UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering and Physical Sciences
School of Engineering

Detection and Segmentation of Fauna in Seafloor Imagery For Biomass Estimation

DOI: To be Issued

by

Jennifer Louise Walker
BSc Computer Science
ORCiD: 0000-0002-1449-9012

A thesis for the degree of
Doctor of Philosophy

July 2023

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences
School of Engineering

Doctor of Philosophy

Detection and Segmentation of Fauna in Seafloor Imagery For Biomass Estimation

by Jennifer Louise Walker

Machine learning based image processing is sensitive to variation caused by hardware and observation conditions, making the use of machine learning with marine imagery particularly difficult when transferring knowledge between datasets. There is also a considerable gap between the outputs of machine learning systems and useful biological information for marine conservation purposes. This thesis investigates the effects of physics based image normalisation and augmentation methods on the transferability of an object detection and segmentation system between two distinct datasets taken at different altitudes from the seafloor with different camera and lighting systems. Scale, colour, and lens distortion correction methods are investigated, along with augmentation methods including linear contrast, motion blur, and noise, and more advanced distorting methods such as elastic distortions and piece-wise affine transformations. A set of experiments for each combination of independent variables has been carried out, finding a clear improvement when using scale correction. When applying to low altitude datasets only there is an increase in average performance from 62.2% to 68.6%, and when transferring knowledge from high to low altitude datasets, there is an increase in performance from an average of 26.5% to 44.1% when using scale normalisation. Colour normalisation also had a large impact, when applied to low altitude data showing an increase in performance from 56.6% to 74.1%, and when transfering from high altitude to low altitude datasets showing an increase in performance from 32.7% to 38.0%. The impacts of lens distortion correction and various augmentation methods were found to be less significant. This thesis goes on to demonstrate the use of segmentation results for biomass estimation through a simple polynomial relationship between segment size and length of an individual, and previously well established Length Weight Relationships (LWRs). The resulting method is fully scalable to larger datasets with no additional human effort required, a vast improvement on the current labour intensive biomass estimation methods used.

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. None of this work has been published before submission

Signed:.. Date:..................

Acknowledgements

Thanks to the Ocean Perception research group at the University of Southampton, my research buddy Amy Parkes, and of course to my loving partner and constant support Lucas Noyau. This research would not have been possible without the funding and support from both the Leverhulme Trust and the NEXUSS Centre for Doctoral Training.

Definitions and Abbreviations

ROV	Remotely Operated Vehicle
AUV	Autonomous Underwater Vehicle
LWR	Length Weight Relationship
$SSLR$	Segment Size to Length Relationship
CNN	Convolutional Neural Network
SVM	Support Vector Machine
MLP	Multi-Layer Perceptron
MAE	Mean Absolute Error
MSE	Mean Squared Error
CRF	Conditional Random Field
IOU	Intersection Over Union
MAP	Mean Average Precision
$\mathcal{L}$	Low Altitude Dataset
$\mathcal{H}$	High Altitude Dataset

Chapter 1

Introduction

1.1 Background and Motivation

Benthic imaging systems are developed in order to gain a greater understanding of habitat and population distributions on the seafloor. Acoustic approaches such as bathymetry and side-scan sonar imaging are useful tools for mapping large habitat features, however visual imagery remains the most reliable non-invasive method for identifying fauna and small scale features, and is the focus of this thesis. Visual imagery must be collected at a relatively close distance off the seafloor, commonly less than 10 metres, compared to bathymetry and side-scan sonar, which can be collected at greater than 100 metre altitude. Therefore it must be collected by an underwater vehicle such as a towed vehicle, a remotely operated vehicle (ROV) or an autonomous underwater vehicle (AUV). Visual imagery datasets collected by such vehicles may then be analysed in a variety of ways to extract useful information about the habitat and populations present.

During robotic deployments, several thousands of millimetre resolution images are gathered along robot trajectories that are kilometres in length. These show different substrates, geological features, and individual fauna that need to be characterised in order to describe the seafloor ecosystem. Manual analysis of such images is common, but is time consuming and scales poorly to large scale marine monitoring endeavours, and therefore automated analysis is an area of much interest for the field of marine environment monitoring [6]. Automated analysis of such images is a challenge due to many factors. Firstly, many developments in automated analysis rely on the availability of example images and assigned labels that an algorithm can try to reproduce. Typically these labelled training datasets need to be large enough to represent the different types of targets and scenes that need to be described, as machine learning algorithms do not generalise well to data outside the training domain. In marine imagery the amount of available labelled training data is severely

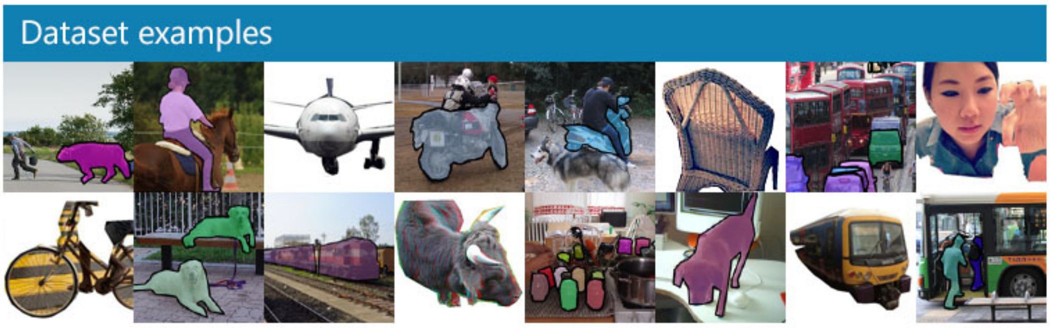


FIGURE 1.1: Examples of labelled images from the COCO dataset [1]

limited compared to terrestrial imagery datasets such as COCO [1] with over 200,000 labelled images, Pascal VOC [7] with over 10,000 labelled images, and Cityscapes [8] with 20,000 labelled images. These datasets contain object classes such as people, cars, houses, and trees, and are not directly useful for marine imagery analysis. Examples from the COCO dataset are shown in Figure 1.1.

Efforts have been made to create larger marine datasets for improved machine learning algorithm training, including BIIGLE 2.0 [9], Squidle+ [10], and FathomNet [11].

It is often down to the teams wanting to interpret marine imagery to generate their own training datasets, which poses a large time and energy investment that for most applications is not justifiable. Secondly, the marine environment introduces many challenges for visual imaging systems, such as increased light absorption underwater, and varying levels of visibility due to water turbidity. Furthermore, variations to the imaging system and the vehicle it is mounted on pose additional challenges. The use of artificial light is required in deep water, and variation to how this is set up has large impacts on the appearance of images. The camera specifications vary between imaging systems, providing differences in lens distortion and resolution. Finally, the vehicle's target altitude off the seafloor, and ability to maintain that altitude in rugged territory, has a large impact on the spatial resolution and appearance of objects within images.

Because we have different imaging systems and varied environments, we need to have a thorough understanding of how these things impact the information we gather. We also need to establish best practises to ensure we get as much information out of the data as possible and understand how reliable the information is. It is also important to investigate whether training data taken from one system can be effectively applied to data gathered by another system, or under a different set of conditions.

This thesis investigates methods to automatically analyse AUV imagery, and goes on to investigate the transferability of such a system to images taken by a different imaging system to those it was trained on. In order to achieve this, the thesis will investigate the use of standard data augmentation techniques that have been shown to be effective in terrestrial image interpretation. This will be combined with methods of data normalisation that are more specific to the underwater imaging domain, including using optical models to correct for underwater image distortions, and use available meta-data such as imaging altitude to correct for colour degradation cause by light absorption and non-uniform illumination effects. The investigations will be carried out on a dataset gathered in the North East Pacific Ocean at approximately. 700-800 metres deep, that consists of imagery gathered by two AUVs with very varied specifications.

This thesis goes on to demonstrate how object detection and segmentation results may be used to automatically estimate biomass on a per class basis, which has been identified as a key ecological indicator for marine research. [12] This novel method uses previously identified Length to Weight Relationships (LWRs), and newly calculated Segment Size to Length Relationships (SSLRs) to form a simple polynomial relationship between the segment size and estimated biomass of an individual. This method provides a fully automated scalable solution to biomass estimation as machine learning methods already exist, and are continually developed and improved, that provide automated object segmentation.

1.1.1 Image Acquisition Methods

There are a variety of visual benthic image acquisition methods, including static camera systems installed on the seafloor, towed camera systems, ROVs and AUVs.

Static camera systems are less prone to motion blur, and are not required to traverse terrain in the same way that moving camera systems are. The main drawback of such a system is that it only captures information in one static location, so is more often used for temporal monitoring for a very small area of interest such as at a deep ocean observatory such as MARS at Station M [13], and PAP-SO at the Porcupine Abyssal Plane [14].

Towed camera systems are the simplest and often the cheapest of the moving imaging system options, however it also has the least control in terms of navigation and positioning, and must be towed at a high enough altitude off the seafloor to ensure it won’t collide with rugged terrain.

ROVs provide the most control and flexibility, being connected to a manned vessel on the surface by a data and power cable [15]. Human operators of ROVs can often see live video feeds from the vehicle, and can operate the vehicle’s actuators and sampling equipment such as robotic manipulators.

AUVs require no human intervention during operation except for deployment and recovery, having autonomous control systems on board [16; 17; 18; 19; 20]. They can be programmed with a target route beforehand, or can use simple decision making algorithms to determine their route in situ. As such, AUVs can be used to collect imagery across a large area of the sea floor with relatively little human interaction. This method of imagery acquisition is used for the experiments in this thesis as the resulting datasets are large, consisting of hundreds to thousands of images from a single deployment, giving a valuable opportunity for automated analysis.

Collecting images from the seafloor is only the first step in gaining a greater understanding of habitat and population distributions. The images need to be interpreted in order to extract useful information.

1.1.2 Image Interpretation

Useful information that can be extracted from images include population counts, substrate identification, percentage cover of coral colonies, species diversity and distributions, and biomass distributions.

Common methods for assigning manual labels to images include, but aren't limited to; whole image annotation, point labels on objects of interest, and classification of randomly selected point labels. When making size or biomass estimates, length weight relationships are often used, requiring the labeller identify some key length per morphotype, such as the length of many fish, or the distance from the centre to the end of an arm of a sea star. There are many software packages for manual labelling of images, but ultimately manual labelling is a time consuming task that isn't scalable to large scale datasets. In practice, a small subset of the images is used, either through random sampling or stratified sampling, to create a more manageable dataset, before using aggregate statistics to represent the larger dataset. The statistical significance of the resulting information is lower than what could be achieved when using more of the available images, and information on the distribution is lost. Rare classes of information, be that fauna species, substrate type, geological feature, or something else, may not appear within the subset of images. Furthermore, arbitrary decisions on how to subsample, and how many images to subsample, may have unseen effects on the research outcomes.

Machine learning algorithms, although they often require labelled data to train on, are a scalable solution that can be applied to a much larger number of images without increasing the human input needed. However, training machine learning algorithms with marine imagery datasets faces many challenges.

