), COCO (Lin et al., 2014), Pascal VOC (Everingham et al., 2015)), satellite (SpaceNet (Van Etten et al., 2018)) or autonomous driving (KITTI (Geiger et al., 2013)), and are essential for their training, limits the direct application of these techniques to the marine domain. However, machine learning, especially modern deep learning techniques, can be still useful to achieve a better understanding



(A) ae2000f (B) Autosub6000

(C) TUNA-SAND (D) Sirius

FIGURE 1.1: Camera equipped AUVs. (A) ae2000f AUV equipped with the SeaXerocks camera system (Thornton et al., 2016). (B) Autosub6000 AUV equipped with the BioCam camera system (West et al., 2020). (C) TUNA-SAND AUV equipped with a visual camera system (Nishida et al., 2013). (D) Sirius AUV equipped with a visual camera system (Williams et al., 2010).

of large datasets that humans cannot inspect thoroughly, and this is the topic that will be explored in this thesis.

For machine learning based image interpretation, the following three topics have been intensely studied: (1) image classification, where a single or a few attribute labels are given to each image; (2) object detection, where each object of interest is detected and located in images; and (3) image segmentation, where a class is attributed to individual pixels in the images. Although all three categories are potentially useful for seafloor image interpretation, this work focuses on learning features, or latent representations, to efficiently describe images, and the application of these representations to seafloor image classification. Generally, classification is achieved using supervised machine learning models that require a training dataset where the data inputs are annotated using the same labelling scheme as the desired prediction targets, or outputs. This requires single or multiple attribute labels to be manually given by humans. However, although the large training datasets needed to train deep learning classifiers are typically not available for seafloor image classification tasks, other forms of potentially useful data are available for robotically collected images. For example, AUVs used for seafloor imaging surveys are typically equipped with position measuring systems for autonomous

navigation, and so the geolocation of each image observation is generally available. In addition, other environmental parameters such as water temperature, salinity and pH are often also measured. Since these metadata are potentially correlated with attributes of the observed seafloor, it important to consider how these relationships could be leveraged for machine learning based seafloor interpretation.

This thesis investigates how commonly gathered metadata can be used to guide learning for image classification to significantly reduce, and where possible eliminate completely the need for human annotations. Figure 1.2 illustrates the concept of leveraging metadata in image representation learning that will be explored. The objective here is training a function $f(\cdot)$ that maps the images $\boldsymbol{x}$ to a set of descriptive features, known as their latent representations $\boldsymbol{h}$. In this simple example, the images can be split into three classes that humans might be interested in, i.e. house, camper van and car. When $f(\cdot)$ is trained only on images themselves so that similar appearance images have similar latent representations, three clusters corresponding to the three classes would appear in the latent space. However, if $f(\cdot)$ is trained considering 'the number of beds in the illustrated target' metadata, the house and camper van classes, which are slightly similar in their appearance, would be mapped more closely in the latent representation space. In another case, if 'the maximum speed the target can travel at' metadata are leveraged, the camper van class would be mapped more closely to the car class in the latent representation space. The preferable $f(\cdot)$ depends on the application, i.e. the former is useful for selecting places to live, and the latter is useful for considering how to get from one place to another. An important point here is that the modified image latent representations can be predicted without their corresponding metadata once $f(\cdot)$ is trained. The metadata is leveraged to help identify important image features that it should learn for the target applications only during training. Once training is complete, this information is embedded in $f(\cdot)$, and when an image is presented, the function will look for these prioritised image features and map the images to the appropriate location in the latent representation space without the need to provide the metadata. This example illustrates how metadata can be used to introduce specific domain knowledge into machine learning.

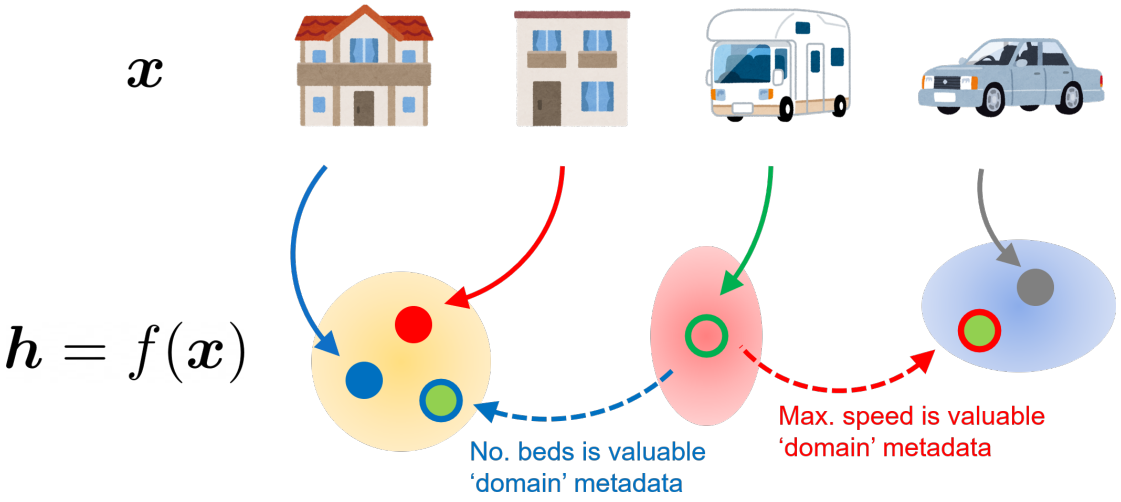


FIGURE 1.2: Simple example of metadata leveraging. $f(\cdot)$ is the function that maps images $\boldsymbol{x}$ to their latent representation $\boldsymbol{x}$. When three classes, i.e. house, camper van and car, are included in the set of $\boldsymbol{x}$, an ideal $f(\cdot)$ maps $\boldsymbol{x}$ into three clusters corresponding to the three classes, considering only their appearance. If metadata 'number of beds' attributes are given as valuable metadata, $f(\cdot)$ which maps the representations of house class and camper van class close together can be regarded as more useful.

The metadata obtained with seafloor images by AUVs can be leveraged in the same way as this simple example; however, there are limited examples of research that attempts to implement this idea. For example, Rao et al. (2014, 2017) use depth metadata to aid learning for seafloor imagery. However, the method is tailored to depth and does not consider how other types of metadata could be used. Since various types of potentially useful metadata are available, this PhD work aims to develop a versatile framework that allows a wide range of metadata to be flexibly leveraged in seafloor visual image interpretation.

1.2 Problem statement

A major challenge for automating seafloor image interpretation, is the sensitivity of the appearance of the seafloor in images to the underwater environment and observation conditions used when the images were taken. The strong wavelength dependent attenuation of light in water is sensitive to the turbidity and composition of the medium it passes through, and the hardware needed to take images in water at oceanic depths is complex and is not standardised. In addition to having cameras in bespoke sealed pressure-resistant housings that can introduce variable degrees of image distortion, seafloor imaging requires submersibles equipped with illumination sources that can maintain low altitudes (typically a few metres) over natural terrains, where protrusions are often large relative to the imaging altitude and the ability to maintain a target range differs significantly between platforms. Although the mechanisms that degrade underwater image appearance are researched and can be partially compensated, these combined factors reduce the consistency of how objects and scenes on the seafloor look in images, which in turn limits the transferability of learning across different marine image datasets.

Other limitations include the high cost of obtaining seafloor images, the significant domain expertise needed to accurately annotate seafloor imagery and the relatively small size of the marine imaging community compared to more general applications (e.g. face identification, autonomous driving and medical imaging). These factors mean that comprehensive datasets suitable for training machine learning models are unlikely to be built in the foreseeable future. At the same time, the large input of human effort needed to train effective deep learning models using traditional supervised learning methods is unlikely to be justified to train models on a per-dataset basis in most monitoring applications. These problems motivate investigations into how the amount of human input needed can be reduced without degrading the quality of automated interpretation outputs.

1.3 Hypothesis, aim and objectives

This thesis aims to develop a method to efficiently train deep learning models on a per dataset basis and so enable seafloor image interpretation that is of practical use in marine science. The hypothesis is that gains in learning efficiency can be achieved by introducing domain knowledge into the deep learning training process. By expressing domain knowledge through relationships of the metadata gathered together with seafloor imagery, this knowledge can be leveraged while reducing the reliance on human input to train machine learning models. For complex natural environments, however, domain knowledge typically does not set absolute conditions, and so it is necessary to constrain learning without over-exerting assumptions about patterns that are expected but not always observed in the data. To implement and validate this hypothesis, the following objectives are set:

- Develop effective representation learning methods for seafloor imagery that leverage metadata in order to loosely constrain learning based on relevant domain knowledge. The low-dimensional latent representations generated by this approach will allow for important and discriminate information about a seafloor image dataset to be preserved in a way that allows for efficient interpretation.
- Develop unsupervised and semi-supervised methods that use these latent representation spaces for large scale interpretation of seafloor image datasets. By developing methods that can be deployed on a per dataset basis with realistic levels of human input, this will allow semantic maps to be generated to help scientists rapidly understand vast seafloor scenes in timeframes that are relevant for planning during ongoing scientific expeditions.
- Validate the effectiveness of the above methods using established metrics on seafloor image datasets that are representative of different types of surveys used in marine monitoring and conservation. The advantage posed by the methods developed in this thesis will be systematically verified through comparison with alternative processing methods at various stages of the learning and interpretation pipeline.

1.4 Contributions

The contributions of this thesis are the following:

- Development of a self-supervised representation learning method that leverages a soft assumption about metadata to guide the training of a deep learning autoencoder.
- Development of a contrastive learning based self-supervised representation learning method that makes a hard assumption on the similarity of images based on their metadata relationships.
- Proposal of seafloor interpretation applications that require only little or no supervision by humans for gaining an insight into trending patterns in the images of the dataset.
- Proposal of a pipeline for semi-supervised seafloor imagery interpretation by efficiently aligning the seafloor image representations with human interests.

1.5 Outline

The remainder of this thesis is structured as follows:

Chapter 2 presents a literature review of the methods for computer-aided interpretation of seafloor visual imagery. At the beginning of the chapter, several important characteristics of seafloor imagery that form a barrier for applying machine learning techniques to this domain are described. Subsequently, noticeable works in this domain, in which machine learning techniques are exploited, are introduced. The chapter also investigates state-of-the-art representation learning methods which are potentially applicable for seafloor imagery learning.

Chapter 3 first introduces a general idea for exploiting metadata in seafloor imagery learning. Then, two image representation learning methods based on two different assumptions; the hard assumption and the soft assumption; are formulated to implement the proposed idea. In the autoencoder based method, the soft assumption is implemented as a novel loss function derived from metadata to regularise training. In the contrastive learning based method, metadata is leveraged to select similar image pairs based on the hard assumption. Subsequently, the seafloor interpretation pipeline exploiting the representation learning outcome is proposed. The pipeline is developed to reduce or even eliminate human efforts for analysing large datasets of seafloor imagery. The unsupervised application, e.g. clustering and representative imagery identification and the semi-supervised pipeline, are proposed.

Chapter 4 validates the performances of the proposed representation learning and the application pipelines on real seafloor imagery datasets. The datasets consist of ~150k seafloor images taken over 16 dives along sparse and dense survey trajectories spanning a seafloor depth range of 20 to 780 metres in shallow coastal waters of Tasmania, Australia, and gas hydrate field off the coast of Oregon, USA.

Chapter 5 summarises the major findings of this thesis.

Chapter 2

Background

In this chapter, some important background for discussing the requirements for computer-aided seafloor imagery interpretation and developing machine learning based interpretation methods are introduced. Section 2.2 introduces machine learning basics which the following discussions in this thesis are based on. The characteristics of the target datasets, i.e. seafloor imagery datasets, are also described. Subsequently, noticeable previous efforts for interpreting seafloor imagery by machine learning are shown. Section 2.3 shows the general concept of representation learning. The section first shows why representation learning is considered crucial in this work, then currently available methods are introduced. The formulations of these methods are also shown for the discussion for the development of the novel methods proposed in this thesis. Section 2.4 presents several important methods for applying machine learning in domains where fewer human annotations are available.

2.1 Machine learning basics

Before the literature review, this section introduces the general concept of machine learning necessary for the following discussion. Machine learning techniques, especially modern deep learning techniques, have been recognised as particularly useful for the domains where semantic interpretation of the mass volume of high-dimensional data (e.g. image, video, audio) is necessary.

2.1.1 Perceptron

The great advantage of today's deep learning techniques is that their artificial neural network architecture can model significantly complex functions for classification, regression and other problems. The complicated modelling is available because it can obtain the mathematical models with a significantly large number of parameters, i.e. weights and biases, using sophisticated optimisation techniques.

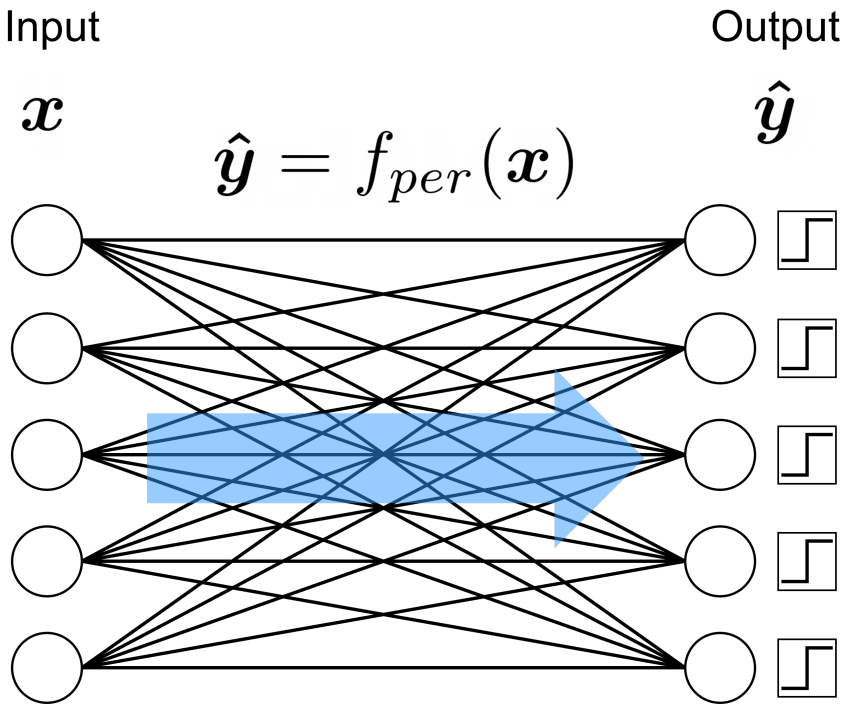


FIGURE 2.1: Overview of a perceptron.

The most fundamental network component for every deep learning architecture is a perceptron $f_{per}(\cdot)$. This component is illustrated in Figure 2.1. The left-side nodes are input nodes. The number of input nodes n corresponds to the dimensionality of the input vector $\boldsymbol{x}$. The right-side m nodes are output nodes or neurons. Each output node has an additive part, shown as a circle, and an activation function f_{act}, shown as a box. The edges between input nodes and output nodes correspond to the weight values w,

which are unique to each edge. The weight value of the edge between the i-th input node and the j-th output node is labelled as w_{ij}. w_{0j}, which corresponds to the biases is also introduce, though it is not shown in Figure 2.1 for simplicity. w_{0j} is always multiplied by -1, so that the independent value from inputs is added to j-th output node as a bias. The value of each output node is derived only from the input nodes: multiplying the weight values corresponding to the edges between the input nodes and the target output node, adding these values together, and inputting the sum into the step function. As a result, the only thing the output nodes share is the inputs, and the output nodes are completely independent of each other. This procedure for deriving the output value of j-th output node $\hat{y}_j$ can be shown as follows:

$$\hat{y}_j = f_{act}(\sum x_i w_{ij}). \tag{2.1}$$

The output vector $\hat{\boldsymbol{y}}$ can be described as $\hat{\boldsymbol{y}} = [\hat{y}_1, \hat{y}_2, ..., \hat{y}_m]^T$.

In the classification problem, a perceptron is optimised so that the output $\hat{\boldsymbol{y}}$ becomes close to the ground truth vector $\boldsymbol{y}$. Both $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$ have the same dimensionality that corresponds to the number of classes that exist in the target dataset. When input data x belongs to the j-th class in the potential classes, the ground truth vector $\boldsymbol{y}$ becomes a m-dimensional one-hot vector where only the j-th element is one and the others are zero. The perceptron should be optimised so that $\hat{\boldsymbol{y}}$ becomes close to $\boldsymbol{y}$ as much as possible. When the step function outputs 1 if the input is 0 or higher and outputs 0 in other cases, w_{ij} is updated by gradient descent as follows,

$$w_{ij} \leftarrow w_{ij} - \eta \left(\hat{y}_j - y_j\right) \cdot x_i, \tag{2.2}$$

feeding all available sets of x and y repeatedly until convergence. η is a learning rate that controls how much to change the weights at each iteration. A large η value tends to make the training process unstable, but a small η value possibly requires more iteration steps until convergence.

2.1.2 Deep learning

As shown in Equation 2.1, the perceptron f_{per} is a linear function, so only linear planes in the input vectors' n-dimensional space can appear as the boundary of each class, so that it can precisely classify only linearly separable classes. The idea of the multilayer perceptron method is to connect multiple perceptron units one after another, so that the output nodes of one perceptron unit become the input nodes of the next perceptron.

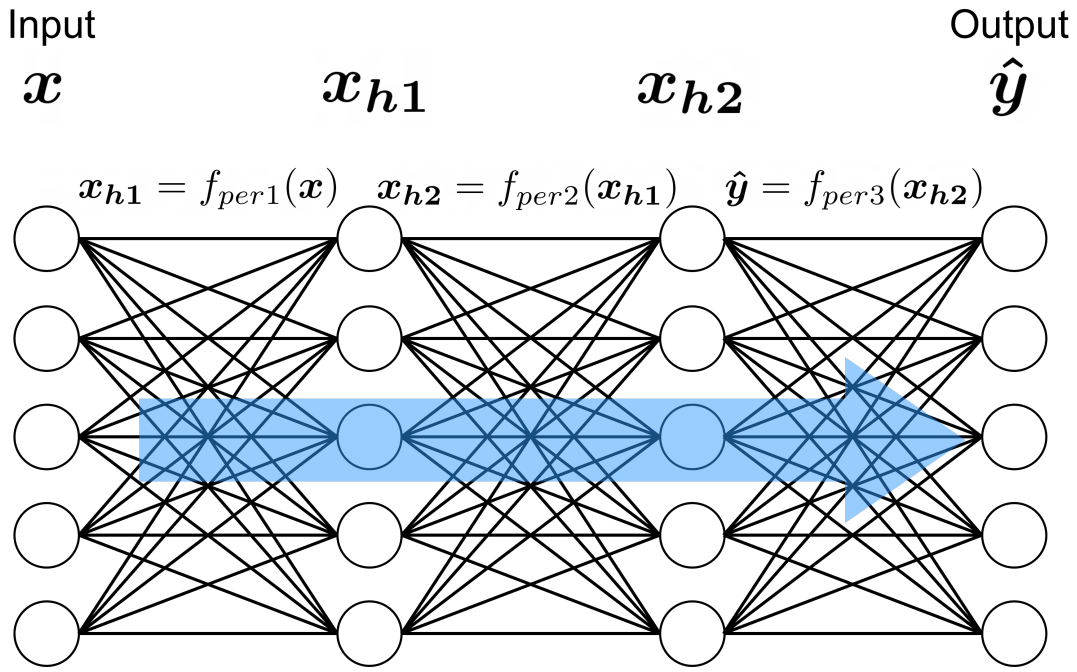


FIGURE 2.2: Overview of a multilayer perceptron.

Figure 2.2 shows the overview of a multilayer perceptron. This example shows the structure of a 4-layer multilayer perceptron where three perceptron networks, i.e. f_{per1}, f_{per2} and f_{per3}, are connected. When an input vector $\boldsymbol{x}$ is given, the first layer perceptron f_{per1} maps $\boldsymbol{x}$ to its output vector $\boldsymbol{x_{h1}}$, then the second layer perceptron f_{per2} maps $\boldsymbol{x_{h1}}$ to $\boldsymbol{x_{h2}}$. Finally, the third and final layer perceptron f_{per3} maps $\boldsymbol{x_{h2}}$ to $\hat{\boldsymbol{y}}$, which corresponds to the output vector of the whole multilayer perceptron. Every perceptron in a multilayer perceptron includes a differentiable nonlinear function as its activation function such as sigmoid, hyperbolic tangent or Rectified Linear Unit (ReLU). As a result, the whole multilayer perceptron $f_{mlp}(\boldsymbol{x}) = f_{per1} \circ f_{per2} \circ f_{per3}(\boldsymbol{x})$ becomes also a differentiable nonlinear function. Thanks to its stacked structure and nonlinear activation functions appearing at every layer, a multilayer perceptron can express complex nonlinear class boundaries so that it can achieve higher performance for non-linear separable data. Modern deep learning architectures such as AlexNet are introduced based on the same basic idea that a more deeply and widely stacked multilayer perceptron can express more complex class boundaries.

For classification problems, the cross entropy loss is commonly used as the objective function of multilayer perceptron, which is defined against y as follows:

$$L_{ce} = -\sum y_i \cdot \log(\hat{y}_i) = -\sum y_i \cdot \log(f_{mlp}(\boldsymbol{x})). \tag{2.3}$$

Since L_{ce} is still differentiable, the weight parameters of a multilayer perceptron can be optimised by gradient descent. In a multilayer perceptron, the partial derivative values

for each weight value, which are necessary for gradient descent, are not calculated analytically nor numerically. Instead, they are found by backpropagation, where each partial derivative value is derived efficiently from the output vector and intermediate terms calculated for obtaining it (e.g. $\boldsymbol{x_{h1}}$, $\boldsymbol{x_{h1}}$ in Figure 2.2) following the chain rule at each iteration.

Multilayer perceptrons can be optimised against any type of differentiable loss function by backpropagation. For regression problems where $\boldsymbol{y}$ is given as a standard vector which should be predicted as precisely as possible, the following Mean Squared Error (MSE) loss is used:

$$L_{mse} = -\sum (y_i - \hat{y_i})^2 = -\sum (y_i - f_{mlp}(\boldsymbol{x}))^2. \tag{2.4}$$

While other models generally need gradient matrices analytically solved for efficient optimisation, multilayer perceptron does not require them. Any differentiable function can be used as a loss function and optimised, and so the loss function can be designed flexibly depending on the problems.

The basic idea of deep learning is stacking perceptron layers more deeply so that more complex functions can be expressed. In addition, when input data $\boldsymbol{x}$ is 2 or 3-dimensional tensors, rather than vectors, convolutional layers are introduced. The layers of a multilayer perceptron are called the fully connected layers since each node is connected to all nodes of the previous and next layers. This structure is unreasonable for image data because the pixels at faraway positions are not strongly correlated, so it is redundant to consider all connections. Instead, in convolutional layers, 3-dimensional filters are applied to the input data and their intermediate representations. The features observed in this filter size are learned intensively at the layer. The next convolutional layer can learn the features that appear in the broader areas in the original image by applying convolutional filters to the previous convolutional outputs. Pooling, where only the maximum or average value of each filtering result is kept as the output value, and others are dropped, are often applied after the convolution so that only informative parts are preserved. Therefore, Convolutional Neural Network (CNN), in which convolutional layers are applied besides fully connected layers, is significantly efficient for extracting low-dimensional vectors which preserve important information of original high-dimensional data. AlexNet (Krizhevsky et al., 2012), where 5 convolutional layers and 3 fully connected layers are applied, is the first CNN that is successfully applied for general image interpretation tasks. Later more efficient CNNs with more complex structures such as ResNet (He et al., 2016) and GoogLeNet (Szegedy et al., 2015) appear and are widely used for image interpretation applications.

2.1.3 Evaluation metrics

This section introduces some important methods for evaluating the performance of machine learning.

For classification problems, the confusion matrix is generally used for visualising the relationship between the classifier's predictions $\hat{y}$ and ground truth annotations $\hat{y}$. Figure 2.3 shows an example of a confusion matrix for classification problems on a dataset that consists of the data from three classes A, B and C. It is a square matrix where all the possible classes are allocated in both the horizontal and vertical directions, which corresponds to the predictions and ground truths, respectively. The element of the confusion matrix at (i, j) shows the number of input data whose ground truth class is i but predicted as j by the classifier. The elements on the main diagonal correspond to the numbers of correctly classified samples. In this example, the matrix reveals that most samples from the three classes were classified correctly, but two samples of class C were misclassified as A. For a small number of classes, the confusion matrix is useful for understanding the overall performance of classifiers.

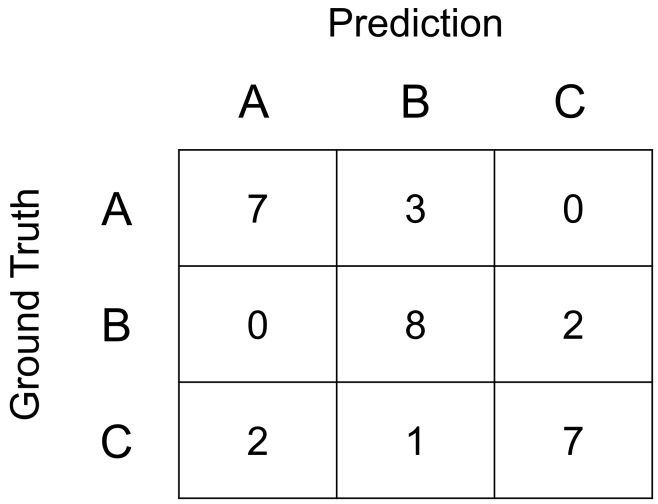


FIGURE 2.3: Example of a confusion matrix.

For evaluating the performance of multiple classifiers, comparable evaluation metrics are essential. Several important metrics for evaluating classification performance are introduced here. For simplicity, the metrics are defined on a binary classification problem, where the classifier should determine whether each sample belongs to 1 ('Positive') or 0 ('Negative'). The confusion matrix can be drawn in Figure 2.4. 'True' and 'False' mean that the classification result of the corresponding matrix element is correct and incorrect, respectively.

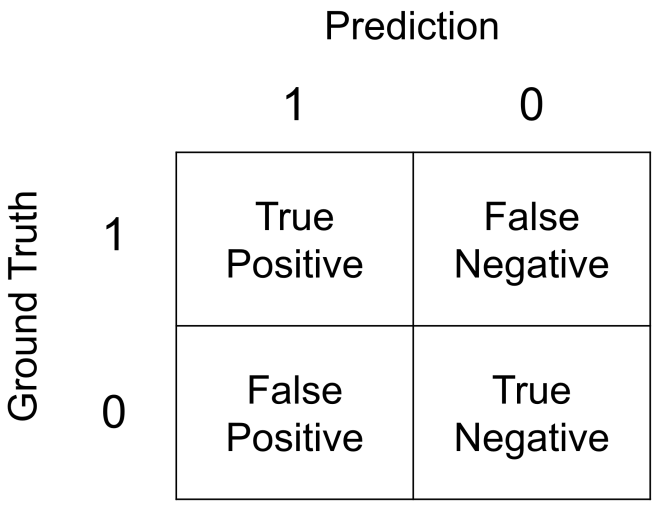


FIGURE 2.4: Confusion matrix of binary classification.

The elements on the main diagonal of Figure 2.4 are correct, and those off the diagonal are wrong, as well as the previous confusion matrix example (Figure 2.3). The classification *Accuracy* can be defined as the sum of the number of True Positive (*TP*) and True Negative (*TN*) samples divided by the total number of the samples:

$$Accuracy = \frac{\#TP + \#FP}{\#TP + \#FP + \#TN + \#FN}. \tag{2.5}$$

Precision is the ratio of correctly predicted positive samples to the total number of positive predicted samples and is defined as follows:

$$Precision = \frac{\#TP}{\#TP + \#FP}. \tag{2.6}$$

Recall is the ratio of correctly predicted positive samples to the total number of actually positive samples and is defined as follows:

$$Recall = \frac{\#TP}{\#TP + \#FN}. \tag{2.7}$$

As their definitions show, *Precision* and *Recall* are inversely related to some extent; if *#FP* increases (meaning that a broader definition of a target class is obtained by the classifier), then *#FN* is likely to decrease, and vice versa. Adequate classifiers should show high scores on both of them, and so the following F_1-score is defined as a harmonic mean of *Precision* and *Recall* for evaluation:

$$F_1 = 2\frac{Precision \times Recall}{Precision + Recall} = \frac{\#TP}{\#TP + \frac{(\#FN + \#FP)}{2}}. \tag{2.8}$$

For multiclass classification problems, these metrics can be derived against each class and can be used for per-class performance evaluation. For overall performance evaluation, F_1 can be derived by counting all *TP*, *FN* and *FP* samples. This F_1 value is called micro averaged F_1. When the target dataset is imbalanced, i.e. the number of samples in each class is not even, the micro averaged F_1 is largely affected by the classification result against the majority classes. When all classes in the dataset can be considered as evenly important, macro averaged F_1, which is the average of the F_1-score of each class, can be considered as more appropriate since it does not ignore the minority classes.

Besides classification problems where classifiers predict the class $\hat{y}$ of each sample in the same categorisation scheme as the ground truth y, clustering is also commonly used for automating big data analysis. In clustering problems, no information on the ground truth of each sample is given to the algorithms during the training, so they cannot predict classes. Instead of class prediction $\hat{y}$, clustering algorithms output c, which is the cluster that the input sample is likely to belong, against each sample. Therefore the evaluation metrics for classification problems cannot be applied. Normalised Mutual Information (NMI) (Estévez et al., 2009) is one of the most widely used metrics, which is defined based on information entropy $H(\cdot)$ of the set of the ground truth labels Y and the assigned cluster labels C as follows:

$$NMI(Y, C) = 2\frac{H(Y) - H(Y|C)}{H(Y) + H(C)}. \tag{2.9}$$

$H(Y|C)$ is a conditional entropy of the ground truth labels within each cluster. A NMI score is bounded between 0 (no mutual information) and 1 (perfectly correlated). 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 number of clusters found using a clustering algorithm is often different from the number of ground truth classes. In this case, NMI is a favourable metric since it does not require the targets to have the same number of clusters or categories.

2.2 Interpreting seafloor imagery

Determining the distribution of habitats, substrates and infrastructures are tasks that lie at the core of seafloor survey. Determining the distribution of habitats, substrates and infrastructures are tasks that lie at the core of seafloor survey. Computer vision and machine learning techniques can be efficient solutions for these tasks. However, the implementation is not straightforward because of several special characteristics known about underwater imaging. These characteristics often form barriers to their widespread use in real-world survey scenarios. This section states the important characteristics of robotically collected underwater imagery in section 2.2.1. Then the previous efforts which have been made for extracting features from seafloor imagery necessary for machine learning based processing are shown in section 2.2.2. Subsequently, supervised learning based approaches and unsupervised learning based approaches for seafloor image interpretation are introduced in section 2.2.3 and section 2.2.4, respectively.

2.2.1 Characteristics of underwater images

Underwater camera systems are mostly composed of commercially available camera modules, which can capture Red Green Blue (RGB) colour or greyscale images with high-dynamic-range, and powerful strobe lights. Therefore, the collected images are compatible with modern computer vision and machine learning techniques. However, these images have properties that are not common in other domains that need to be considered:

1. **Colour and geometry distortion**
   Different wavelengths of light attenuate at different rates in water, causing underwater images to look blue-green compared to the true colour of observed targets. The relatively low imaging altitudes (typically less than 10 m) and wide angle lenses often used to maximise area cover result in large relative range differences within an image due to terrain profiles and between images due to vehicle dynamics, which change the hue of images. Between datasets and platforms, there are additional sources of variability, including different water column properties that affect the wavelength dependence of light attenuation, and the use of artificial light sources with different wavelength profiles. In addition to colour degradation, the variable range causes spatial inconsistencies that distort the shape and size of observed targets.

2. **Small footprint**
   Light rapidly attenuates in water, and so powerful artificial light sources are needed to obtain visual images in most applications. The range at which images can be obtained is limited to approximately 10 m for most setups, which

constrains the footprint of a single frame to edge lengths of a similar magnitude. Since many patterns of interest (e.g. substrates, habitats, infrastructure) exist on far larger spatial scales, multiple images need to be taken along trajectories to capture these broader scale patterns.

3. **Metadata reference**
   Most images of the seafloor are gathered by robotic platforms or fixed observatories, and so various types of metadata, e.g. georeference (latitude, longitude and depth), water temperature, salinity and pH, are typically available. As for georeference, since Global Navigation Satellite Systems (GNSS) cannot be used underwater, most mobile robotic platforms have navigational suites that fuse data from an Attitude and Heading Reference System (AHRS), Doppler Velocity Log (DVL) and depth sensor with acoustic positioning systems such as an Ultra Short BaseLine (USBL). Georeferencing is typically achieved with a relative accuracy of approximately 1 % of distance travelled, and absolute accuracy of approximately 1 % of depth (Paull et al., 2014). Stationary systems have similar absolute position accuracy.

4. **Imbalanced Class Distribution**
   Seafloor substrates and habitats can change over spatial scales larger than the extents observed during most robotic imaging surveys. Furthermore, there are many types of benthic communities, geological features and infrastructures that are sparsely distributed, making subsea datasets highly susceptible to skewed class membership (Bewley et al., 2015a; Mahmood et al., 2018).

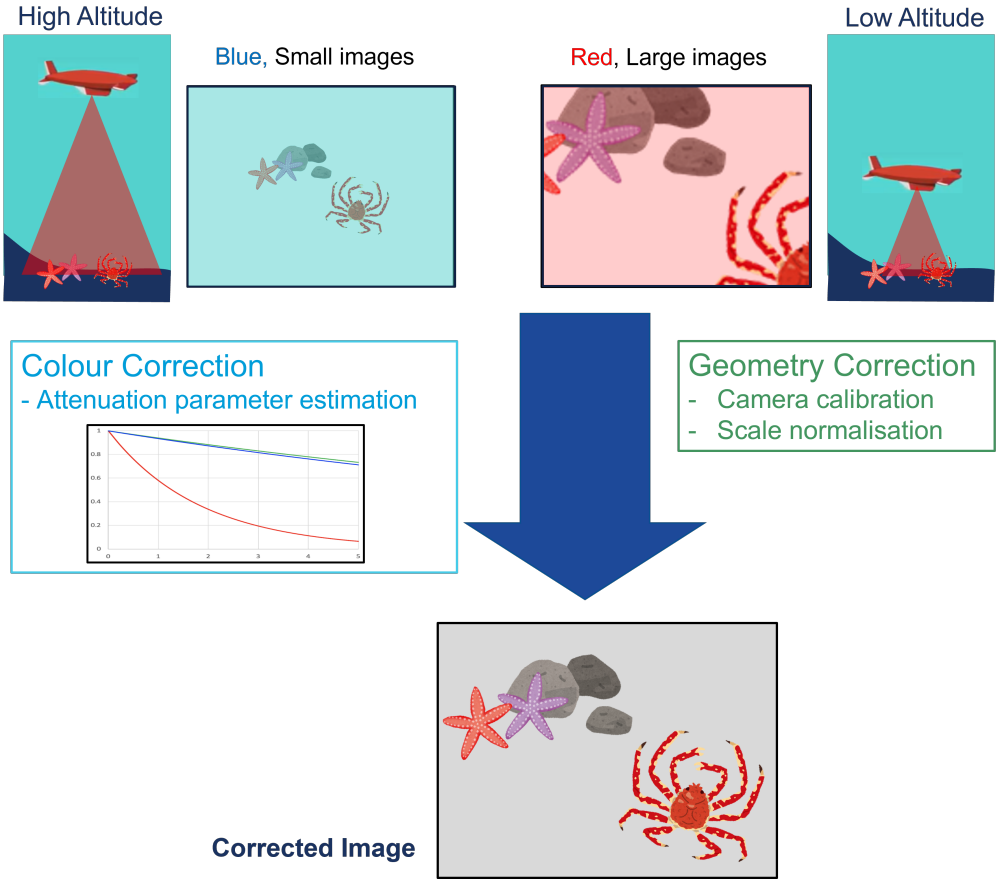


FIGURE 2.5: Colour and geometry distortion in underwater images. Since red lights are absorbed rapidly compared with blue and green lights, the images captured at higher altitude from seafloor look bluer and darker. This colour distortion can be corrected by colour channel and range dependent attenuation parameters. Geometry distortion caused by imaging hardware and altitude fluctuation can also be corrected by camera parameters and altitude.

Imbalanced class distributions, colour and geometry distortions can degrade learning performance (Krawczyk, 2016; Walker et al., 2019). The problem of small footprints can potentially be solved if pixel-order accurate georeferencing can be achieved, as artefact-free photomosaics can be generated and cropped to form image patches for processing. However, for seafloor imaging applications, position estimates contain non-negligible uncertainty compared to the resolution and footprint of obtained imagery. Although techniques such as simultaneous localisation and mapping are available (Mahon et al., 2008), the need for artificial strobes and the limited energy available on robotic platforms limits the relative overlap that can be achieved between images. This makes generating pixel-order accurate photomosaics more challenging to obtain than with satellite and aerial drone imagery, which typically have lower resolution, larger image footprints with greater overlap and accurate position information. These points favour the use of single image frames for automated interpretation of underwater imagery since these contain fewer artefacts.