Firstly, there aren't many large scale labelled datasets suitable for training a generalisable system. Many large scale image acquisition programmes have been

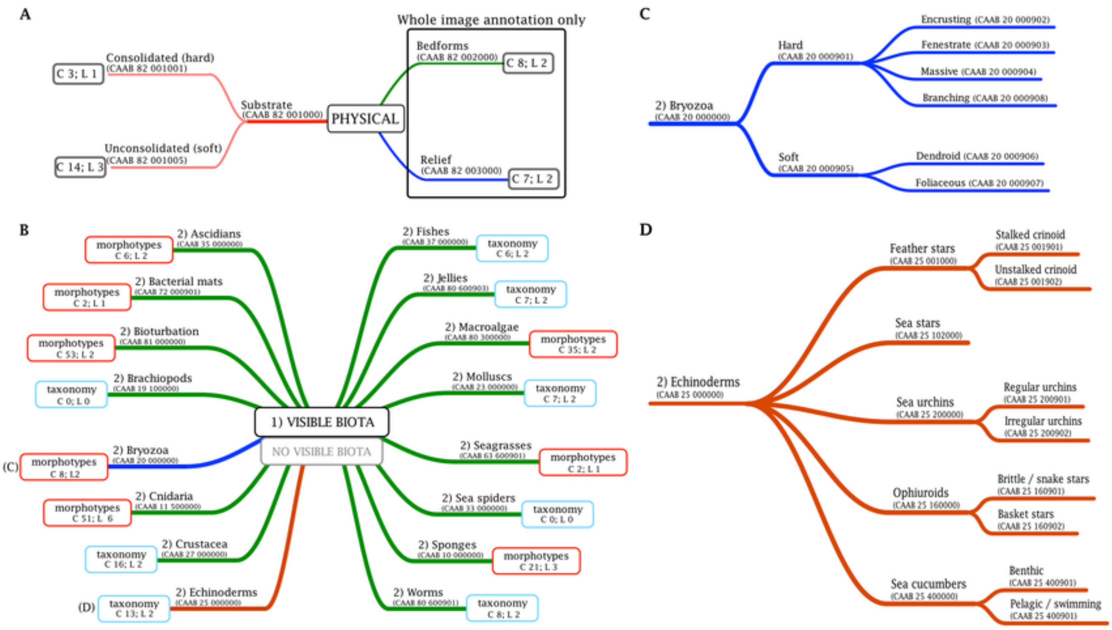


FIGURE 1.2: An overview of the coarsest levels of the CATAMI classification scheme [2]

carried out, providing many images, such as the Australian integrated marine observation system (IMOS) national AUV marine monitoring scheme [21]. Unfortunately, only a small subset of images from a small subset of such schemes are labelled in a way that is useful for machine learning training. One example is FathomNet [11], which uses video imagery from ROVs. Another example is BENTHOZ-2015 [22], a large scale dataset consisting of 9,874 images, with 50 randomly selected points labelled in each image. This dataset is specific to shallow water off the coast of Australia, and consists of 148 substratum and biological classes. Many of these classes are highly specific to the shallow water reefs in Australia, and aren't generalisable to other areas. Schema have been developed to try to encompass all the classes of interest such as VARS [23], developed for video annotation, and CATAMI [2], developed for marine image annotation in Australia. There are many classes within these schema, with complex hierarchies, and for a worldwide general marine dataset, many labelled examples would be needed for every class. An overview of the CATAMI classification scheme is shown in Figure 1.2, the full hierarchy of labels is not shown, as it contains over 280 classes.

The field of deep-sea AUV benthic imagery is currently limited to small area and vehicle specific datasets and often using different label formats hindering the ability to collate these experiments into a larger dataset. These individual labelled datasets often consist of a small number of images, as manual labelling is costly in terms of time and effort, with a small number of researchers working on each dataset.

Secondly, the appearance of marine imagery varies greatly with environmental factors and hardware setups. The behaviour of light in water, and interacting with particles suspended in it, poses a challenge for underwater imaging, and introduces a large amount of variance to the appearance of objects depending on their distance from the camera and the water conditions. Hardware setups and observation conditions vary greatly too, varying in their distance off the seafloor, resolution, lens distortion, and artificial lighting setups; notably the only light source when surveying in the deep sea [24].

1.2 Problem Statement

Automated image analysis using machine learning is commonly used with Autonomous Underwater Vehicle (AUV) collected data, given the high volume of images captured and the limited time for manual annotation by experts. For effective machine learning, target imagery must be captured within the same bounds as the training data to ensure similarity in visual characteristics and distribution, enabling accurate predictions on new, unseen data. However, standardisation across AUV imagery is difficult due to differences in hardware, survey techniques, and environmental factors.

This thesis aims to improve our understanding of how this variation limits the information we can obtain from automated image analysis and how to overcome these differences. Specifically, the thesis addresses the problem of using physics-based image normalisation and data augmentation methods to generalise across hardware setups and environmental conditions in an automated object detection and segmentation machine learning system. The study uses data from the Adaptive Robotics 2018 Falkor expedition, including surveys conducted over multiple days with two AUVs of differing specifications.

The thesis assumes that physics-based image normalisation and data augmentation methods effectively mitigate the impact of differences in hardware and environmental conditions. Additionally, the use of the Falkor data provides a sufficient level of variation to capture the differences that exist in AUV imagery. The thesis employs mean Average Precision (mAP) as a metric to measure the performance of the machine learning algorithm, justifying its use. The ability to predict useful biological statistics, such as biomass estimation, is also investigated through traditional Length Weight Relationships and a novel Segment Weight Relationship method. The thesis assumes that manually annotated lengths and traditional Length Weight Relationships are relatively accurate, and the correlation between manual and automated systems is used to demonstrate the feasibility of the approach.

This thesis aims to address the challenges of standardising AUV imagery for machine learning-based image analysis. By exploring the effectiveness of physics-based image normalisation methods and data augmentation techniques, it seeks to overcome the differences in hardware setups and environmental conditions. Using the Adaptive Robotics 2018 Falkor expedition data, with variations in AUV imagery, this research investigates traditional and novel methods for estimating biological statistics. Ultimately, this work has the potential to enhance our ability to understand and manage marine ecosystems through efficient analysis of AUV imagery.

1.3 Contributions

The main contribution of this thesis to the field is a set of in depth experiments into machine learning algorithms for the identification of fauna in benthic imagery, and the transferability of these algorithms to other imaging systems. These experiments use data collected using two AUVs with differing target altitudes and image resolutions. Having data collected using two different imaging systems in the same geographical area allows for a unique opportunity to analyse the cross vehicle transferability of algorithms.

Augmentation and normalisation methods have been selected based on the context of marine imagery, using image meta data and the physics of light's behaviour in water to normalise for colour and scale, and a contextual understanding of AUV camera positioning and seafloor rugosity to inform augmentation technique selection such as affine and piece-wise transformations.

These experiments use metrics that are standard in the field of machine learning, and go on to present context specific performance metrics using Length Weight Relationship (LWR) regression. LWRs are commonly used in marine imagery analysis to estimate biomass based on a key length measurement such as from the tip of the head to the tail of many fish morphotypes, or the width of the shell of crustaceans, or the length from the centre to the end of an arm for echinoderms. The LWRs used in this thesis are from the open source database Fishbase [5], and are largely calculated using a Bayesian approach by Froese et al. [25]. By calculating the relationship between such key length measurements and the size of segments, the estimation of biomass may now be scaled to larger automatically analysed datasets. Rather than measuring the performance of Mask RCNN based solely on classic machine learning metrics, looking instead at the performance based on the context specific final outcomes, ie. population count and biomass estimation, we assess the suitability of Mask RCNN to the specific context of marine imagery and population monitoring.

1.4 Outline

This thesis is structured as follows:

Chapter 2 outlines existing literature and established methods in the areas of computer vision for object detection, classification, and segmentation, and of visual benthic imagery.

Chapter 3 presents the methodology used in the experiments presented in this thesis, covering the neural network architectures used, data augmentation and normalisation methods including physics based approaches, the metrics used to assess the neural networks' performance, and the regression model used to relate length weight relationships to segment sizes.

Chapter 4 presents the datasets collected and analysed as part of this thesis. It describes the acquisition hardware used, the labelling method, and the resulting population counts for each class.

Chapter 5 presents an in depth experiment to the effects of data augmentation and data normalisation practices for benthic imagery on the performance of Mask RCNN.

Chapter 6 presents an in depth experiment into the transferability of Mask RCNN to data captured by a different acquisition system to the training data. Transferability from low altitude images to high altitude images and vice versa is investigated, and the impact of different data augmentation and normalisation practices on this transferability is assessed.

Chapter 7 presents a novel automated biomass estimation method that can use the output of the Mask RCNN network to form these estimates. It makes use of newly calculated relationships between segment size and line length per morphotype, and known Length Weight Relationships.

Chapter 8 discusses in further detail the possible implications of this research, how it differs from other findings in the literature, and other topics such as class imbalance, unseen classes, and the the unattainable idea of an ideal universal marine imagery dataset.

Chapter 9 concludes this thesis, summarising the findings of the experiments, and their implications for the wider field. Suggested best practices for automated marine imagery analysis are discussed, and areas requiring further investigation addressed.

Chapter 2

Literature Review

2.1 An Overview of Automated Image Analysis

Image analysis is the extraction of quantifiable information from images that are otherwise just matrices of pixel values. This comes in many forms, including image clustering and image classification, where the entire image, or part of it, is given some value based on the pixel values within a region. More detailed image analysis involves identifying regions within the image, either through semantic segmentation where every pixel is assigned a class, or object detection, where individual instances of objects are identified. Approaches that involve some manually identified ground-truth, and a model that aims to recreate that relationship between input images and the given labels, are supervised learning methods. Other approaches that don't make use of manual labels are unsupervised methods, instead using similarities between images to cluster them, or using encoding and decoding to extract important features, for example.

This thesis focuses on supervised learning methods for identifying objects within images, as it has the potential for extracting key marine statistics such as population counts and biomass distributions, useful for researchers and conservationists. This approach involves having human experts identify objects in a set of images which is then used for training and testing the performance of a model. This is often shown in the form of input variables X and their target output variables Y. The model is then optimised to best recreate the target output.

Many modern approaches involve some form of neural network as advances in computational ability and in neural network architecture design have produced impressive gains in performance. The next section covers how neural networks function. More traditional approaches are also explored in this literature review, many of which have the advantage of lower computational costs, but often don't perform as well as neural networks and deep learning approaches.

2.2 Classical Approaches

Image analysis can be done using basic image features such as brightness, colour, and texture. One such approach is thresholding, where images are translated into binary black and white, with the category of each pixel depending on the intensity of the pixel value [26]. This approach is especially useful for text analysis where thresholding clearly separates the text from the background, or in other very basic semantic segmentation tasks where the subject is sufficiently different to the background in colour or intensity.

Other basic approaches include canny edge detection, where the patterns of pixel values are analysed and edges of high contrast are identified [27]. This approach identifies the edges of objects with a high contrast compared to the background, and therefore can be used to segment objects, but is highly sensitive to noise, such as a noisy marine environment with complex backgrounds and varying amounts of contrast between objects and their environment. A combination of edge detection and pixel grouping algorithms can be used to achieve a simple semantic segmentation algorithm [28].