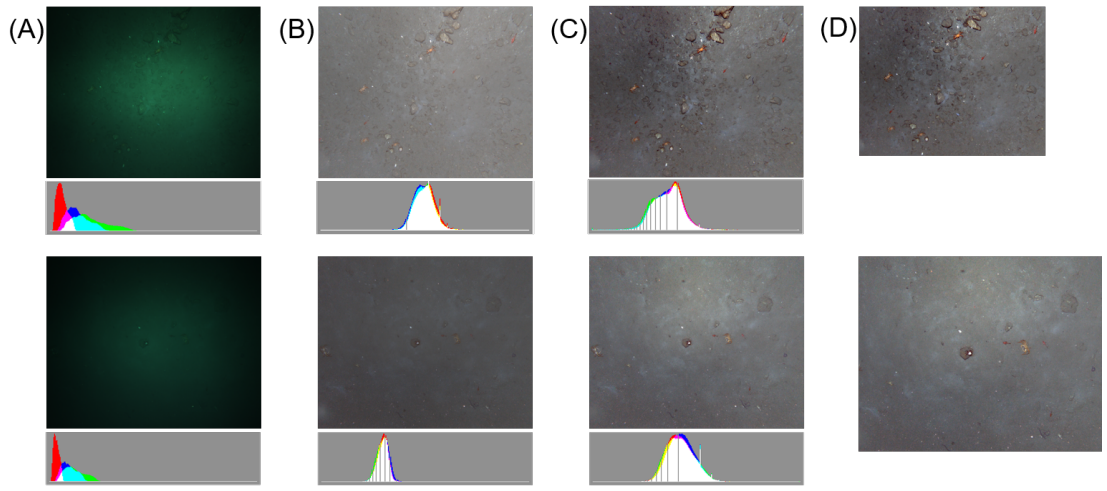


FIGURE 2.6: Colour and geometry correction of seafloor images. Seafloor images captured at 5.1 m (top, 5 mm/pixel) and 7.0 m (bottom, 7 mm/pixel) altitude for (A) raw, (B) pixel-wise normalisation, (C) attenuation correction and pixel-wise normalisation and (D) undistortion and rescaling to a constant spatial resolution (6 mm/pixel, equivalent to 6.0 m altitude).

There have been many studies investigating computational and physically grounded principles to compensate for the artefacts caused by colour and geometry correction. Figure 2.5 shows the process overview for correcting these distortions. As for 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 underwater images appear blue and green (Jaffe, 1990) (Figure 2.6A). Seafloor images captured at low altitudes (Figure 2.6A top) are also brighter than the images captured at high altitudes (Figure 2.6A 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 2.6B). However, pixel-wise normalisation cannot correct colour variations caused by altitude differences within a dataset. One way of compensating for these variations is by grouping the images so that the images captured at a similar altitude range belong to the same group, then apply the pixel-wise colour correction on a per-group basis. However, this method requires enough images in each group, and the images which can not be assigned to any altitude range group can not be corrected consistently. Bryson et al. (2013) proposed a more 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:

$$\mu_x(u,v,\nu,d) = a(u,v,\nu)\exp(-b(u,v,\nu)d) + c(u,v,\nu). \tag{2.10}$$

The indices $[u, v]$ specify each pixel's location in the image frame, ν is the colour channel, d is the 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 practical scenarios. This 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. Figure 2.6C shows the result of the proposed colour correction. Compared to Figure 2.6B, the brightness between images taken at different altitudes is more uniform. For geometry correction, the 3d information needed to fully compensate for scale effects within an image frame is not always available. However, scale can be approximately corrected considering imaging altitude. Geometric distortions are also corrected using lens calibration data.

2.2.2 Feature engineering

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. Since visual images are significantly high-dimensional for conventional machine learning algorithms and possibly redundant, it would be preferable if each image is expressed in a low-dimensional representation without losing its intrinsic information. The method for obtaining these representations 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. This section describes feature engineering techniques applied in the previous studies, especially before the deep learning techniques were applied to this domain.

Manually engineered feature descriptors have been investigated by several groups for efficient image representation (Bewley et al., 2015b; Rao et al., 2017; Steinberg et al., 2011; Beijbom et al., 2012; Kaeli and Singh, 2015; Neettiyath et al., 2020). In Beijbom et al. (2012) and Neettiyath et al. (2020), 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) (Ojala et al., 2002) and Sparse coding Spatial Pyramid Matching (ScSPM) (Yang et al., 2009) have also been applied to identify spatially invariant patterns that appear at different scales within images of the seafloor (Bewley et al., 2015b; Rao et al., 2017). In Kaeli and Singh (2015), accumulated histograms

of oriented gradients from image keypoints were used to describe seafloor images for the purpose of clustering and anomaly detection. However, all of the descriptors mentioned above require manual tuning of parameters to effectively describe the datasets they are applied to. For seafloor imagery interpretation, the most proper feature descriptor would differ depending on the targets, imaging hardware and imaging conditions. These manual feature engineering techniques cannot deal with the influx of a large volume of new datasets in this domain.

2.2.3 Supervised learning

A large proportion of automated classifiers have used a combination of hand-picked features chosen based on expert knowledge of the application domain or through a reward-based selection process (Beijbom et al., 2012; Neettiyath et al., 2020) as described in section 2.2.2. In Beijbom et al. (2012) the authors apply a Support Vector Machine (SVM) to texture- and colour-based features designed to classify seafloor images into different substrates types for reef ecology surveys. In Inglada (2007), hand-picked geometric features are combined with an SVM for the classification of satellite images. In Neettiyath et al. (2020), a similar approach is applied for seafloor mineral prospecting. Spatial invariant features such as Local Binary Patterns (LBP) (Ojala et al., 2002) and Spatial Pyramid Matching (SPM) (Yang et al., 2009) have also been effectively applied to classification problems (Bewley et al., 2015b; Rao et al., 2017). However, these types of features require manual tuning of parameters, or feature engineering, to efficiently describe each independent dataset. Furthermore, a separate classification process is required, which also requires parameter tuning. As such, these feature engineering based methods often require expert knowledge of both the data and application domain and have limited versatility when applied to multiple datasets.

Recently, deep learning techniques have been applied to the classification problem in this domain. A key advantage of deep learning techniques is that both the latent representation of data and classification can be simultaneously optimised in a single end-to-end training process. This avoids the need for costly and potentially subjective feature engineering and reduces the need for parameter tuning, making deep learning techniques a compelling choice. In Mahmood et al. (2018), the ResNet (He et al., 2016) deep learning CNN is used to classify images of coral into nine separate classes, achieving higher classification resolution than prior studies and demonstrating the ability of deep learning to effectively model class boundaries used in scientific taxonomy. However, to work effectively, deep learning classification techniques typically require a large number of annotated examples of each class. Although several labelling platforms tailored to seafloor imagery exist (Bewley et al., 2015a; Langenkämper et al., 2017), the sensitivity of images to environmental and acquisition conditions, the complexity of annotation

schemes and the comparatively small size of each environmental monitoring community mean that large-scale label repositories such as those in terrestrial imaging (Deng et al., 2009b), satellite imagery (Van Etten et al., 2018) and autonomous driving (Geiger et al., 2013) do not yet exist. Furthermore, for sub-sea imaging, most groups gather images using custom built imaging hardware, where in Langenkämper et al. (2020), the authors reported that even small differences in sub-sea imaging hardware potentially limits learning transferability and distorts deep learning classifier outputs.

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

2.2.4 Unsupervised learning

Unsupervised learning techniques have great potential for image interpretation in environmental monitoring because they do not require annotations and so can be efficiently trained and applied on a per-dataset basis. As with any automated image analysis, feature engineering is crucial for effective interpretation. In Steinberg et al. (2011), LBP (Ojala et al., 2002) features derived from greyscale images, 3d rugosity and colour are applied to seafloor image clustering. The authors later applied ScSPM (Yang et al., 2009) as a more generic approach to describe seafloor images (Steinberg, 2013). In Friedman et al. (2011), the non-parametric Bayesian clustering technique used in Steinberg et al. (2011) and Steinberg (2013) is extended to incorporate annotations made during active learning (Settles, 2009) for seafloor imagery. In Kaeli and Singh (2015), the accumulated histogram of oriented gradients from keypoints are used to describe each image, and this is applied to clustering and anomaly detection. More recently, Shields et al. (2020) used unsupervised clustering results generated from visual images as labels for supervised learning of seafloor bathymetric datasets.

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

2.3 Representation learning

For introducing machine learning techniques to seafloor imagery interpretation, obtaining low-dimensional representations of raw images is necessary for any type of algorithm. Instead of feature engineering (section 2.2.2), where human efforts for identifying features which properly explain original images, several dimension reduction techniques, such as Principal Component Analysis (PCA) (Wold et al., 1987), where human supervising is not required, exist. Deep learning based representation learning methods such as autoencoder and contrastive learning are also actively studied and achieve great success in obtaining image representations. All of the techniques introduced in this section can be categorised as self-supervised learning methods, where only data themselves are necessary, and human annotations are not required for training.

2.3.1 Concept

In the classification problem, the machine learning pipeline should output the class prediction value $\hat{y}$ which the input $\boldsymbol{x}$ belongs to. Figure 2.7 shows the overview of a typical classification pipeline where a conventional classifier is applied. For conventional classifiers, e.g. logistic regression, k-nearest neighbours, support vector machine, random forest and Gaussian process classifier, the original data $\boldsymbol{x}$ are too high-dimensional in

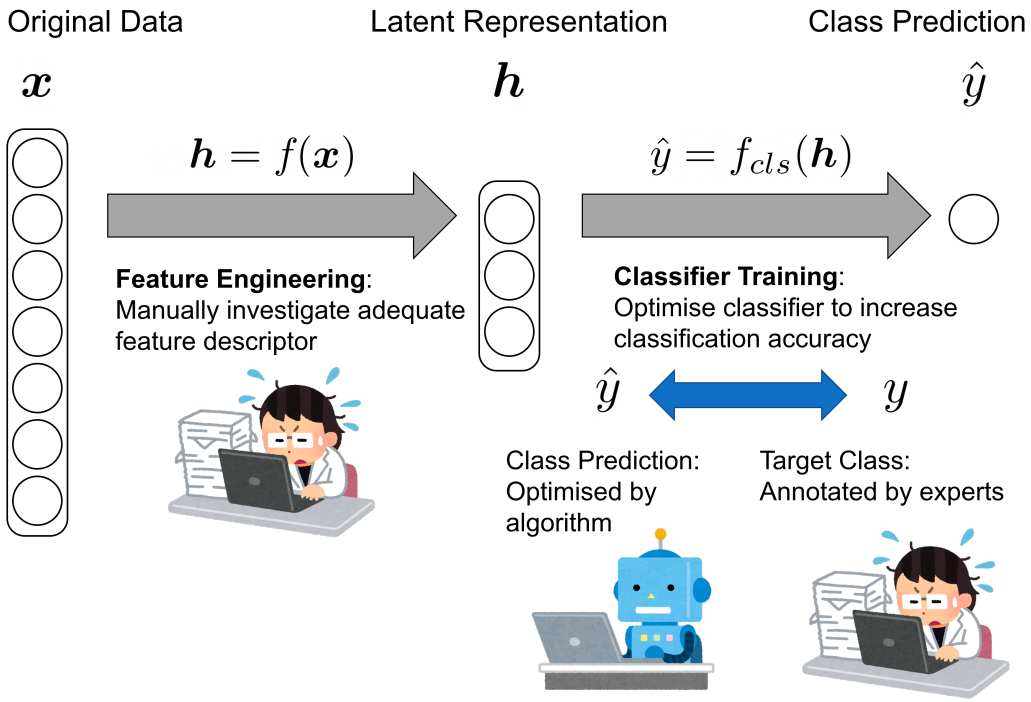


FIGURE 2.7: Overview of supervised classification with feature engineering.

most practical applications, and so their features or latent representations $\boldsymbol{h}$ are passed to the classifier $f_{cls}(\cdot)$ as the inputs. Previously, the function $f(\cdot)$, which maps original data $\boldsymbol{x}$ to their corresponding latent representations $\boldsymbol{h}$, was determined by 'feature engineering'. In this process, the experts who sufficiently understand the characteristics of the domain empirically investigate the proper feature descriptor for the data and design $f(\cdot)$. The classifier $f_{cls}(\cdot)$ is optimised so that the class prediction $\hat{y}_i$ for ith data point $\boldsymbol{x}_i$, which is estimated from $\boldsymbol{h}_i$, equals its ground truth class y_i. Ground truth class values y should be manually annotated to each data point of x, so great human efforts are also required here. The necessity of feature engineering and annotating prevents machine learning techniques from being applied to many domains where available human expert resources are limited.

Recently, so-called deep learning techniques are recognised as powerful, especially for high-dimensional data classification problems. Figure 2.8 presents the overview of deep learning based classification. The key characteristic of a deep learning classifier f_{dlc} is that it can directly predict the classes $\hat{y}$ from their original high-dimensional inputs $\boldsymbol{x}$. Thus, the complex feature engineering process is not necessary for deep learning. In comparison with feature engineering, where both domain data knowledge and general data processing knowledge are required, annotating does not require specific skills if the target dataset and annotating scheme are general. However, compared with a conventional classifier f_{cls}, a deep learning classifier f_{dlc} has a considerably larger number of parameters (weights) that should be optimised. Therefore, a large number of sets of $\boldsymbol{x}$ and y are required as a training dataset for preventing the parameters from

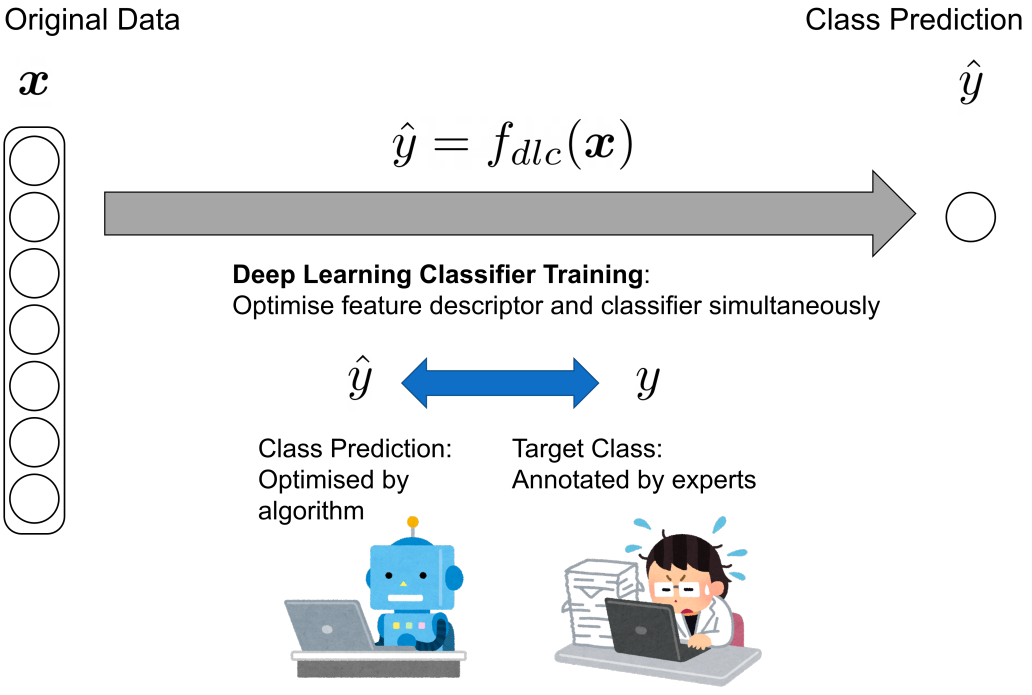


FIGURE 2.8: Overview of supervised classification with deep learning.

over-fitting to the specific examples contained in the dataset. This is a serious disadvantage for non-general applications where gathering a large amount of annotated data is difficult, including seafloor imagery interpretation.

For the domains in which preparing large scale datasets is infeasible, applying deep learning techniques for obtaining latent representations $\boldsymbol{h}$ instead of $\hat{y}$ is possibly efficient. In other words, automatically finding the function $f(\dot{})$ that maps original data $\boldsymbol{x}$ to their latent representations $\boldsymbol{h}$. The methods for implementing this idea are categorised as representation learning. Figure 2.9 shows a classification pipeline where representation learning is applied. Since deep learning based self-supervised representation learning techniques, e.g. autoencoder and contrastive learning, where annotations are not required for optimising $f(\cdot)$, have been developed, the obtained latent representations $\boldsymbol{h}$ are not affected by the biases caused by the manual feature engineering processes and annotating. Once the subjective representations $\boldsymbol{h}$ are obtained, various types of machine learning techniques, e.g. classification, clustering, content based retrieval, can be applied for automatic interpretation of original data $\boldsymbol{x}$.

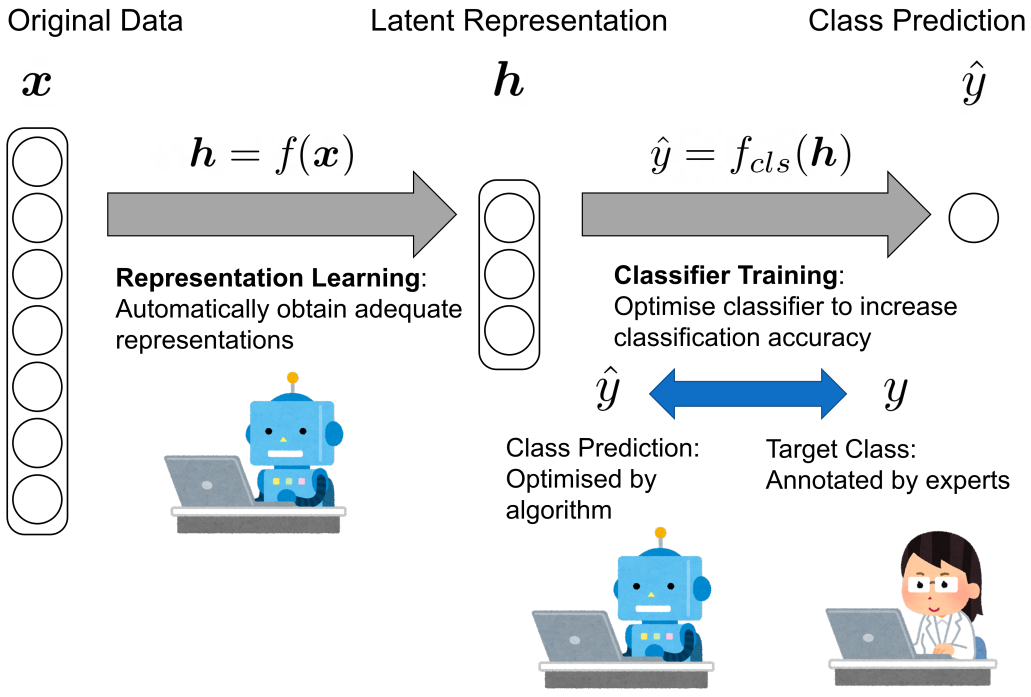


FIGURE 2.9: Overview of supervised classification with representation learning.

2.3.2 Dimension reduction

There exist several noticeable dimension reduction techniques for embedding high dimensional data into low dimensional data without a manual process like feature engineering.

PCA (Wold et al., 1987) is one of the most popular and established techniques where a set of d_{org}-dimensional vectors can be embedded into d-dimensional vectors (d is arbitrary and smaller than d_{org}), keeping the distinctiveness of each original vector as much as possible in the set.

t-distributed stochastic neighbour embedding (t-SNE) is another important technique of dimension reduction. It embeds high-dimensional vectors into 2 or 3 dimensional vectors, keeping the relative distance between each vector in the original space as much as possible. For a set of d_{org}-dimensional vectors $\boldsymbol{x}$, the probability p_{ij} which represents the similarity between ith vector $\boldsymbol{x}_i$ and jth vector $\boldsymbol{x}_j$ ($i \neq j$) in the set is defined as follows. First, the conditional probabilities $p_{j|i}$ are defined as

$$p_{j|i} = \frac{\exp\left(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2 / 2\sigma^2\right)}{\sum_{k \neq i} \exp\left(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_k\right\|^2 / 2\sigma^2\right)}, \tag{2.11}$$

for different i and j, and $p_{i|i} = 0$. σ is a perplexity parameter. Then the joint probabilities p_{ij}, which correspond to the similarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are defined as

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2}. \tag{2.12}$$

The latent representation vectors $\boldsymbol{h}$ are embedded in the d-dimensional spaces, considering this similarity. d for t-SNE is particularly set to two or three since this technique is mostly applied for data visualisation. The joint probabilities q_{ij} correspond to the similarity of two latent representations $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ are defined as

$$q_{ij} = \frac{\left(1 + \left\|\boldsymbol{h}_i - \boldsymbol{h}_j\right\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \left\|\boldsymbol{h}_k - \boldsymbol{h}_l\right\|^2\right)^{-1}}. \tag{2.13}$$

q_{ij} is defined based on a Student t-distribution with a single degree of freedom i.e. $\left(1 + \left\|\boldsymbol{h}_i - \boldsymbol{h}_j\right\|^2\right)^{-1}$. To force $\boldsymbol{h}$ to have a similar neighbouring relationship to original $\boldsymbol{x}$, the Kullback-Leibler divergence between two affinity matrix P and Q, whose elements are p_{ij} and q_{ij}, respectively, are computed as

$$L_{tsne} = \mathrm{KL}(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{2.14}$$

For the optimisation, L_{tsne} is minimised by an improved gradient descent based technique. The gradient of L_{tsne} is given by

$$\frac{\delta C}{\delta h_i} = 4 \sum_j \left(p_{ij} - q_{ij}\right) \left(h_i - h_j\right) \left(1 + \left\|h_i - h_j\right\|^2\right)^{-1}, \tag{2.15}$$

and a momentum term is used for accelerating the optimisation.

2.3.3 Autoencoder

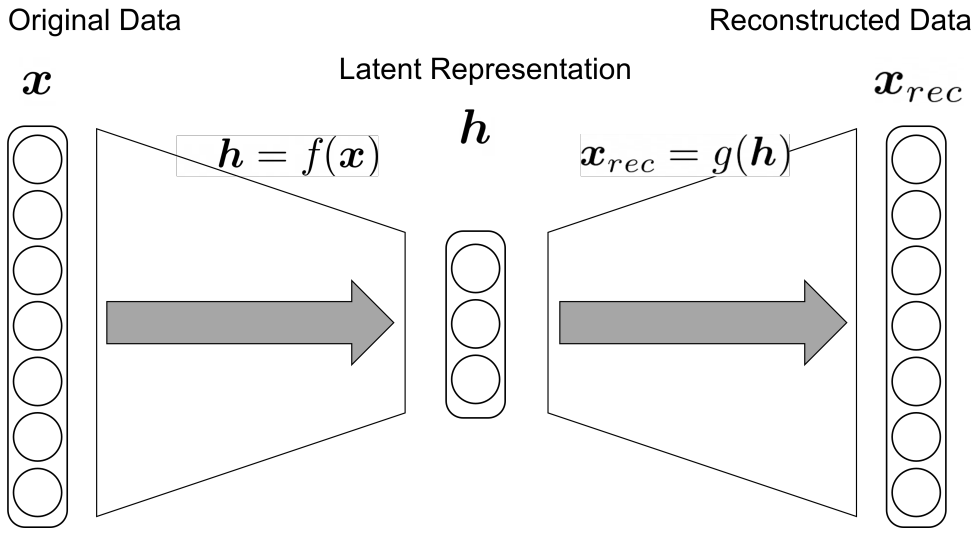


FIGURE 2.10: Overview of an autoencoder.

The autoencoder is a variation of the artificial neural network that is useful for representation learning. Figure 2.10 shows an overview of the autoencoder. It consists of two parts; an encoder $f(\cdot)$ and a decoder. The encoder $f(\cdot)$ maps original data $\boldsymbol{x}$ into a latent representation $\boldsymbol{h}$ of lower dimensionality and can be expressed as $\boldsymbol{h} = f(\boldsymbol{x})$. The decoder $g(\cdot)$ is expressed as $\boldsymbol{x}_{rec} = g(\boldsymbol{h})$, and reconstructs $\boldsymbol{x}_{rec}$ 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 $\boldsymbol{x}_{rec}$ can be measured as the mean squared error. Given n samples in a dataset, the autoencoder's objective function can be formulated as MSE loss between $\boldsymbol{x}$ and $\boldsymbol{x}_{rec}$ as follows:

$$L_{rec} = \frac{1}{n} \sum_{i=1}^{n} \left\|\boldsymbol{x}_i - \boldsymbol{x}_{rec\,i}\right\|^2, \tag{2.16}$$

where w denotes the parameters of the encoder and decoder, respectively. The biggest advantage of the autoencoder is that the neural networks used can be trained without the need for expert annotations. Since $\boldsymbol{x}_{rec}$ 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. For seafloor imagery learning applications, autoencoders are applied partly to learn mid-level features in visual imagery after extracting low-level features with ScSPM in Rao et al. (2017). Convolutional autoencoders are applied in Flaspohler et al. (2017) for unsupervised representation learning from seafloor imagery and shows that they outperform hand-designed features in discovering characteristic patterns.

To enhance the unsupervised representation 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

$$L_{all} = (1-\lambda)L_{rec} + \lambda L_{clust}, \tag{2.17}$$

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:

$$L_{clust} = \mathrm{KL}(P\|Q) = \sum_i \sum_k p_{ik} \log \frac{p_{ik}}{q_{ik}} \tag{2.18}$$

$$q_{ik} = \frac{\left(1 + \|\boldsymbol{h}_i - \boldsymbol{\mu}_k\|^2\right)^{-1}}{\sum_{k'} \left(1 + \|\boldsymbol{h}_i - \boldsymbol{\mu}_{k'}\|^2\right)^{-1}} \tag{2.19}$$

$$p_{ik} = \frac{q_{ik}^2 / f_k}{\sum_{k'} q_{ik'}^2 / f_{k'}}, \tag{2.20}$$

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

2.3.4 Contrastive learning

The recent development of contrastive learning concepts have demonstrated significant performance gains in self-supervised representation learning (Jing and Tian, 2020; Le-Khac et al., 2020). The main idea behind contrastive concepts is to simultaneously provide similar and dissimilar image pairs during training, where similar pairs are mapped close to each other in the representation space, and dissimilar pairs are mapped far apart. These concepts require a binary prior that describes whether the image pairs provided during training are expected to be similar or not.

The triplet loss L_{trp} (Hadsell et al., 2006) is the loss function for contrastive learning where the contrastive concept is straightforwardly implemented. It requires a similar pair $[\boldsymbol{x}_i, \boldsymbol{x}_j]$ and a dissimilar pair $[\boldsymbol{x}_i, \boldsymbol{x}_k]$. From their latent representations $\boldsymbol{x}_i$ $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$, the triplet loss is defined as follows:

$$L_{trp} = \sum max(0, \|\boldsymbol{h}_i - \boldsymbol{h}_j\|^2 - \|\boldsymbol{h}_i - \boldsymbol{h}_k\|^2) + m. \tag{2.21}$$

m is a margin parameter that should be kept between each dissimilar pair in latent space. Figure 2.11 illustrates the overview of the triplet loss based contrastive learning. The triplet loss intends to push the dissimilar sample $\boldsymbol{h}_k$ outside of the neighbourhood by a margin while keeping similar samples $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ within the neighbourhood.

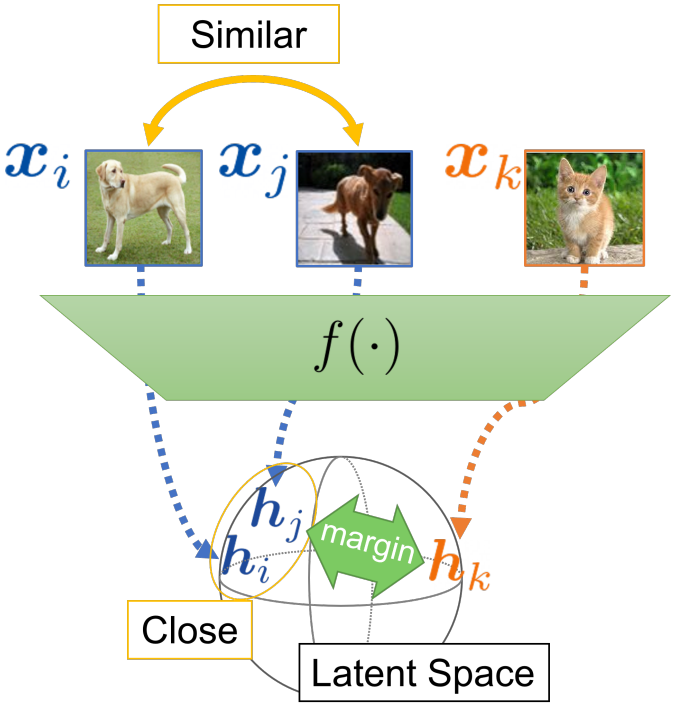


FIGURE 2.11: Overview of contrastive learning.

In Chen et al. (2020a), a method to generate similar and dissimilar pairs without any direct human input is developed using data augmentation. The proposed Simple framework for Contrastive Learning of visual Representations (SimCLR) applies random data augmentations to artificially generate similar image pairs, which are then contrasted with dissimilar pairs where different images are used. The method demonstrated significant gains in performance compared to supervised training using the transfer learning approach (Tan et al., 2018). The formulation of SimCLR is introduced in section 3.3.1 in detail.

2.4 Annotation effort reduction

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

2.4.1 Transfer learning

Transfer learning allows supervised learning models to be trained using a relatively small number of annotations in the target dataset by making use of much larger annotated datasets from a different domain. Several frameworks have been proposed to implement this concept (Tan et al., 2018). Network-based transfer learning has been applied in many application domains including medical (Shin et al., 2016), satellite (Yao et al., 2016), and seafloor imaging (Zurowietz et al., 2018). This approach works by reusing networks that have been pre-trained using large, generic datasets (e.g. ImageNet (Deng et al., 2009a), COCO (Lin et al., 2014), Pascal VOC (Everingham et al., 2015)) that consist of hundreds of thousands to more than ten million labels as an initial model. Though the number of dataset specific annotations needed depends on the domain, number of classes and data augmentation methods used, previous studies on satellite (Marmanis et al., 2015) and medical imagery (Shin et al., 2016) have required several hundreds of domain specific labels for effective use.

In Zurowietz and Nattkemper (2020) a pipeline to make seafloor imagery datasets transferable for inference on images from other datasets is proposed for the segmentation of marine organisms. The work proposes how to reduce scale variance across multiple datasets, which is highlighted as an important consideration for seafloor imagery. However, considering the other distortion factors mentioned in section 2.2.1, this method only partially extends the transferability, and no general method that can deal with all the possible distortion factors has been proposed.

2.4.2 Prioritised labelling

Images in a dataset do not have equal value for training machine learning algorithms. In Lapedriza et al. (2013), the authors demonstrate that training data selection can have a significant impact on supervised learning, where CNNs trained on a well selected subset of annotations can outperform CNNs trained using a larger number of annotations. In Paul et al. (2016), annotation efforts are prioritised using k means clustering to estimate the entropy of each sample, showing significant gains in performance compared to random selection.

In active learning (Settles, 2009), the learner interacts with human annotators by iteratively proposing data samples that it considers will most efficiently improve performance. Several strategies have been proposed to achieve this. Most approaches prioritise unlabelled samples that have the highest estimated uncertainty, or are predicted to have the biggest impact on the model. However, the heuristics used to suggest samples can only be calculated after the initial subset has been analysed by the algorithm. Although the initial subset can impact subsequent learning performance, its selection falls outside of the scope of most active learning techniques (Settles, 2009; Li et al., 2019).

In Zurowietz et al. (2018), an autoencoder is used to locate objects of interest in an unsupervised manner. The method highlights these regions to human experts in order to facilitate efficient use of time for manual segmentation. The approach leverages the assumption that interesting objects are relatively rare in the original seafloor image datasets they are applied to. Regions with a high autoencoder reconstruction loss value are considered likely to include targets of potential interests, and these regions are flagged for prioritised annotation by humans. Active learning is also applied for seafloor image interpretation in Friedman et al. (2011) and Shields et al. (2020), where the authors implemented this with ScSPM as the feature descriptor.

2.4.3 Group labelling and label extrapolation

Group-based labelling (Dai et al., 2012; Wigness et al., 2015) is a technique that assigns annotations to subgroups of clustered data in order to reduce the human annotation effort. An advantage of this approach is that it can be applied to datasets with no labels by using unsupervised clustering methods to generate the groups. However, determining the annotation for a cluster of images can be more complex than per-sample based annotation, especially when unsupervised cluster decision boundaries are not aligned with the desired class boundaries, resulting in conflicted human annotations. In Tian et al. (2007), the authors modified Gaussian mixture model based clustering to find clusters with high intra-cluster similarity since the samples in these clusters are considered to be more informative than others. Although these techniques have shown significant improvement in learning efficiency, the underlying assumption is that effective clustering can be achieved.

Predictive pseudo-labelling (Lee, 2013) reduces human effort by first training a classifier on a small subset of data that requires fewer annotations than the target dataset. An advantage of this over group labelling is that annotators consider individual images. After initial training, the classifier predicts labels for the remaining data, and these pseudo-labels are used together with the original annotations to fine-tune a classifier. Li et al. (2019) reports that SVM and Random Forest classifiers outperform CNNs when generating pseudo-labels from an initial annotated subset. Wu and Prasad (2017) used pseudo-labelling to improve the classification performance for a hyperspectral satellite

image dataset, demonstrating effective application of this approach to unstructured environmental monitoring data, where random subsets were used for initial training. The use of prioritisation methods for subset selection has not previously been investigated.

Chapter 3

Method

This chapter presents novel representation learning methods for seafloor imagery and their applications, where domain knowledge and metadata are exploited. Section 3.1 proposes a general idea for leveraging metadata in seafloor imagery learning. The soft assumption and the hard assumption, where the metadata is used differently for representation learning, are introduced. Section 3.2 formulates an autoencoder based representation learning, where the soft assumption is implemented as a novel loss function derived from metadata to regularise training. Section 3.3 shows a contrastive learning based method, where metadata is leveraged to select similar image pairs based on the hard assumption. Section 3.4 shows the unsupervised learning applications of the proposed representation learning techniques, e.g. clustering and representative image identification. Section 3.5 presents the methods for aligning the acquired representations with human interests. Content based retrieval and semi-supervised learning are introduced as the applications.

3.1 Concept overview

As introduced in chapter 2, many efforts have been made for computer-aided seafloor imagery interpretation. Like the other image learning domains, modern CNN is considered competitive for acquiring a discriminative model. However, for training CNN, a significant number of annotated images are required, a potentially serious drawback for real-world applications. Representation learning introduced in section 2.3, i.e. autoencoder and contrastive learning, can train a CNN without human annotations. The trained CNN can encode images to latent representations that preserve only the important information of originals. Since the obtained representations can be applied to various types of machine learning applications, seafloor interpretation can be achieved more flexibly and accurately through representation learning.

When applying general representation learning techniques to seafloor visual imagery learning, the special characteristics of seafloor imagery mentioned in section 2.2.1 should be considered. Though some of them form barriers for applying machine learning techniques, metadata availability can be a great advantage for acquiring more efficient CNN models, since metadata such as georeference and water condition conveys potentially useful information to discriminate the attributes of the observed images. For example, 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 a spatial scale larger than the footprint of a single image frame. In other words, similar features possibly appear in two images if they are captured at physically close locations. Figure 3.1 illustrates the basic idea of leveraging horizontal location metadata for representation learning. Seafloor images are collected along the platform's trajectory, so many pairs of images which are collected within a close physical distance are included in a dataset.

To leverage this domain knowledge in representation learning, the following assumption is made on image similarity, i.e. how much the images share the appearance features in common:

Soft assumption
Two images tend to be more similar when they are captured within a closer physical distance than two that are far away. More generally, the similarity between two images correlates with the similarity of metadata between them.

Since metadata is assumed to be loosely correlated with image appearance, this assumption is referred to as the soft assumption in the rest of this thesis. On the other hand, more deterministic clues would potentially help efficient training. For example, in supervised learning, human annotations are generally given as binary values rather than continuous values that show the belongingness possibility to each candidate class.