Semantic segmentation can also be achieved through region growing algorithms [29], where initial regions are defined by point labels or automatically based on basic image features like colour and intensity, and regions with similar features are added, expanding regions to segment entire images. More complex region based approaches use splitting and merging of regions to achieve better semantic segmentation results [30].

2.3 Support Vector Machines

Support Vector Machines (SVMs) are a machine learning technique that involves projecting data points into a large dimensionality space maximising the distance between points belonging to different classes, and estimating the class of a new data point by where in that space it gets projected onto [31]. SVMs are binary linear classifiers, where the input data is a list of features defined as X and the target label is a binary value y.

SVMs have performed well on machine learning tasks, but are often out performed by deep neural networks on more complex tasks. In their simplest form they are used for binary classification, but in more complex structures they are capable of tasks such as semantic segmentation [32]. Their uses in marine image processing are discussed later in this chapter, but they are not the focus of this thesis.

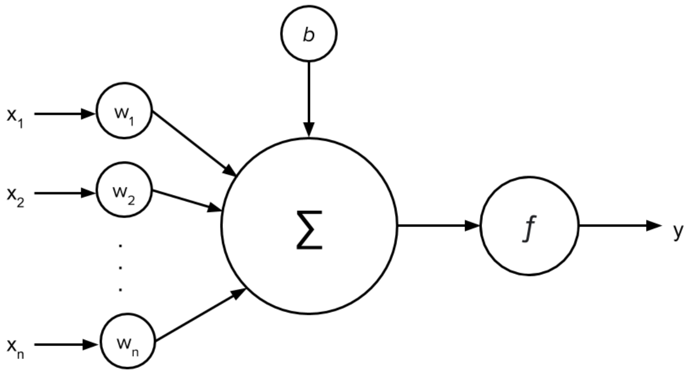
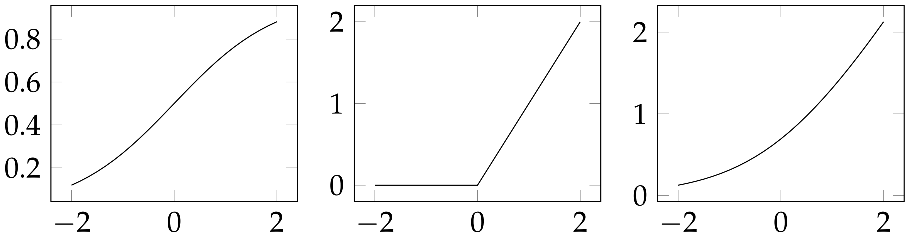


FIGURE 2.1: A diagram of a single neuron with n inputs, x_1 to x_n, showing a weight applied to each input, those values being summed, a bias being added, and an activation function applied, producing the output y.



(A) Sigmoid

(B) ReLU

(C) Softplus

FIGURE 2.2: Graphs of three different activation functions.

2.4 Neural Networks

Neural networks are composed of neurons [33] arranged into what are referred to as layers or perceptrons [34]. Each neuron takes weighted inputs from either the input data in the case of the input layer, otherwise neurons from the previous layer, and applies an activation function which determines the output of the neuron. This activation function can take many forms, including sigmoid, reLU, and softplus, for example. These activation functions are shown in Figure 2.2. A single neuron is demonstrated in Figure 2.1, where x_1 to x_n represent input values, w_1 to w_n represent the weights assigned to each of those inputs, b represents the bias value applied, f represents the activation function, and y represents the output value of this neuron. This can be mathematically defined as

$$f((\sum_{i=1}^{n} w_i x_i) + b) = y \tag{2.1}$$

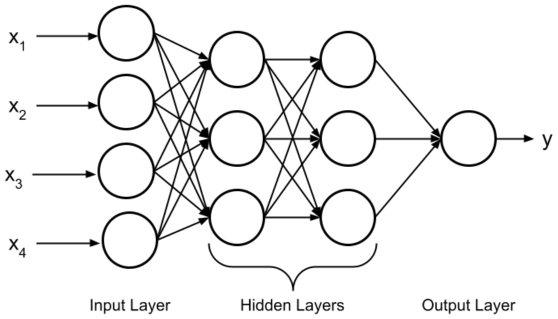
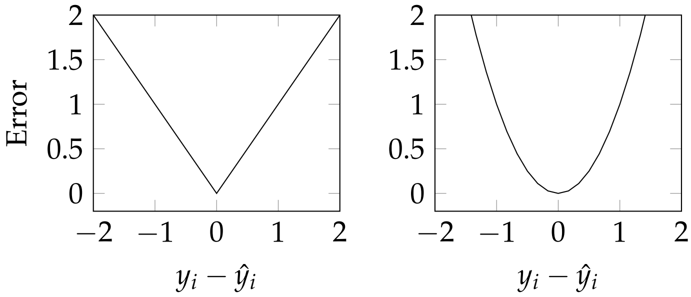


FIGURE 2.3: A diagram of a Multi Layer Perceptron (MLP)

When arranged into a series of layers, as shown in Figure 2.3, neurons can be formed into Multi Layer Perceptrons (MLPs), that take input data x and form output y. A single neuron can perform very simple separation problems, but when combined in an MLP can model much more complex problems [35]. In this MLP, the final output may be a numerical value in the case of regression, or a categorical output in the case of classification. For image analysis, classification is far more common practice as it's useful for categorising images or the pixels or objects within them.

Training a neural network involves a set of input data, here labelled X, and their corresponding target output values, here labelled Y, and some measure of the distance between the actual neural network output and the target output values, referred to as a loss function. The weights connecting the neurons in the network, w_1 through to w_n in Figure 2.1, are then adjusted using gradient descent, with the aim of moving towards a lower loss value with each training epoch. The training of an individual neuron, for example, would involve passing the input values $X = \{x_1, x_2, ..., x_n\}$ through the neuron, applying the weights, the bias, and the activation function, to reach some output value y. This output would then be compared with the target output for the given input, $Y = \{y_1, y_2, ..., y_n\}$, using some loss function to quantify how similar they are, with a smaller loss value meaning they are more similar, and a higher loss value meaning they are more different. The weights w_1 through w_n are then adjusted a small amount in the direction that reduces this loss value, making the output y more similar to the target output. For a single neuron this process is quite simple, however for multiple layers this wasn't feasible until the invention of back propagation [36] allowing for the weights of neurons in earlier layers to be updated via gradient descent.



(A) Absolute Error

(B) Squared Error

FIGURE 2.4: Graphs showing two different regression loss functions

Loss functions vary by type, with the two most common types being regression functions and categorical functions. They also vary by how they calculate the error or difference between the actual output and target output. For example, regression loss functions compare two numerical values, with two common functions being Mean Absolute Error (MAE) and Mean Squared Error (MSE) shown in Figure 2.4 and calculated like so;

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.2}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.3}$$

The mean absolute error is linear in how it treats differences in values, whereas the mean squared error gives relatively higher weight to larger errors. This can result in different behaviour when training a machine learning algorithm, as the mean squared error will penalise an algorithm that is correct most of the time but wildly off on a small number of validation samples more so than the mean absolute error would.

Categorical loss functions also vary in the way they calculate the error. Binary loss functions apply when only two categories are present, if there are more than two classes then multi-class categorical loss functions are required. The most common multi-class categorical loss function is categorical cross entropy, calculated like so;

$$CCE = -\sum y_i * \log \hat{y}_i \tag{2.4}$$

where the output y and the target output $\hat{y}$ are arrays of the same length as the number of classes. For example y may be $[0,1,0,0]$ where the sample in question belongs to the second class, and $\hat{y}$ may be $[0,0.7,0.2,0.1]$ where each value represents the predicted probability of the sample belonging to each class.

Gradient descent in neural networks has the potential to reach very good solutions to problems, but faces some challenges.

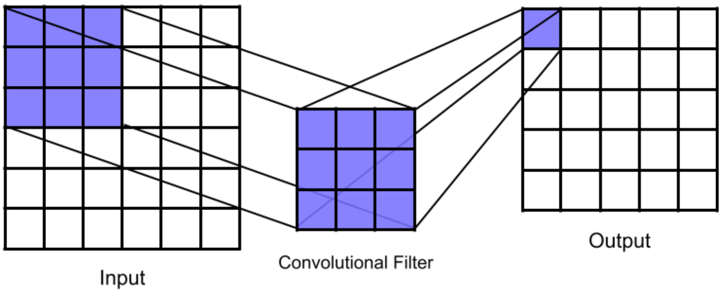


FIGURE 2.5: A diagram of a Convolutional Filter

- Local minima occur in the solution space, where a lower global minima may be achieved when starting from another set of weights, applying noise to the gradient descent process, or through processes such as simulated annealing [37].

- Vanishing gradient occurs with very deep neural networks consisting of many layers, where the gradient tends towards 0 and deep layers can't be trained. A range of solutions to this problem have been proposed and are in use, one being the ReLU activation function where the gradient is always 1 or 0 so many gradients multiplied together will also always be 1 or 0, rather than an increasingly small number as more values between 0 and 1 are multiplied together [38]. Another solution is skip layers used in residual neural networks passing gradient information directly between layers [39].

- Over fitting is when a network so closely models the training data that it fails to model the underlying relationship between the inputs and outputs, and fails to generalise to unseen data. This can be overcome in a variety of ways. One such way is including drop out where some neurons are ignored at random in each training epoch, which reduces over-fitting by creating thinned out networks that differ from each other in terms of architecture, with the final result being a combination of all of these thinned out networks [40]. Another is data augmentation where the input data is adjusted in some way, creating slight variations in the training data while maintaining the validity of the target output [41].

2.4.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a form of neural network with shared weights, meaning some of the weighted inputs are shared between neurons. Shared weights can represent more complex structures in neural networks. In the case of CNNs these shared weights are in the form of convolutional filters, an NxN sliding window that is applied across a 2D slice of data, as shown in Figure 2.5. Basic

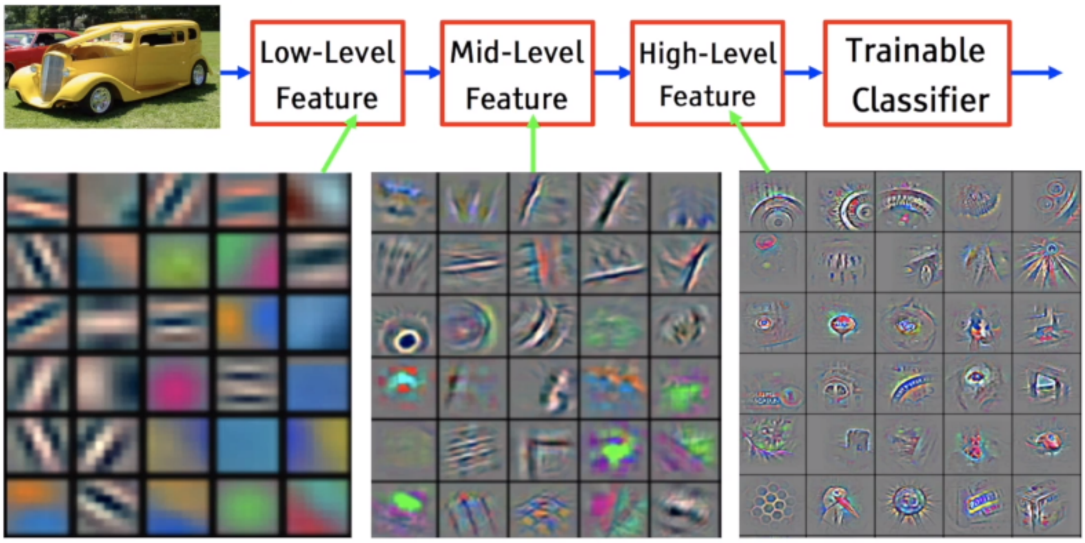


FIGURE 2.6: Learnt features in a CNN, published by medium.com in July 2020

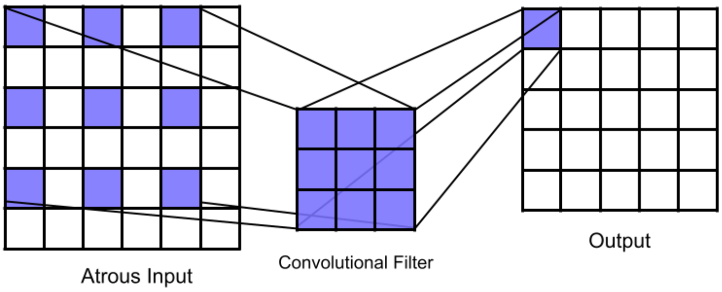


FIGURE 2.7: A diagram of an Atrous Convolutional Filter with a stride of 1

convolutional filters can identify edges in different directions, textures, and other simple small scale features. Because convolutional filters are used across the entire image, with the same weights each time it is applied, these features are identified wherever they appear in the image, they are not location specific. When assembled into a network with subsequent layers of convolutional filters, later layers can identify much more complex features, as shown in Figure 2.6.