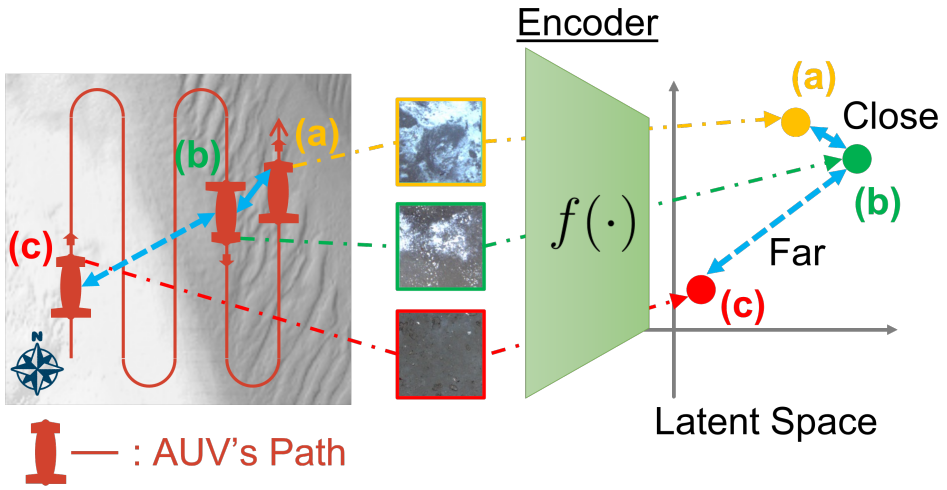


FIGURE 3.1: Overview of leveraging horizontal location metadata for representation learning. Since observation targets such as sediments exist over a spatial scale larger than the footprint of a single image frame, two images captured within close distance are likely to be similar, so that their latent representation would be similar.

This approach controls training more forcefully; however, overfitting would likely occur if many exceptions or errors are included in a training dataset. The hard assumption for considering this approach in representation learning is defined as follows:

Hard assumption

Two images captured within a certain close physical distance must be similar, and vice versa. More generally, two images must be similar if the metadata attached to them are similar enough.

The basic idea of leveraging domain knowledge in this thesis is refining existing representation learning techniques by formulating these assumptions in their training. Since the two assumptions here are exclusive, they are individually formulated in two different representation learning methods in the following sections. The soft assumption based method is presented in section 3.2, and the hard assumption based method is shown in section 3.3.

3.2 Representation learning with soft assumption

Figure 3.2 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 has 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, the soft assumption, e.g. 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, can be effective. This section introduces a novel metadata regularised representation learning method. 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 validation result of the proposed method is shown in section 4.2.1.

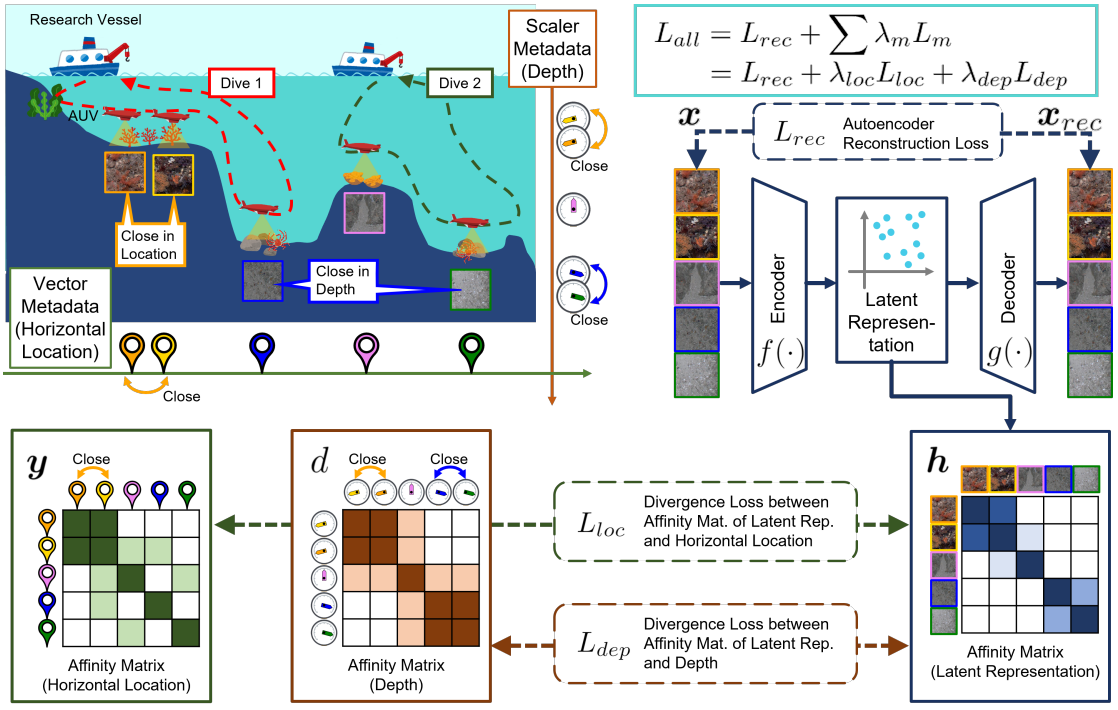


FIGURE 3.2: Overview of representation learning method with soft assumption. The method regularises the representation space of an autoencoder by embedding images taken at similar horizontal locations, or separated but with similar depths, in nearby regions of the latent representation space. This is 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.

3.2.1 Vector and scalar regularisation

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 tensor $\boldsymbol{h}$ ($\boldsymbol{h}=f(\boldsymbol{x})$), 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, a loss function of the following form is minimised:

$$L_{all} = L_{rec} + \sum \lambda_m L_m. \tag{3.1}$$

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.

AUVs typically measure their horizontal location, depth and altitude for basic navigational functionality. This metadata can be leveraged to regularise autoencoder training by formulating equation (3.1) as follows:

$$L_{all} = L_{rec} + \lambda_{loc} L_{loc} + \lambda_{dep} L_{dep}, \tag{3.2}$$

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 the implementation, AlexNet (Krizhevsky et al., 2012) and its inverted architecture are used as the encoder and decoder, respectively, where any type of neural network can be used to construct autoencoders in a similar way.

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, a loss function that has a similar structure to the loss function of t-SNE (Maaten and Hinton, 2008) (equation (2.11) - (2.14) in section 2.3.2) is introduced.

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 the proposed loss function, $\boldsymbol{y}$, 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 $i \neq j$ as:

$$p_{j|i} = \frac{\exp\left(-\left\|\boldsymbol{y}_i - \boldsymbol{y}_j\right\|^2 / 2\sigma_{loc}^2\right)}{\sum_{k \neq i} \exp\left(-\left\|\boldsymbol{y}_i - \boldsymbol{y}_k\right\|^2 / 2\sigma_{loc}^2\right)}, \tag{3.3}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}, \tag{3.4}$$

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:

$$q_{ij} = \frac{\left(1 + \left\|\boldsymbol{h}_i - \boldsymbol{h}_j\right\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \left\|\boldsymbol{h}_k - \boldsymbol{h}_l\right\|^2\right)^{-1}}, \tag{3.5}$$

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:

$$L_{loc} = \mathrm{KL}(P \| Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{3.6}$$

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. Equation (3.3) - (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 $\boldsymbol{x}_{emb}$.

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:

$$r_{j|i} = \frac{\exp\left(-(d_i - d_j)^2 / 2\sigma_{dep}^2\right)}{\sum_{k \neq i} \exp\left(-(d_i - d_k)^2 / 2\sigma_{dep}^2\right)}, \tag{3.7}$$

$$r_{ij} = \frac{r_{j|i} + r_{i|j}}{2n}, \tag{3.8}$$

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

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.

3.2.2 Mini-batch sampling considering metadata

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, a sampling method that balances the number of images that are nearby and far away in each metadata space is introduced. 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 Jing and Tian (2020). 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.

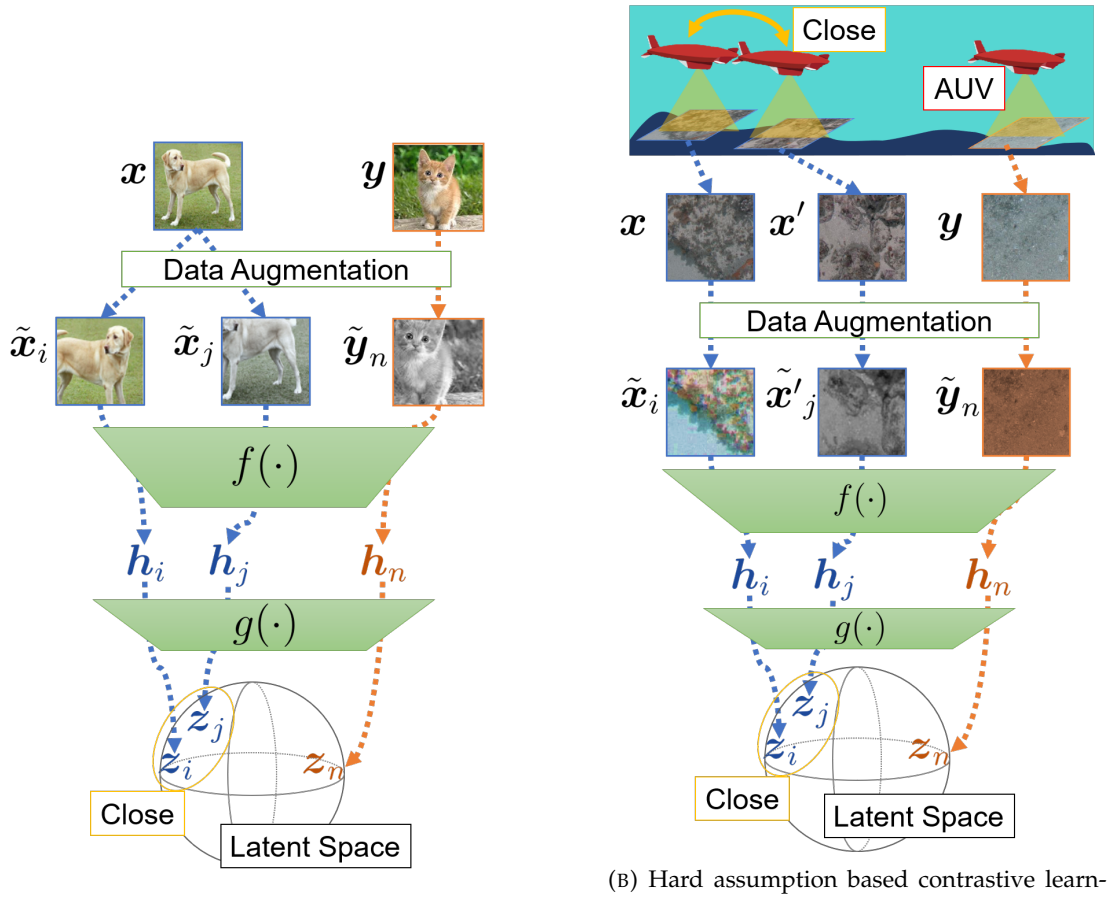
3.3 Representation learning with hard assumption

In section 3.2, the autoencoder based image representation learning technique, where the georeference metadata are leveraged based on *Soft assumption* in section 3.1, is introduced. As described in section 3.1, *Hard assumption* where the similarity of two images is more strongly assumed, would be also useful for seafloor image representation learning. This section investigates whether georeference information can also be leveraged to improve the latent representations generated in contrastive learning (Chen et al., 2020a). Unlike the modified autoencoder loss functions used in section 3.2 where location information can be used to loosely regularise learning, the binary similarity condition that is imposed in contrastive learning forces a much stronger constraint on the latent representations that get generated. In order to validate this similarity assumption, the proposed method takes advantage of the fact that AUVs capture images that often overlap and have footprints that are generally smaller than the patch size of habitats and substrates on the seafloor.

The following subsections give a formulation of state-of-the-art modern contrastive learning approaches such as SimCLR (Chen et al., 2020a), and introduce a novel contrastive learning method for efficient representation of spatially contiguous georeferenced imagery. The validation result of the proposed method is shown in section 4.2.2.

3.3.1 Contrastive learning

SimCLR learns representations by maximising agreement between differently augmented images generated from the same original image. The learning framework, illustrated in Figure 3.3a, consists of four parts; data augmentation, base encoder $f(\cdot)$, projection head $g(\cdot)$ and a contrastive loss function. Data augmentation transforms each image $\boldsymbol{x}$ in the target dataset randomly to artificially generate two correlated images, $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}_j$, where random cropping, colour distortions and Gaussian blur augmentations are applied in this order. The base encoder $f(\cdot)$ is a CNN that extracts representation vectors from the augmented images. The method allows any CNN to be used for $f(\cdot)$, where Chen et al. (2020a) found this approach to be most effective on deeper and wider ResNet (He et al., 2016) architectures. $\boldsymbol{h}_i \in \mathbb{R}^d$ is a feature vector extracted from $\tilde{\boldsymbol{x}}_i$ by the base encoder ($\boldsymbol{h}_i = f(\tilde{\boldsymbol{x}}_i)$). The projection head $g(\cdot)$ is a two layer multilayer perceptron (MLP) to obtain $\boldsymbol{z}_i \in \mathbb{R}^{d'}$ ($\boldsymbol{z}_i = g(\boldsymbol{h}_i)$). The dimension d' of the MLP output are smaller than the dimension d of the base encoder since the contrastive losses defined in lower-dimensional spaces are more efficient for representation learning. A minibatch of n^* original images are taken into consideration at each iteration, so $2n^*$ augmented images including n^* similar pairs are sampled. For a similar pair, other $2(n^* - 1)$ augmented images ($\tilde{\boldsymbol{y}}_n$ in Figure 3.3a) can be regarded as dissimilar examples within the minibatch. The Normalised Temperature-scaled Cross Entropy loss function



(A) SimCLR

(B) Hard assumption based contrastive learning

FIGURE 3.3: Overview of SimCLR and the proposed hard assumption based contrastive learning. The two methods apply different conditions to generate similar pairs of images to implement contrastive learning. In SimCLR (a), similar image pairs [$\tilde{\boldsymbol{x}}_i$, $\tilde{\boldsymbol{x}}_j$] are generated by applying different random augmentations to the same image $\boldsymbol{x}$. The proposed method (b) generates similar pairs [$\tilde{\boldsymbol{x}}_i$, $\tilde{\boldsymbol{x}'}_j$] using different images that were taken from physically nearby locations, $\boldsymbol{x}$ and $\boldsymbol{x}'$. The large range of variability captured in the generated similar pairs allows for robust CNN training.

(*NT-Xent*) (Sohn, 2016; Wu et al., 2018; Oord et al., 2018) between the similar pair $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}_j$ is defined as

$$\ell_{i,j} = -\log \frac{\exp\left(\text{sim}\left(\boldsymbol{z}_i, \boldsymbol{z}_j\right)/\tau\right)}{\sum_{k=1}^{2n^*} \mathbb{1}_{[k\neq i]} \exp\left(\text{sim}\left(\boldsymbol{z}_i, \boldsymbol{z}_k\right)/\tau\right)}, \tag{3.9}$$

where sim$(\cdot)$ denotes cosine similarity, $\mathbb{1}_{[k\neq i]} \in \{0,1\}$ is the indicator function which is 1 if $k \neq i$, and τ is the temperature parameter. The total minibatch loss can be written as,

$$\mathcal{L} = \frac{1}{2n^*} \sum_{k=1}^{n^*} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]. \tag{3.10}$$

The parameters of the base encoder $f(\cdot)$ and the projection head $g(\cdot)$ are updated by a stochastic gradient descent (SGD) optimiser with linear rate scaling (Goyal et al., 2017).

SimCLR can efficiently train CNNs using large unannotated image datasets, where the latent representations derived from the original images $\boldsymbol{x}$ were shown to outperform other state-of-the-art methods in the benchmark classification tasks. It was further shown that fine-tuning of SimCLR trained CNNs can achieve more accurate classification with two orders of magnitude fewer labels than conventional supervised training methods.

3.3.2 Integrating georeference information

A limitation of SimCLR is that the variety of possible image appearances is limited by the types of augmentation used, and only features intrinsic to each image can be efficiently extracted. However, when applied to practical semantic interpretation, people are typically interested in correlating images that show a greater degree of variability than can be described by algorithmic augmentation alone. It can be predicted that the performance of downstream interpretation tasks will benefit if a greater variety of appearances can be integrated into the similar pairs during CNN training. The hard assumption based contrastive learning method proposed in this paper allows great variability to be introduced into the similar image pairs by leveraging the georeference information associated with each image. It can be argued that the level of variability between images taken nearby will exhibit a level of variability that is more representative of that seen across similar habitats or substrates than augmentation alone.

Figure 3.3b shows the overview of the proposed method. Each similar image pair $[\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}'_j]$ is generated from two different images, where $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}'_j$ are generated from $\boldsymbol{x}'$, which is a different image to $\boldsymbol{x}$ but is taken of a physically nearby location. For each image $\boldsymbol{x}$ captured at the 3d georeference of $(g_{east}, g_{north}, g_{depth})$, $\boldsymbol{x}'$ is randomly selected at each iteration from the images which satisfy the following criteria:

$$\sqrt{(g'_{east} - g_{east})^2 + (g'_{north} - g_{north})^2 + \lambda(g'_{depth} - g_{depth})^2} \leq r, \tag{3.11}$$

where $(g'_{east}, g'_{north}, g'_{depth})$ is the 3d georeference of $\boldsymbol{x}'$, λ is the scaling factor for depth direction. Introducing $\lambda > 1$ allows the depth difference between images to be weighted so that the nearby images with large depth gap are not selected, where values of $\lambda < 1$ tend to ignore differences in depth. This flexibility is introduced because the relative impact depth has on image appearance can vary across different application, where for example shallow water application typically have a stronger correlation due to the variable influence of sunlight reaching the seafloor than deep-sea applications. To identify an image pair, the distance r needs to be larger than the distance between adjacent

images taking into account variability in the acquisition interval, and smaller than the patch size of substrates and habitats so that paired images are likely to be similar in appearance. In practise, a small value is advantageous since the similarity assumption is likely to be violated near patch boundaries as r increases. The lower limit for r should also be conservatively set since restricting pairs to only its nearest neighbour means that the same pairing is more likely to be selected multiple times during training, which does not generate any additional information compared to the original SimCLR.

Once $\boldsymbol{x}'$ is selected, the same types of random data augmentation used in SimCLR are applied to each image to obtain the similar pair $[\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}'_j]$.

In this implementation, the hard assumption is leveraged only in data loading and augmentation parts, and the main parts for training, such as the loss function, is the same as the original SimCLR contrastive learning method. Practically, this is preferable because other contrastive learning methods such as SimCLRv2 (Chen et al., 2020b), BYOL (Grill et al., 2020), and Barlow Twins (Zbontar et al., 2021), can be also used in the implementation straightforwardly.

3.4 Unsupervised applications of representation learning

3.4.1 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. For high-dimensional datasets, clustering algorithms often degrade their performance or are even unable to be solved. Both of the two representation learning methods proposed in section 3.2 and section 3.3 can set the number of dimensions of the latent representation to an arbitrary value. Therefore, if the dimensionality of the latent representation $\boldsymbol{h}$ is small enough, clustering techniques can be applied directly without any further dimensional reduction. The non-parametric Bayesian method described in Blei et al. (2006) is potentially preferable for a newly collected image dataset where the number of classes included is usually unknown since it automatically determines the appropriate number of clusters simultaneously during the processing. The clustering result on a real-world seafloor imagery dataset is demonstrated in section 4.3.1.

3.4.2 Representative image identification

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

In the proposed pipeline, k means clustering is applied to the obtained latent representation to identify densely populated regions. The number of clusters should be large enough to avoid missing small classes. As long as this condition is satisfied, the outputs are not strongly sensitive to small differences in k as the clusters attempt to evenly represent the different regions of the latent space. In this work, $k = \lceil k_e/10 \rceil \times 10$ is used, where k_e is a number of clusters estimated by the elbow method (Satopaa et al., 2011). The value of k is k_e rounded up to the nearest ten. Next, a subset of images for prioritised annotation are selected by taking $\lfloor M/k \rfloor$ or $\lceil M/k \rceil$ images from each cluster so that the total number of images is M. This generates a training class distribution that follows the cluster distribution, which eases the class imbalance problem as long as effective clustering is achieved. The way samples are chosen from within each cluster can also affect learning. In Paul et al. (2016), it is assumed that the samples close to the cluster boundaries are important as they have a greater effect on classification

decision boundaries. This assumption is reasonable if the boundaries of clustering and classification are comparable, but in situations where class boundaries are ambiguous, like in many environmental monitoring applications, it is possible that variability in the annotations will degrade learning performance.

In this study, it is assumed that the samples provided for training should represent the variability within each cluster in order to deal with situations where the clustering resolution is not sufficient to resolve class boundaries. Two approaches are implemented to achieve this. The first approach uses k means clustering and randomly samples data from within each cluster so that each cluster in the acquired latent representation is evenly represented in the training data. A more structured form of latent space representation, which is implemented using hierarchical k means clustering, is also investigated. This approach is originally proposed in Nister and Stewenius (2006) where a multi-stage clustering process is introduced. The first stage explores the dominant patterns in the whole dataset, and the following stages attempt to select a representative set of samples from within each cluster. This approach has also been applied to extract representative data in text clustering problems (Gowda et al., 2016). In this work, it is considered to be important to guarantee that samples are selected from dense regions of the latent representation, and so after the first k means clustering, $\lfloor M/k \rfloor$ or $\lceil M/k \rceil$ sub-clusters are generated within each cluster and samples that are closest to each sub-cluster centroid are selected so that the total number of samples is M. This representative image identification is performed on a real-world dataset is in section 4.3.2.

3.5 Efficient alignment with human interests

3.5.1 Content based retrieval

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 $\mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 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 space, georeference information is unnecessary for this application once the autoencoder has been trained. Predicting the performance of the two metrics is difficult for features learnt by an autoencoder since the interpretation of their meaning is non-trivial. In practical cases, adequate metrics should be empirically investigated. The performance results with these similarity metrics are shown in section 4.4.1.

3.5.2 Semi-supervised learning

Data augmentation (Shorten and Khoshgoftaar, 2019) plays an important role in reducing the risk of overfitting during CNN training. Since the features in most images of the seafloor and of land can be considered invariant to rotation and flipping (Cheng et al., 2016), these augmentations are applied randomly during the training process, together with random shift operations to account for uncertainty in position. These transformations are applied with different parameters (i.e. rotation angle and offset) that are randomly assigned every time an image is fed into the model during training. Weighted sampling is also applied at each epoch to balance the number of samples in each class. Data augmentation is not applied to colour and scale distortions since it can be consistently corrected taking into account illumination and turbidity conditions and lens distortions.

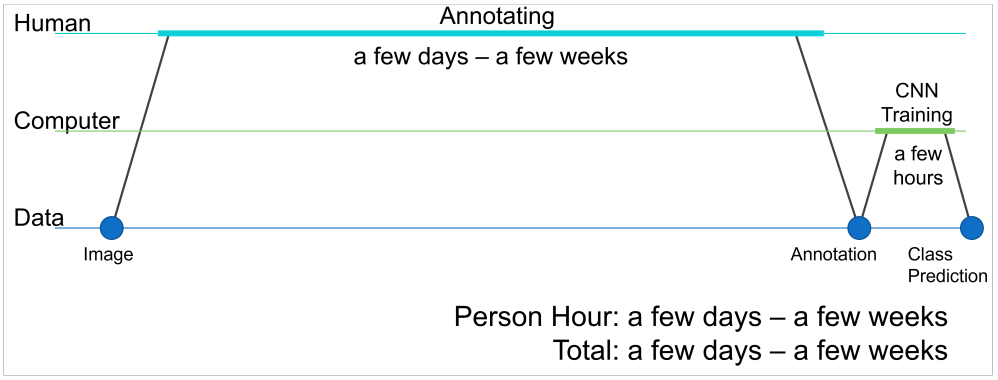
Pseudo-labels are predicted for each unseen image based on its location relative to annotated samples in the acquired latent space. Although the clustering results used to identify images for prioritised annotation can be used for this purpose, the decision boundaries of clusters and classes are not necessarily aligned. Therefore, different approaches are investigated to estimate class decision boundaries, comparing the performance of nearest neighbour (1-NN), Random Forest and SVM (Friedman et al., 2001)

with linear and Radial Basis Function (RBF) kernels as methods capable of expressing varying degrees of complexity of class boundaries in the latent space.

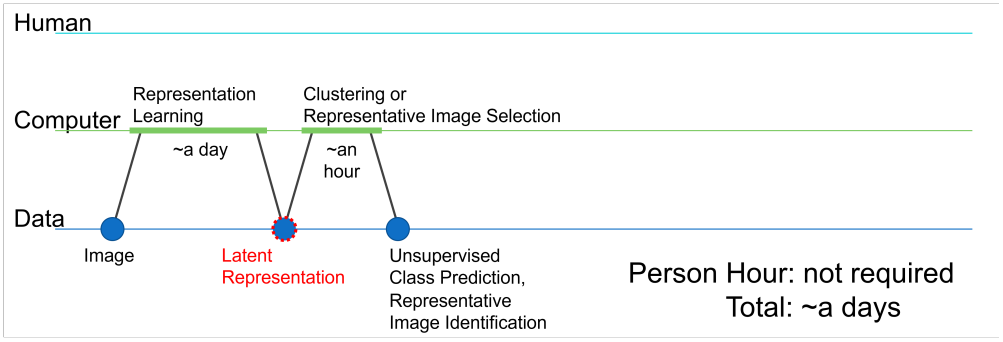
Although the original pseudo-labelling implementation for deep-learning applies a single winner takes all class label to unseen data (Lee, 2013), recent research has demonstrated that taking the uncertainty of each pseudo-label into consideration can improve downstream classification accuracy (Iscen et al., 2019; Arazo et al., 2020). Class boundaries in environmental monitoring data are often ambiguous and so to address uncertainty near class decision boundaries, probabilistic pseudo-labelling using a Gaussian Process classifier (Williams and Rasmussen, 2006) is implemented to predict class conditional probability distributions for each sample in the latent space.

Both the annotations and pseudo-labels assigned to the remaining images are used to train CNNs, where for probabilistic pseudo-labelling, the conditional probability distributions are applied to the softmax loss of CNN training in order to describe the pseudo-label uncertainty. The suitability of these classifiers for pseudo-labelling is determined through validation against human annotations. The validation result of the proposed method is shown in section 4.4.2 and 4.4.3.

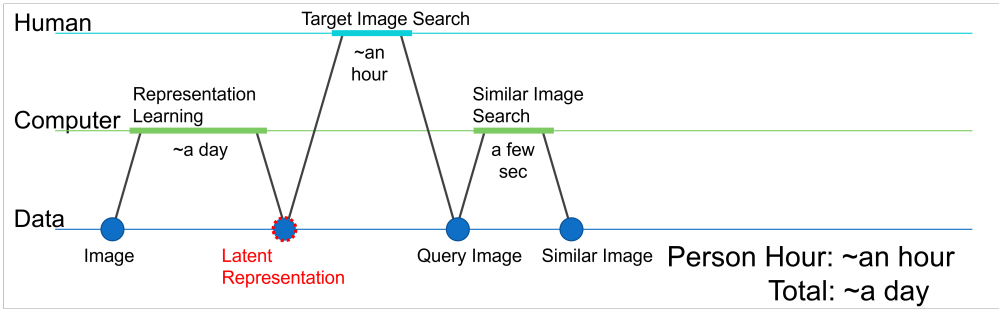
This semi-supervised learning approach can significantly speed up the whole interpretation process. In Figure 3.4, the time required for supervised learning (Figure 3.4a) and the proposed semi-supervised learning (Figure 3.4d) are estimated. Since deep learning CNNs require a large number of annotations for supervised learning, the total time required highly depends on annotating process. When tens to hundreds of thousands of images are included in the target dataset, annotating enough images would take a few days at least. On the other hand, the proposed semi-supervised learning pipeline can select the images for annotation efficiently from the dataset, and so fewer annotations are required. As a result, the required person hours is far fewer than the supervised learning pipeline, and the classification result can be obtained within a significantly short time. The required time estimations of the unsupervised representation pipeline described in section 3.4 and content based retrieval introduced in section 3.5.1 are shown in Figure 3.4b and Figure 3.4c, respectively.



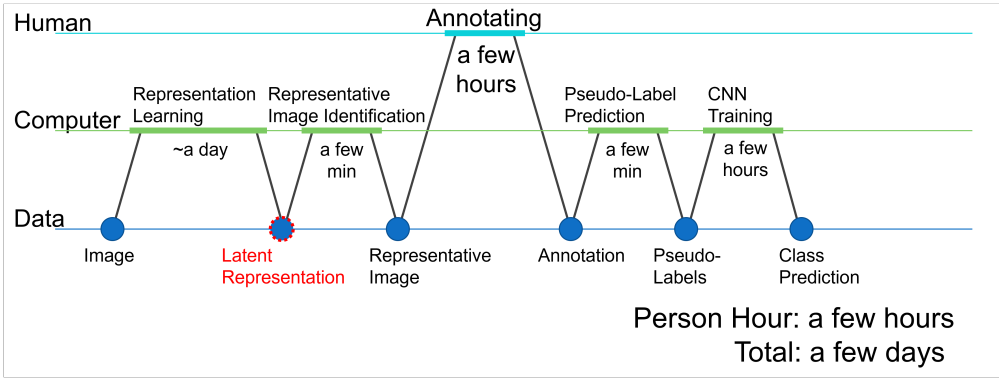
(A) Supervised learning



(B) Clustering and representative image selection



(C) Content based retrieval



(D) Semi-supervised learning

FIGURE 3.4: Swim lane chart of deep learning based supervised learning pipeline (A), clustering and representative image selection (B), content based retrieval (C) and semi-supervised pipeline (D).

Chapter 4

Result

In this chapter, the seafloor interpretation methods proposed in chapter 3 are evaluated. Three real-world seafloor datasets are used for training CNNs by the two representation learning methods, and the performance is evaluated in terms of the classification with the obtained representations. Also, the proposed applications where the obtained representations are exploited are demonstrated. Section 4.1 describes the three real-world seafloor imagery datasets used for the experiments in the following sections. Section 4.2 shows the experiment results for validating the proposed two representation learning methods (e.g. the soft assumption based method and the hard assumption based method introduced in sections 3.2 and 3.3, respectively). Section 4.3 presents the unsupervised interpretation results, e.g. clustering and representative image identification, based on the acquired seafloor image representations. Section 4.4 demonstrates the results of efficient alignment of the acquired representations with human interests.

4.1 Description of datasets

The proposed representation learning techniques and their applications target real-world seafloor imagery. For evaluating their performance in practical scenarios, three real-world seafloor imagery datasets, i.e. Southern Hydrate Ridge dataset, Tasmania dataset and Hippolyte Rocks dataset, are used in the following experiments. This section describes the details of these datasets.

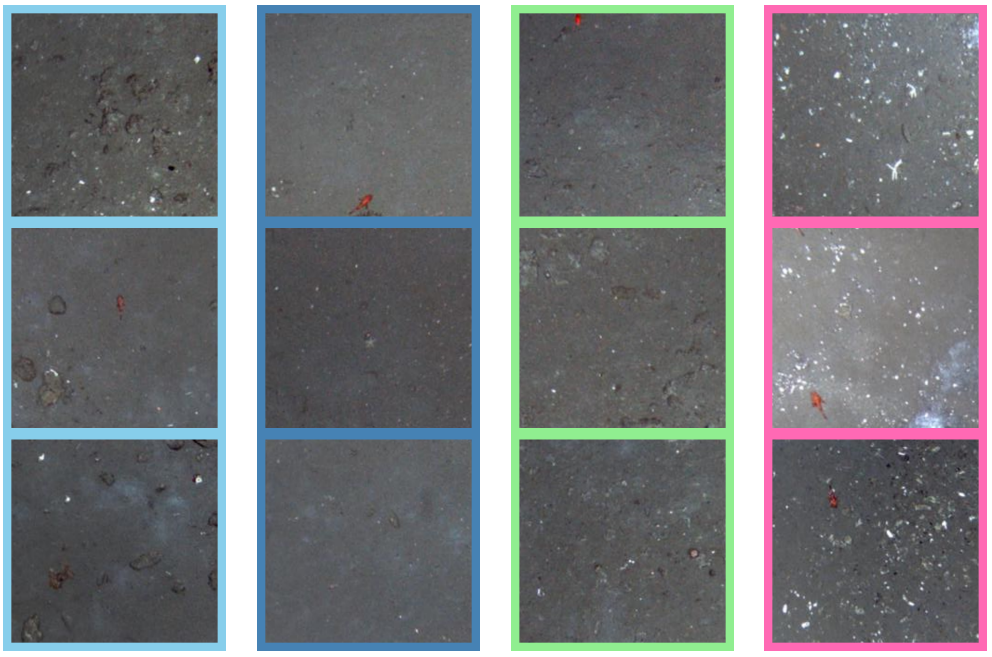
4.1.1 Southern Hydrate Ridge dataset

TABLE 4.1: Description of the Southern Hydrate Ridge dataset

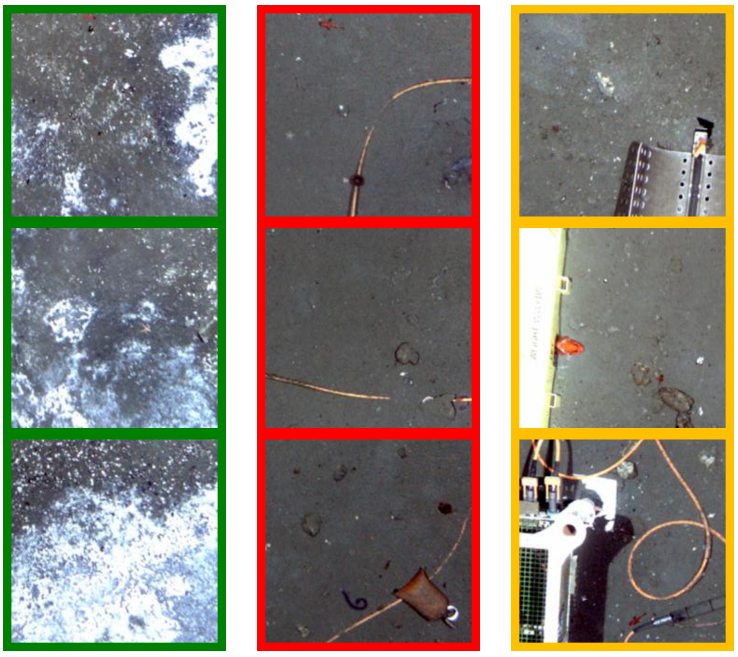
Vehicle	ae2000f
Trajectory Type	Dense Grid
Altitude	5.0 - 7.0 m
Ave. Velocity	0.7 m/s
Camera FoV	68 $\times$ 57 deg
Camera Resolution	1280 $\times$ 1024
Frame Rate	0.25 Hz
Spatial Resolution (Rescaled)	10 mm/pixel
Location	Southern Hydrate Ridge, US
Latitude	44.5683 - 44.5715 $^\circ$ N
Longitude	125.1455 - 125.1506 $^\circ$ W
Seafloor Depth	765 - 785 m
Year	2018
No. of Dives	4
No. of Images	62,875
No. of Annotations	18,740
No. of Classes	7 (See Figure 4.1)

Southern Hydrate Ridge dataset is a seafloor visual imagery dataset 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, US. 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 4.1 gives an overview of the dataset, and Figure 4.2a 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 $\times$ 227, overlapping partially). The original images are first scaled so that they have a constant spatial resolution of 10 mm/pixel. The total number of patches for



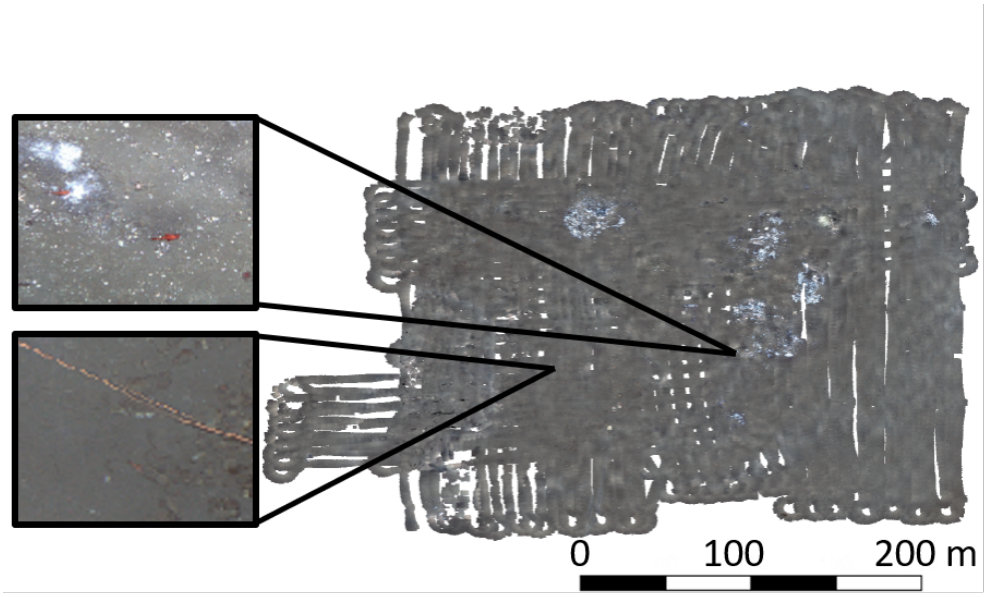
Label	'Rock'	'Sand'	'Carbonate'	'Shell Fragment'
Count	7,660 (40.9 %)	6,781 (36.2 %)	2,014 (10.7 %)	1,151 (6.1 %)



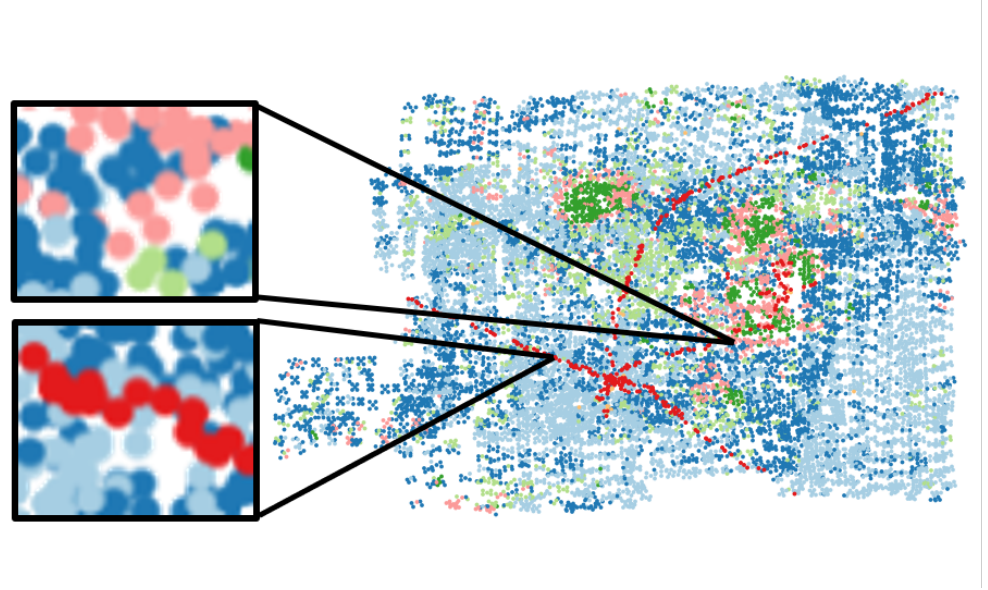
Label	'Bacterial Mat'	'Cable'	'Artificial Object'
Count	751 (4.0 %)	344 (1.8 %)	39 (0.2 %)

FIGURE 4.1: Class example images together with the number of expert human annotations in each class (Southern Hydrate Ridge dataset)

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,



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



(B) Distribution of ground truth category. The same colour scheme is used as in Figure 4.1.