Atrous, or strided, convolutional filters are much like the standard convolutional filters described above, but instead of being a solid NxN grid, they are spaced out, as shown in Figure 2.7.

Conditional Random Fields are a discriminative model based on Markov Random Fields [42]. They are used to consider neighbouring examples when forming a prediction, and therefore are useful in image analysis where the context of an object, or the pixels surrounding it, are of great importance in identifying it. They have been used as a stand alone machine learning technique for semantic segmentation, and have been used in conjunction with other methods to improve segmentation mask outputs.

2.5 Whole Image or Point Annotations

Whole image annotation involves assigning a single class label to each image. Point annotation involves picking a point in the image and some surrounding patch of pixels, and treating that as a stand alone image for annotation. This is achieved by translating the image into some set of features or a point in a feature space, and then dividing that feature space along some boundaries to form categories. The way in which the given feature space is built varies, and some classic computer vision techniques for this include Scale Invariant Feature Transforms (SIFT), bag of visual words, colour descriptions such as colour histograms, and texture descriptors such as Gabor and Haar Wavelets. More modern CNN approaches extract features from the images through layers of neurons as described above.

Point annotations can be useful for gathering statistics such as substrate identification, but, whether using random or stratified points, is not sufficient for statistics such as population count, percentage cover, or biomass estimation.

2.6 Semantic Segmentation

Rather than assigning a single label to each image, semantic segmentation involves assigning a label to each pixel or region of pixels within an image. Greedy graph-based approaches have been used [43] for computational efficiency, and for unsupervised learning where target labels don't exist. This approach does not assign a semantic class label to each segment. SVMs have been used for semantic segmentation [44], using a 5x5 sliding window across the input image, and extracting basic colour and texture features from that 25 pixel patch. The development of CNNs has allowed for large performance gains in semantic segmentation, and the use of Fully Convolutional Networks (FCNs) allowed for the adaptation of whole image classifiers into semantic segmentation classifiers by replacing the final fully connected layer with a convolution layer [45]. More recent work includes DeepLab developed by Chen et al. [46] the use of deep CNNs to extract score maps, followed by interpolation up to the original image size, and then CRFs to classify pixels. They then improved on this with the development of DeepLabv3+ [47], making use of multiple scales of atrous convolutions to improve on segmentation at varying scales.

The training of semantic segmentation systems from sparse labels such as per-image classes is a common approach to overcome a lack of training data, with approaches such as super pixels and region growing. Kolesnikov et al. presented their Seed, Expand, Constrain (SEC) system [48], with a novel loss function that measures the performance in each of the three given areas. Seeding is the allocation of a class at a given location with no measure of size or boundary, expanding is the ability to expand from a seed location to a reasonable size for the object, and constraining is the ability to determine a precise boundary. The combination of these three principles results in relatively well performing networks from sparse labels.

Semantic segmentation provides the information required for percentage coverage statistics, however fails to identify the boundaries between different instances of the same class, so can not be used for population counts or biomass estimates.

2.7 Object Detection

A limiting factor in semantic segmentation is that there's no distinction between individual instances of the same class in the image. More complex systems that implement object detection are required to classify distinct individuals.

Object detection provides the information needed for population counts, biodiversity measures, and if followed up with size estimation per instance, biomass estimates.

There are two types of approach to this problem, one stage and two stage object detection algorithms.

2.7.1 One Stage Object Detection

One stage object detection requires a predetermined number of objects to detect per image and is a single neural network from end to end, making it faster to train than more complex two stage object detection methods. One stage systems are more commonly used where computational cost needs to be minimised, such as in the field. Two such networks are YOLO [49] and DetectNet [50].

2.7.2 Two Stage Object Detection

Two stage object detection involves firstly a region proposal network, and then a second network that refines and classifies the proposed regions, allowing the prediction of a varying number of objects in each image, but taking longer to train than the simpler one stage detection methods. The improved accuracy of these systems is better suited to processing images after collection where computational cost isn't as limited, and higher accuracy can be prioritised. For this reason, this thesis focuses entirely on two stage object detection methods.

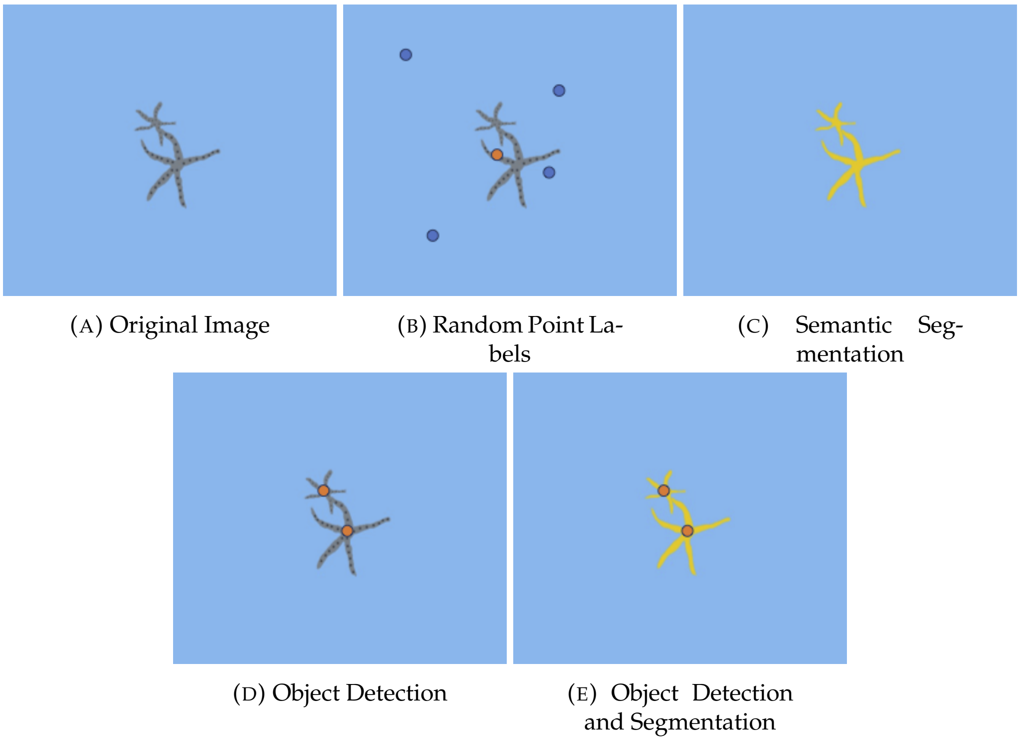
A popular two stage object detection system combines a region proposal network with a following convolutional neural network to classify the object in the proposed region, thus called R-CNN [51]. In the first stage, a selective search algorithm generates potential object bounding boxes or regions of interest (RoIs) based on appearance and texture similarities. In the second stage, RoI pooling converts the proposed regions into fixed-size feature maps, allowing efficient processing by subsequent network layers.

The CNN extracts high-level features from each region, encoding them into feature vectors. These feature vectors are passed through fully connected layers for object classification, using a softmax function to determine class probabilities for each region.

R-CNN also predicts more accurate bounding box coordinates by refining the initial proposals through regression layers. Non-maximum suppression is applied to remove redundant bounding box predictions and retain the most confident and non-overlapping detections.

This architecture has been adapted to improve both efficiency and accuracy, resulting in the development of Fast R-CNN[52]. Fast R-CNN, an evolution of R-CNN, addresses some of the limitations of its predecessor by introducing a streamlined approach. It eliminates the need for selective search by directly taking the entire image as input and performing RoI pooling on shared convolutional feature maps. This significantly speeds up the computation time compared to the multi-stage R-CNN. Fast R-CNN also introduces a single-stage training process that jointly learns the classification and bounding box regression tasks, leading to improved performance.

Building upon the advancements of Fast R-CNN, the object detection architecture was further improved to create Faster R-CNN[53]. Faster R-CNN addresses the computational bottleneck of region proposal generation in Fast R-CNN by introducing a Region Proposal Network (RPN). The RPN operates on shared convolutional feature maps, generating region proposals directly, which eliminates the need for separate selective search or external proposal methods. By sharing computation with the subsequent stages, Faster R-CNN achieves significant speed improvements. Additionally, Faster R-CNN introduces a unified network that jointly learns the region proposal and object classification tasks end-to-end. This integration allows for more efficient training and improved accuracy.



(A) Original Image

(B) Random Point Labels

(C) Semantic Segmentation

(D) Object Detection

(E) Object Detection and Segmentation

FIGURE 2.8: Comparison of labelling methods on two intersecting sea stars, deomnstrating the difference between random point labels, semantic segmentation, object detection, and object detection with semantic segmentation

2.7.3 Object Segmentation

Once objects have been identified, and often with a bounding box surrounding their location, a step further is to then segment each detected image. This allows for the boundary of the object to be more finely classified, and for the size of the object to be determined, at least in terms of pixels in the image, if not physical measurements.

Mask R-CNN is an extension of the Faster R-CNN model that incorporates an additional stage to generate binary masks alongside the bounding box refinement and object classification stages. It utilises the same region proposal network architecture as Faster R-CNN but introduces an extra branch dedicated to mask prediction. This branch outputs a pixel-level mask for each proposed region, allowing for precise segmentation of objects within the regions. By combining region proposals, bounding box refinement, classification, and mask prediction, Mask R-CNN provides a comprehensive framework for object detection and instance segmentation [3].

Figure 2.8 shows examples of each of these labelling methods to better illustrate the differences between them.

Lateef et al. carried out an extensive review of modern image analysis techniques in 2019, focusing on semantic segmentation but also categorising approaches such as object detection and segmentation, and video frame segmentation [54]. This review

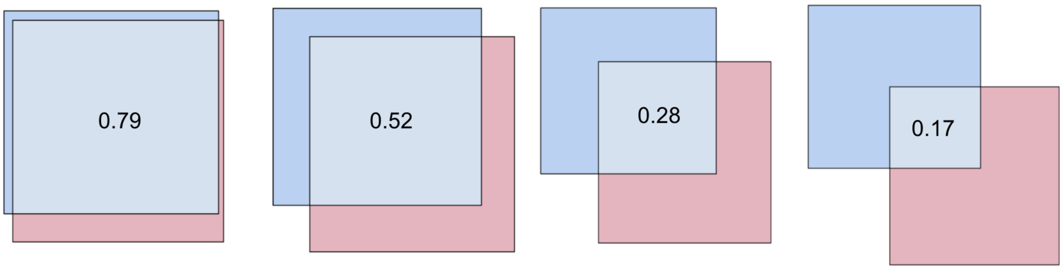


FIGURE 2.9: Example IoU scores on intersecting boxes

covers most of the more modern techniques described in this section, and separates popular architectures into clearly defined categories.

2.8 Metrics for Success

An important aspect of training machine learning algorithms is how we measure the similarity between the target output and the algorithm output. It is also important when we are comparing two approaches to understand what metric is being used to measure success. For regression models, this involves using measurements such as mean squared error. For semantic segmentation, the proportion of correctly classified pixels is often used to evaluate performance. In the case of object detection and segmentation, there are a lot of factors to consider.

2.8.1 Intersection over Union

The first step in assessing the performance of an object detection system is to measure the similarity between an object in the groundtruth dataset and an object in the predicted output. A popular measurement for this is the Intersection over Union (IoU) score, defined as the intersecting area between the groundtruth and estimate divided by the union of the two areas. Example IoU scores are shown in Figure 2.9.

2.8.2 Mean Average Precision Score

To understand how the Mean Average Precision (mAP) score is calculated, we first need to understand precision-recall curves. The precision is defined as $\frac{TP}{TP+FP}$, and recall is defined as $\frac{TP}{TP+FN}$, where TP is the sum of the true positives, where a prediction was made that matched a groundtruth object, FP is the sum of false positives, where a prediction was made where no such object exists in the groundtruth, and FN is the sum of false negatives, where predictions were not made but an object did exist in the groundtruth dataset. The certainty threshold indicates how certain the neural network's prediction needs to be to include the object in the prediction outputs. The precision recall curve is the plot of the precision and recall at different certainty thresholds. The average precision is defined as the area under this curve, resulting in a score between 0 and 1, with 1 being the best possible score. This score is calculated for each test example, and the mean of these scores is the mAP score.

2.9 An Overview of Visual Benthic Imagery

Visual benthic imagery provides the means to observe habitats and fauna on the sea floor. This section outlines the key challenges specific to imagery taken in a marine environment rather than in a terrestrial environment, and current uses for benthic imagery for habitat and population mapping.

2.9.1 Environmental and Hardware Caused Variance

As touched on in Chapter 1, a major motivation for this thesis is to overcome the marine imagery specific challenges presented when automating image analysis. In part, these difficulties arise from the large variance caused by environmental and hardware differences.

One challenge specific to the marine environment is the effect of light attenuation through water. Visible light attenuation in water is much higher than in air, and is much higher for longer wavelength light waves such as red, than shorter ones such as blue and green, meaning the appearance of objects and their colour varies depending on the distance from the light source and the camera.

The above mentioned light attenuation challenge is only intensified when taking into account water turbidity in natural environments. Water turbidity is the measurement of water clarity, and depends on the amount of light scattered by particles suspended in the water. In natural benthic environments the turbidity varies greatly depending on the amount of sediment picked up by currents, the movement of fauna, or even the propellers of the AUV collecting imagery, as well as by marine snow.

As demonstrated by the points above, the distance between the light source, object being photographed, and the camera, have a large impact on the appearance of objects in the images taken. The camera and lighting mounting positions on the AUV will directly affect this. Furthermore, the lighting intensity, colour, and direction have a heavy influence. Finally, the camera itself has a large impact on the appearance of objects, depending on the lens distortion, resolution, colour sensitivity, and so on.

Due to non-uniform lighting, with the centre of marine images often more well lit, and the edges often much darker, objects may appear differently even within the same image depending on their location in that image. The effect this may have on object detection algorithms has been investigated by Schoenning et al. [55], finding that lighter morphotypes are harder to identify in the over saturated centre of images, and darker morphotypes are harder to identify in the dark edges of images.

To summarise, a large amount of variance is introduced to how identical objects would appear under different environments, with different imaging hardware, and at different locations relative to the camera.

2.9.2 Uses for Marine Ecology Monitoring

Conservation typically involves understanding the spatial distribution, physical characteristics and diversity of species in a region. In order to derive this sort of information from images, it is necessary to convert from collections of pixel values to real world dimensions and meaningful labels.

Traditionally this has been a manual task, for example when Seiler et al. [56] manually annotated the lengths of a specific species of ocean perch, Helicolenus percoides, in AUV imagery, in order to determine both the occurrence, length, and biomass distribution, but also information pertaining to habitat preferences.

Morris et al. [57] also manually annotated their images. They annotated the location and lengths of a large range of morphotypes in images taken at the Porcupine Abyssal Plane. This study used images taken by the Autosub6000 AUV, images taken by a towed camera system, and data collected by trawling, in order to make a comparison between the different data acquisition methods and the resulting measures of diversity and abundance in the surveyed area. This study aimed to understand the influence of sloped mounds on the distribution of different species. Benoist et al. [58] also aimed to do this, and they used AUV images stitched together into tiles of approximately 7.3m^2 to assess faunal density, diversity, and composition. They also investigated the area coverage required to make a statistically sound image based survey, and suggests the use of number of individuals found rather than the area covered by an AUV to define a sampling unit.

Another example using manually annotated marine imagery is the study by Thornton et al. [59], who used 3D image reconstructions from AUV images to model the population distributions and biomass distributions at two deep-sea vent sites in the Iheya North field in the East China Sea. Unlike the previous two studies mentioned which investigated the impact of slopes, this study looked into the impact of hydrothermal vents on the distribution of different species. In this study, the biomass estimates were based on physical samples of each major morphotype being collected, cleaned, ground down into a fine powder, and incinerated to determine their organic carbon content.

Howell et al. [60] used a combination of trawl data, video from towed cameras, and photographic transects, to model the aggregations of deep-sea sponges in the North Atlantic. These datasets were also manually annotated to extract single points per sponge to use a presence-background modelling approach, and the gathered information was used to assess what drives the distribution of the sponges, and how that relates to current Marine Protected Areas (MPAs) designed to protect them, and inform future changes to MPAs to better achieve this.

Another important conservation effort involves understanding the impact of deep-sea mining, and assessing the rate of recovery for an impacted eco-system. Simon-Lledo et al. [61] carried out an AUV imagery based survey of the Clarion Clipperton Zone in order to assess possible damages caused by polymetallic nodule mining, and the suitability of the protected areas meant to mitigate the damage and allow re-population of affected areas. Each individual greater than 10mm was labelled with as specific a morphotype label as possible, and their biovolume was estimated using the generalised volumetric method developed by Benoist et al. [62]. Biovolume has also been estimated from micro-photography for meiso-fauna, harpacticoids and nematodes, using a highly manual process that relies on a high contrast between the background and the individuals in question, that is not applicable to photography in a natural environment with natural substrate backgrounds [63].

To summarise, many studies have applied marine imagery to support marine ecology and conservation efforts. Each of these examples involves labelling the images, either with single points to look at abundance and distribution, or including size information to extend this to look at biomass, to then gain useful biological statistics from the images. Each of these examples also uses this information in conjunction with other information, such as geographical location or measured environmental variables such as temperature or salinity, to address scientific hypotheses. Studies like this enable us to gain a better understanding of the impacts of MPAs, sloped regions, hydrothermal vents, and more.

2.10 Automated Analysis of Benthic Imagery

Efforts to automate the analysis of benthic imagery are increasingly common with the increasing volume of benthic data collected, and with the improving performance of computer vision approaches. Both classic computer vision techniques, such as basic colour and texture features and SVMs, and more modern deep learning approaches have been investigated, and a summary of some notable automated benthic imagery analysis is presented below. The growth in this area is well investigated and presented by Blaschke et al. who specifically look at Geographic Information System (GIS) object based image analysis for remote sensing, and the growth in research in this area over the past few decades [64].

Combining datasets collected in the same locations to form time series data has been achieved via survey planning and aligning imagery datasets to compare exact locations, allowing for analysis of how an area is changing over time [65].

2.10.1 Habitat Classification

As mentioned above, the coarsest form of automated image analysis is whole image annotation. This is used in the marine imagery context for habitat labelling, categorising an image based on the substrate and objects present in the image.

Features based on grouping pixels into regions of similar colour and texture are used to determine image complexity, and therefore predict habitat types where smaller particle grain sizes like sand produce less image complexity than larger particle grain sizes such as pebbles and boulders [66]. These features are also used to estimate the seafloor heterogeneity by analysing the local variability in complexity. Finally, this method is also used to estimate seafloor coverage percentage by classifying the groups of pixels with a Random Forest binary classifier.

Automated habitat classification based on labelled examples has been investigated, making use of colour and salience features [67]. A combination of feature selectors and classification methods are used by Shihavuddin et al. in their method, suggesting users of the method select the feature selectors and classification methods best suited to their datasets [68].

2.10.2 Point Classification

The classification of a given point in an image, or a given patch of an image, is another common form of annotation in benthic imagery. Investigations have been carried out into extracting features from the patch of an image surrounding the point to be classified, followed by an SVM to classify the point based on said features [69; 70; 71]. The feature extraction methods investigated primarily involve classic computer vision techniques such as bag of visual words, colour, and texture features.

In the case of Mahmood et al.[72], point classification has been used in a grid pattern on a larger image mosaic to perform coarse semantic segmentation to identify where coral is present.

Sun et al. [73] use deep learning networks originally trained on terrestrial datasets before being fine-tuned to marine imagery, to identify morphotypes in patches of images.

2.10.3 Semantic Segmentation

Semantic segmentation is most useful in the field of marine imagery when labelling contiguous areas belonging to a specific class, such as coral or areas of substrate. Biological statistics of note that can be gained from these labels include seafloor coverage of species in the surveyed area. A comparative study of classic semantic segmentation techniques was carried out by Wahidin et al. in 2015, but this did not include deep learning methods or neural networks [74]. The use of an ensemble of SVMs and basic features was investigated by Blanchet et al. in 2016 [75]. Also in 2016, Schoening et al. used an unsupervised learning technique to do binary segmentation between a background and a foreground for the analysis of polymetallic nodules, analysing their size, density, and distribution [76]. King et al. carried out a study into the use of more modern semantic segmentation techniques for the segmentation of coral in 2019 using neural networks. [77] Mizuno et al. have made coral coverage estimates using CNNs using an array of cameras on a towed vehicle [78].

Efforts to overcome the challenge of sparse training data have been made in the area of semantic segmentation of marine environments, where Alonso et al. [79; 80] have developed a method to train a semantic segmentation system with point labels through the use of super pixels and region growing algorithms.