FIGURE 4.2: Overview of the 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 a characteristic distribution.

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.

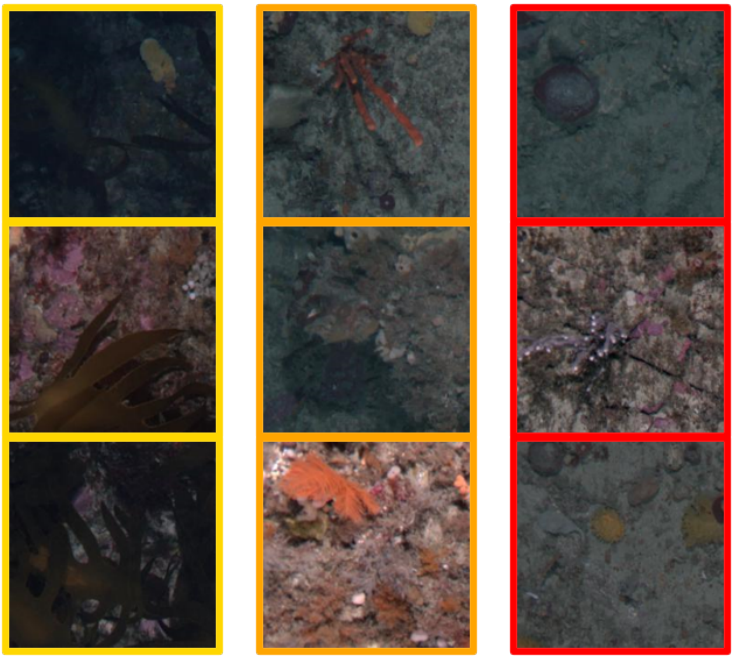
For Southern Hydrate Ridge dataset, ground truth annotations were generated using SQUIDLE+ (Bewley et al., 2015a) by experts for 18,740 (approx. 30 %) image patches randomly selected from the original 62,875 image patches. Figure 4.2b shows the spatial distributions, the numbers and the examples of each 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 4.1, it is noticeable that these categories form larger patterns than the footprint of images; thus, the proposed metadata (georeference) utilisation is assumed to be effective. In this experiment, only the dominant label is given to each image patch based on the individual annotator's judgement. Although this complicates the quantitative evaluation of performance, the relative performance between different conditions of the proposed representation learning can be used to verify how effective the methods developed in this paper are for semantic interpretation.

4.1.2 Tasmania dataset

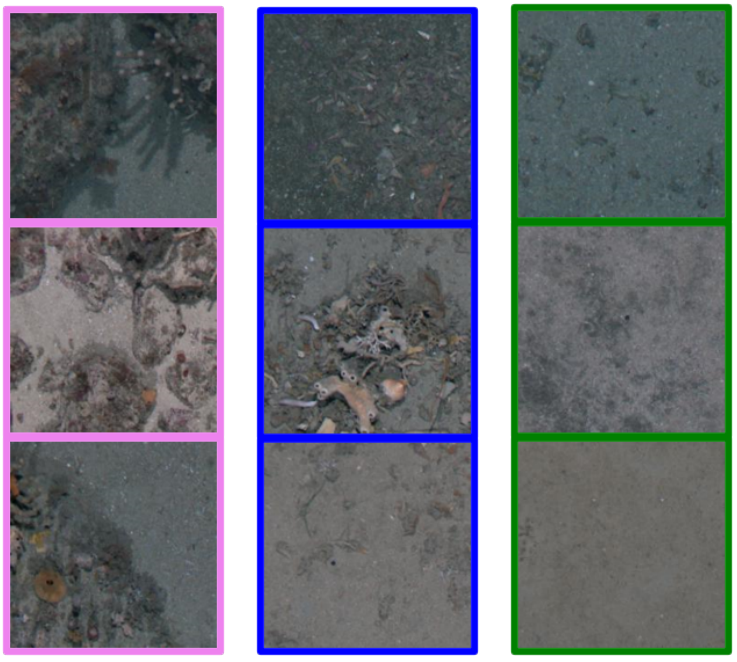
TABLE 4.2: Description of the Tasmania dataset

Vehicle	Sirius AUV
Trajectory Type	Sparse
Altitude	1.0 - 3.0 m
Ave. Velocity	0.5 m/s
Camera FoV	42 × 34 deg
Camera Resolution	1,360 × 1,024
Frame Rate	1 Hz
Spatial Resolution (Rescaled)	2 mm/pixel
Location	East Coast of Tasmania, Australia
Latitude	42.9113 - 43.1289 ° S
Longitude	147.9646 - 148.0555 ° E
Seafloor Depth	28 - 96 m
Year	2008
No. of Dives	12
No. of Images	86,772
No. of Annotations	5,369
No. of Classes	6 (See Figure 4.3)

Tasmania dataset consists of 86,772 seafloor images taken by the Australian Centre for Field Robotics' Sirius AUV from a target altitude of 2 m. The dataset contains habitat and substrate distributions as shown in Figure 4.3, including kelp (A), a registered essential ocean variable, and rocky reefs (B,C,D), which can form habitats for various conservation targets such as coral and sponges (Moltmann et al., 2019b). 5,369 randomly selected images are annotated by human experts into 6 classes. 50 images randomly selected from the 6 classes (total of 300 images) are used for validation and M=[40, 100, 200, 400, 1000] images selected from the remaining 5,069 annotated images are used for training in the following experiments. The georeference information of each image is determined based on the stereo SLAM pipeline described in Mahon et al. (2008) and Johnson-Roberson et al. (2010). The original resolution of the images is 1,360 × 1,024, where the average distance between adjacent images is approximately 0.5 m, so some images partly overlap each other. Prior to analysis, each image in the dataset is re-scaled to a resolution of 2 mm/pixel based on the imaging altitude. Randomly cropped 224 × 224 regions of the images are used for training, where validation is performed on the same sized regions cropped from the centre of the images.



Label	A	B	C
Count	531 (9.9 %)	1,084 (20.2 %)	903 (16.8 %)



Label	D	E	F
Count	598 (11.1 %)	1,568 (29.2 %)	685 (12.8 %)

FIGURE 4.3: Class example images together with the number of expert human annotations in each class (Tasmania dataset)

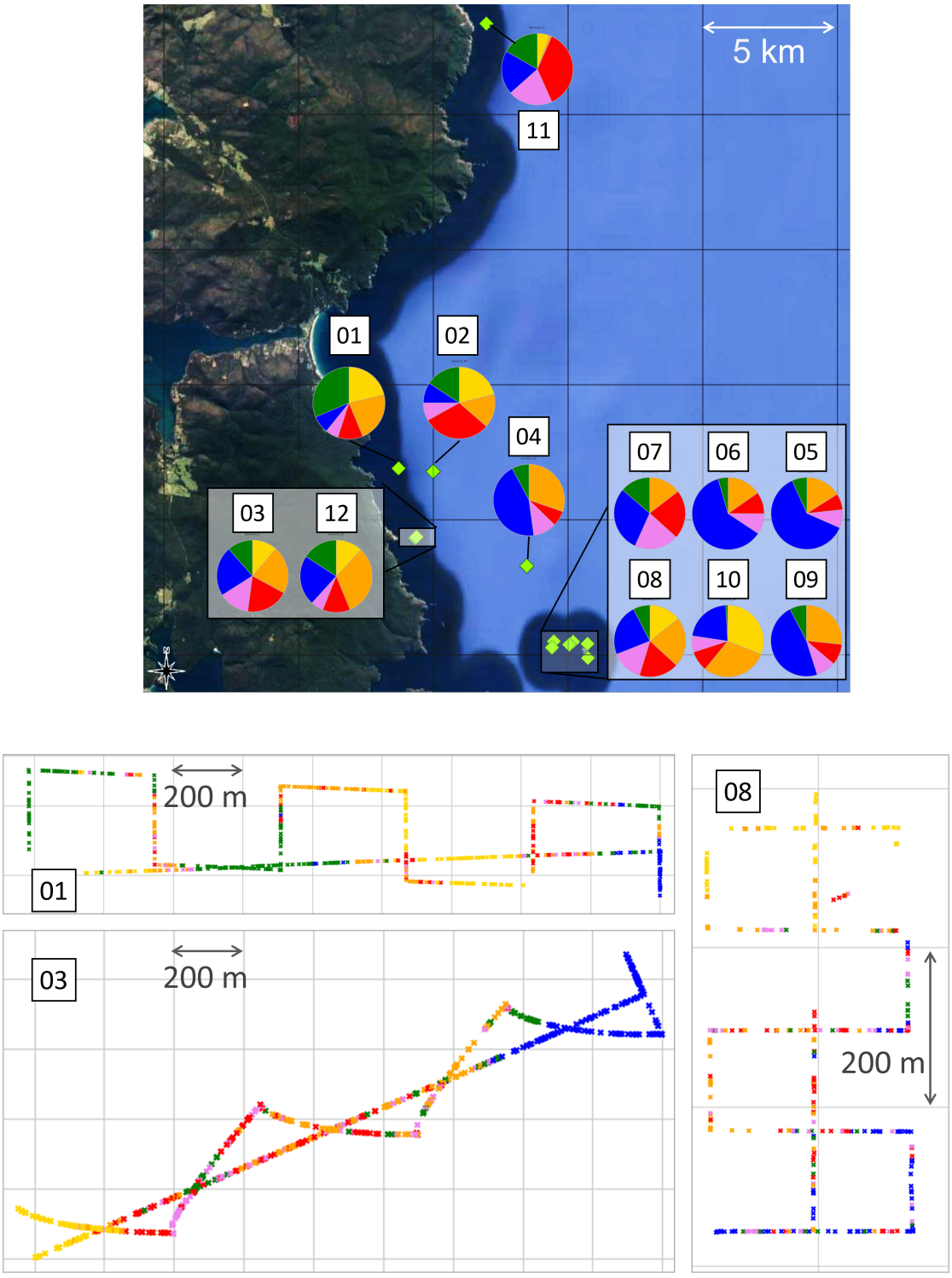


FIGURE 4.4: Overview of the Tasmania dataset. The images were gathered through 12 AUV deployments. The start points of each deployment are shown as green dots. The pie charts show the class distributions according to ground truth. The same colour scheme is used as in Figure 4.3, which shows example images of each class. The survey paths of Dives 01, 03 and 08 are shown with the human annotated class distributions at the bottom.

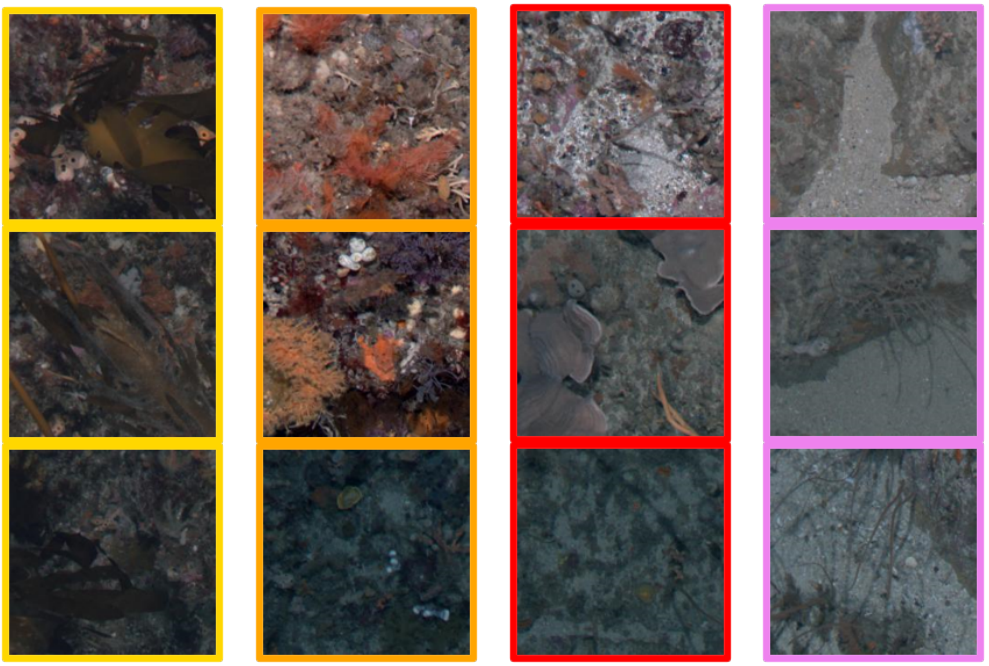
4.1.3 Hippolyte Rocks dataset

TABLE 4.3: Description of the Hippolyte Rocks dataset

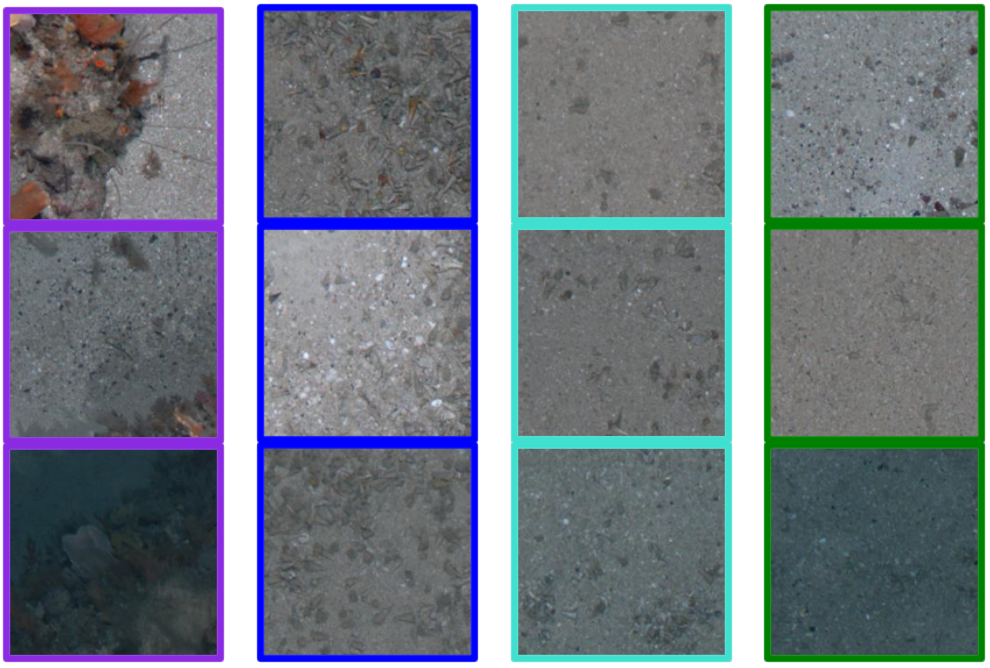
Vehicle	Sirius AUV
Trajectory Type	Sparse
Altitude	1.0 - 3.0 m
Ave. Velocity	0.5 m/s
Camera FoV	42 × 34 deg
Camera Resolution	1,360 × 1,024
Frame Rate	1 Hz
Spatial Resolution (Rescaled)	2 mm/pixel
Location	Hippolyte Rocks, Australia
Latitude	43.1136 - 43.1289 ° S
Longitude	148.0358 - 148.0555 ° E
Seafloor Depth	28 - 96 m
Year	2008
No. of Dives	6
No. of Images	32,097
No. of Annotations	2,221
No. of Classes	8 (See Figure 4.5)

Hippolyte Rocks dataset consists of 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 4.3. Hippolyte Rocks dataset is a subset of Tasmania dataset, corresponding to Dives 07 - 08 in Figure 4.4, but a different class categorisation scheme shown as Figure 4.5 is applied.

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 (Moltmann et al., 2019a). The original resolution of the images is 1,360 × 1,024. Each image in the 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 Figure 4.5, where these are used to validate the performance of the proposed method. Figure 4.6a 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. Figure 4.6b shows the depth distribution of annotated images in each class together with 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.



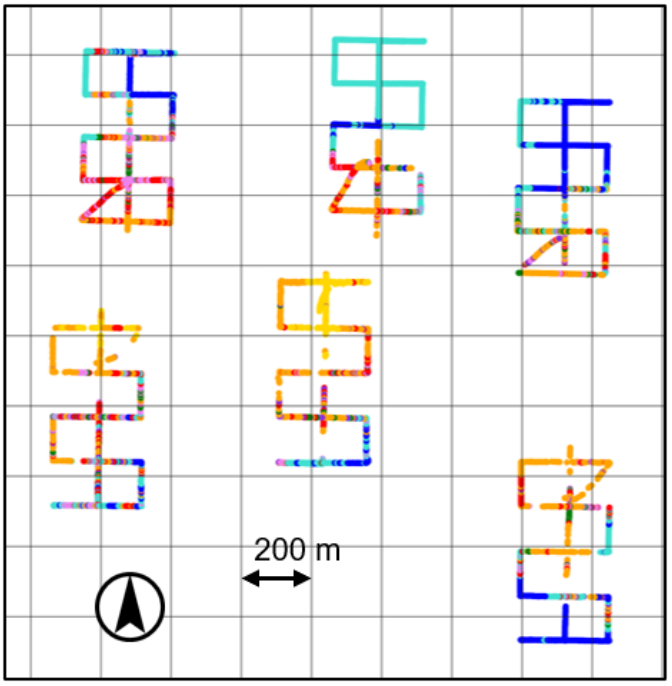
Label	A	B	C	D
Count	152 (6.8 %)	449 (20.2 %)	285 (12.8 %)	157 (7.1 %)



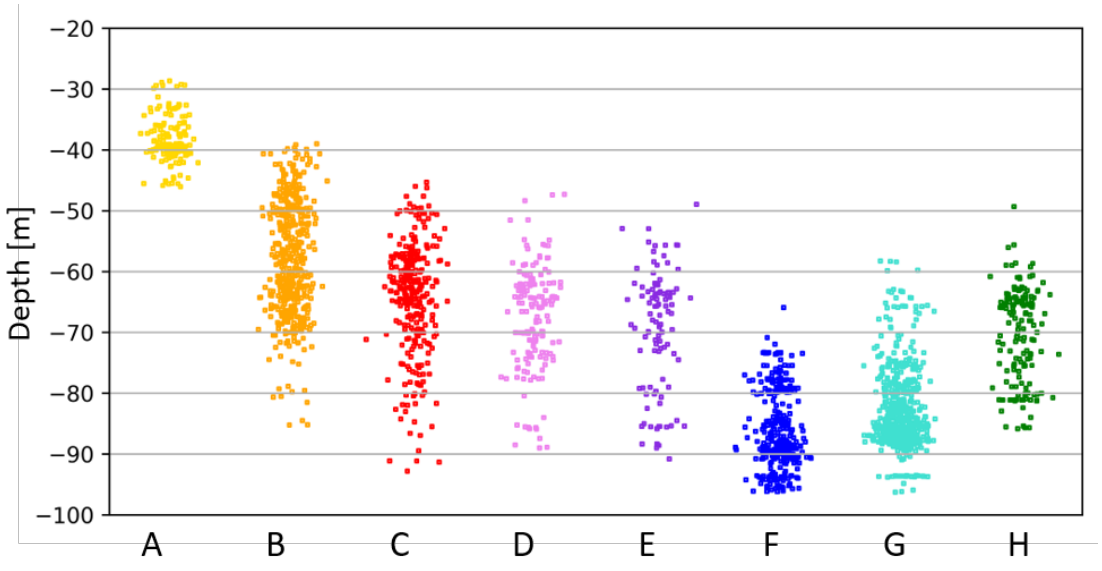
Label	E	F	G	H
Count	101 (4.6 %)	315 (14.2 %)	605 (27.2 %)	157 (7.1 %)

FIGURE 4.5: Class example images together with the number of expert human annotations in each class (Hippolyte Rocks dataset).

Horizontal location and depth estimates for each image are generated based on the Simultaneous Localisation and Mapping (SLAM) pipeline described in Mahon et al. (2008). When georeference metadata is used for the soft assumption based autoencoder



(A) Horizontal distribution.



(B) Class vs depth.

FIGURE 4.6: Overview of the Hippolyte Rocks dataset. The horizontal spatial distribution of the human annotated classes are shown in (A) and the depth distribution of each class is shown in (B), where the same colour scheme has been used throughout the figure.

regularisation proposed in section 3.2.1), georeference errors smaller than σ_{loc} in equation (3.3) or σ_{dep} in equation (3.7) 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 (Paull et al., 2014). 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.

4.2 Representation learning

This section investigates the effectiveness of domain knowledge introduction to seafloor image representation learning by the proposed methods. The representation learning performances are basically evaluated based on the classification accuracy achieved using the acquired representations. In section 4.2.1, the soft assumption based autoencoder performance is compared with a standard CNN based autoencoder to investigate the performance improvement achieved by the regularisation with the novel loss function proposed in section 3.2. In section 4.2.2, the proposed hard assumption based representation learning is compared with the original SimCLR method.

4.2.1 Autoencoder performance validation (soft assumption)

Method and dataset

This section evaluates the representation learning performance of the soft assumption based autoencoder proposed in section 3.2. Hippolyte Rocks dataset introduced in section 4.1.3 is used in the experiment.

Learning configuration

To investigate the effectiveness of the proposed regularisation, the autoencoder is trained (i) without regularisation, (ii) with L_{loc}, (iii) with L_{dep}, (iv) with both L_{loc} and L_{dep} on the dataset. AlexNet (Krizhevsky et al., 2012) 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. 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 equation (3.3) is set to 10.0 m, and σ_{dep} in equation (3.7) 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 equation (3.2) 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.

Evaluation protocol

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). These classifiers are commonly selected in seafloor classification problems (Rigby et al., 2010; Stephens and Diesing, 2014). 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 averaged) is used for performance evaluation, since all classes are considered as equally important. 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, content based retrieval, and use in observation-aware path planning methods.

Result

Table 4.4 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 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.

Figure 4.7 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. This characteristic is observed in the result, where configuration (iv) leads to an 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.

TABLE 4.4: F_1-scores (macro averaged) 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.

Discussion

In conclusion, the experiment result shows that the soft assumption based autoencoder proposed in section 3.2 is efficient in learning representations of seafloor imagery. The result and following discussion in this section can be summarised as follows:

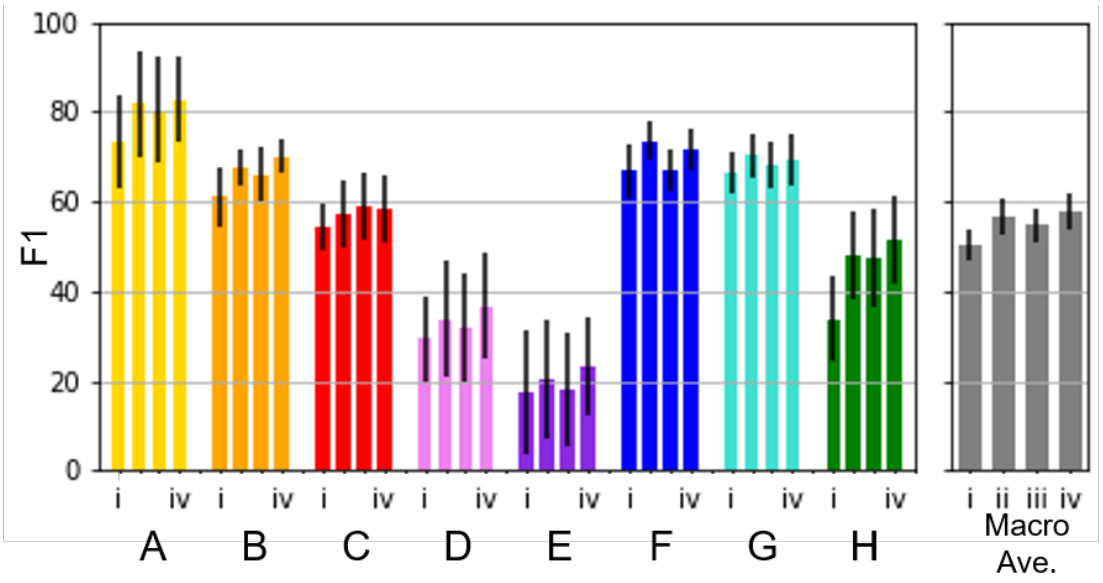


FIGURE 4.7: Per-class F_1-scores and their macro average for (i) no regularisation, (ii) horizontal location regularisation, (iii) depth regularisation and (iv) horizontal location and depth regularisation. R-SVM is used as the classifier in this plot.

- 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.

4.2.2 Contrastive learning performance validation (hard assumption)

Method and dataset

In this section, the representation learning performance of the hard assumption based contrastive learning proposed in section 3.3 is evaluated. Tasmania dataset introduced in section 4.1.2 is used for the training and validation.

Learning configuration

The proposed hard assumption based contrastive learning method can by used to train any type of CNN. The well established ResNet18 (He et al., 2016) is used for benchmarking in this experiment. The latent representation $\boldsymbol{h}$ and $\boldsymbol{z}$ dimensions are set to d=512 and d'=128, respectively. A minibatch size of $n^* = 256$, learning rate of 3.0×10^{-4}, weight decay of 1.0×10^{-4}, temperature τ=0.07 in equation (3.9) was used for all experiments in this section. The thresholds of closeness and depth scaling factor in equation (3.11) were set to r=1.0 m and λ=1.0, respectively. The value for r is conservative compared to the expected substrate and habitat patch size in the surveyed region, and was chosen to yield 2 to 4 nearby images based on the average distance between images (see Table 4.2). This minimises the probability of non-similar image pairs being selected near patch boundaries and the likelihood of duplicate pairs being selected during training. The value of λ was chosen to evenly treat horizontal and vertical displacement between images.

Other than the method for generating similar image pairs, identical parameters were used for the proposed method and the original SimCLR method to allow for comparison. Both methods are trained on the all 86,772 images in the dataset. The performance of the proposed method against transfer learning is also benchmarked, using ResNet18 that was pre-trained on ImageNet.

Evaluation protocol

CNNs trained using three different approaches (ImageNet, SimCLR, the proposed hard assumption based method) are evaluated following the protocol used by Chen et al. (2020a). Once the CNNs are trained, the latent representations they generated are analysed using different classifiers; a linear logistic regression, a non-linear Support Vector Machine with a Radial Basis Function kernel (SVM with RBF), and a fine-tuned CNN classifier. The logistic regression and SVM with RBF are both trained on the latent representation space output $\boldsymbol{h}$ of ResNet18 after CNN training. For fine-tuning, a minibatch size of 256, an Adam optimiser with learning rate of 3.0×10^{-4} and no weight decay were chosen. The macro averaged F_1-score over 6 classes determined from the independent validation set is used to compare the classification accuracy of each training method. All experiments are repeated ten times in each configuration, where the

standard deviation (SD) of scores is shown alongside the mean value to describe variability. The classifiers are trained on M=[40, 100, 200, 400, 1000] images, which are selected so that all six classes are equally represented. This requires significant human effort to determine the relevant classes and identify images corresponding to each class, which is not practical for most field survey scenarios. Other data selection strategies (i.e. random selection and hierarchical k-means clustering based selection proposed in section 3.4.2) with the same representation learning setup are examined in section 4.4.3.

Result

Table 4.5 shows the macro averaged F_1-scores of each CNN training and classifier configuration on the class-balanced subsets of the annotated images. The results show that the proposed hard assumption based learning has the best performance for all values of M, with the linear classifier (C1) showing the best performance for M=40, and the SVM classifier (C2) best for all other M values. The latent representations generated using the proposed method achieves an average 7.4 % and 5.2 % increase in performance compared to the best performing ImageNet and SimCLR trained configurations.

Among the ImageNet pre-trained CNN (A$*$), the CNN fine-tuned using M images (A3) achieves the highest accuracy for all M with an average performance gain of 6.6 %. This is owed to the capacity of CNNs to simultaneously optimise representation learning and classification during training. In A1 and A2, the lower level feature extractor optimised on ImageNet is not updated. The inferior performance compared to A3 indicates that the latent representations generated using ImageNet are suboptimal for the seafloor images used in this work, failing to describe their useful distinguishing features.

In contrast, for SimCLR and the proposed method trained CNNs (B$*$ and C$*$), the fine-tuned scores for the ResNet18 classifier are lower than the scores of linear and SVM classifier. This shows that the constraining effect of contrastive learning in SimCLR or the proposed method training generates highly optimised latent representations. Since conventional fine-tuning does not maintain this constraining effect, it degrades performance, achieving a similar level of accuracy as fine-tuning of ImageNet pre-trained CNN in A3 for larger values of M. This finding is in contrast to the results of Chen et al. (2020a), where fine-tuning of CNNs trained using SimCLR significantly outperform linear classifiers applied to latent representation space for generic terrestrial image datasets analysed. A possible reason for the difference in behaviour is the relatively high dimensionality of $\boldsymbol{h}$ (d=512) compared to the small number of classes (6) in the dataset considered in this paper, combined with the continuous transition of image appearance across the class boundaries, both of which are different to terrestrial benchmark datasets, which typically have a larger number of classes with discrete boundaries, both of which can make the latent representation more sensitive to the constraining effect of contrastive learning. Figure 4.19a shows representative configurations from Table 4.5. The proposed representation learning method with a SVM

TABLE 4.5: CNN training method comparison on class balanced training subset

Config. Label	CNN Training	Classifier	Number of Annotations (M)				
			40	100	200	400	1000
A1	ImageNet	linear	54.9±4.7	61.6±2.8	63.0±2.2	67.5±2.2	67.4±2.1
A2	ImageNet	SVM	47.0±4.9	55.3±4.9	60.2±2.3	66.2±1.1	69.7±1.1
A3	ImageNet	Res18	58.9±2.6	65.5±2.7	68.2±2.5	71.2±1.7	73.8±1.3
B1	SimCLR	linear	62.5±2.7	65.2±2.8	67.1±1.2	69.2±2.2	71.8±1.0
B2	SimCLR	SVM	62.4±2.7	66.9±1.8	69.2±1.8	71.8±1.4	74.1±1.0
B3	SimCLR	Res18	53.4±4.4	61.3±2.2	65.5±2.0	68.9±2.7	72.4±0.9
C1	Proposed	linear	**63.8**±2.9	67.8±2.4	71.4±1.4	72.9±1.8	74.9±1.0
C2	Proposed	SVM	61.7±2.5	**70.1**±2.4	**74.5**±1.4	**75.8**±1.4	**78.3**±1.1
C3	Proposed	Res18	53.6±5.3	62.8±2.2	66.2±2.9	69.5±1.9	73.2±1.3

The CNNs are trained using three different methods (Supervised Learning by ImageNet, SimCLR and Proposed). The latent representations ($\boldsymbol{h}$) extracted from the M annotated images by each CNN are used for logistic regression classification (linear) and SVM (with RBF) training. Also the CNNs are fine-tuned on the same subsets of images. The M images are selected so that all 6 classes in the dataset are evenly described. The classifiers are trained 10 times with different random seeds, and mean and SD values of F_1-scores (macro averaged) are shown. The best score for each M is shown as bold.

classifier (C2) outperforms all other configurations except for B2 when M=40. Having said this, the best performance for M=40 is achieved by the proposed method with a linear classifier (C1) as can be seen in Table 4.5. When the CNNs are fine-tuned, transfer learning with ImageNet (A3) outperforms fine-tuned SimCLR and proposed method (B3 and C3).

Discussion

The experiment result shows that the hard assumption based contrastive learning in section 3.3 is efficient in learning representations of seafloor imagery. The proposed method outperforms existing state-of-the-art contrastive learning (SimCLR) and transfer learning for downstream supervised classification tasks using an equivalent CNN architecture (ResNet18). On an ideal, class balanced training dataset, the SVM with RBF kernel trained on the representations acquired by the proposed method shows an average of 5.2 % and maximum of 7.7 % improvement compared to the accuracy scores of SimCLR for M=[40, 100, 200, 400, 1000] annotations. Compared to ResNet18 trained by transfer learning, an average improvement of 7.4 %, a maximum of 9.2 % is achieved.

4.3 Unsupervised interpretation for scene understanding

In this section, the unsupervised seafloor interpretation pipelines proposed in section 3.4 are examined on the real-world seafloor imagery datasets. In section 4.3.1, the clustering pipeline proposed in section 3.4.1 is performed and evaluated on Southern Hydrate Ridge dataset (section 4.1.1). In section 4.3.2, the representative image selection proposed in section 3.4.2 is demonstrated on Tasmania dataset (section 4.1.2) and Hippolyte Rocks dataset (section 4.1.3).

4.3.1 Clustering

Method and dataset

In this section, the clustering pipeline proposed in section 3.4.1 is performed on the latent representations of Southern Hydrate Ridge dataset (section 4.1.1) acquired by the soft assumption based autoencoder proposed in section 3.2.

Learning configuration

To evaluate the effectiveness of the novel aspects of the proposed soft assumption based representation learning, the autoencoder is trained on the dataset with/without colour attenuation correction, rescaling and the metadata regularisation. Since the depth is less likely to correlate with the image appearances at the depth range of Southern Hydrate Ridge dataset (i.e. 765 - 785 m, see Table 4.1), only horizontal location is used as metadata for the soft assumption based regularisation. For defining the affinity matrix in horizontal location metadata space, Student's t-distribution instead of Gaussian distribution in equation (3.3) is used for computational efficiency. In mini-batch sampling, all images are selected from the neighbourhoods of the first randomly selected image. 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 n^* is fixed as 256, and an Adam optimiser (Kingma and Ba, 2014) is used. When training without the 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. 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 latent representation shows 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 4.5) these metric are used since no standard methods are available that can overcome these limitations.

Representation learning result

The internal evaluation metrics corresponding to each training condition, labelled C_1 to C_9, are shown in Table 4.6. The latent representations $\boldsymbol{h}$ are normalised in each dimension as standard scores. Table 4.6 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 4.8 illustrates the distribution of ground truth in the latent representation space $\boldsymbol{h}$ using t-SNE visualisation (Maaten and Hinton, 2008). Figure 4.8a and 4.8b are for autoencoders trained without/with the georeference regularisation, respectively (corresponding to C_4 and C_8 in Table 4.6). The most distinguishing characteristic of the resulting representation is that the distribution corresponding to 'Cable' forms an obvious cluster at the centre of Figure 4.8b with clear separation from other categories, while it is widely distributed in 4.8a 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 4.8b than in Figure 4.8a, as reflected by the improved evaluation metrics in Table 4.6.

TABLE 4.6: Evaluation results of the proposed feature learning and clustering.

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

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.