In some cases, semantic segmentation has been used to perform population counts, such as in one study by Schoening et al. [55], where the assumption is that individuals don't overlap with one another to such an extent as to obscure the count. Unsupervised approaches to segmentation have been investigated, for example by Steinberg et al. [81] who used Bayesian models to cluster images and segments within the images. Unsupervised approaches to these problems are incredibly valuable as the cost of manual labelling is so high, however they are limited in terms of performance and in the information output by them - for example with no prior knowledge trained into the system about what different classes are, the output cannot apply correct class labels. Segmentation in lab environments has also been investigated to segment camouflaging cuttlefish [82]. This method is unlikely the generalise well to natural environments and natural background substrates, and is based in very traditional computer vision techniques, using textons and SVMs.

2.10.4 Object Detection

Object detection is used for population counts and population distribution mapping, as it involves identifying all occurrences of a class within the given images. A well established architecture called Inception V3, developed by Szegedy et al. [83], has been used by Piechaud et al. [84] to detect epifauna in benthic imagery. Mandal et al. [85] have used the Faster R-CNN architecture [53] to detect fish in frames of underwater video from GoPro cameras in shallow waters. Zurowietz et al. used Mask R-CNN for machine assisted image annotation, labelling individuals in images with a circle identifying the boundaries of the object [86]. Zurowietz et al. went on to investigate the transferability of knowledge for object detection in images from different vehicles, finding that when vehicles are within half or double the distance off the seafloor from each other that knowledge transfer performs well, and that beyond that gap knowledge transfer falters but still improves performance compared to when no knowledge transfer is performed [87]. They also suggest, based off their results, that the labelled dataset for training should be taken at half the distance from the seafloor as the target dataset for the best results.

2.10.5 Other forms of analysis

Features extracted from AUV imagery have been used to compare AUV datasets collected from the same geographical location in order to match image locations [88]. This is particularly useful for time based surveys, assessing differences over any period from days to years, although the study shows a higher match rate on datasets taken 12 hours apart compared to datasets taken 2 years apart due to environmental changes in the area.

Imagery has also been used in conjunction with other data, for example by Rao et al. [89] who used a combination of AUV imagery and ship based bathymetry measurements to automatically classify habitat types for the purpose of insitu deployment planning.

Another approach to reducing the human effort required to analyse AUV imagery is to automatically identify images of interest, or salient images, for manual analysis. Johnson-Roberson et al. investigated such an approach [90] using colour and texture features to automatically select salient images, and found this approach generalised to new environments.

2.11 Summary

This chapter has given an overview of current computer vision techniques and marine imagery analysis, and the overlap between the two. Different computer vision approaches suit different aspects of marine imagery analysis, such as whole image annotation for habitat classification, and object detection for population counts.

With a lack of large consolidated datasets for marine imagery, compared to their terrestrial counterparts, the application of machine learning to marine imagery has many hurdles to overcome. Over fitting to a specific training dataset, limiting the generalisability to new unseen data, is a major concern. Despite this, we see many examples of automated imagery analysis being used to estimate ecological variables such as biodiversity. Clearly, from the literature presented, there is a great amount of improvement possible for the generalisability of automated benthic imagery analysis, and such an improvement would greatly benefit the benthic ecological community.

Chapter 3

Methodology

In the field of machine learning, where accuracy scores are often the sole focus, and its use in marine imagery where the amount of labelled data to assess such metrics, and the reliability of these labels, comes into question - this thesis aims to provide insight into how best to tackle inter and intra-vehicle learning, and the effects not only on classic machine learning metrics, but on the biomass estimates generated from this method.

The following section presents the experiment design, and later sections go into more depth on the individual methods utilised in these experiments.

3.1 Experiment Design

For the experiments in this thesis, one machine learning architecture was selected and used, Mask R-CNN, as it is implemented as an open source repository for accessibility, has performed well on test datasets such as COCO[1], and as it is an object detection and segmentation system, it allows for the extraction of more useful values such as biomass estimates than other possible approaches like labelling points or whole images. A second machine learning system, DeepLab V3+[46], is described in the following section that was used for preliminary studies, but ultimately not included in this thesis as it is a semantic segmentation system better suited to estimating statistics such as coral percentage coverage, as opposed to identifying distinct individual instances of a class, due to its inability to discern overlapping individuals.

The data augmentation and normalisation techniques selected were chosen for their marine specific applications, for the removal of artefacts commonly found in marine imagery such as vignetting, or for the simulation of marine imagery variation such as marine snow. While investigating each of these techniques independently of each other would be of interest, it is the combination of these techniques that poses the most interest. One hypothesis is that through a combination of normalisation and augmentation techniques, it is possible to achieve greater transferability between vehicles of differing altitudes. For example, some augmentation techniques such as elastic distortion are hypothesised to perform differently for scale normalised images than for scale varying images, as in the normalised images the elastic distortions are all at the same scale. For this reason, these experiments present the results of every single combination of independent variables investigated.

In machine learning experiments, it is common practice to consider the impact of random aspects on the results and account for their variation. In the case of the Mask R-CNN architecture, certain aspects remained unseeded to ensure that the introduced variations are properly addressed. It is important to note that small performance differences observed under specific seeded conditions may not hold consistently across different seeds. To mitigate this, each experiment was repeated seven times, and the analysis focused on the three best-performing runs. This approach ensures a comprehensive evaluation and provides a robust understanding of the model's performance.

3.2 Convolutional Neural Networks

The automated image processing investigated in this thesis uses CNNs, as they constitute the current state of the art in both semantic segmentation and object detection.

3.2.1 DeepLab V3+

Preliminary experiments were carried out and published using DeepLab V3+, analysing the impact of physics based corrections and data augmentation on the segmentation accuracy, presented and published at the Underwater Technology 2019 conference [91]. The DeepLab V3+ architecture is a semantic segmentation architecture, and the study presented investigated estimating coral coverage, a goal well suited to semantic segmentation. The studies presented in this thesis, however, focus on the final goal of estimating biomass for morphotypes with distinct individual instances visible in AUV imagery, so this approach was not continues and will not be discussed further in this thesis.

3.2.2 Mask R-CNN

The results presented in this thesis show the impact of physics based corrections and data augmentation on the object detection and segmentation accuracy of Mask

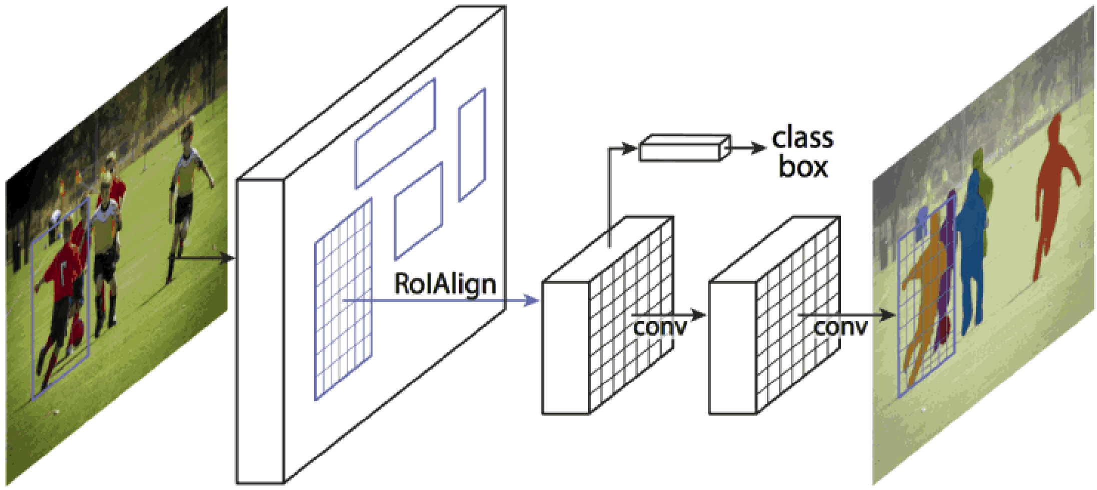


FIGURE 3.1: Mask R-CNN Architecture Summary [3]

R-CNN [3], based on an implementation published on github by Matterport [92], selected due to it being open source and being well supported by the machine learning community. Mask R-CNN is based off of Region-based Convolutional Neural Networks (R-CNN), building on Fast R-CNN [52] and Faster R-CNN [93], as described in more detail in Chapter 2. The network consists of a Region Proposal Network (RPN) followed by a Convolutional Neural Network (CNN) for feature extraction, classification, and segmentation. It achieved a relatively high mAP score, compared to other architectures, of 60% on the COCO dataset when using an IoU threshold of 50%. A further breakdown of results at different IoUs, and the results of different machine learning architectures, were published by He et al. [3]. Figure 3.1 shows a high level summary of the Mask R-CNN architecture.

3.3 Data Augmentation

Data augmentation is a widely employed technique in machine learning to mitigate overfitting and augment the training dataset without requiring additional manual labels. In the experiments conducted for this thesis, a Python data augmentation library called imgaug was utilised [94]. imgaug provides a comprehensive set of tools and functions to apply a diverse range of augmentation transformations to images, such as random rotations, translations, scaling, and mirroring.

Due to the size of the training dataset in these experiments, data augmentation is used in all of the experiments, with the base set of augmentations matching current literature in the area, using flipping, rotating, Gaussian blur, and JPEG compression

[87]. Gaussian blur is the application of a Gaussian convolutional filter across an image, where the value in the centre of the convolutional filter, or kernel, is the highest value, having the highest input on what the new pixel value at that position will be, with the surrounding pixels having a smaller value the further from the centre of the kernel they appear, creating a blurring effect. It is called Gaussian blur due to the Gaussian distribution of values in the convolutional filter.

JPEG compression creates small artefacts in the image due to the loss of information in the compression process. JPEG compression is incredibly common in image processing, and the input data for the keras implementation of Mask R-CNN being investigated is JPEG, so a certain level of JPEG compression is applied to all the training data by default. By varying the amount of compression, and therefore the amount of information lost, the system should be robust when handling unseen data that has been compressed to different levels of information loss.

The effect of extra augmentations are investigated, with the addition of pixel value addition, multiplication, salt and pepper noise, motion blur, and linear contrast transformations.

Pixel value addition adds the same value to every pixel in an image. This can make an image appear brighter with positive numbers, or darker with negative numbers. Pixel value multiplication is similar but applies a multiplication function to every pixel in an image. This makes an image appear brighter with values over 1, and darker with values under 1.

Salt and pepper noise is the addition of white and black pixel patches over an image. The size, or granularity, of these patches, and their density both greatly affect the appearance of this augmentation, and are controlled by parameters to the function. With dense small patches of only white patches, the effect looks similar to snow, and with larger less dense patches it can simulate an object in the image being obscured or at the edge of an image.

Motion blur, similar to the previously defined Gaussian blur, is the application of a convolutional filter, however the distribution is not a normal Gaussian curve, but a directional distribution. The resulting image has the appearance of being caught in motion. This simulates the natural motion blur that occurs when a vehicle is moving too fast for the shutter speed of the camera, or if an object pictured is moving at a high speed relative to the camera.

Linear contrast transformations can either increase or decrease the contrast of an image, by either increasing the differences in pixel values or decreasing them. This effect is very similar to that of multiplication.

A further final set of augmentations is investigated that have been grouped as elastic augmentations, including elastic transformations, piece-wise affine transformations, and perspective transformations. All of these augmentation types are hypothesised to simulate varying terrain and vehicle roll and pitch, with elastic and piece-wise affine transformations simulating bumps and dips in the terrain, and perspective transformations simulating roll and pitch by viewing the scene from a slightly different angle.

Elastic transformations move pixels around the image using displacement fields. Two parameters, alpha and sigma, control the strength and the smoothness of the displacement respectively. The visual effect of this augmentation can vary from a water like effect to a noisy pixellated effect depending on the parameters used.

Piece-wise affine transformations cause local distortions in a grid pattern, applying randomly selected affine transformations to each section of the grid.

Perspective transformations are applied to the entirety of the image and are designed to simulate a change in perspective, such as pitch and roll of a vehicle. This effect is achieved by selecting 4 random points in a defined area near each corner, and transforming the image such that they are the four corners of the newly augmented image.

In the experiments presented in this thesis, the augmentations used by default in all experiments are flipping, rotating, Gaussian blur, and JPEG compression. When referring to extra augmentations, this is referring to the group of addition, multiplication, motion blur, linear contrast, and salt and pepper noise. When referring to elastic augmentations, this is referring to the group of elastic transformations, piece-wise transformations, and perspective transformations. The parameters used for each of these augmentations are listed in Table 3.1.

TABLE 3.1: Parameters for data augmentation methods. Unless specifically stated, 0.1 is used as the frequency parameter for all augmentation methods.

Augmentation	**Parameters**
Flipping	Horizontal & Vertical Frequency 0.5
Rotation	0.25 Frequency for Each 90° Angle
Gaussian Blur	Sigma Values 0.01 to 0.7
JPEG	Values 1 to 5
Add	Values -10 to 10
Multiply	Values 0.75 to 1.25
Motion Blur	K Values 3 to 10
Linear Contrast	Alpha Values 0.5 to 1.5
Salt & Pepper	p Value 0.05
Elastic Transformations	Sigma Values 4 to 6, Alpha Values 0 to 7
Piece-wise Transformations	Scale Values 0 to 0.05
Perspective Transformations	Scale Values 0 to 0.02

3.4 Data Normalisation

Where data augmentation introduces artificial variance to the training data, data normalisation takes the opposite approach of removing unwanted variance. It is known that environmental factors such as water turbidity and distance between the camera and the object of importance, in this case the sea floor, have a large impact on the visual appearance of underwater imagery, and this section covers the various methods to normalise this variance that were investigated in this thesis.

3.4.1 Colour Normalisation

Two methods of colour variance normalisation are investigated here, basic mean and standard deviation corrections per colour channel, carried out in the first stage of the machine learning algorithm by default, and a pixel-wise statistics correction method described below.

Pixel-wise statistics correction uses a greyworld assumption often used in traditional image correction, and the version used in this study was developed by Bryson et al. [88]. This assumption is that the average pixel in a well balanced image should be a neutral grey colour. This is ordinarily applied on an image by image basis, shifting the distribution of pixel values to centre on the specified mean value and standard deviation. The method investigated here, on the other hand, uses this grey world assumption across the entire dataset based on pixel positions, for example the mean of all pixel values at coordinate [0, 0] should be a neutral grey and the values should be shifted to achieve this, likewise with all pixel values at coordinate [0, 1], and [0, 2] and so on, for all pixel coordinates in the images.

This method doesn't require any metadata be attached to the images, and has the effect of removing vignetting from the images, the effect of brighter pixels in the centre of images and darker pixels in the corners and around the edges. Vignetting is common in deep sea imagery due to the artificial lighting, and removing this effect removes an aspect of variance to the visual appearance of fauna captured in the centre of images and in the edges of images.

This approach doesn't take into account variance in altitude of the vehicle, so variance introduced by this is not corrected for with this method.

It can be mathematically defined as;

$$I_y(u,v,\lambda) = m(u,v,\lambda) * I_x(u,v,\lambda) + n(u,v,\lambda) \tag{3.1}$$

$$m(u,v,\lambda) = \sqrt{\frac{\sigma_y^2}{\sigma_x^2(u,v,\lambda)}} \tag{3.2}$$

$$n(u,v,\lambda) = \mu_y - m(u,v,\lambda)\mu_x(u,v,\lambda) \tag{3.3}$$

where μ_y and σ_y^2 are the target mean and variance, and $\mu_x(u,v,\lambda)$ and $\sigma_x^2(u,v,\lambda)$ are the mean and variance of pixel intensities across the images at pixel location (u,v) for the channel λ. $I_x(u,v,\lambda)$ is the input intensity at pixel location (u,v) for the channel λ, and $I_y(u,v,\lambda)$ is the resulting normalised intensity at pixel location (u,v) for the channel λ. $m(u,v,\lambda)$ is the calculated scale factor, and $n(u,v,\lambda)$ is the calculated offset, to get from $I_x(u,v,\lambda)$ to $I_y(u,v,\lambda)$.

Although not investigated in this study, a method for further investigation is altitude based correction. It builds on the previously described pixel-wise statistics method, and was also developed by Bryson et al[88], taking into account altitude information attached to the images. This method corrects for altitude variance and water turbidity by fitting the given pixel values to an attenuation over distance function, approximating the light attenuation through the water for the given dataset.

It is similarly mathematically defined as;

$$I_y(u,v,\lambda,d) = m(u,v,\lambda,d) * I_x(u,v,\lambda,d) + n(u,v,\lambda,d) \tag{3.4}$$

$$m(u,v,\lambda,d) = \sqrt{\frac{\sigma_y^2}{\sigma_x^2(u,v,\lambda,d)}} \tag{3.5}$$

$$n(u,v,\lambda,d) = \mu_y - m(u,v,\lambda,d)\mu_x(u,v,\lambda,d) \tag{3.6}$$

where the additional parameter d is the distance from the camera to the seafloor, often recorded by an altimeter on board the AUV. It is possible to assign distance bins, for parameter d, to group images for this calculation.

In cases where the depth variable is known not only for each image, but also for each pixel in each image for example in 3D reconstructions, each pixel or region of pixels may be corrected for their depth independent of the rest of the image, demonstrated by Bodenmann et al. [95].

3.4.2 Scale Normalisation

A further effect of changing altitudes is scale variance, with images taken at different distances off the sea floor. By using altitude information in conjunction with camera opening angle information, images can be normalised for scale.

The spatial scale of the image, assuming no pitch or roll, a flat surface, and an accurate altitude measurement, can be calculated by

$$s_i = 2a(tan(\theta/2)) \tag{3.7}$$

where s_i is the spatial scale of the image, a is the altitude, and θ is the opening angle of the camera. This is then used to calculate the spatial scale of each pixel, and the rescaling factor to reach a target pixel spatial scale with

$$s_p = s_i/p \tag{3.8}$$

$$r = s_p/s_t \tag{3.9}$$

where s_p is the estimated spatial scale per pixel, p is the size of the image in pixels, s_t is the target pixel spatial scale, and r is the rescaling factor.

The practical implications of this method are shown in Figure 3.2, with the low altitude data on the left, and high altitude data on the right. It shows not only how the spatial scale is normalised for through up or down scaling, but also how the low altitude data is adjusted to match the high altitude optical resolution.

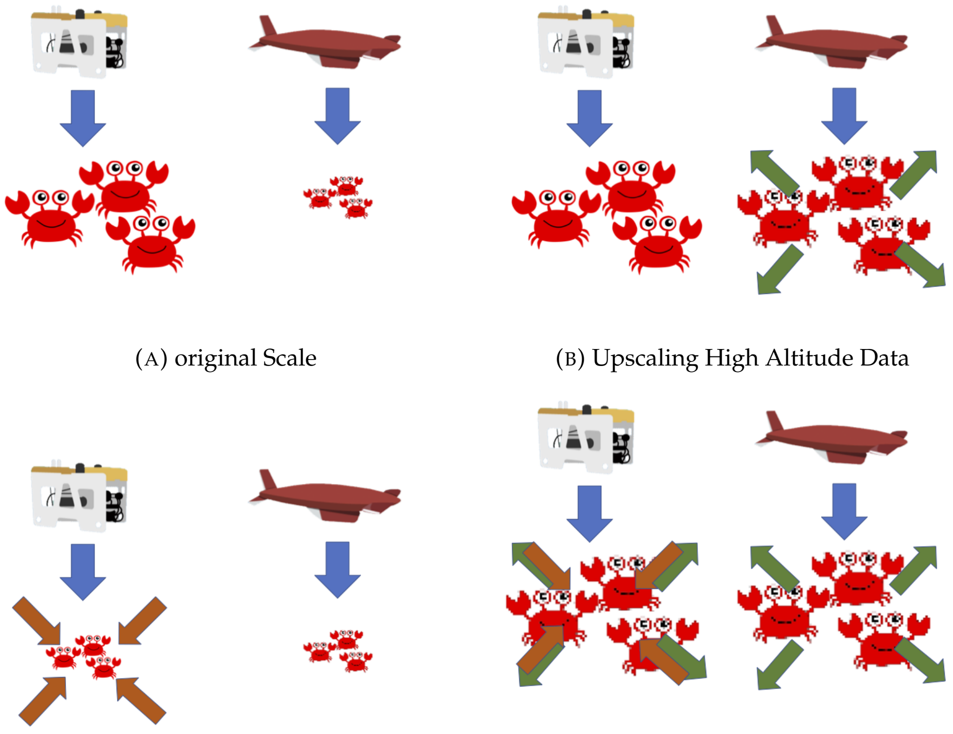
3.4.3 Lens Distortion Normalisation

Differing camera lenses and pressure windows result in differing levels of lens distortion, which when corrected for not only normalises the appearance of objects in centre and at the edges of images, but also normalises the appearance of images taken by cameras with different lens and pressure window configurations.

The python library OpenCV includes a function for correcting lens distortion in images. It corrects for radial distortion and tangential distortion. Radial distortion causes straight lines to appear curved in an image, an effect that becomes more pronounced further from the centre of the image, and is defined as follows, where r represents the radial distance from the centre of the image, and k_1, k_2, and k_3 are the radial distortion coefficients, calculated during camera calibration:

$$x_{distorted} = x(1 + k_1 r_2 + k_2 r_4 + k_3 r_6) \tag{3.10}$$

$$y_{distorted} = y(1 + k_1 r_2 + k_2 r_4 + k_3 r_6) \tag{3.11}$$



(A) original Scale

(B) Upscaling High Altitude Data

(C) Downscaling Low Altitude Data

(D) Upscaling High Altitude Data, and Matching Low Altitude Resolution

FIGURE 3.2: Demonstration of rescaling methods with low altitude data on the left, high altitude data on the right

Tangential distortion causes some areas in an image to appear closer than others, and is defined as follows, where p_1 and p_2 are the tangential distortion coefficients, also calculated during camera calibration:

$$x_{distorted} = x + (2p_1xy + p_2(r_2 + 2x^2)) \tag{3.12}$$

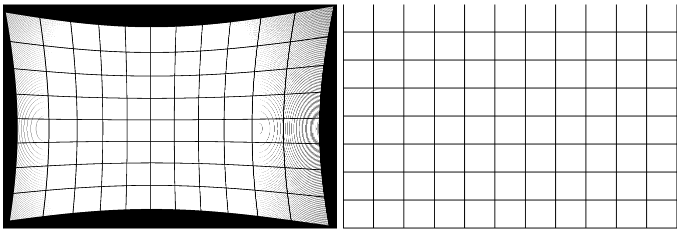
$$y_{distorted} = y + (p_1(r_2 + 2y^2) + 2p_2xy) \tag{3.13}$$

To correct for these two forms of image distortion, we requires the distortion coefficients and a camera matrix, defined as follows:

$$d = (k_1, k_2, p_1, p_2, k_3) \tag{3.14}$$

$$c = \begin{bmatrix} f_x & 0 & 0 \\ 0 & f_y & 0 \\ c_x & c_y & 1 \end{bmatrix} \tag{3.15}$$

where d are the distortion coefficients, and c is the camera matrix containing (f_x, f_y) as the focal length, and (c_x, c_y) as the optical centre. All of these parameters are



(A) With lens distortion

(B) Corrected for lens distortion

FIGURE 3.3: Lens Distortion of AE2000f Images shown with a simple grid

calculated at the camera calibration stage, carried out before studies in the field using a chequerboard pattern.