Clustering result

For evaluating clustering performance, Normalised Mutual Information (NMI) (Estévez et al., 2009) is used, following the previous works which also use imbalanced seafloor datasets for 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). 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).

Table 4.6 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 introduction of the horizontal location based loss 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 4.6) are shown in Figure 4.9. The relationship between the ground truth and this clustering result are shown in Table 4.7. To obtain a better understanding of each identified cluster, a treemap (Bruls et al., 2000) is shown in Figure 4.10, 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 4.11. 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 4.8 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.

TABLE 4.7: Confusion matrix of the clustering result. Rows and columns correspond to the ground truth and the clustering result using C_8 in Table 4.6, 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

Habitat map

Habitat maps are useful as they summarise the geological and ecological patterns observed in a seafloor region. Figure 4.12 shows the habitat map obtained by plotting the semantic clusters generated by the proposed method. Figure 4.12b shows the result with the highest NMI score (C_8 in Table 4.6), and Figure 4.12a is the clustering result for the same pre-processing steps but without the georeference regularisation (C_4 in Table 4.6). Comparison with the distribution of ground truth in Figure 4.2b illustrates that the habitat map in Figure 4.12b can identify areas corresponding to categories such

TABLE 4.8: Precision, recall and F_1-score for the clustering result using C_8 in Table 4.6. The same cluster merging as in Figure 4.11 is applied. The total accuracy across all categories is 0.56.

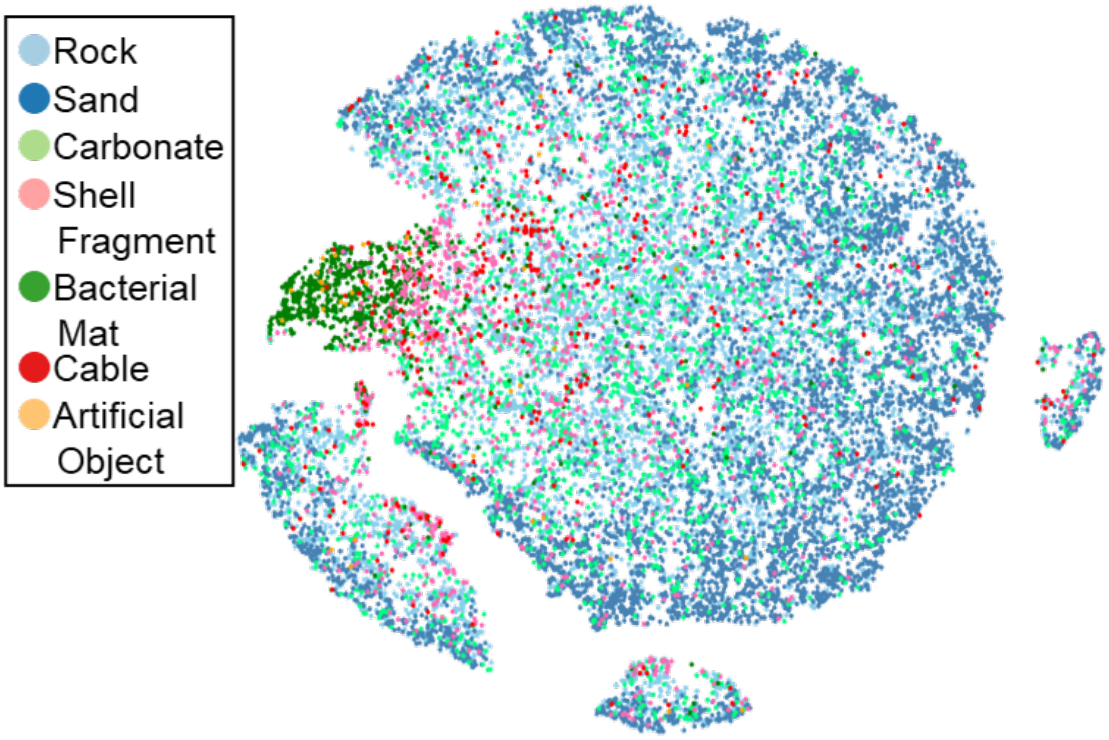
Category	Precision	Recall	F_1-score
Rock	0.68	0.59	0.63
Sand	0.68	0.59	0.63
Carbonate	0.21	0.30	0.25
Shell Fragment	0.41	0.44	0.43
Bacterial Mat	0.61	0.85	0.71
Cable, Artificial Object	0.25	0.63	0.36

as 'Bacterial Mat', 'Shell Fragment', and 'Cable' more effectively than the habitat map in Figure 4.12a. Since these categories have geographic distribution patterns larger than the footprint of an image, the proposed georeference regularisation is effective at extracting the features that are representative of these categories.

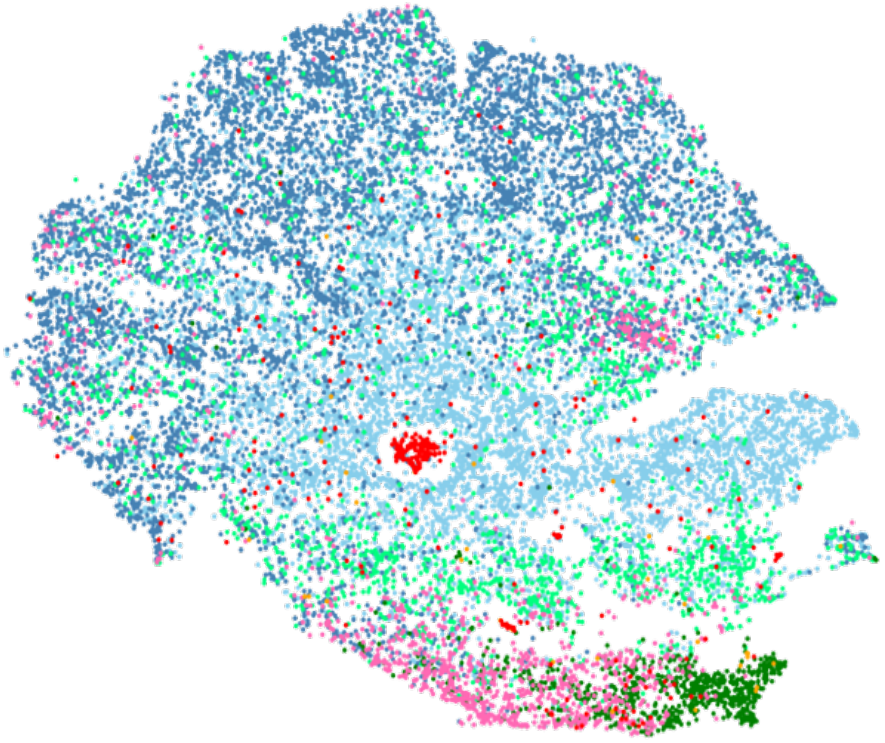
Discussion

The clustering result and the following investigation in this section reveal the effectiveness of metadata introduction to representation learning. The importance of colour correction is also demonstrated. The key findings in this section can be summarised as follows:

- Autoencoders implemented using deep convolutional neural networks form an effective and generic method to learn features in seafloor visual imagery.
- The use of the soft assumption based autoencoder training leads to a factor of two improvement in the retrieval of information from the seafloor images analysed. 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.
- Non-parametric Bayesian unsupervised clustering 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.



(A) Without georeference regularisation (C_4 in Table 4.6)



(B) With georeference regularisation (C_8 in Table 4.6)

FIGURE 4.8: *t*-SNE visualisation of the latent representation $\boldsymbol{h}$ for the ground truth.

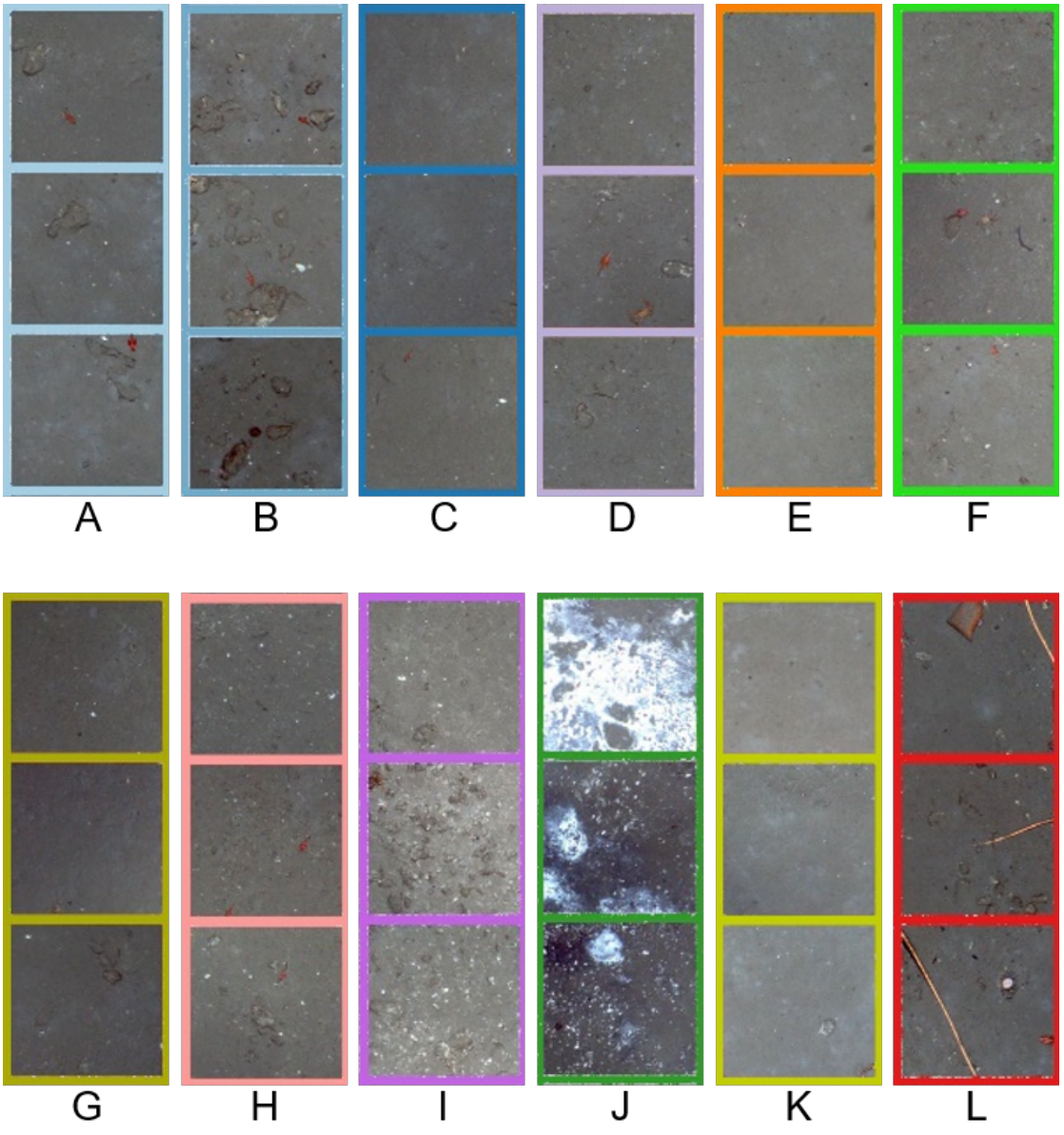


FIGURE 4.9: Representative samples of each cluster (C_8 in Table 4.6).

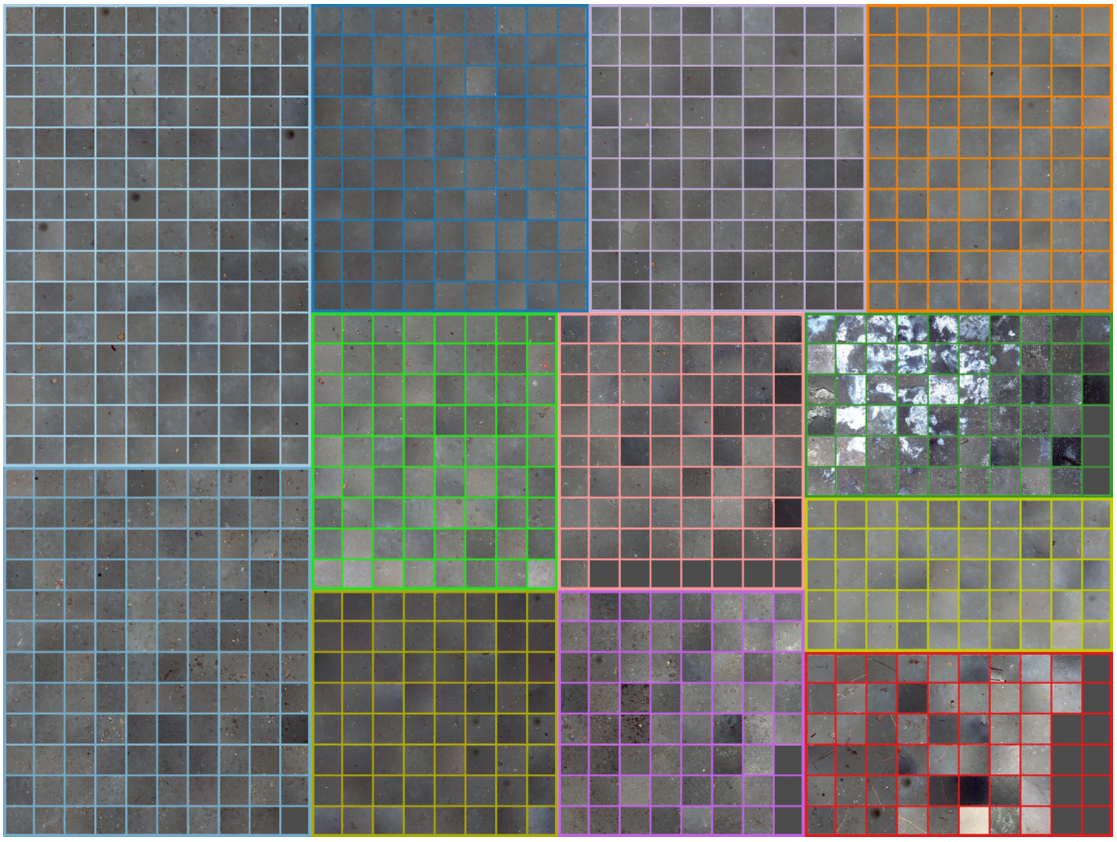


FIGURE 4.10: Visualisation of the sizes of each cluster (C_8 in Table 4.6) using a tree-map representation. The same colours as Figure 4.9 are assigned for each cluster and the areas are proportional to the number of image patches in each cluster.

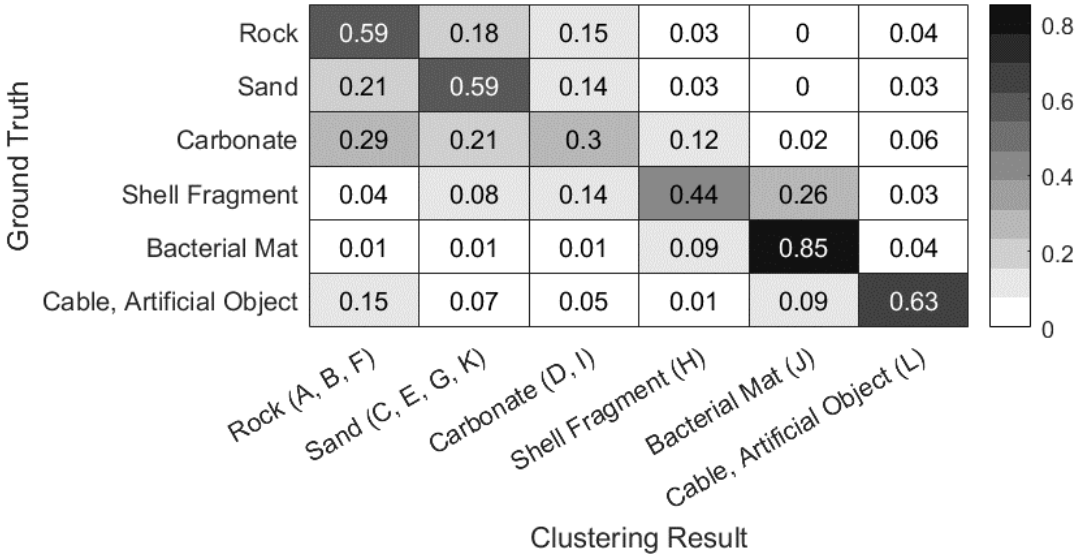
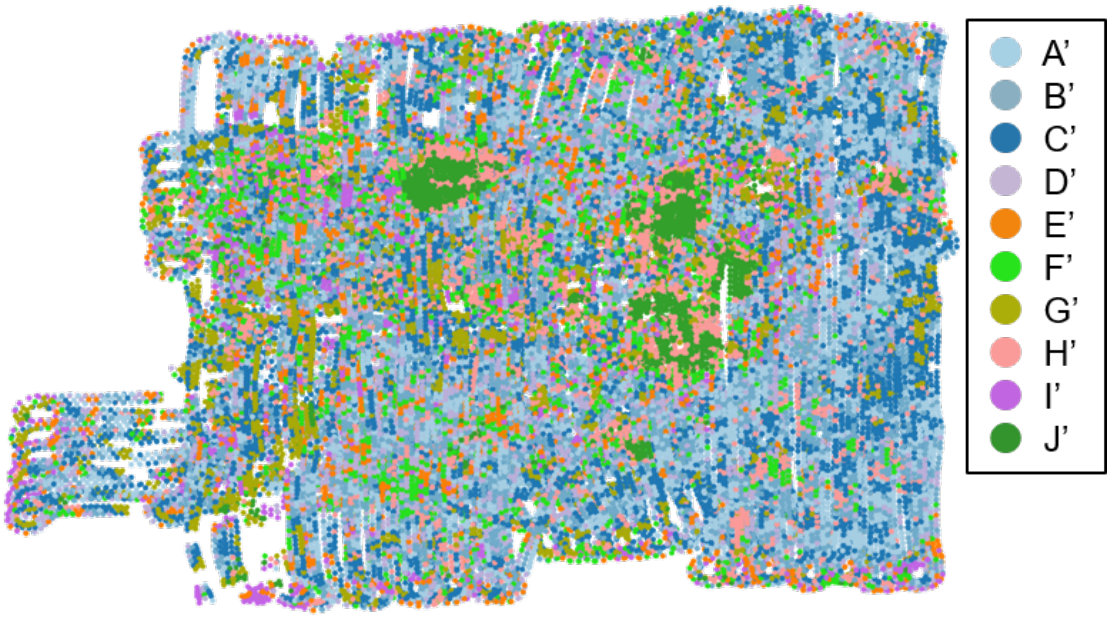


FIGURE 4.11: Confusion matrix between ground truth categories and the unsupervised clustering result using C_8 in Table 4.6. Some clusters and ground truth categories are manually merged based on the appearance of representative images. The values in the matrix are normalised, and diagonal elements correspond to the recall values in Table 4.8.



(A) Without georeference regularisation (C_4 in Table 4.6)

(B) With georeference regularisation (C_8 in Table 4.6)

FIGURE 4.12: Habitat maps based on unsupervised clustering result. The clusters corresponding to 'Bacterial Mat' ('J'), 'Shell Fragment' ('H') and 'Cable' ('L') appear clearly in Figure 4.12b. The results demonstrate that the proposed georeference regularisation enhances clustering performance over wide spatial distributions.

4.3.2 Representative image identification

Method and dataset

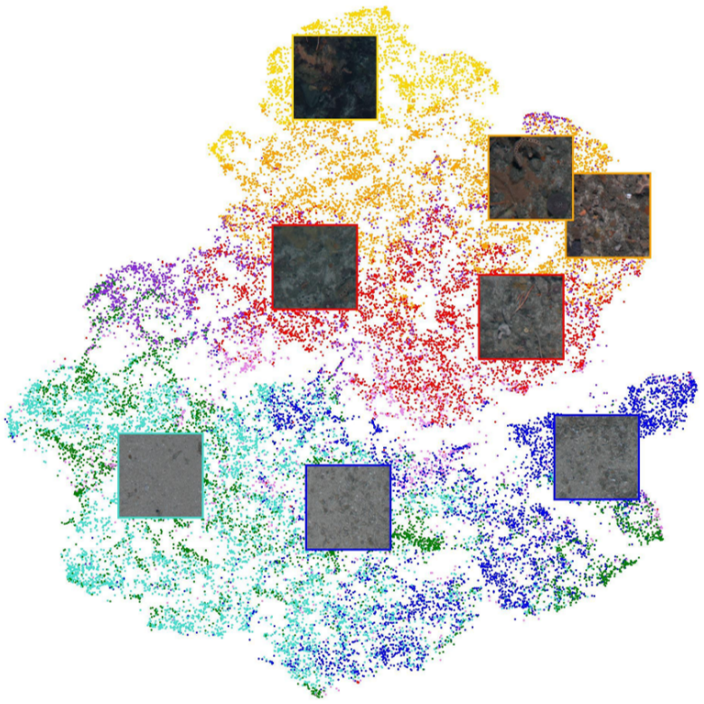
In this section, the representative image identification method proposed in section 3.4.2 is demonstrated. First, the representations of images in Hippolyte Rocks dataset (section 4.1.3) learnt by the soft assumption based autoencoder (section 3.2) is used for the demonstration. The horizontal location and depth are exploited as metadata (corresponding to configuration (iv) in Table 4.4). Then the representative images of Tasmania dataset (section 4.1.2) are identified based on the representations learnt by the hard assumption based contrastive learning (section 3.3).

Result (soft assumption)

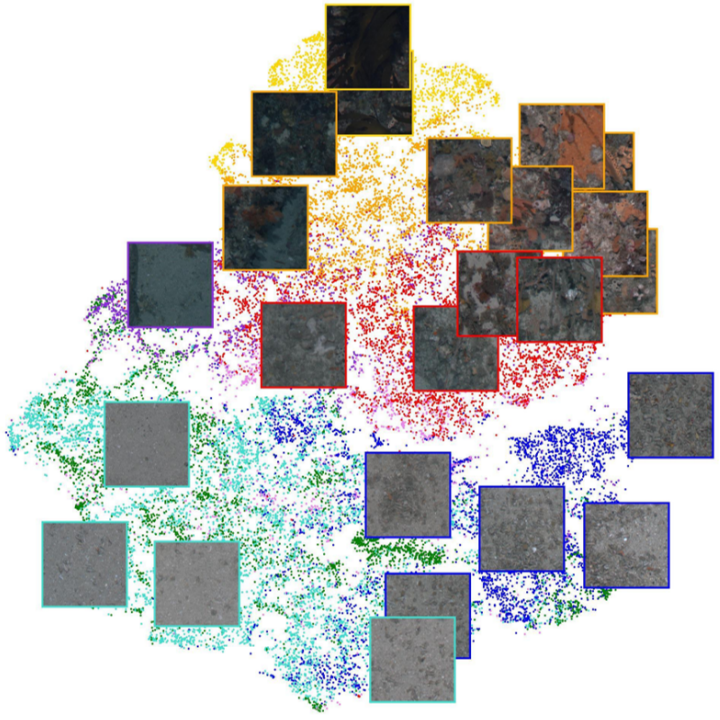
Figure 4.13 shows the automatically selected representative images, overlaid on the representations of Hippolyte Rocks dataset using a t-SNE visualisation (Maaten and Hinton, 2008). In Figure 4.13a, k-means clustering is applied to the acquired latent representations, and the images closest to each of the k centroids are selected as representative images. Here, k=8 which is automatically determined based on the elbow-method (Satopaa et al., 2011) is adopted. In Figure 4.13b, the hierarchical k-means clustering 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.

Result (hard assumption)

Figure 4.14 shows the representative images of Tasmania dataset (section 4.1.2) selected by three different methods, e.g. (a) *Balanced*, where the representatives are selected so that all classes are equally included in number (similar to the training data selection in section 4.2.1), (b) *Random*, where the representatives are randomly selected without any constraint and (c) *H-kmeans*, where the identification method proposed in section 3.4.2 is applied. For representation learning, the hard assumption based contrastive learning (section 3.3) is applied with the same configuration as the experiment in section 3.3. The representatives in each selection strategy are overlaid on the latent representation obtained by the proposed hard assumption based contrastive learning method. t-SNE (Maaten and Hinton, 2008) is applied for visualisation. In this figure, M=30 images are shown for ease of visualisation, where the background points show the image representations that are not selected. The colour of the points and selected image borders



(A) k-means clustering (k=8).



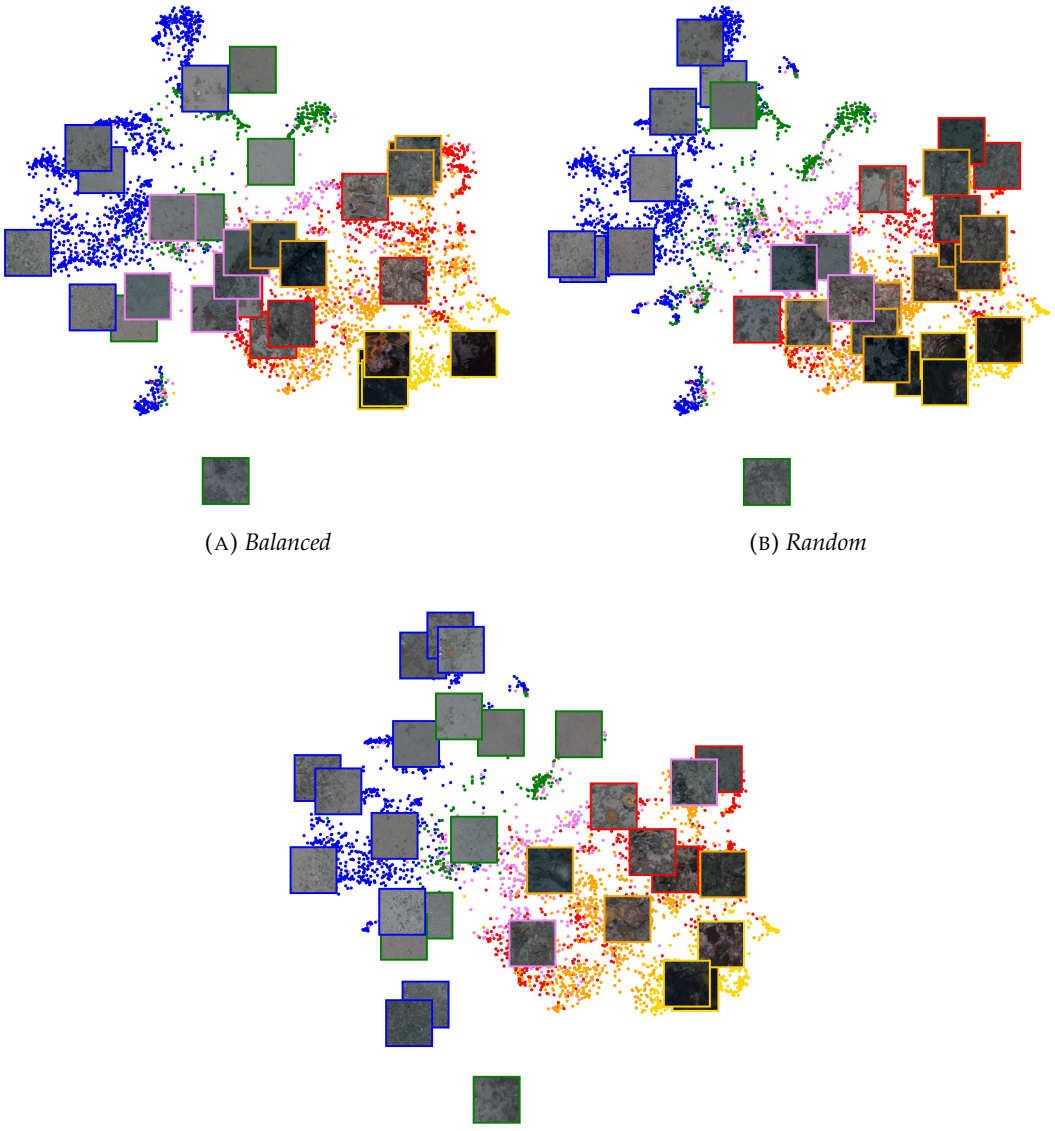
(B) Hierarchical k-means clustering (k=8 and k'=3).

FIGURE 4.13: Representative image identification from the Hippolyte Rocks dataset. The t-SNE latent representation learnt by the proposed method for both horizontal location and depth regularisation, i.e. configuration (iv) in Table 4.4. 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.

illustrate the human class annotation of each annotated image using the same colour key as Figure 4.3. The visualisation shows that *random* selection strategy fails to select images from the central region of the latent representation space that is relatively sparsely populated. On the other hand, *H-kmeans* selects images evenly from the different regions of the latent representation. The *balanced* selection strategy also fails to sample images from several regions. This is because they are mapped to different regions of the latent space as more densely populated regions that have the same class. These undersampled regions of the latent space can be easily confused by a classifier, where the final assigned class will depend on the distribution of nearby training samples.

Discussion

The representative image identification can be applicable to the latent representations obtained by both the soft assumption based autoencoder and the hard assumption based contrastive learning. 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 in real seafloor survey scenarios.



(A) *Balanced* (B) *Random*

(C) *H-kmeans*

FIGURE 4.14: Comparison of representative image identification strategy on the Tasmania dataset. $M = 30$ images are selected by (a) *Balanced*, (b) *Random* and (c) *H-kmeans* strategy. The selected images are shown on *t*-SNE visualisation of latent representations obtained by the hard assumption based contrastive learning. While *Random* sampling fails to select from the centre area, *H-kmeans* successfully select the images in the relatively sparse areas so that a more informative training dataset is gained. Similarly, *balanced* fails to sample regions of the latent space where there are more densely populated regions of the same class. In these situations, class assigned by the classifier will depend on the class of training examples that happen to be nearby. The same colour scheme as Figure 4.3 is applied.

4.4 Efficient alignment with human interests

In this section, the pipelines for aligning the representation learning with human interests proposed in section 3.5 are evaluated on the seafloor imagery datasets. In section 4.4.1, the content based retrieval pipeline proposed in section 3.5.1 is evaluated on Southern Hydrate Ridge dataset (section 4.1.1). In section 4.4.2, the soft assumption based autoencoder (section 3.2) is trained on Southern Hydrate Ridge dataset (section 4.1.1) and the learnt representations are used for the proposed semi-supervised pipeline (section 3.5.2). In section 4.4.3, the hard assumption based contrastive learning (section 3.3) is performed on Tasmania dataset (section 4.1.2), then the outputs are exploited for the proposed semi-supervised pipeline.

4.4.1 Content based retrieval

Method and dataset

This section evaluates the effectiveness of the proposed domain knowledge introduction to content based retrieval in a seafloor imagery dataset (section 3.5.1). Southern Hydrate Ridge dataset (section 4.1.1) is used for the experiment, and the representations are learnt by the proposed soft assumption based autoencoder (section 3.2) trained in the same configuration as the experiment in section 4.3.1.

The performance of content based retrieval using Euclidean distance and cosine similarity are quantitatively evaluated by taking the average values of top-10 retrieval accuracy, defined as the rate of images retrieved with the same ground truth category as the retrieval image for each ground truth category (Wu et al., 2013). Representation learning is achieved by the autoencoder trained with/without the proposed metadata regularisation loss function and with/without rescaling. These correspond to autoencoders labelled C_4, C_5, C_8 and C_9, respectively in Table 4.6 in section 4.3.1.

Result

The results in Table 4.9 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 learning the representations 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 not need to make such explicit assumptions.

Regarding the similarity metrics, the proposed loss function for soft assumption 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 latent space is challenging, and the results indicate that Euclidean distance and cosine similarity are almost equivalent for the dataset used in this study.

TABLE 4.9: 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 4.6, the proposed georeference regularisation significantly improves the accuracy scores, especially for 'Cable' which has a characteristic spatial distribution.

Condition in Table 4.6	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 4.6 (for Reference)	0.10		0.10		**0.23**		0.22	

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 retrieval can generate target specific distribution maps using the same latent 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 retrieval 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.

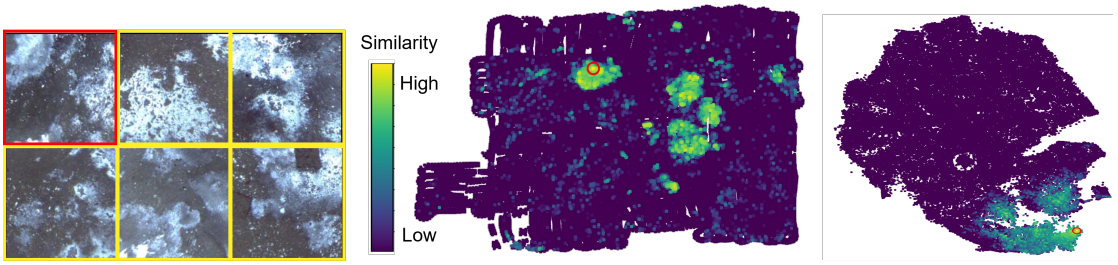
Similarity map

The similarity maps in Figure 4.15 show some results of content based retrieval and the locations of images that have a similar appearance. Figure 4.15a shows the result of a bacterial mat image search. On the whole, the areas with high similarity in the similarity map show similar distributions to the 'Bacterial Mat' in the ground truth (Figure 4.2b). Since the similarity scores vary continuously, the result is useful for analysing small differences between images which are categorised as 'Bacterial Mat'.

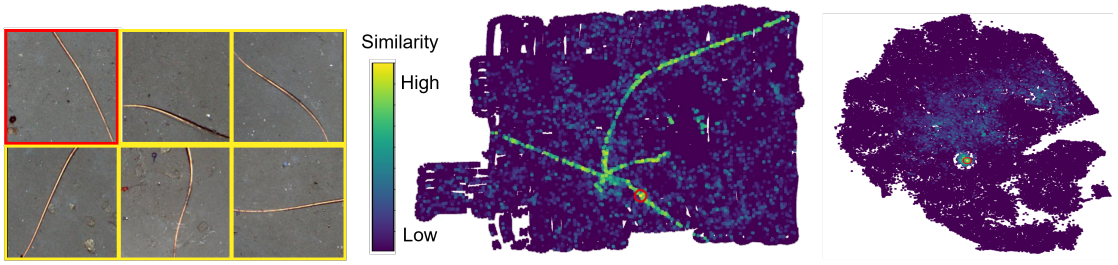
Figure 4.15b shows the result when a typical image of a cable is chosen as a query. The content based retrieval successfully extracts cables deployed in this area, and the similarity map shows the distribution of cables more clearly than the clustering result (Figure 4.12b). Features that are small, 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 4.9). However, relatively minor categories such as these can be effectively found using content based retrieval, where Figure 4.15c and 4.15d show the different distributions in this area. These similarity 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 4.15 are spread widely and are not limited within the neighbouring area of the query images. This fact confirms that the proposed loss function for the soft assumption allows meaningful features to be extracted from the images themselves without over-regularising the results of the content based retrieval. Looking more closely at Figure 4.15d 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 retrieval may be useful to reduce the effort required for manual annotation by filtering out candidate images that are more likely to contain the targets of interest.

Discussion

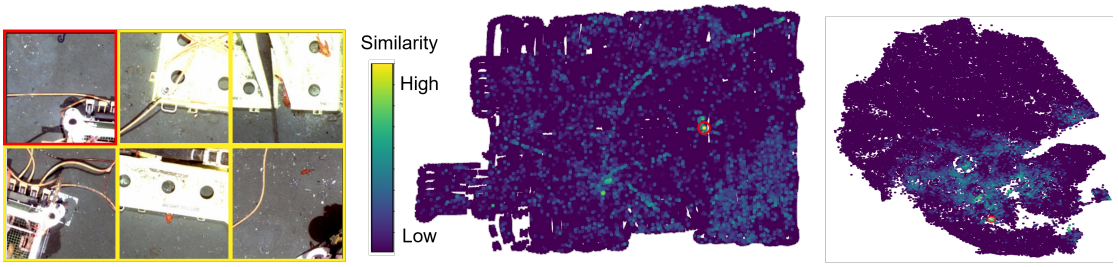
The content based retrieval results show that the retrieval accuracy is significantly improved by exploiting horizontal location in autoencoder training. Also, it is revealed that correction for spatial scale and distortion of images prior to representation learning improves the performance against artificial 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.



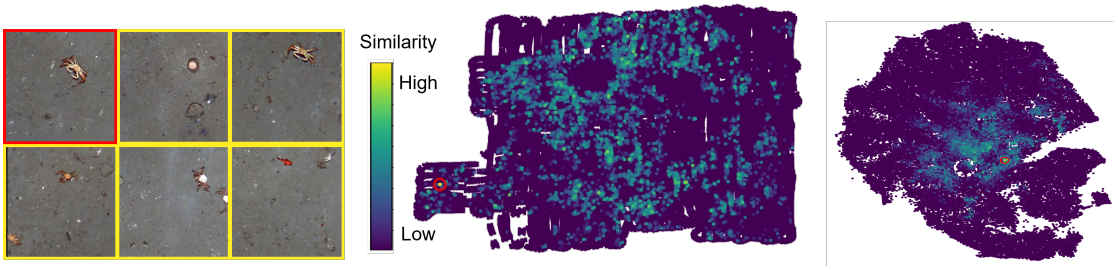
(A) Bacterial mat



(B) Cable



(C) Infrastructure



(D) Crab

FIGURE 4.15: Content based retrieval result. Red frame images are query and yellow frame images are Top 5 similarity images. The maps on the middle show the similarity distributions between the queries and all the other images in the dataset. The red circles in the similarity maps show the location of the query images. The plots on the right show similarity values corresponding to all images in the dataset in the latent space, visualised by *t*-SNE.

4.4.2 Semi-supervised learning (soft assumption)

Method and dataset

This section evaluates the proposed semi-supervised interpretation (section 3.5.2). The representations are learnt by the proposed soft assumption based autoencoder (section 3.2) trained in the same configuration as C_9 in Table 4.6 in section 4.3.1 and 4.4.1, i.e. horizontal location is exploited as metadata and both the light attenuation correction and rescaling is applied to the images. The Southern Hydrate Ridge dataset (section 4.1.1) is used for the training and test. The 'Cable' class and 'Artificial Object' class are merged into a single 'Artificial Object' class in this experiment, so there are six ground truth classes in total.

Classification with conventional classifiers

First, the performance of conventional (non-CNN) classifiers is investigated in order to generate effective pseudo-labels from a small subset of annotated examples. Five well established classifiers; k-NN with k = 1 (1-NN), Random Forest, SVM with two kernel types (Linear and RBF) (Friedman et al., 2001) and Gaussian Process (Williams and Rasmussen, 2006), are applied to the latent space mapped by an autoencoder with the soft assumption based regularisation that has been trained on all available image patches. The results are compared to those with a standard convolutional autoencoder without the proposed metadata based regularisation. To evaluate the performance with a small number of annotations, an adjusted cross-validation is applied. First, half (i.e. 9370) of the annotated image patches are randomly selected as a test dataset, preserving the class distribution of the entire dataset. Then M images are selected from the remaining patches based on random selection, k means based selection, and the proposed hierarchical k means based selection. Following the equation defined in section 3.4.2, $k = 20$ is used for both k means and hierarchical k means based selection. In k means based selection, $M/20$ images are selected randomly from each cluster. In hierarchical k means based selection, the second stage k means is applied to each cluster to find $M/20$ subcluster centroids, and the images closest to each centroid are selected for annotation. Training and testing are executed ten times for each configuration with M = 20, 40, 100, 200, 400, 1000 and 9370. When M is 9370, all available training images are selected and so the sampling method used becomes irrelevant.

Table 4.10 shows the mean and SD of the F_1-scores for a ten-time cross-validation with each configuration (A1 - A20). The data selection strategy has a greater impact on performance than the choice of classifier, with all classifiers benefiting significantly from hierarchical k means prioritisation. The relative gains in accuracy compared to random selection are especially large for small values M (20, 40 and 100), confirming the importance of the data selection strategy when training with a small number of annotations. For all values of M, the combination of the soft assumption based autoencoder

TABLE 4.10: F_1-scores (macro averaged) mean and SD (%) of the classification result with conventional classifiers

Config. Label	Rep. Learning	Data Selection	Classifier	Number of Annotations (M)						
				20	40	100	200	400	1000	9370
A1	with	random	1-NN	31.8±9.1	40.1±3.0	44.0±4.5	47.6±3.3	48.9±3.1	50.6±1.5	54.0±0.5
A2	soft		Random Forest	27.6±6.8	38.3±4.1	43.0±4.5	48.7±4.9	51.8±3.3	56.4±1.4	61.0±0.3
A3	assumption		Linear SVM	33.8±9.6	43.5±3.9	48.2±4.4	52.6±3.4	54.3±2.5	56.3±1.9	60.0±0.5
A4			RBF SVM	31.6±7.6	42.4±3.7	48.4±3.7	54.4±3.3	57.4±3.7	60.2±0.8	**63.3**±0.7
A5			Gaussian Process	29.6±5.2	39.1±6.2	46.1±3.9	47.2±5.2	52.5±2.2	56.9±2.0	***63.2***±0.6
A6		*k* means	1-NN	41.6±5.1	46.2±5.2	47.2±4.3	50.8±1.7	51.0±1.9	52.5±1.1	54.0±0.5
A7			Random Forest	33.7±5.7	43.1±5.2	49.2±4.0	54.7±1.2	56.9±1.6	59.1±0.7	61.0±0.3
A8			Linear SVM	43.2±5.4	47.9±5.8	51.5±4.2	56.6±1.3	57.1±1.4	59.9±1.0	60.0±0.5
A9			RBF SVM	42.0±6.0	50.7±5.3	***55.1***±4.4	***59.2***±1.5	***60.5***±1.3	***62.4***±0.8	**63.3**±0.7
A10			Gaussian Process	42.1±6.8	45.8±6.8	51.8±2.5	55.1±1.5	57.2±1.6	60.0±0.8	***63.2***±0.6
A11		H-*k* means	1-NN	46.9±7.2	48.6±4.3	48.9±2.9	52.2±2.4	52.3±1.7	53.0±0.8	54.0±0.5
A12			Random Forest	42.1±7.4	47.9±3.9	51.8±2.5	55.8±1.5	57.6±1.5	59.3±0.9	61.0±0.3
A13			Linear SVM	***47.4***±8.1	50.9±4.7	53.6±3.0	56.8±1.6	58.3±1.2	60.8±0.9	60.0±0.5
A14			RBF SVM	**48.0**±8.3	**54.8**±2.3	**56.9**±2.0	**60.1**±1.0	**61.0**±1.0	**62.7**±0.7	**63.3**±0.7
A15			Gaussian Process	44.5±7.7	***51.4***±3.8	***55.1***±2.3	56.1±2.1	59.5±1.2	61.2±1.1	***63.2***±0.6
A16	without	H-*k* means	1-NN	25.5±1.3	30.5±1.5	33.2±1.0	33.8±1.2	35.6±1.4	36.6±0.8	38.3±0.5
A17	soft		Random Forest	24.4±1.7	29.0±3.0	32.0±1.6	33.6±2.2	35.6±1.1	39.1±0.8	41.1±0.4
A18	assumption		Linear SVM	10.0±5.6	8.3±4.5	6.0±3.4	8.5±8.5	6.7±2.6	10.9±3.1	34.9±0.7
A19			RBF SVM	21.7±3.4	28.2±2.6	29.6±4.0	35.0±1.8	38.3±1.5	42.0±0.9	44.9±0.6
A20			Gaussian Process	9.7±0.0	9.7±0.0	9.7±0.0	10.3±1.4	14.9±1.3	18.9±0.8	21.5±0.3

Classifiers are applied to the latent representations obtained by the soft assumption based autoencoder and a standard convolutional autoencoder. Three data selection strategies and five classifiers are validated on the Southern Hydrate Ridge dataset with different numbers of annotations (M). The combination of the soft assumption based autoencoder, hierarchical k means (H-k means) based data selection and SVM with RBF kernel performs the best for all M. When $M = 9370$, all images other than the test images are selected as the training images regardless of the data selection strategy. Bold and bold italics indicate the best and next best performer for each value of M.

pre-training and hierarchical k means based data selection with a RBF kernel SVM (configuration A14) performs the best among the tested cases. The linear kernel SVM and Gaussian Process generally perform better than 1-NN and Random Forest, where the linear SVM tends to be better for small values of M and Gaussian Process better for larger M.

The standard deep learning autoencoder (configuration A16 - A20) is far less effective than the autoencoder based on the soft assumption. This is an expected result since the experiment result in section 4.3.1 has already shown that the autoencoder achieves poor clustering performance without the georeference regularisation, and the underlying assumption behind the data selection strategies investigated here is that effective clustering can be achieved. The same trend as Table 4.10 is observed for three aerial imagery datasets, where the results are shown in Appendix A. This demonstrates that the proposed location guided latent representation learning and representative image selection are effective for environmental applications across different types of georeferenced image datasets and domains.

Classification with CNN (CNN architecture comparison)

This section evaluates the proposed semi-supervised learning pipeline's performance using CNNs. The M training images and test images are selected in the same way as the

experiment for Table 4.10. When M is smaller than the total number of available training data, data augmentation, pseudo-labelling (PL) or probabilistic pseudo-labelling (PPL) are applied so that the number of training images at each epoch is the same as the total number of the training images to allow for fair comparison of the results.

The proposed semi-supervised training method can be applied to any CNN architecture. Here, The impact of using the following three well established CNN architectures are investigated on classification accuracy: AlexNet, ResNet18 and ResNet152 (He et al., 2016). The accuracy of each configuration is evaluated based on the mean F_1 (macro averaged).

Each CNN is pre-trained using ImageNet, where experiments are performed with all layers and only last layer training on the Southern Hydrate Ridge dataset following the network-based transfer learning process described in Tan et al. (2018). Since AlexNet is used as the basic architecture of the autoencoder implemented in this work, its encoder part can be regarded as an AlexNet classifier where the weight values have been optimised to describe all the available images in the target dataset through latent representation learning. The performance of the soft assumption based autoencoder pre-trained CNN is compared to traditional ImageNet pre-trained CNNs to assess the effectiveness of embedding additional information acquired from metadata.

The following parameters are experimentally determined: Mini-batch sizes of 128 samples are used for AlexNet (all layer and final layer training) and ResNet18 (all layer training), 32 for ResNet18 (final layer training) and 16 for ResNet152, Adam (Kingma and Ba, 2014) is used as the optimiser and the learning rate is set to 1e-5 except for ResNet18 (final layer training) where it is set to 1e-4, and the number of training epochs is 50 for all configurations.

Table 4.11 shows the results for configuration B1 to B8. As expected, the accuracy improves when a larger number of annotations are used to train each CNN architecture. Overall, B4, which corresponds to AlexNet pre-trained with the soft assumption where only the last layer is trained on the Southern Hydrate Ridge dataset, shows the best performance except for when $M = 40$ and 9370. The performance gap between B4 and B8, where all the layers are trained on the Southern Hydrate Ridge dataset, is potentially caused by overfitting due to high model flexibility of B8. Though B8 outperforms B4 for $M = 40$, the difference in performance here is marginal. B8 shows a similar level of accuracy to B5, where ImageNet is used for pre-training instead of the the soft assumption based autoencoder, indicating that the advantage of the pre-training is lost when all the layers are trained. When $M = 9370$, B6, corresponding to the case where all the layers of an ImageNet pre-trained ResNet18 are trained on the Southern Hydrate Ridge dataset, shows the best accuracy. This suggests that ResNet18's deeper architecture and use of residual blocks allows for better performance than AlexNet when a sufficient number of training examples is available. However, B4 is the best option

TABLE 4.11: F_1-scores (macro averaged) mean and SD (%) of the classification result with CNN Trained by standard supervised learning, active learning and the proposed-supervised learning.

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

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

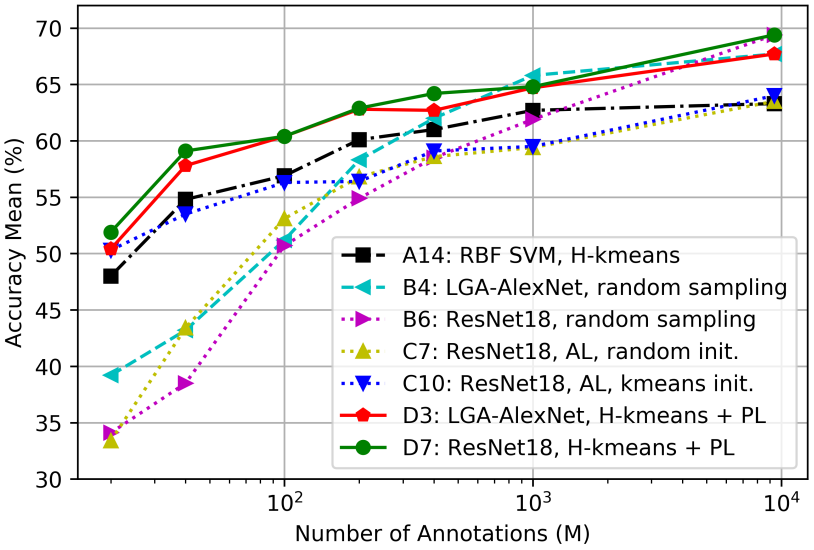
overall for $M \leq 1000$, which is significant since this work focuses on efficient training with a small number of annotated examples.

The comparison between B1 to B3 (last layer only) and B5 to B7 (all layer) indicates that training only the last layer limits the performance of each architecture for large values of M, indicating that there is a significant difference between the low-level and mid-level features of ImageNet and the Southern Hydrate Ridge dataset.

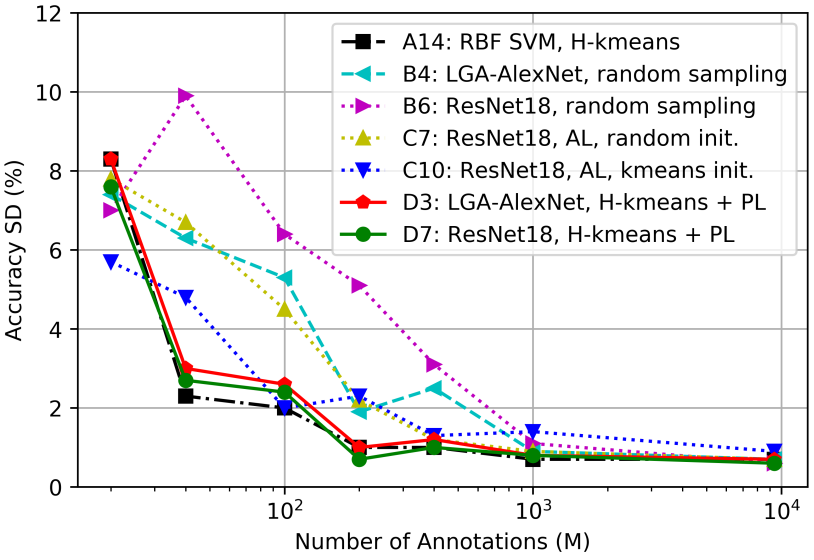
In the proposed pipeline, the number of training examples can be considered large due to the use of pseudo-labels. Therefore, B6 is chosen to be investigated, as it demonstrated the best capacity for learning among B1 to B8, and B4 is also examined since it is the most efficient learner for $M \leq 1000$.

Classification with CNN (Active learning comparison)

Active learning methods attempt to improve learning efficiency by training classifiers on a subset of annotated samples, and proposing which samples should be annotated next based on their prediction uncertainty (Settles, 2009). CNNs are well suited to this iterative process of prediction and prioritised annotation as their outputs are already conditional probabilities against labels and so uncertainty metrics can be easily derived. Common strategies for uncertainty based prioritisation include Least Confidence (LC)



(A) Number of Annotations (M) - F_1 (macro-average) mean.



(B) Number of Annotations (M) - F_1 (macro-average) SD.

FIGURE 4.16: Comparison of classification performance investigated in section 4.4.2. Mean and SD of F_1 (macro-average) values against each M are shown. Representative configurations are chosen from Table 4.10 and Table 4.11. The proposed pipeline (D3 and D7) outperforms others for $M \leq 400$. It is notable that the proposed pipeline always outperform a conventional classifier (A14, RBF SBM) by several percent.

sampling, margin sampling and entropy based sampling, all of which have previously been demonstrated to be effective for seafloor imaging applications (Friedman et al., 2011).

Conventional active learning starts the iterative training process with a randomly selected subset of samples. However, its performance is sensitive to this initial selection and so whether an initial selection of samples nearest to the centroids of the k means

clusters in the latent space obtained by the soft assumption based autoencoder improves their performance is investigated. Subsequent batches of samples (20 when $M \leq 1000$ or 1000 when $M > 1000$) are selected based on the active learning query strategies and iteratively added to the subset of annotated samples for training. A training epoch of 10 was chosen so that the total number of epochs is comparable to the standard supervised learning results (B1-8) and proposed methods (D1-D8).

In this experiment, two different CNN architectures (AlexNet and ResNet18) are assessed, and compare the performance of three well established active learning iterative sampling techniques (LC sampling, margin sampling and entropy based sampling). The active learning process is initialised using two different initial subset selection methods. The first initial subset is randomly sampled, corresponding to traditional active learning workflows. This is compared to active learning initialised by a k means centroid based sample initialisation method that takes advantage of the latent representations learnt during pre-training.

Configuration C1 to C12 in Table 4.11 show the accuracy scores for CNNs trained using the different configurations for active learning. Comparing the proposed method pre-trained AlexNet (C1 to C3) with their transfer learning counterpart (B4), the active learning has reduced accuracy. However, for ResNet18, the accuracy increases when active learning is applied (B6 and C7 to C9). It is noticeable that for larger M (particularly $M = 9370$), active learning degrades performance, possibly due to overfitting of CNN weights at an early phase of the iterative learning process, causing them to remain trapped in local minima. This is because the CNN is trained sequentially on discrete subsets of data, where the stored weights are used to initialise the optimisation of the next subset to limit the total number of training epochs required (Zhou et al., 2017). Although this issue of overfitting is potentially mitigated by resetting the CNN weights between each training subset (Gal et al., 2017), this requires a large number of training epochs, making it impractical for use in domains that require per-dataset training.

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

Classification with CNN (Data selection strategy comparison)
Four data selection strategies; k means, hierarchical k means, and hierarchical k means

with pseudo-labelling or probabilistic pseudo-labelling, are validated in this section. The previous section already confirmed that hierarchical k means based data selection is effective for small values of M when combined with conventional non-CNN classifiers. In order to allow for fair comparison, the number of training samples used by the CNN at each training epoch is fixed to the total number of available labelled training image patches (i.e. 9370 in this experiment). For configurations where all available labelled image patches are used in the training (i.e. all pseudo-label and probabilistic pseudo-label configurations and where $M = 9370$ without pseudo or probabilistic pseudo-labelling), each original labelled training image patch is used once, and these samples are individually subjected to data augmentations that randomise orientations, flipping and position offsets at each training epoch before being used by the CNN. For configurations where the number of labelled image patches used in the training is less than available labelled training image patches (i.e. $M < 9370$ with no pseudo or probabilistic pseudo labels), the selected original images are sampled multiple times (i.e. approximately $9370/M$ times for this experiment) so that a fixed number of labelled training samples are provided to the CNN, where each sample is subjected to random data augmentation before being used by the CNN at each training epoch. In Lee (2013), pseudo-labels are determined by k means clustering result, corresponding to 1-NN in Table 4.10. However, Table 4.10 shows that RBF kernel SVM consistently achieves better performance when estimating class decision boundaries and so RBF kernel SVMs are used to assign predictive pseudo-labels in this work. Although the Gaussian Process classifier described in section 3.5.2 did not perform as well as the RBF kernel SVM, the prediction uncertainty may be useful for CNN training and so experiments are also performed using these outputs as probabilistic pseudo-labels.

Configuration D1 to D4 in Table 4.11 shows the performance metrics for each data selection strategy with the soft assumption pre-trained AlexNet CNN with last layer training. D5 to D8 show the same comparison for ImageNet pre-trained ResNet18 CNN with all layer Southern Hydrate Ridge training. For both AlexNet and ResNet18, the combination of hierarchical k means and pseudo-labelling achieves the best performance for $M \leq 200$. Comparing the cases with pseudo-labelling (D3 and D7) to the cases without (D2 and D6) shows that pseudo-labelling consistently improves the classification performance, where D7, which applies hierarchical k means and pseudo-labelling to ResNet18, performs the best for $M \leq 200$ among all the configurations in Table 4.11. The accuracies achieved by D7 with $M = 20, 40, 100$ are similar to the metrics achieved for B1 to B4 with $M = 200, 400, 1000$, which have an order of magnitude more annotations. In particular, B6 and D7 use the same CNN architecture, showing that gains in learning efficiency can be attributed to the semi-supervised training method, resulting in a significant reduction in human effort to achieve a similar level of classification accuracy. Although the efficiency gains diminish as the number of human annotations available for training increases, the proposed method never degrades the CNN's performance for an equal number of annotations. Another way to look at this

is that the largest gains in learning efficiency are achieved when there is only a small amount of human effort available for annotation tasks, where D7 with 40 prioritised annotations reaches 85 % of the accuracy achieved by the best performing supervised CNN, B6, trained using 9370 human annotations, which represents just 0.4 % of the human effort. The data also shows that the the combination of hierarchical k means and pseudo-labelling improves the repeatability between experiments under the same conditions, which is an important attribute for practical application of automated data interpretation.

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

CNN and conventional classifier comparison

Figure 4.16 compares the performance metrics of several representative configurations in Table 4.10 and Table 4.11. The mean and SD values of the scores from ten repeat trials for each configuration are shown in Figure 4.16a and Figure 4.16b, respectively. The result under configuration A14 are shown as this is the best performing conventional (i.e. non-CNN) classifier. For the CNN classifiers, configurations B4, B6, C7, C10, D3 and D7 are shown to demonstrate the effectiveness of the proposed pipeline compared to other data selection strategies (random selection and active learning).

Overall, the CNNs trained with proposed pipeline (D3 and D7) outperform the conventional classifier (A14) and the best performing CNN trained using active learning (C7 and C10), except for $M = 1000$. The outputs of the A14 form the inputs to train D3, where the soft assumption based autoencoder is used for pre-training the AlexNet CNN. The improvement in performance shows that the CNN does not merely replicate the class boundaries found in the annotations and the pseudo-labels, but learns more general boundaries that discriminate the classes more accurately. ResNet18 (D7) shows better performance than AlexNet (D3) when trained using the same outputs of A14, indicating an ability to more accurately model complex class boundaries. This was generally the case for all random selected training data and the proposed pipeline. Comparing $M = 1000$ and $M = 9370$, the conventional classifier's accuracy is not significantly improved even though almost 10 times the number of annotations are used

for training. On the other hand, the CNNs achieve statistically significant increases from $M = 1000$ to $M = 9370$ in all cases. This supports the common understanding that deep learning CNNs are a better option than conventional classifiers when large training datasets are available, and that conventional classifiers are a reasonable option when only a small number of annotations are available for training.

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

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

Per-class performance investigation

So far the macro-averaged F_1-score has been used as a metric to compare the overall performance of different classifiers. This is appropriate when all classes in a dataset are assumed to be equally important. However, there are applications where this is not the case, and in these scenarios it is more valuable to consider performance on a per-class basis. Figure 4.17 and Figure 4.18 compare the per-class confusion matrices for M values of 20, 40, 100 and 1000 for configurations B2 and D7. These represent the outputs of the best performing network, ResNet18, trained using standard transfer learning and the proposed semi-supervised pipeline, respectively. The values in each confusion matrix are normalised by the number of ground truth annotations so that the diagonal elements correspond to the recall value of each class. The confusion matrices corresponding to the trials with the closest F_1-score (macro-average) to the mean of ten repetitions (Table 4.11) are chosen for each value of M. The values of zeros for $M = 20$ and 40 in Figure 4.17 suggest no images corresponding to 'Artificial Object' were selected in the random selection used for training and so predictions could not be made effectively for this class. On the other hand, Figure 4.18 shows that all six classes

in the dataset are predicted for all M, illustrating the advantage of using hierarchical k means based data selection to avoid minority classes from being overlooked even when the total number of annotated images is small.

Comparing the habitat maps generated using the classification results to the ground truth annotations (Figure 4.2b) shows that the random data selection (Figure 4.17) requires a larger number of training samples M to capture the different spatial distribution patterns of each class. Using the proposed semi-supervised training method (Figure 4.18) results in more consistent performance from a per-class perspective, and provides a better approximation of the ground truth class distribution patterns even for small values of M. The consistent performance for different numbers of input training data is an important attribute for practical application since the annotation resource available for different datasets is likely to vary. These points favour the proposed method over random sampling approaches that are more sensitive to the number of available annotations, and require larger amounts of training data to achieve similar performance.

Discussion

The proposed semi-supervised pipeline with the soft assumption based autoencoder for representation learning is examined in this section. The experiment results on Southern Hydrate Ridge dataset consisting of more than 18,000 human annotations demonstrate that:

- The proposed semi-supervised learning pipeline can achieve classification accuracy equivalent to naively trained CNNs with an order of magnitude fewer human annotations (i.e. tens to hundreds, as opposed to thousands). The results demonstrate improvements in accuracy by a factor of 1.2 to 1.5 when a hundred or less annotations are used, where the largest gains in learning efficiency are achieved with small numbers of annotations. The method also reduces the statistical variability between independent trials under the same learning configurations to approximately 0.6 of that when random sampling is used. The proposed method reaches 85 % of the accuracy achieved by the best performing naively trained CNN (trained using 9370 human annotations) with just 40 prioritised annotations, which represents 0.4 % of the human effort.

- The strategy to select data for human annotation affects final classification performance. Introducing structure to prioritise annotation effort using hierarchical k means in the latent space obtained by the soft assumption based autoencoder and assigning pseudo-labels with a RBF kernel SVM to identify class boundaries improves the classification performance of CNNs by an average of 1.34 times compared to naive dataset annotation when 100 or less annotations are used. A similar gain in performance is seen when the soft assumption based autoencoder

with k means selection is used to initialise active learning, with a 1.25 factor improvement compared to equivalent randomly initialised active learning setups.

- The proposed method makes more efficient use of human effort than traditional active learning based techniques tested, and is less prone to overfitting, achieving a factor 1.12 and 1.22 improvement in performance for AlexNet and ResNet18 respectively when compared to randomly initialised active learning across all values of M.

- CNN architectures are able to generalise class boundaries of interest to humans even when pseudo-labels are assigned to all data in a training set. The resulting CNN is able to improve the relative classification accuracy by an average of 6.4 % compared to the classification accuracy of the pseudo-labels themselves.

- The performance of conventional classifiers for pseudo-label generation is significantly improved using k means based selection compared to random selection when generating subsets of data for annotation. A factor of 1.30 improvement in classification accuracy is achieved for prioritised subsets with a hundred samples or less.

- Implementation of annotation effort prioritisation strategies relies on effective unsupervised clustering performance for seafloor images, where the use of georeferencing information by the soft assumption based autoencoder compared to an equivalent autoencoder that only uses information in images resulted in an improvement in classification accuracy by a factor of 1.4 to 8.9 (average 3.1) for the configurations tested in this work.

In addition, it is shown that the method generalises to other environmental monitoring domains, where the use of georeference information in the the soft assumption based autoencoder and prioritised image annotation strategy improve the classification accuracy for three different aerial image datasets when compared to traditional autoencoders that use only a reconstruction loss and random sampling strategies for supervised learning.

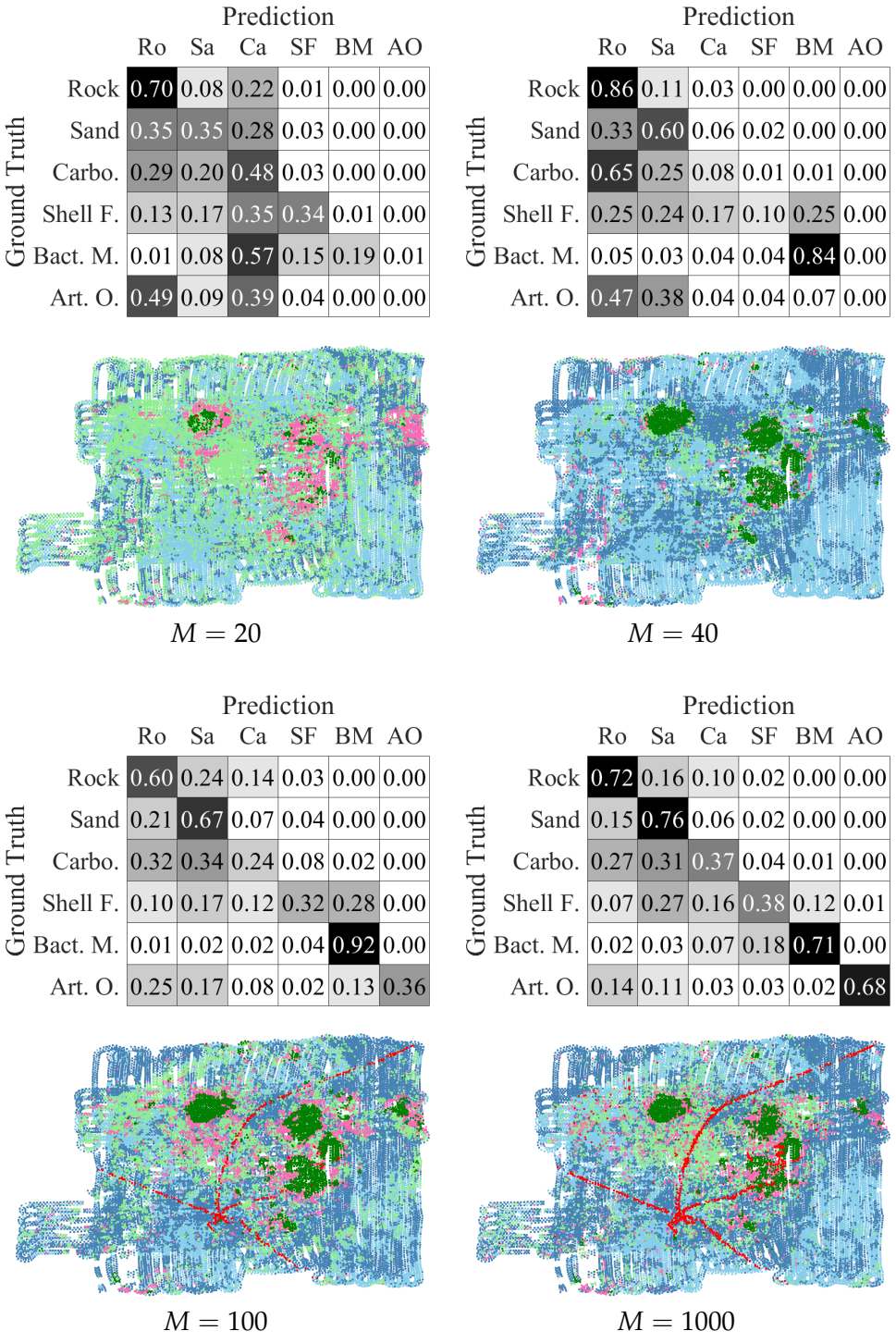


FIGURE 4.17: Confusion matrices and habitat maps predicted by ResNet18 trained using the random data selection (configuration B6 in Table 4.11). This corresponds to conventional good practise, using a CNN pre-trained on the ImageNet annotation dataset and fine-tuning all the layers using randomly sampled annotated images with data augmentation. The results show that for a values of $M = 20$ the 'Artificial Object' and 'Bacterial Mat' class that contain the fewest samples are not efficiently learned, and even for $M = 40$, 'Artificial Object' is not recognised. The confusion matrix shows that even with $M = 1000$, there is still significant confusion when classifying 'Carbonate' and 'Shell Fragment'.

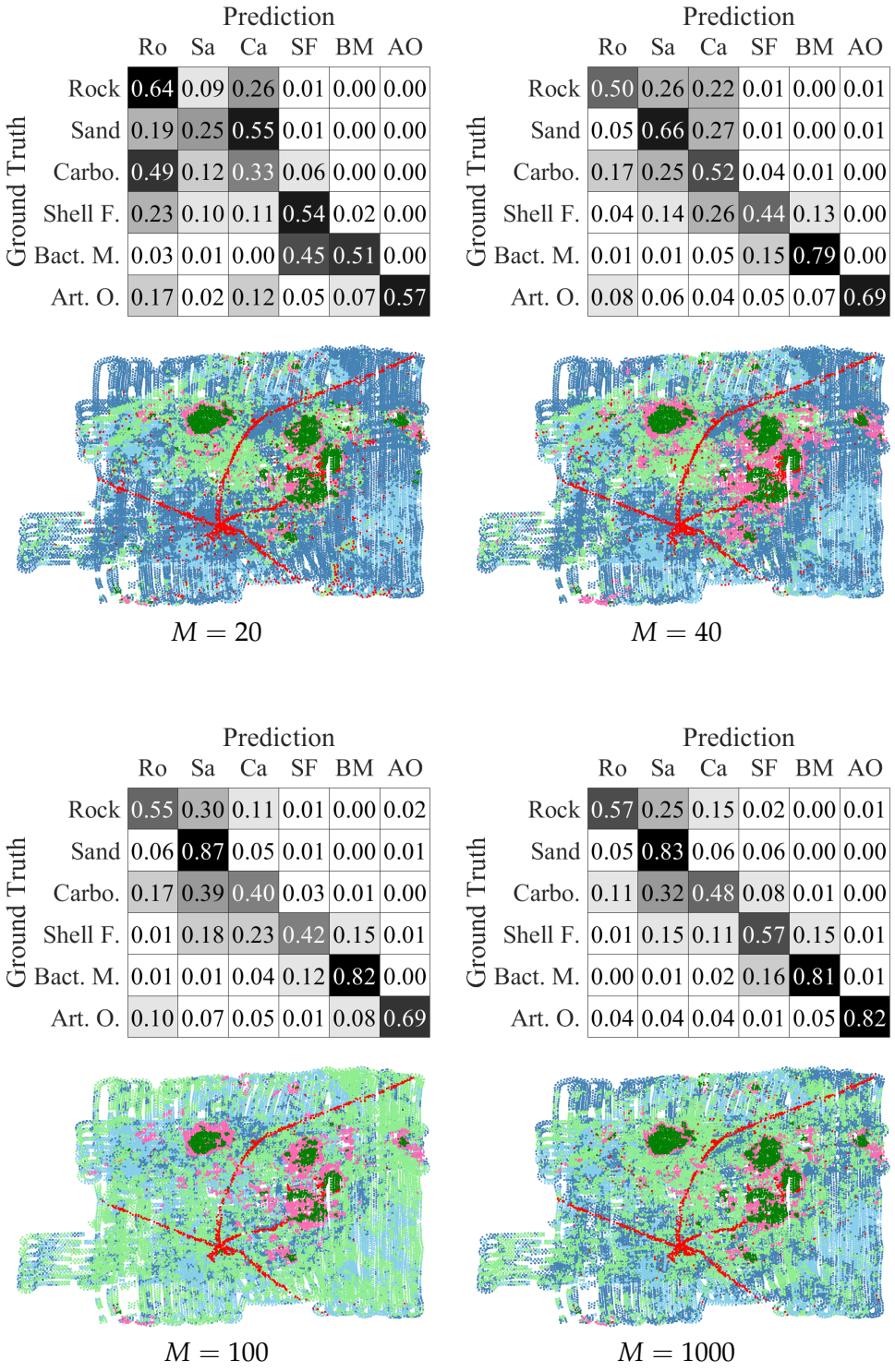


FIGURE 4.18: Confusion matrices and habitat maps predicted by ResNet18, which was trained using the proposed semi-supervised method, using the hierarchical k means based data selection and pseudo-labelling (configuration D7 in Table 4.11). Compared to Figure 4.17, the results show improved learning efficiency, especially for small values of M, where the 'Artificial Object' and 'Bacterial Mat' class are efficiently learned using just 20 human annotations, despite these being the classes with the smallest number of data samples. The performance with $M = 100$ shows similar performance to when the same CNN architecture is trained using an order of magnitude more annotations from randomly selected data (i.e Figure 4.17).

4.4.3 Semi-supervised learning (hard assumption)

Method and dataset
In this section, the proposed semi-supervised interpretation pipeline (section 3.5.2) is investigated with the hard assumption based contrastive learning (section 3.3) for representation learning. Tasmania dataset (section 4.1.2) is used for the training and test.

While the training data is sampled in the class balanced manner in the experiment in section 3.3, the experiments in this section compares the performance using *random* and *H-kmeans* based selection strategies. In addition, the effectiveness of pseudo-labelling with linear logistic regression (PL-linear), a non-linear SVM with RBF (PL-SVM) for CNN fine-tuning is investigated. The training data selection strategies can both realistically be implemented in field survey scenarios since they do not assume any prior knowledge of the datasets, and the images that require annotation can be rapidly identified in a fully unsupervised manner.

Data selection method comparison
The performance using the *random* and *H-kmeans* training data selection strategies, both of which do not need prior human input to understand the datasets, are shown in Table 4.12 for different values of M. The different CNN training methods shows the same trend as the previous results with *balanced* training data selection (Table 4.5). When the classifiers are trained on the latent representations, the proposed hard assumption based method outperforms SimCLR and ImageNet pre-training, achieving and average performance gains of 6.3 % and 20.0 %, respectively across all M. As previously observed, fine-tuning SimCLR and proposed method trained CNNs degrades their performance. However, the pseudo-labelling introduced in (E7, E8, F7, F8) mitigates this effect by using a larger number of images for fine-tuning, which avoids the problem of overfitting that can occur when only a small number of images are used in fine tuning. This effect is strongest for small values of M=40, 100, where performance gains of 13.1 % and 8.0 % are achieved for both the proposed method and SimCLR compared to equivalent configurations that do not use PL.

Figure 4.19b shows representative configurations in Table 4.12. The configurations with the proposed method (F$*$) outperform their counterparts with the SimCLR (E$*$) except for the case where M=40 where E4 performs better than F4. In general, the use of *H-kmeans* improves performance compared to equivalent *random* configurations, achieving performance gains of 13.1 % and 5.7 % respectively for M=40, 100. Although the gain in performance reduces for larger M, for the proposed hard assumption based representation learning *H-kmeans* selection always improves performance compared to equivalent *random* configurations for all values of M. An important observation is that the proposed representation learning achieved the best performance for all values of M for both the *balanced* and *H-kmeans* selection strategies.

A comparison between Table 4.5 and Table 4.12 shows that the proposed representation learning with *H-kmeans* performs better than with the *balanced* selection strategy for all values of M, with gains of 3.1 % and 2.7 % for small values of M=40, 100, and averaging a performance gain of 1.6 %.

From a practical perspective, the proposed representation learning with M=100 *H-kmeans* machine prioritised annotations and the SVM-RBF classifier (F4), and PL-SVM fine tuning (F7) achieves the same accuracy as state-of-the-art transfer learning (i.e. D5 M=1000) using an order of magnitude fewer human annotations. The method also achieves the same accuracy as state-of-the-art contrastive learning approaches (i.e. E3 M=400) using a quarter of the annotations, where prior works rely on random data annotations and do not propose a data selection strategy. Being able to perform accurate classification with a relatively small number of labels (i.e. 0.1 % of the entire dataset) can be considered as an important development since providing 100 annotations represents a level of human effort that can be justified for most application in the field. It is also shown that for applications that can justify a larger amount of human effort (i.e. $M = 1000$), the proposed representation learning outperforms conventional transfer learning (D5) and contrastive learning (E3) by 8.5 % and 7.5 % respectively. In addition to the demonstrated performance gains, the use of the proposed representation learning consistently improves performance over alternative configurations for all conditions tested in this work, and machine guided annotation *H-kmeans* benefits performance for all configurations where $M<400$, and although the performance gains diminish for larger M, it never leads to significant performance reduction. The results indicate that these approaches can robustly improve the performance of CNNs for seafloor image interpretation.

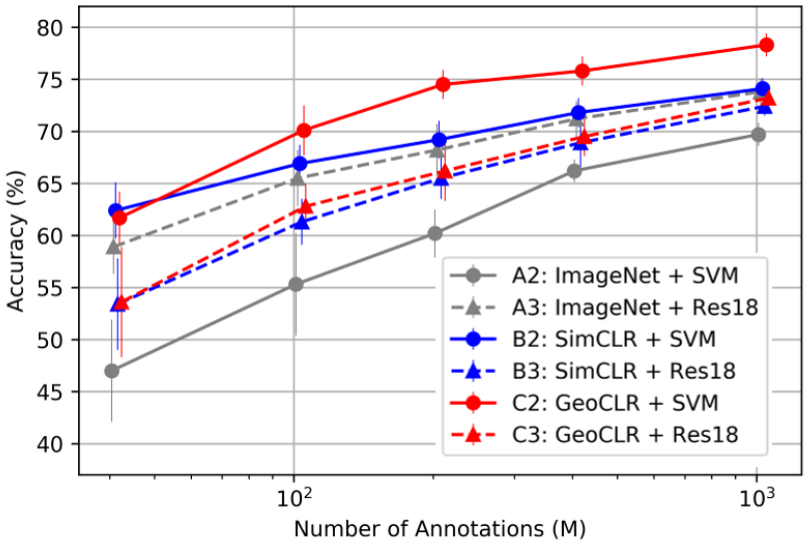
TABLE 4.12: Data selection method comparison

Config. Label	CNN Training	Classifier	Data Selection	Number of Annotations (M) 40	100	200	400	1000
D1	ImageNet	linear	random	45.4±5.8	57.0±3.2	61.3±2.8	63.3±2.9	67.5±2.2
D2	ImageNet	linear	H-*k*means	49.5±5.6	58.0±4.7	64.4±3.4	66.5±2.6	68.8±1.8
D3	ImageNet	SVM	random	35.5±4.4	50.2±3.7	58.4±3.3	63.7±1.4	67.9±1.0
D4	ImageNet	SVM	H-*k*means	43.0±3.4	57.5±2.8	63.7±1.0	67.0±1.4	69.7±1.3
D5	ImageNet	Res18	random	55.5±3.1	63.2±3.2	67.0±2.0	69.7±2.4	72.8±2.3
D6	ImageNet	Res18	H-*k*means	51.8±5.7	64.1±2.0	67.6±2.7	73.1±1.3	74.1±1.7
D7	ImageNet	Res18	PL-linear	51.9±5.7	62.2±4.5	68.6±1.8	70.9±2.0	71.9±2.3
D8	ImageNet	Res18	PL-SVM	46.4±5.4	58.9±3.4	67.1±1.6	69.9±1.7	72.6±2.1
E1	SimCLR	linear	random	55.0±4.0	63.8±3.0	66.3±2.1	67.7±3.2	71.2±1.1
E2	SimCLR	linear	H-*k*means	61.9±2.6	66.5±1.6	68.2±1.4	69.3±2.7	69.5±1.9
E3	SimCLR	SVM	random	47.2±5.5	64.5±2.4	68.8±1.6	72.0±2.2	73.5±0.7
E4	SimCLR	SVM	H-*k*means	58.0±2.0	67.6±1.5	70.9±1.3	71.7±1.8	73.7±1.5
E5	SimCLR	Res18	random	49.5±7.5	60.3±2.2	64.5±2.4	67.7±1.8	71.5±2.3
E6	SimCLR	Res18	H-*k*means	56.4±3.3	65.2±2.2	66.2±1.7	69.3±2.0	70.1±1.2
E7	SimCLR	Res18	PL-linear	64.3±2.2	68.8±1.4	69.5±1.7	70.5±1.7	72.8±1.3
E8	SimCLR	Res18	PL-SVM	63.1±2.8	69.8±1.8	70.6±1.1	72.7±1.1	72.9±0.8
F1	Proposed	linear	random	58.9±5.0	67.8±2.7	70.8±1.7	72.7±2.5	75.1±1.4
F2	Proposed	linear	H-*k*means	**65.8**±2.9	70.5±1.7	72.8±2.0	73.0±2.1	74.6±2.5
F3	Proposed	SVM	random	53.2±5.9	68.8±3.1	72.9±2.2	75.5±1.0	77.5±1.2
F4	Proposed	SVM	H-*k*means	55.3±4.2	71.8±1.6	**74.6**±1.5	**76.6**±1.2	**79.0**±1.0
F5	Proposed	Res18	random	49.5±7.9	60.3±3.8	65.2±1.7	69.0±3.0	73.2±1.9
F6	Proposed	Res18	H-*k*means	56.5±3.4	65.5±1.4	66.8±1.9	70.9±1.3	73.9±1.7
F7	Proposed	Res18	PL-linear	64.2±2.5	71.7±2.3	72.7±1.6	73.5±1.4	75.7±1.6
F8	Proposed	Res18	PL-SVM	63.7±2.1	**72.0**±2.1	72.5±1.5	74.3±1.0	75.2±1.3

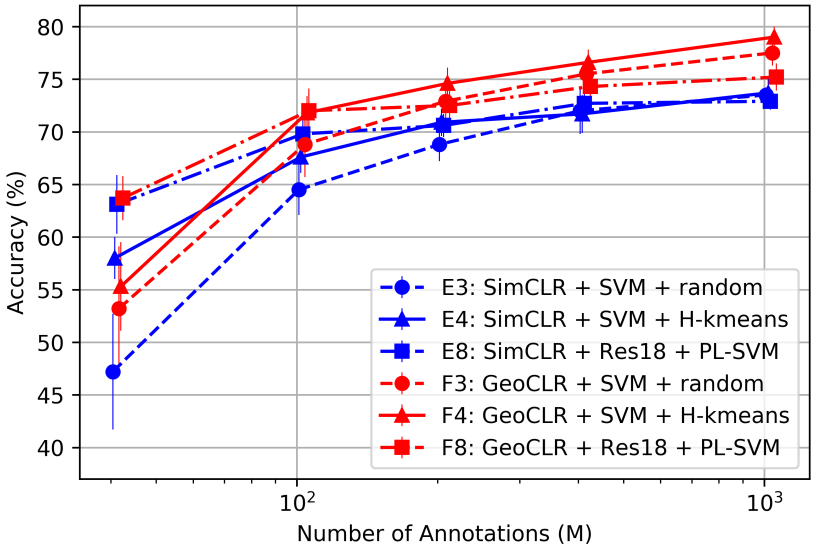
The same CNNs as Table 4.5 where different data selection strategies (*random* and *H-kmeans* are used in the downstream classification task. In contrast to the *balanced* selection strategy shown in Table 4.5, these selection strategies do not require prior analysis by humans and so are available for analysis of data as it get collected in the field. The same classifiers (linear, SVM, fine-tuned ResNet18) are investigated. For fine-tuning the CNNs, pseudo-labels generated by linear classifier ($*$-7) or SVM ($*$-8) are used. The classifiers are trained 10 times with different random seed, and mean and SD values of F_1-scores (macro averaged) are shown. The best score for each M is shown as bold.

Estimating relative habitat class proportion

Determining seafloor habitat class distributions is a fundamental task for marine monitoring and conservation. Here the proposed hard assumption based contrastive learning method is applied to estimate the relative proportion of habitat classes and map their physical distribution. Figure 4.20 shows the relative proportion of different habitat classes estimated for M=[40, 100, 200, 400, 1000] machine prioritised annotations for each of the 12 dives in the Tasmania dataset. These are compared to the relative proportions for each dive where all human annotations have been used (i.e. average 450 annotations per dive) which can be regarded as the ground truth here. The equivalent number of annotations per dive for the proposed method average approximately 3 annotations per dive for M=40 to approximately 83 per dive for M=1000. The results show that the estimated proportions approach the ground truth distributions for all dives, with the expected result that performance increases as a larger number of annotations are used for classifier training. The estimated proportions are poor for several of



(A) CNN training method comparison for class-balanced training (Table 4.5)



(B) Data selection method comparison for SimCLR and proposed representation learning (Table 4.12)

FIGURE 4.19: Representative configurations from (a) Table 4.5 and (b) Table 4.12. (a) When the CNNs are trained on the class-balanced subsets, the proposed representation learning with a SVM classifier (C2) outperforms all other configurations except for B2 when M=40. The best performance for M=40 is achieved by the proposed representation learning with a linear classifier (C1). (b) In general, the use of *H-kmeans* improves performance compared to equivalent *random* configurations, and the proposed representation learning outperform their counterparts with the SimCLR except for $\{M$=40, E4$\}$.

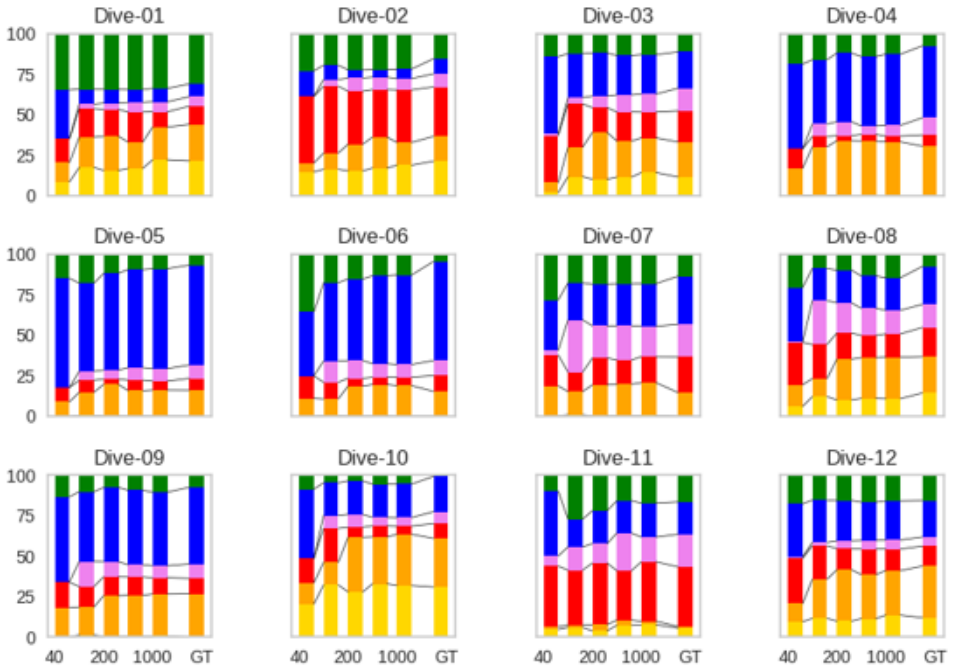
the dives with M=40 when using the F4 SVM classifier (Figure 4.20a), whereas the F8 fine-tuned with pseudo-labels generated by the SVM is generally more robust, approximating the ground truth class proportions better for the same number of training examples (Figure 4.20b). This indicates that the SVM classifier (F4) may be overfitting the latent representation space generated by the proposed representation learning when the number of annotations available is small, where this effect is mitigated by providing a larger number of training examples through pseudo-labels. However, there are some exceptions (Dives 06 and 07) to this and so the outputs with M=40 should be treated with caution where validation data is not available. On the other hand, for $M \geq 100$ both methods (i.e. F4 and F8) perform robustly for all dives, with F4 outperforming F8 and providing more stable estimates for different values of M. This is due to the fact that the latent representation space remains the same regardless of M as no CNN re-training takes place.

Habitat map

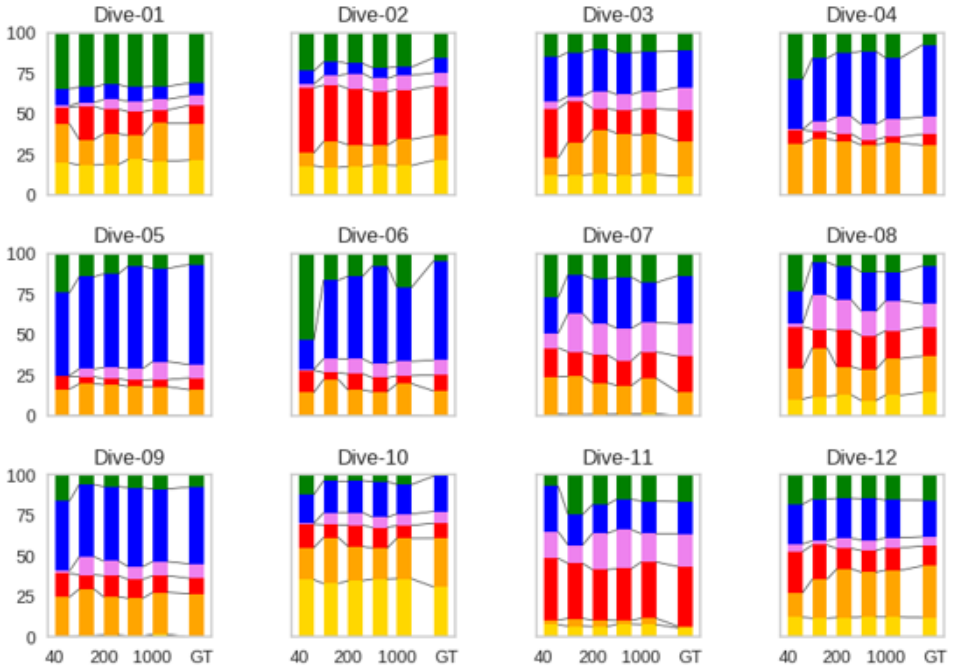
The physical distribution of habitats is important for conservation since it influences the distribution of organisms near the seafloor. It is also important for understanding ecosystem health as benthic habitats such as kelp (seen here) and coral are classified as essential ocean variables.

The proposed method allows efficient estimation of habitat maps based on the 3d location where each classified image was taken. Here, the horizontal distributions of the classes, the depth profiles versus image index, and the class versus depth distributions are shown in Figures 4.21, 4.22 and 4.23 for three dives (01,03 and 08) which were chosen as representative cases. The figures show habitat maps generated using the proposed representation learning for $\{M$=100,F8$\}$, $\{M$=1000,F4$\}$ in Table 4.12 and the ground truth labels.

The results show that both $\{M$=100,F8$\}$ and $\{M$=1000,F4$\}$ configurations closely approximate the ground truth horizontal and vertical habitat class distributions, capturing the continuous spatial transitions between Kelp (A), Low Relief Reef (C), High Relief Reef (B) to Screw Shell Rubble (E) or Sand (F). The class vs depth distributions show that the larger values of M provide a better approximation of vertical class distribution, which is an expected result. However, for classes that exist in a limited depth band (e.g. Kelp (A), Screw Schell Rubble (E)) both values of M capture this trend.

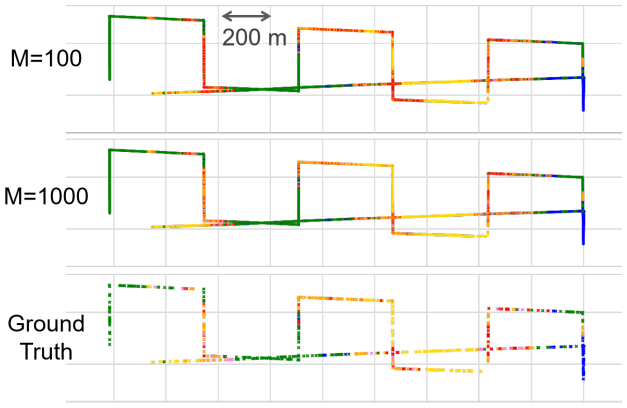


(A) F4 in Table 4.5 (SVM on latent representation $\boldsymbol{h}$)

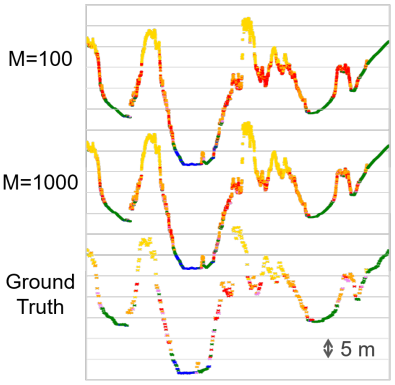


(B) F8 in Table 4.5 (fine-tuned on pseudo-labels generated by SVM)

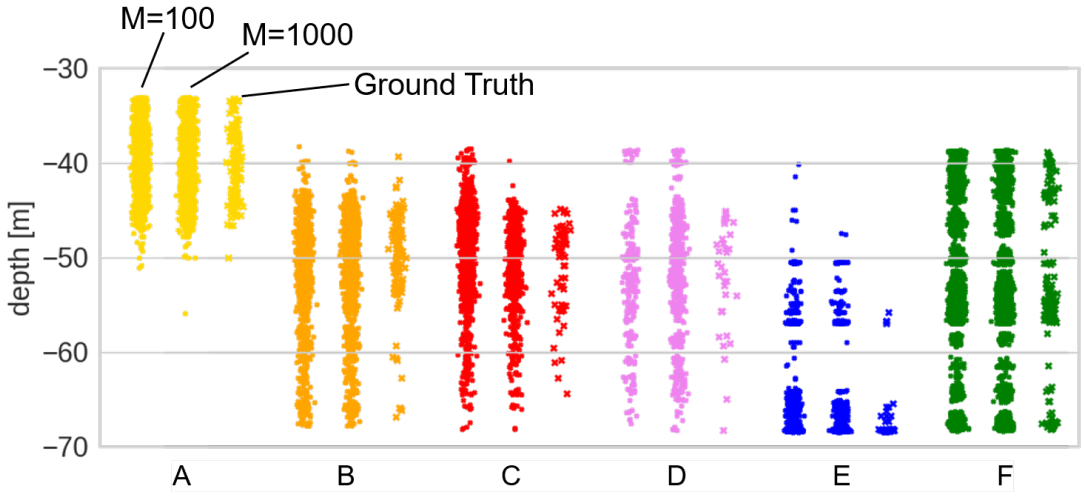
FIGURE 4.20: Class distribution estimated for each dive using the proposed hard assumption based representation learning. The same colour scheme is used as in Figure 4.3. The estimated distributions approaches the ground truths when larger number of annotations are used for classifier training. The use of pseudo-labels is generally favourable for a small number of annotations (i.e. M=40, though this is not always the case. For M>100, F4 performs better than F8 and provides more stable estimates of class distribution as the same latent representation space is used for all M.



(A) Horizontal distribution

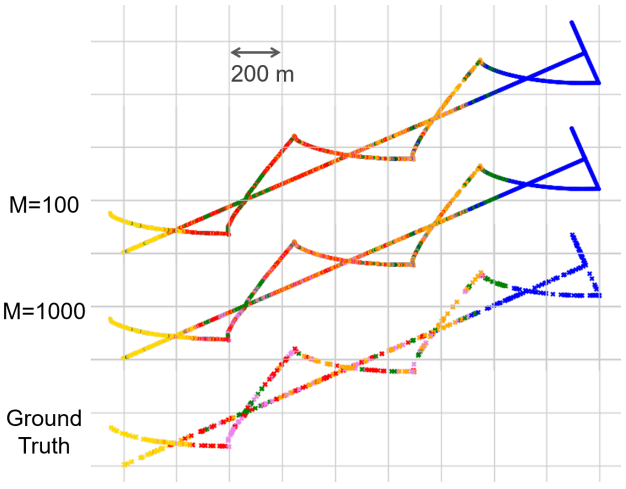


(B) Depth profile vs. image index

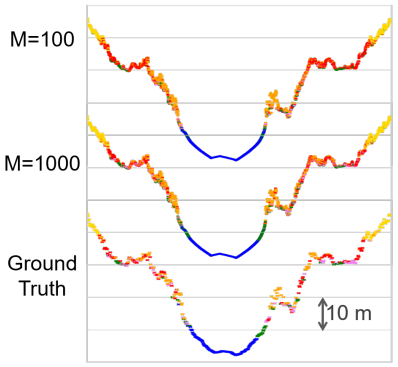


(C) Class vs. Depth

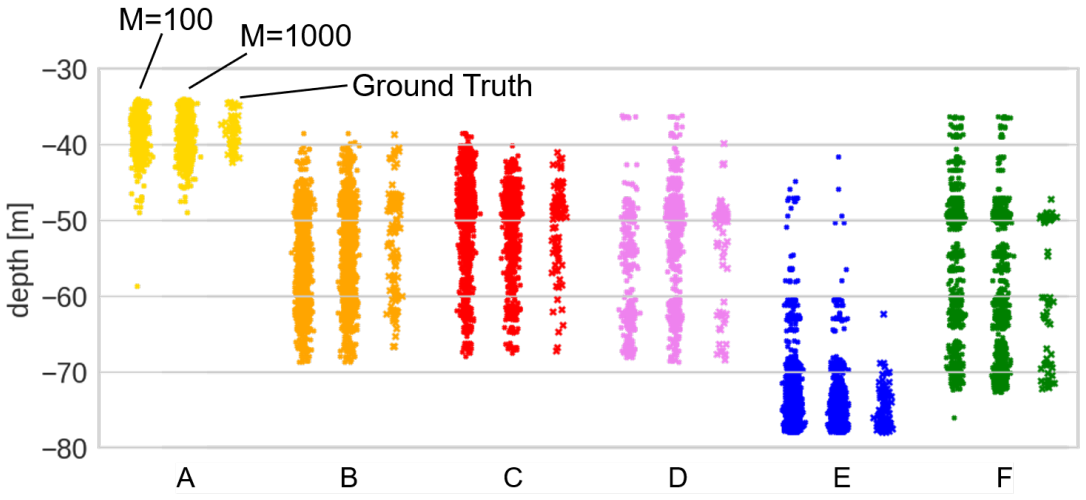
FIGURE 4.21: Class distribution of Dive-01 with M=100 annotations by F8, M=1000 annotations by F4 in Table 4.12 and ground truth. The same colour scheme as Figure 4.3 is applied.



(A) Horizontal distribution

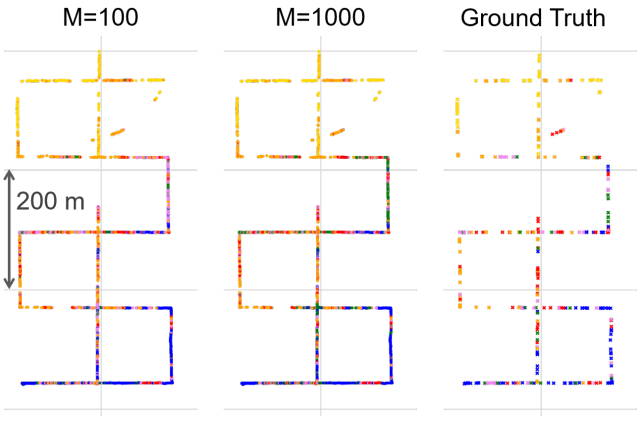


(B) Depth profile vs. image index

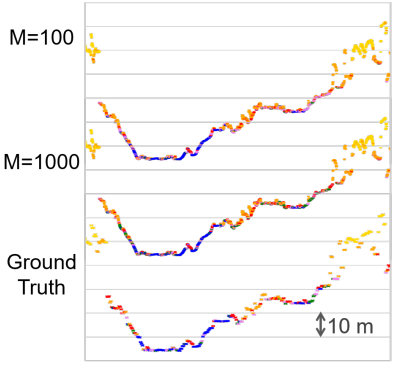


(C) Class vs. Depth

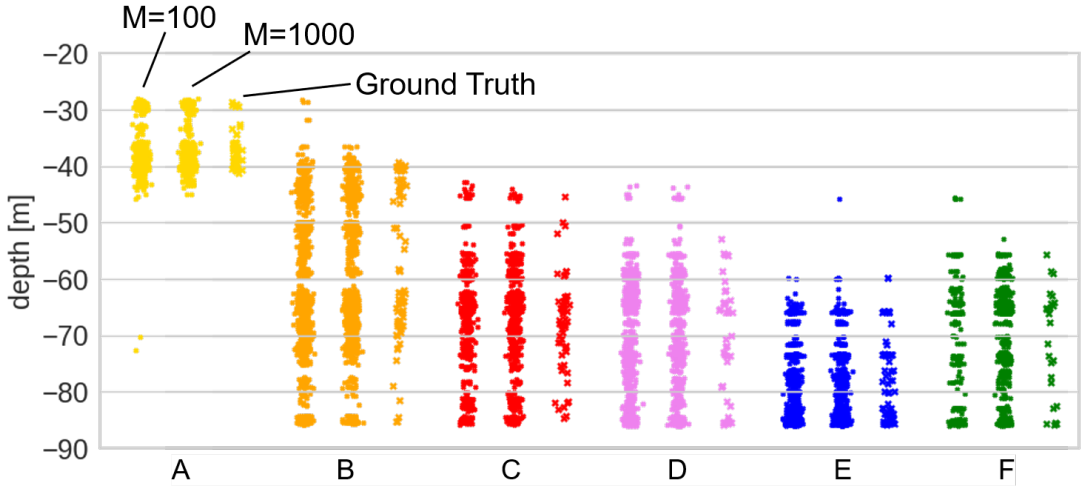
FIGURE 4.22: Class distribution of Dive-03 with M=100 annotations by F8, M=1000 annotations by F4 in Table 4.12 and ground truth. The same colour scheme as Figure 4.3 is applied.



(A) Horizontal distribution



(B) Depth profile vs. image index



(C) Class vs. Depth

FIGURE 4.23: Class distribution of Dive-08 with M=100 annotations by F8, M=1000 annotations by F4 in Table 4.12 and ground truth. The same colour scheme as Figure 4.3 is applied.

Discussion

This section examined the performance of the proposed semi-supervised pipeline with the hard assumption based contrastive learning. The experiments on the Tasmania dataset that includes more than ~86k images and ~5k annotations show that:

- The representations extracted by the hard assumption based contrastive learning are useful for identifying representative images for prioritised human annotation in a fully unsupervised manner. This can improve the performance and efficiency of human effort for classification, where selecting a prioritised training dataset using *H-kmeans* clustering increases the classification accuracy by an average of 4.9 % and maximum of 14.1 % compared to random selection, where the performance gains are more significant for small numbers of M.
- The proposed unsupervised representative image identification results in better performance than providing class-balanced annotated examples in a supervised manner. The machine driven *H-kmeans* selection strategy achieves an average of 1.6 % and maximum of 3.1 % increase in accuracy compared to the class-balanced selection strategy for an equivalent number of annotations, where greater gains are achieved for small numbers of annotations. This indicates that it is more informative to provide training data that evenly describes the latent representation space generated during self-supervised training, than it is to provide training data that evenly describes the targets that are of final interest to humans.
- The combination of the hard assumption based contrastive learning and *H-kmeans* achieves the same accuracy as state-of-the-art transfer learning using an order of magnitude fewer human annotations, and state-of-the-art contrastive learning approaches using a quarter of the labels. This allows the proportion of habitat classes and their spatial distribution to be accurately estimated (> 70 %) annotating only 0.1 % of the images in the dataset. This is significant as providing approximately 100 annotations represents a level of human effort that can be justified for most field applications. For applications where a greater level of human effort is available, it is shown that with 1000 annotations, the proposed hard assumption based contrastive learning outperforms conventional transfer learning and contrastive learning by 8.5 % and 7.5 % respectively, achieving a classification accuracy of 79 %. The combination of the hard assumption and *H-kmeans* never degraded performance compared to equivalent alternative configurations in the experiments described in this paper.

Chapter 5

Conclusions and future work

This PhD thesis has been concerned with leveraging domain knowledge for machine learning based seafloor image interpretation. Two novel representation learning techniques for seafloor imagery have been proposed, and their effects have been evaluated on real-world seafloor image datasets. Interpretation pipelines that eliminate human effort, or require significantly less than conventional supervised learning methods, have also been proposed and examined. A summary of the contributions and findings of chapters 3 and 4 is given in section 5.1, with some further insights in section 5.2 and potential future areas for investigation and application described in section 5.3. Section 5.4 shows the list of authored publications during this PhD programme.

5.1 Conclusions

The contributions of this thesis arise from the development of two novel representation learning methods for seafloor visual imagery that leverage metadata commonly generated during robotic imaging surveys. The following is a summary of the four principal contributions in this PhD thesis.

Seafloor image representation learning based on soft assumptions (proposed in section 3.2 and evaluated in section 4.2.1)
A novel self-supervised representation learning method that uses soft assumptions about metadata relationships to guide the training of a deep learning autoencoder is proposed in section 3.1. The soft assumption is introduced by defining a novel loss function based on the Kullbuck-Leibler divergence between the affinity matrices of each image in the latent space and the metadata spaces. This approach has been generalised to deal with any combination and number of vector and scalar metadata values, as described in section 3.2.1. The method uses contrastive batch sampling, introduced in section 3.2.2, to ensure that a range of metadata relationships are considered at each training step, and a Student's t-distribution is introduced to avoid over constraining the latent representations. The experiment in section 4.2.1 on a real-world seafloor image dataset consisting of ~32k images (section 4.1.3) shows that combining multiple sources of metadata regularisation can outperform single metadata regularisation using the proposed method. Compared to an equivalent convolutional autoencoder that does not consider metadata, regularising learning using depth and horizontal location information improves the performance of five classifiers operating on the latent representations by an average of 10.9%, with a R-SVM classifier showing the largest gain in performance at 15.1% (see Table 4.4).

Seafloor image representation learning based on hard assumptions (proposed in section 3.3 and evaluated in section 4.2.2)
A novel contrastive self-supervised representation learning method is developed that imposes hard assumptions on the relationships between metadata and the similarity of images in section 3.1. A novel similar image pair selection method is developed where 3d georeference information is used as the relevant metadata to implement contrastive learning. The experiment in section 4.2.2 on a real-world seafloor image dataset consisting of ~87k images (section 4.1.2) shows that the downstream classification accuracy is improved by an average of ~5.2 % compared to a state-of-the-art SimCLR representation learning method, which generates similar image pairs through data augmentation without considering where the images were gathered. Compared to traditional transfer learning approaches, an average improvement of 7.4 %, is achieved (see Table 4.5).

Unsupervised seafloor image interpretation methods (proposed in section 3.4 and demonstrated in section 4.3)
To show the advantage of representation learning in real seafloor survey scenarios, unsupervised seafloor imagery interpretation pipelines that make use of the acquired image representations are proposed.

The first application is clustering (section 3.4.1) where a non-parametric Bayesian clustering method, that can automatically determine the number of clusters in a target dataset, is directly applied to the obtained low-dimensional latent representation space. This allows important insight about patterns in the target seafloor images to be gained without any human supervision. The experiment in section 4.3.1 shows that clustering with the proposed soft assumption based metadata guided representation learning achieves almost double the performance of a standard convolutional autoencoder that doesn't consider metadata, based on the NMI score for human generated ground truth classes (see Table 4.6).

The next application is representative image identification presented in section 3.4.2. This method identifies the images that are most representative of the target dataset based on hierarchical *H-kmeans* clustering. Section 4.3.2 shows how this can be applied to the latent representations obtained using both the soft assumption based autoencoder (see Figure 4.13) and the hard assumption based contrastive learning (Figure 4.14). The method allows representative images of large datasets with imbalanced class distributions to be automatically identified in a fully unsupervised way. The combination of representative images with maps of semantic cluster distributions can help achieve efficient understanding of underwater scenes in a fully unsupervised manner. This can be useful for planning operations since the unsupervised outputs can be rapidly generated without the need for any human input.

Efficient alignment of latent representations with human interests (proposed in section 3.5 and evaluated in section 4.4)
The cluster boundaries generated through unsupervised interpretation are not guaranteed to match the semantic boundaries of interest to human experts. To efficiently align boundaries in the acquired representations with human interests, a semi-supervised learning method is proposed in section 3.5.2. The key elements in the pipeline are the use of unsupervised representative image identification and pseudo-labelling to efficiently obtain labels for training a classifier. The representative images annotated by human experts are algorithmically proposed in a fully unsupervised way. This efficiently guides the human annotation effort, and the number of representative images proposed can be flexibly set to match the availability of human effort. The proposed semi-supervised method can be applied to both the soft assumption based autoencoder and the hard assumption based contrastive learning.

The experiment in section 4.4.2 evaluates the semi-supervised pipeline with the soft assumption based method on a real-world seafloor image dataset consisting of ~63k images (section 4.1.1). The pipeline achieves a classification accuracy equivalent to naively trained CNNs that use randomly selected images for annotation, using an order of magnitude fewer human annotations (i.e. tens to hundreds, as opposed to thousands). The results also demonstrate improvements in accuracy by a factor of 1.2 to 1.5 when a hundred or less annotations are used compared to conventional supervised learning methods using an equivalent CNN architecture, where the largest gains in learning efficiency are achieved with small numbers of annotations. Unlike alternative methods that attempt to reduce the number of human annotations needed such as active learning approaches, the proposed method does not require multiple interactions with human annotators, which is a significant disadvantage of active learning methods that need humans to work around the schedule of a machine. The method is also less sensitive to initialisation conditions since sequential learning methods do not address how initial annotation subsets should be sampled.

The advantages of the semi-supervised method are also shown to improve the learning efficiency on three aerial imagery datasets, where the results are shown in Appendix A. This demonstrates that the proposed location guided latent representation learning and representative image selection strategies are effective for environmental applications across different georeferenced imaging domains.

The experiment in section 4.4.3 evaluates the semi-supervised pipeline with the hard assumption based contrastive learning on a real-world seafloor image dataset consisting of ~87k images (section 4.1.2). The proposed pipeline improves the performance and efficiency of human effort for classification, where selecting a prioritised training dataset using *H-kmeans* clustering increases the classification accuracy by an average of 4.9 % and maximum of 14.1 % compared to random selection, where the performance gains are more significant for small numbers of annotations given by humans.

The F_1-score achieved in this work ranges from 60% to 72 %. Although the necessary level of accuracy, and the metrics themselves differ between applications, Purkis et al. (2019) reported 80 % to 90 % accuracy scores for satellite remote sensing based shallow sea habitat mapping using highly tailored processing algorithms for specific habitat classes, and similarly Zelada Leon et al. (2020) reported accuracy levels of 60 % to 70 % scores in side scan sonar based deep sea habitat mapping. An advantage of the method proposed in this research is that the feature learning aspect does not require separate algorithms to be developed for the different applications.

5.2 Further insight

The two different representation learning methods presented in this thesis have different characteristics that can influence their applicability to different datasets and application domains. This section describes these characteristics to help judge the suitability and limitations of each method. Since the soft assumption based method loosely regularises autoencoder training using a modified loss function derived from metadata values, it can consider multiple metadata sources at once and minimises the risk of overfitting the metadata through the use of a Student's t-distribution and maintaining the reconstruction loss. Therefore, this method is suitable for datasets where various types of metadata are available, but where the correlation between the metadata and image appearances is not clear. Examples of applications where this behavioural property might be useful include those where metadata characteristics might vary at a faster temporal scale or different spatial scale than seafloor appearance. e.g. water column pH, temperature, chemical composition near hydrothermal vents or cold seeps. On the other hand, the hard assumption based method exploits the metadata more strictly for selecting similar image pairs. This is useful where the assumptions are known to be valid, e.g. using 3d geolocation where the geological and habitat features are known to vary over spatial scales larger than the footprint of a single image frame). However, efficient training relies on the fact that the metadata used to identify similar image pairs strongly correlates with the similarity of image appearance. As long as this condition is satisfied, the hard assumption method can be applied to train any CNN architecture, which is an advantage over the autoencoder based method since a decoder architecture needs to be implemented. As such, the hard assumption method can take advantage of the rapid advance in CNN architectures more readily than the autoencoder based method.

Regarding the relative effectiveness of the two methods, although the datasets and the validation methods are not identical, the results of experiments in sections 3.2 and 3.3 show that the hard assumption based representation learning can be more effective when similar metadata relationships are used and these relationships show good correlation with image appearance. For the analysis performed in these sections, the datasets partially overlap, and although the soft assumption based method achieves an F_1-score of ~58 % (as shown in Table 4.4), the hard assumption method achieves ~78 % for an expanded dataset which has a ~50 % overlap in the analysed images (as shown in Table 4.5).

In general, deeper CNN architectures, larger minibatch sizes and epochs are known to provide accuracy gains. However, the availability of these gains is determined by the necessary computational power where, for applications in the field, access to high-performance computer networks is limited. The proposed representation learning techniques do not increase the computational requirements, and although the autoencoder

based soft assumption requires a decoder to be implemented, in principal the methods can be used with any CNN architecture, where this can be chosen to match the available computer resources. The desktop workstation used in the experiments in chapter 4 used a commercial grade GPU (NVIDIA TITAN RTX with 24 GB VRAM) and was capable of processing the datasets, including the replicate experiments, within a few days, and a single run can be computed in the order of a few hours. The processing time is short enough to generate results in timeframes relevant to assist planning and interpretation between AUV dives during multi-day field expeditions.

5.3 Future work

This section summarises future work and potential applications that could be investigated following on from this PhD research.

Visual-acoustic multimodal data learning for seafloor surveys
In typical seafloor survey scenarios, acoustic data measured at the same location as images, i.e. bathymetry and backscatter, are often available. Previous works (Rao et al., 2014, 2017; Shields et al., 2020) have looked at combining visual images and acoustic measurements to extend image cluster or class predictions to regions of acoustic bathymetry that have not been imaged, and this has proven to be effective for regional scale (hundreds to thousands of km^2) seafloor habitat interpretation. One potential approach for leveraging acoustic measurement is learning vector-form representations of acoustic measurements with the proposed soft or hard assumption based method and using the obtained representations for representation learning of visual imagery by that same method. Values such as acoustic reflection intensity could be used as metadata inputs and may show strong correlation with visual image appearance. However, it is important to consider the spatial correlation between the acoustic data and the images they are linked to, since both the resolution and footprint of the measurements is likely to differ, and consistent underwater localisation is a challenge.

While the metadata exploited in the proposed methods take the form of vectors or scalars, where the similarity in metadata space can be straightforwardly derived, acoustic seafloor maps can form larger dimensioned tensors, i.e. n-dimensional arrays. Therefore, extending the proposed representation learning methods to be applied to tensor-form data is considered one of the potential future works. Modern multimodal learning concepts can also be applied.

Another potential application is to use the methods described in this work on acoustically derived images. This concept is demonstrated in a co-authored conference publication 'Autonomous Identification of Suitable Geotechnical Measurement Locations using Underwater Vehicles', which applies the soft assumption based method to learn features from patches of laser-derived bathymetry data, where depth deviation from the mean of the patch is linearly mapped to the colour intensity. A similar approach could be adopted for direct classification and clustering of acoustic bathymetry and acoustic backscatter intensity. A potential area for investigation may be to use visual image clustering or classification results as metadata to guide representation learning of acoustically derived images.

Temporal analysis for environmental monitoring
In section 4, the experimental results show that a georeference, i.e. horizontal location and depth, is useful metadata for representation learning on seafloor imagery datasets.

Since seafloor substrates and habitats have unique spatial distributions that extend beyond the footprint of a single image frame, using georeference information as metadata can capture relevant image features. The spatial distributions of habitats and substrates can change gradually over time. However, as long as the images captured in nearby regions are observed at a time interval that is shorter than this rate of change, the images are possibly similar. This tendency can be exploited for image representation learning in the same manner as the proposed metadata leveraging to estimate the potential similarity between images. For environmental monitoring, temporal analysis plays a significant role as well as spatial analysis, and so multi-year observation data of similar places is often gathered. To interpret such datasets, considering both spatial and temporal relationships between images may pose potential advantages over processing temporal datasets in independent time slices.

Over-horizon communication

Due to the large size of images and the limited available communication bandwidth between an AUV and its operators during missions, the collected images cannot normally be seen by humans until the AUV is physically recovered for data extraction. With AUVs now being developed with increased mission durations of weeks to months, the discrete and sequential nature of data gathering and extraction introduces significant latency between data acquisition and the insight it provides to scientists and operators. Such insights can be useful during the deployment, so that AUV missions can be adaptively replanned based on human decisions without physical recovery. The latent representations acquired by the proposed methods are much lower in their dimensions than the original images, so they can be useful for transmitting seafloor status data efficiently compared to sending their original images. Furthermore, the unsupervised representative image identification proposed in this thesis can also be applied to efficiently select which images to transmit to build understanding of the observations over the limited bandwidths available, rather than sending a stratified or randomly sub-sampled selection of images observed.

Application to other imaging domains

Although this work has focused on the analysis of georeferenced seafloor imagery, the methods developed can be applied to other imaging domains. The methods have been shown to improve the learning efficiency when applied to georeferenced aerial imagery (see Appendix A), and when applied to holographic images of suspended particles in a co-authored journal paper 'Unsupervised feature learning and clustering of particles imaged in raw holograms using an autoencoder' that is currently under review, and also to suspended marine particles imaged using an Underwater Vision Profiler (UVP) in a co-authored paper tentatively entitled 'Efficient classification of marine particles using self-supervised learning' that is currently being prepared for submission to a journal.

5.4 Authored publications

The following outputs have been published, submitted or presented based on the work done in this PhD.

Journal article

- Takaki Yamada, Adam Prügel-Bennett, and Blair Thornton. Learning features from georeferenced seafloor imagery with location guided autoencoders. Journal of Field Robotics 38.1: 52-67, 2021.
- Takaki Yamada, Miquel Massot-Campos, Adam Prügel-Bennett, Stefan B. Williams, Oscar Pizarro, and Blair Thornton. Leveraging Metadata in Representation Learning With Georeferenced Seafloor Imagery. IEEE Robotics and Automation Letters, vol. 6, no. 4, pp. 7815-7822, 2021.

Journal article (in review)

- Takaki Yamada, Miquel Massot-Campos, Adam Prügel-Bennett, Oscar Pizarro, Stefan B. Williams and Blair Thornton. Guiding Labelling Effort for Efficient Learning with Georeferenced Images.
- Takaki Yamada, Adam Prügel-Bennett, Stefan B. Williams, Oscar Pizarro, and Blair Thornton. GeoCLR: Georeference Contrastive Learning for Efficient Seafloor Image Interpretation.

Conference

- Takaki Yamada, Miquel Massot-Campos, Emma Curtis, Oscar Pizarro, Stefan B. Williams, Veerle A.I. Huvenne, and Blair Thornton. Metadata Enhanced Feature Learning for Efficient Interpretation of AUV Gathered Seafloor Visual Imagery. Marine Geological and Biological Mapping (GEOHAB), 2021.
- Takaki Yamada, Miquel Massot-Campos, Adam Prügel-Bennett, Stefan B. Williams, Oscar Pizarro, and Blair Thornton. Leveraging Metadata in Representation Learning with Georeferenced Seafloor Imagery. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021.

Invited talk

- Takaki Yamada. Automatic Classification of Seafloor Imagery based on Machine Learning (in Japanese), 5th Subsea and Seafloor Engineering Forum ZERO, 2021.

In addition to these, the following co-authored publications have made use of the methods developed in this PhD for different applications.

Journal article

- Zonghua Liu, Thangavel Thevar, Tomoko Takahashi, Nicholas Burns, Takaki Yamada, Mehul Sangekar, Dhugal Lindsay, John Watson, and Blair Thornton. Unsupervised feature learning and clustering of particles imaged in raw holograms using an autoencoder. Journal of the Optical Society of America A, in press.

Conference

- Jose Cappelletto, Blair Thornton, Adrian Bodenmann, Takaki Yamada, Mehul Sangekar, David White, Justin Dix, and Darryl Newborough. Predicting locations for making geotechnical measurements with Autonomous Underwater Vehicles. IEEE Oceans, 2021.
- Jenny Walker, Takaki Yamada, Adam Prügel-Bennett, and Blair Thornton. The effect of physics-based corrections and data augmentation on transfer learning for segmentation of benthic imagery. IEEE Underwater Technology (UT), 2019.

Appendix A

Results for aerial imagery dataset

The proposed georeference leveraging latent representation learning and representative image selection method is designed for domains where georeferenced image datasets exhibit patterns occurring on spatial scales larger than the size of each image patch read by a CNN. Here, we apply our method to land cover classification based on aerial imagery. This domain shares challenges with seafloor image classification, where both domains have recently seen the increased use of mobile robotic imaging platforms that have reduced the cost of gathering data. However, the cost of annotating images in a way that is suitable for environmental monitoring still requires significant domain expertise. To demonstrate the versatility of the proposed method, experiments are carried out on three varied aerial image datasets.

Dataset

This appendix shows the validation results of the semi-supervised pipeline proposed in section 3.5.2. Instead of seafloor image datasets, aerial image datasets are used, since georeference is also available as metadata possibly related to image appearances. Aerial image datasets from three different regions (Mountain, Island and Urban) of Sweden are used to test the versatility of our method. Table A.1 shows the description of the datasets. The Mountain dataset consists of images of the area surrounding Vindelfjällen Nature Reserve, which is one of the largest protected areas in Europe (see Figure A.1). Six classes are observed in this area, where 'Wetland' and 'Other Non-vegetated' (corresponding to alpine peaks) are unique to this dataset in our experiments. The region also has areas of 'Water'. The Island dataset is of Gotland island, which consists of four classes, including large regions of farmland ('Arable' class), as shown in Figure A.2. The Urban dataset consists of images around the city of Stockholm (see Figure A.3). This dataset consists of six classes, where the 'Artificial' class is used to describe the city and other built up areas suburbs, where this class is unique to this dataset in our experiments. The dataset also contains some 'Arable' and 'Water' regions. All datasets have 'Coniferous', 'Deciduous' and 'Other Vegetated' areas, although their appearances and distribution patterns are different between the datasets.

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

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

Learning

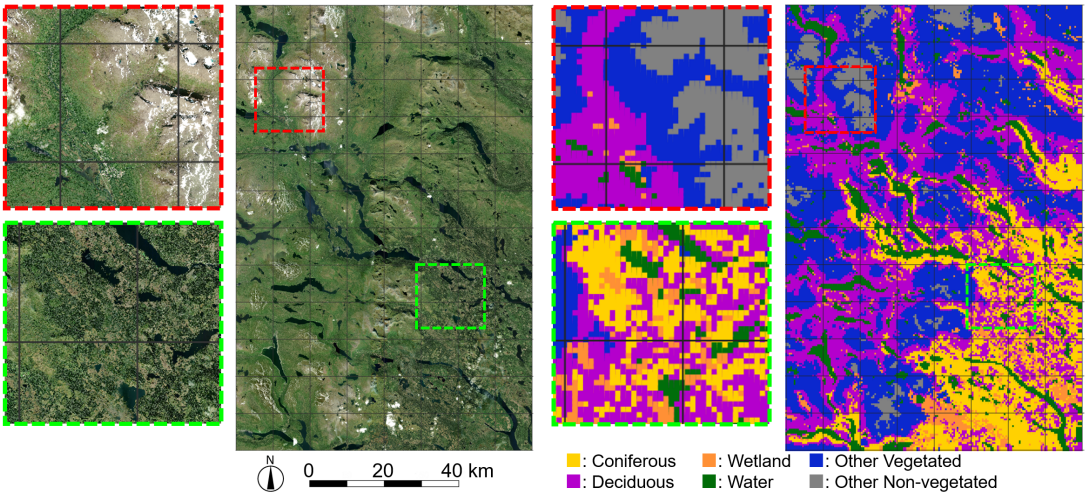
The training configuration follows the same procedure as the experiment in section 4.4.2. Two AlexNet based convolutional autoencoders: with/without the soft assumption based, are trained on the three aerial image datasets introduced in the previous section. For the soft assumption based training, the ratio between the spatial resolution (image patch size) and the distance threshold of the loss function is set to be the same as the seafloor dataset experiment described in the section 4.4.2.

Discussion

Table A.2 shows the F_1 accuracy (macro-average) scores for the Mountain/Island/Urban datasets and the mean and standard deviation values of the three datasets. The results show a similar trend to the experiment on the seafloor dataset. The soft assumption based autoencoder (A1-A15) performs better than the standard autoencoder (A16-A30) when the same data selection method and classifier are applied, showing that georeference information is being effectively leveraged. When using the soft assumption based autoencoder, the proposed hierarchical k means based data selection also outperforms random and k means based data selection when using the same classifier. This improvement in performance is larger for smaller M values, revealing that the proposed method is also effective for classification of aerial imagery with fewer annotations. This trend is not observed when the standard autoencoder is used. This is thought to be because the selection of images automatically chosen for prioritised annotation is less effective when using the standard autoencoder as opposed to the soft assumption based autoencoder.

TABLE A.1: Description of aerial imagery datasets

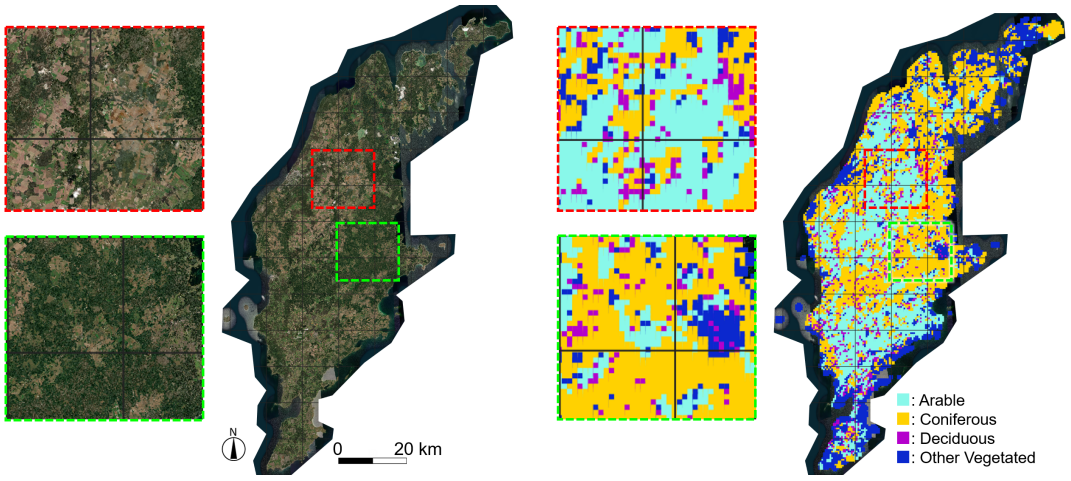
	Mountain	Island	Urban
No. of Image Patches	46,200	15,128	47,961
Resolution [m/pixel]	2.0	2.0	1.0
Imaged Area [km^2]	9,520	3,120	2,470
No. of Classes	6	4	6
Latitude [°N]	65.01 to 66.09	56.91 to 58.00	59.14 to 59.60
Longitude [°E]	14.79 to 16.53	17.96 to 19.35	17.45 to 18.36
Lat. × Lon. Edge Lengths [km]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



(A) Mosaiced aerial imagery of the Mountain dataset

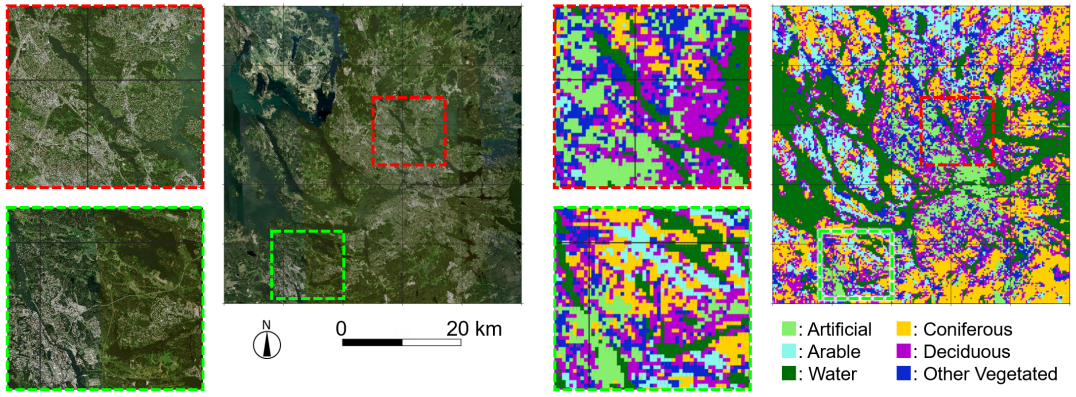
(B) Ground truth classes for the Mountain dataset

FIGURE A.1: Mountain dataset showing the area surrounding Vindelfjällen Nature Reserve in Sweden. Six classes are observed in this area, where 'Wetland' and 'Other Non-vegetated' (corresponding to alpine peaks) are unique to this dataset in our experiments. The dataset also has 'Water', 'Coniferous', 'Deciduous' and 'Other Vegetated' regions, where these classes are shared across the different datasets studied in this work. The figure shows that the spatial distributions of the shared classes are different to their distributions in the Island (Figure A.2) and Urban (see Figure A.3) datasets.



(A) Mosaiced aerial imagery of the Island dataset (B) Ground truth classes of the Island dataset

FIGURE A.2: Island dataset showing Gotland island in Sweden, which consists of four classes, including large regions of farmland ('Arable' class) that dominate the open areas. The dataset also has 'Coniferous', 'Deciduous' and 'Other Vegetated' regions, where these classes are shared across the different datasets studied in this work. The figure shows that the spatial distributions of the shared classes are different to their distributions in the Mountain (Figure A.1) and Urban (see Figure A.3) datasets.



(A) Mosaiced aerial imagery of the Urban dataset (B) Ground truth classes of the Urban dataset

FIGURE A.3: Urban dataset showing the area surrounding Stockholm in Sweden. The 'Artificial' class is used to describe the city and other built up areas, where this class is unique to this dataset in our experiments. The dataset also has 'Arable', 'Water', 'Coniferous', 'Deciduous' and 'Other Vegetated' regions, where these classes are shared across the different datasets studied in this work. The figure shows that the spatial distributions of shared classes are different to their distributions in the Mountain (Figure A.1) and Island (see Figure A.2) datasets.

TABLE A.2: F_1-Score (Macro-Average) Mean and SD (%) of the Classification Result on Three Aerial Imagery Dataset

Config. Label	Rep. Learning	Data Selection	Classifier	Number of Annotations (M)						
				20	40	100	200	400	1000	7500
A1	with soft assumption	random	1-NN	43.8±4.2 42.0/45.3/44.0	49.3±3.0 49.6/47.9/50.4	53.4±2.4 54.5/50.9/54.8	55.4±1.5 56.8/52.5/57.0	57.6±0.9 60.3/53.7/58.8	59.4±0.8 63.0/55.0/60.3	62.1±0.4 66.9/56.4/63.0
A2			RF	42.7±6.2 39.0/45.4/43.8	49.4±5.0 48.5/49.0/50.7	55.8±3.2 55.8/53.2/58.4	57.9±2.3 59.0/53.2/61.5	60.8±1.2 63.2/55.1/64.2	63.3±0.9 66.6/56.8/66.4	66.6±0.3 71.2/58.8/69.8
A3			L-SVM	46.3±4.5 45.6/46.8/46.4	52.1±3.6 52.5/50.2/53.6	58.6±2.3 60.3/55.5/59.9	61.1±1.4 62.7/58.2/62.5	63.3±1.0 65.0/59.7/**65.4**	65.3±0.5 66.7/61.9/67.3	66.9±0.4 68.1/63.4/69.2
A4			R-SVM	42.8±4.7 39.0/46.7/42.8	51.1±3.5 50.4/51.3/51.6	58.5±2.2 59.8/**55.9**/59.9	61.4±1.4 62.9/**58.5**/62.8	**63.8**±1.0 65.6/**60.3**/65.3	**65.7**±0.5 67.5/**62.3**/67.4	69.0±0.3 70.7/65.3/**70.9**
A5			GP	44.0±4.1 42.9/45.1/44.0	50.1±3.2 51.4/48.0/50.9	55.1±2.7 56.5/51.9/57.0	57.6±2.0 59.0/52.7/60.9	60.5±1.3 63.5/54.4/63.7	63.2±1.0 67.2/56.1/66.3	68.1±0.4 72.9/60.5/70.8
A6		*k*means	1-NN	47.1±4.3 46.5/47.4/47.4	51.1±2.4 53.3/49.7/50.3	54.0±2.1 56.6/50.3/55.1	55.9±1.4 58.9/52.0/56.9	57.1±1.1 60.7/53.0/57.8	59.0±0.7 62.3/54.4/60.3	61.7±0.4 66.3/56.2/62.7
A7			RF	45.3±5.7 42.8/47.9/45.1	51.4±3.0 51.5/50.9/51.8	56.0±1.9 57.0/52.4/58.5	58.4±1.7 59.9/54.0/61.2	60.6±1.2 63.3/55.0/63.6	63.1±0.9 66.2/56.9/66.3	66.4±0.4 71.2/58.7/69.3
A8			L-SVM	49.1±4.5 47.3/50.0/50.0	54.4±2.3 56.1/52.1/54.9	58.6±1.9 60.9/54.7/60.1	61.4±1.3 63.7/57.3/63.2	63.3±1.0 65.4/59.6/65.1	64.8±1.0 66.5/60.6/67.4	66.4±0.4 68.0/62.3/69.0
A9			R-SVM	45.6±4.6 42.9/48.3/45.7	53.4±3.5 53.9/**53.3**/52.9	58.0±2.5 59.2/55.2/59.5	61.3±1.4 63.1/57.6/63.1	63.3±1.1 65.9/58.8/65.1	65.2±0.9 67.2/60.8/**67.6**	68.3±0.5 70.6/63.6/70.8
A10			GP	47.6±4.4 47.4/48.1/47.4	51.9±3.0 53.9/50.8/51.0	56.1±1.8 58.6/51.8/57.9	58.5±1.7 61.0/53.8/60.6	60.4±1.2 63.3/54.5/63.4	63.0±0.8 66.3/56.4/66.4	67.7±0.4 72.5/60.1/70.5
A11		H-*k*means	1-NN	50.7±3.0 50.8/49.9/51.4	52.7±2.4 55.8/49.2/52.9	56.6±1.6 60.9/51.3/57.5	58.2±1.1 61.6/53.9/59.1	59.1±0.7 63.5/53.8/60.0	60.3±0.6 64.5/54.8/61.4	62.4±0.4 67.4/56.4/63.3
A12			RF	49.0±3.6 47.1/49.3/50.7	52.4±2.9 51.5/52.1/53.6	57.9±1.7 59.3/54.3/60.0	59.9±1.7 60.4/56.6/62.8	62.1±1.2 64.8/56.7/64.6	63.8±0.7 67.7/56.9/66.7	66.9±0.4 72.3/58.9/69.6
A13			L-SVM	**51.8**±3.1 **52.5**/50.2/**52.7**	**55.0**±2.2 **58.0**/50.5/**56.4**	**59.8**±1.6 **63.0**/54.1/**62.3**	61.8±1.4 **65.0**/57.2/63.1	63.2±0.8 65.9/58.7/65.0	65.0±0.7 66.9/61.0/67.0	67.1±0.3 69.0/63.5/68.9
A14			R-SVM	50.8±3.3 51.4/**52.1**/48.9	53.8±2.4 55.5/52.5/53.5	59.3±1.9 62.0/54.4/61.5	**61.9**±1.3 64.8/57.5/**63.3**	63.5±0.7 **66.8**/58.4/**65.4**	65.5±0.5 68.1/61.0/67.4	**69.3**±0.3 71.5/**65.4**/**70.9**
A15			GP	50.9±3.1 51.0/50.1/51.7	53.6±2.6 56.1/50.7/54.0	58.2±2.0 61.7/53.1/60.0	60.5±1.8 63.0/55.8/62.8	62.5±1.2 66.0/56.3/65.0	64.1±0.7 **68.3**/57.0/67.0	68.2±0.4 **73.4**/60.6/70.8
A16	without soft assumption	random	1-NN	42.0±4.7 43.0/42.6/40.4	46.5±3.5 50.0/44.5/45.1	49.7±2.0 53.7/47.3/48.1	51.1±1.2 55.3/48.2/49.9	52.9±0.8 57.2/49.2/52.2	54.6±0.6 59.2/50.6/53.9	57.0±0.3 62.0/52.5/56.5
A17			RF	40.4±5.1 40.6/41.5/39.0	47.0±4.3 49.8/45.6/45.5	51.8±2.4 54.5/50.1/50.8	53.7±2.0 56.1/51.0/54.1	55.9±1.0 59.2/51.9/56.6	58.0±0.8 61.2/52.9/59.8	61.3±0.4 64.6/54.8/64.5
A18			L-SVM	43.4±5.1 44.7/43.8/41.8	47.9±3.8 51.1/46.1/46.5	53.5±2.2 56.4/51.5/52.4	55.9±1.2 58.6/53.4/55.8	58.3±0.9 60.8/55.4/58.6	60.9±0.8 63.6/57.6/61.6	63.0±0.3 65.8/59.1/64.0
A19			R-SVM	41.5±5.8 39.1/42.4/43.0	48.5±3.7 49.8/48.3/47.3	54.2±1.9 57.3/52.9/52.4	56.4±1.0 60.1/54.3/54.8	59.0±0.9 63.0/56.4/57.6	61.3±0.6 64.8/58.2/60.8	64.6±0.2 67.1/61.1/65.7
A20			GP	42.1±4.5 43.5/42.0/40.7	47.1±3.6 50.5/44.7/45.9	51.1±2.2 54.1/49.0/50.2	53.4±1.4 56.5/50.1/53.6	55.7±1.1 59.0/51.3/56.8	58.3±0.9 62.0/53.0/59.8	62.5±0.4 67.0/55.6/64.9
A21		*k*means	1-NN	43.5±4.7 46.7/41.7/42.1	45.7±2.4 50.4/45.2/41.7	49.9±2.0 55.7/47.2/46.8	51.4±1.0 56.5/47.9/49.8	53.0±1.1 57.7/49.6/51.8	53.8±0.8 59.1/49.7/52.7	56.3±0.4 61.4/51.6/55.9
A22			RF	42.6±4.7 44.3/42.7/40.7	46.2±3.3 49.5/47.8/41.2	52.2±1.9 57.5/49.1/49.9	54.2±1.3 58.1/51.2/53.3	56.1±1.2 59.6/52.1/56.5	58.0±0.6 61.3/53.1/59.5	61.0±0.4 64.4/54.5/64.0
A23			L-SVM	44.2±4.4 46.5/43.9/42.2	48.3±2.4 51.7/48.6/44.6	52.7±2.4 56.1/50.9/51.1	55.2±2.2 58.4/53.0/54.4	57.7±1.4 60.5/54.5/58.2	59.4±1.3 61.4/55.6/61.3	60.7±1.1 61.7/57.1/63.3
A24			R-SVM	43.8±4.4 45.5/43.4/42.5	48.6±2.7 51.0/49.3/45.4	52.3±2.6 54.7/51.6/50.7	54.9±2.4 59.8/51.7/53.3	57.6±1.7 61.4/54.2/57.3	59.8±1.4 63.5/55.9/59.9	63.5±0.8 66.9/58.9/64.9
A25			GP	44.0±4.6 47.0/42.3/42.5	46.9±2.7 51.2/47.0/42.6	51.9±2.0 57.7/49.1/48.9	54.1±1.2 58.2/50.5/53.5	56.1±1.1 59.4/52.1/56.8	58.3±0.7 61.7/53.1/60.1	62.2±0.4 66.5/55.2/64.8
A26		H-*k*means	1-NN	45.5±3.2 47.6/44.3/44.5	48.0±2.3 54.6/44.9/44.5	51.8±1.8 58.2/47.5/49.7	53.0±1.2 57.8/49.1/52.2	53.6±0.8 58.7/49.3/52.8	54.8±0.7 59.5/50.8/54.2	57.2±0.3 62.2/52.5/56.7
A27			RF	44.8±3.3 44.3/47.3/42.7	48.3±2.8 54.0/46.6/44.3	51.7±2.1 55.8/48.7/50.6	54.6±1.1 58.3/51.1/54.4	55.9±0.9 58.9/52.1/56.8	58.1±0.6 61.4/53.3/59.7	61.5±0.4 65.7/54.6/64.3
A28			L-SVM	47.5±2.7 49.5/46.2/46.7	49.4±3.2 52.7/48.1/47.4	53.6±2.1 57.9/51.0/52.0	54.7±1.4 57.0/51.3/55.6	56.4±1.3 58.2/53.5/57.4	58.5±1.5 58.9/56.2/60.5	62.7±0.4 66.1/59.1/62.9
A29			R-SVM	46.8±3.5 50.0/45.9/44.6	49.6±3.3 52.7/49.2/47.0	53.6±2.0 57.0/51.4/52.6	55.3±1.6 57.6/53.5/54.8	57.8±1.1 60.9/55.1/57.5	60.2±1.6 63.2/57.4/60.0	65.2±0.3 68.8/61.1/65.6
A30			GP	46.4±3.4 48.1/46.0/45.1	49.1±2.6 55.9/46.5/44.9	52.6±1.8 56.9/49.5/51.4	55.0±1.3 58.4/51.5/55.1	56.3±1.0 59.3/52.5/57.3	58.4±0.5 61.3/53.5/60.4	62.7±0.5 67.2/55.6/65.4

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

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*.

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., and McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

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.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020a). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. (2020b). Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*.

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.

Dai, D., Prasad, M., Leistner, C., and Van Gool, L. (2012). Ensemble partitioning for unsupervised image categorization. In *Proceedings of the 12th European conference on Computer Vision-Volume Part III*, pages 483–496.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, PAMI-1(2):224–227.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009a). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009b). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

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.

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2015). The Pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136.

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.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Gal, Y., Islam, R., and Ghahramani, Z. (2017). Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237.

Gowda, H. S., Suhil, M., Guru, D., and Raju, L. N. (2016). Semi-supervised text categorization using recursive k-means clustering. In *International Conference on Recent Trends in Image Processing and Pattern Recognition*, pages 217–227. Springer.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. (2017). Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., et al. (2020). Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

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.

Inglada, J. (2007). Automatic recognition of man-made objects in high resolution optical remote sensing images by svm classification of geometric image features. *ISPRS journal of photogrammetry and remote sensing*, 62(3):236–248.

Iscen, A., Tolias, G., Avrithis, Y., and Chum, O. (2019). Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5070–5079.

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

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

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., van Kevelaer, R., Purser, A., and Nattkemper, T. W. (2020). Gear-induced concept drift in marine images and its effect on deep learning classification. *Frontiers in Marine Science*, 7:506.

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.

Lapedriza, A., Pirsiavash, H., Bylinskii, Z., and Torralba, A. (2013). Are all training examples equally valuable? *arXiv preprint arXiv:1311.6510*.

Le-Khac, P. H., Healy, G., and Smeaton, A. F. (2020). Contrastive representation learning: A framework and review. *IEEE Access*.

Lee, D.-H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3.

Li, Z., Ko, B., and Choi, H.-J. (2019). Naive semi-supervised deep learning using pseudo-label. *Peer-to-Peer Networking and Applications*, 12(5):1358–1368.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

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.

Marmanis, D., Datcu, M., Esch, T., and Stilla, U. (2015). Deep learning earth observation classification using imagenet pretrained networks. *IEEE Geoscience and Remote Sensing Letters*, 13(1):105–109.

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.

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

Moltmann, T., Turton, J., Zhang, H.-M., Nolan, G., Gouldman, C., Griesbauer, L., Willis, Z., Piniella, A. M., Barrell, S., Andersson, E., Gallage, C., Charpentier, E., Belbeoch, M., Poli, P., Rea, A., Burger, E. F., Legler, D. M., Lumpkin, R., Meinig, C., O'Brien, K., Saha, K., Sutton, A., Zhang, D., and Zhang, Y. (2019b). A global ocean observing system (goos), delivered through enhanced collaboration across regions, communities, and new technologies. *Frontiers in Marine Science*, 6:291.

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

Nishida, Y., Ura, T., Sakamaki, T., Kojima, J., Ito, Y., and Kim, K. (2013). Hovering type auv "tuna-sand" and its surveys on smith caldera in izu-ogasawara ocean area. In *2013 OCEANS-San Diego*, pages 1–5. IEEE.

Nister, D. and Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2161–2168. Ieee.

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.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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

Paull, L., Saeedi, S., Seto, M., and Li, H. (2014). Auv navigation and localization: A review. *IEEE Journal of Oceanic Engineering*, 39(1):131–149.

Purkis, S. J., Gleason, A. C., Purkis, C. R., Dempsey, A. C., Renaud, P. G., Faisal, M., Saul, S., and Kerr, J. M. (2019). High-resolution habitat and bathymetry maps for 65,000 sq. km of earth's remotest coral reefs. *Coral Reefs*, 38(3):467–488.

Rao, D., De Deuge, M., Nourani-Vatani, N., Douillard, B., Williams, S. B., and Pizarro, O. (2014). Multimodal learning for autonomous underwater vehicles from visual and bathymetric data. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3819–3825. 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.

Rigby, P., Pizarro, O., and Williams, S. B. (2010). Toward adaptive benthic habitat mapping using gaussian process classification. *Journal of Field Robotics*, 27(6):741–758.

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.

Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE.

Settles, B. (2009). Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

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

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60.

Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1857–1865.

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.

Stephens, D. and Diesing, M. (2014). A comparison of supervised classification methods for the prediction of substrate type using multibeam acoustic and legacy grain-size data. *PloS one*, 9(4):e93950.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. (2018). A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer.

Thornton, B., Bodenmann, A., Pizarro, O., Williams, S. B., Friedman, A., Nakajima, R., Takai, K., Motoki, K., Watsuji, T.-o., and Hirayama, H. (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.

Tian, Y., Liu, W., Xiao, R., Wen, F., and Tang, X. (2007). A face annotation framework with partial clustering and interactive labeling. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Van Etten, A., Lindenbaum, D., and Bacastow, T. M. (2018). Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*.

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.

West, G., Bodenmann, A., Newborough, D., and Thornton, B. (2020). Resolution and coverage-the best of both worlds in the biocam 3d visual mapping project. *Journal of Ocean Technology*, 15(3):67–76.

Wigness, M., Draper, B. A., and Beveridge, J. R. (2015). Efficient label collection for unlabeled image datasets. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4594–4602. IEEE.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Williams, S. B., Pizarro, O., Webster, J. M., Beaman, R. J., Mahon, I., Johnson-Roberson, M., and Bridge, T. C. (2010). Autonomous underwater vehicle–assisted surveying of drowned reefs on the shelf edge of the great barrier reef, australia. *Journal of Field Robotics*, 27(5):675–697.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Wu, H. and Prasad, S. (2017). Semi-supervised deep learning using pseudo labels for hyperspectral image classification. *IEEE Transactions on Image Processing*, 27(3):1259–1270.

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.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742.

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

Yamada, T., Massot-Campos, M., Prügel-Bennett, A., Williams, S. B., Pizarro, O., and Thornton, B. (2021a). Leveraging metadata in representation learning with georeferenced seafloor imagery. *IEEE Robotics and Automation Letters*, 6(4):7815–7822.

Yamada, T., Prügel-Bennett, A., and Thornton, B. (2021b). Learning features from georeferenced seafloor imagery with location guided autoencoders. *Journal of Field Robotics*, 38(1):52–67.

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.

Yao, X., Han, J., Cheng, G., Qian, X., and Guo, L. (2016). Semantic annotation of high-resolution satellite images via weakly supervised learning. *IEEE Transactions on Geoscience and Remote Sensing*, 54(6):3660–3671.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*.

Zelada Leon, A., Huvenne, V. A., Benoist, N., Ferguson, M., Bett, B. J., and Wynn, R. B. (2020). Assessing the repeatability of automated seafloor classification algorithms, with application in marine protected area monitoring. *Remote Sensing*, 12(10):1572.

Zhou, Z., Shin, J., Zhang, L., Gurudu, S., Gotway, M., and Liang, J. (2017). Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7340–7351.

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).

Zurowietz, M. and Nattkemper, T. W. (2020). Unsupervised knowledge transfer for object detection in marine environmental monitoring and exploration. *IEEE Access*, 8:143558–143568.