Figure 3.3 shows an example of a simple grid under the effect of lens distortion, and after correcting for the distortion.

3.5 Metrics

The metrics used to measure the performance of an algorithm vary in what they prioritise and how they measure success, so have a large impact on capturing the outcome of training a machine learning algorithm. For all the experiments in this thesis the mAP score is calculated using an IoU threshold of 0.5, meaning the intersecting area of a ground truth label and the estimated label must be at least half the total union of those two areas.

The statistical significance of the variance in mAP scores overall is analysed using an ANOVA test, and the impacts of different categories on the results are analysed with Tukey HSD tests. Tukey HSD tests are used to make multiple comparisons between the means of different groups, taking into account the multiple comparisons being made. It is important to note that there are certain assumptions that need to be made for the results of these tests to be valid. The data must be normally distributed and have equal variances between groups. Additionally, the observations must be independent and identically distributed.

3.6 Length Weight Relationship Regression

LWRs are the estimated relationship between the length of a morphotype and its weight, calculated from many measured data points using polynomial regression. LWRs were developed by measuring fish caught either by line fishing or trawling, and therefore rely on the morphotype in question being caught and measured before [96; 97]. Where the calculation of LWRs relies on the individuals being caught and

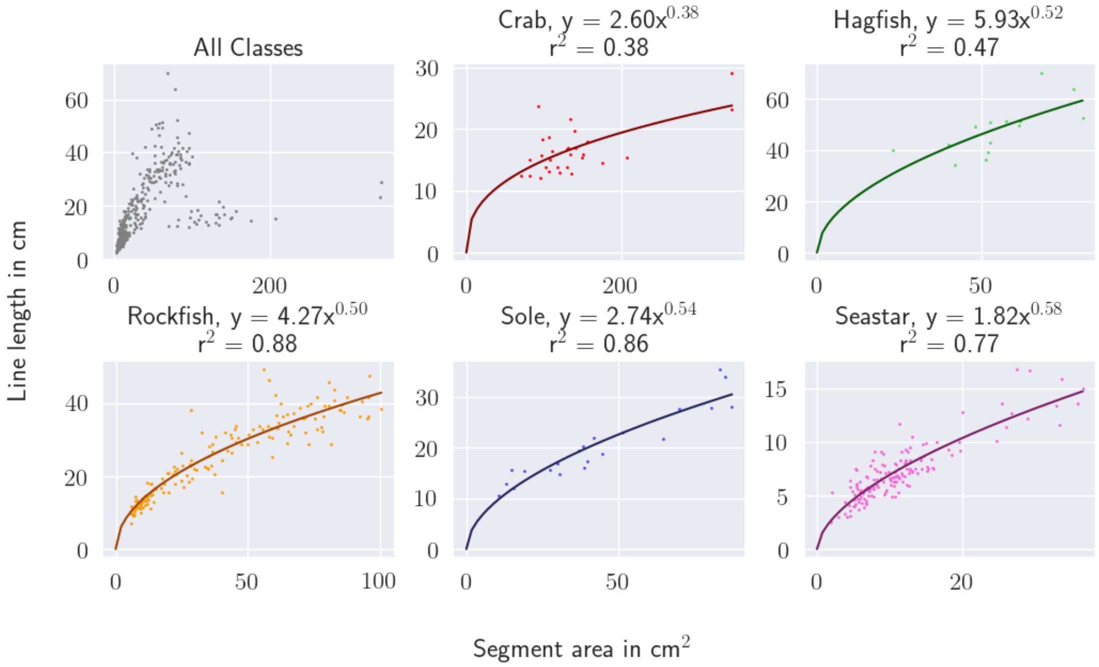


FIGURE 3.4: Estimated segment size to length relationships. The relationships are calculated using regression on measured segment sizes and line lengths, resulting in an estimated polynomial relationship.

measured, LWRs can be applied with non-invasive methods, as the length of an individual can be estimated from imagery. The LWRs used in this study are from the large open source database Fishbase [5].

LWRs have been used with images to estimate biomass before, for example Seiler et al.[56] used length weight relationships and manually annotated AUV images to determine the biomass distribution of the Ocean Perch. It is worth noting, however, that overly bent or curled up fish were excluded from this study. Garcia et al. have used object detection and segmentation to then estimate the length of fish to determine if they are undersized for commercial trawling and should therefore be released [98]. This method involves estimating the internal skeleton position, a complex problem that is specific to the morphotype in question. This method is well suited to the single morphotype problem it was designed for as the internal skeleton differs between different morphotypes, while the method presented in this thesis is more generalisable to multiple morphotypes, better suited for biodiversity estimates, and estimating biomass across multiple morphotypes.

The method developed for this thesis estimates the polynomial relationship between the segment size in cm^2 and the length in cm of the measurement required for the morphotype's specific LWR. The length measurement required differs between morphotypes, for the Rockfish morphotype this length is from the head to the tail, but for the Crab morphotype it is the width of the carapace. In order to estimate this relationship, a set of images were labelled both with length measurements (cm) and

TABLE 3.2: Segment-Size Length Relationship, calculated from collected data for this thesis - cm^2 to cm, Length Weight Relationships, gathered from the large open source database Fishbase [5] - cm to $grams$, and derived Segment-Size Weight Relationships - cm^2 to $grams$

Class Name	SLR	LWR	SWR
Crab	$2.6x^{0.38}$	$0.00036x^{2.92}$	$0.0058x^{1.11}$
Hagfish	$5.93x^{0.52}$	$0.0048x^{2.72}$	$0.61x^{1.41}$
Rockfish	$4.27x^{0.5}$	$0.012x^{3.08}$	$1.08x^{1.54}$
Soles	$2.74x^{0.54}$	$0.0074x^{3.09}$	$0.17x^{1.669}$
Seastar	$1.82x^{0.58}$	$0.00032x^{2.43}$	$0.0014x^{1.41}$

with segments (cm^2). By performing regression on the measured segments and lengths, the segment size to length relationship is calculated, the SLR. Because SLRs allow us to estimate the length based on a segment size, and LWRs allow us to estimate the weight based on a length, a derived segment size to weight relationship (SWR) can be calculated, and the estimation of biomass from segmented images becomes an automated process.

Further development of this method, and relying on such a method for biological measurements and estimates, would require additional morphotypes to be labelled, and for further samples of the given morphotypes to be labelled to increase the sample size and to increase trust in the given relationships.

Figure 3.4 shows the relationships between the length measurements and the segment sizes. There is a stronger relationship between segment size and length measurements for Rockfish, Sole, and Seastars than for Crabs and Hagfish. This may be because of the complex shape of crabs, and due to the burrowing and curling behaviour of Hagfish where their shape and the amount of their body visible above ground vary. Table 3.2 shows these segment-size length relationships (SLRs), the known length weight relationships (LWRs) for each class, and the combined segment-size weight relationship (SWRs) calculated by the combination of these two functions.

Chapter 4

Datasets

This chapter presents the datasets used for the experiments in this thesis, describing their acquisition and characteristics.

4.1 Adaptive Robotics 2018 Expedition

The 2018 Adaptive Robotics Expedition, on the Schmidt Ocean Institute's RV Falkor, went to the South Hydrates Ridge off the coast of Oregon in the North East Pacific Ocean. This expedition, lead by Professor Blair Thornton, was a technical trial into the use of multiple AUVs, one ROV, and on ship measurements in parallel for marine surveying with on ship data processing and analysis.

Figure 4.1 shows the route the RV Falkor took during the expedition. Figure 4.2 shows the full 3D mosaic reconstruction of the surveyed area.

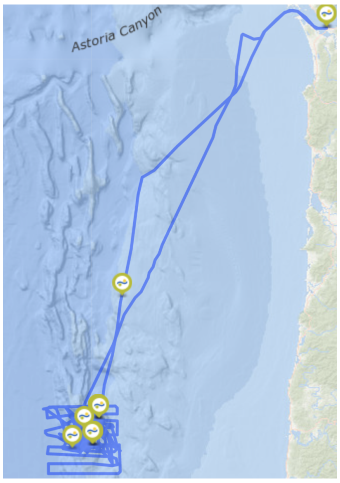


FIGURE 4.1: Map of the Adaptive Robotics 2018 Expedition

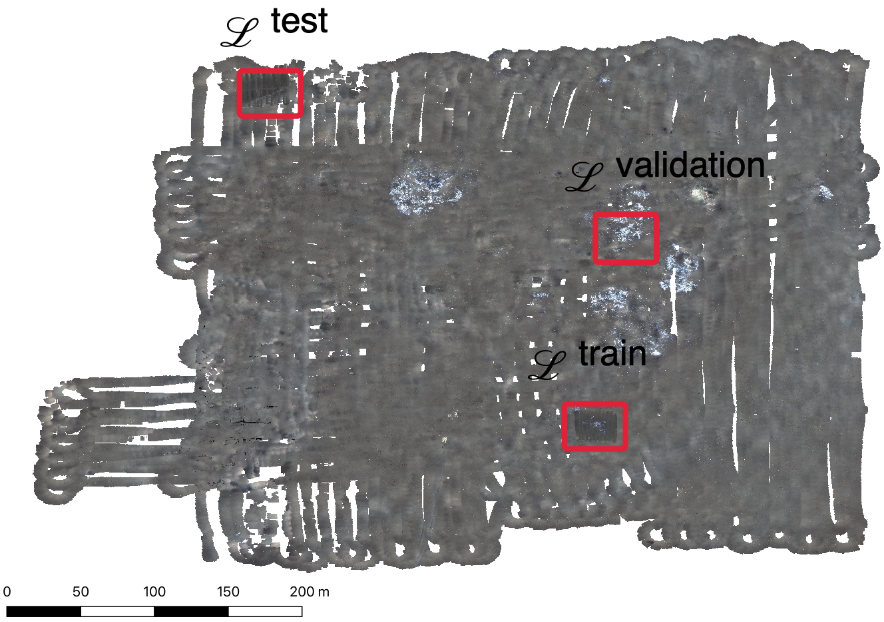
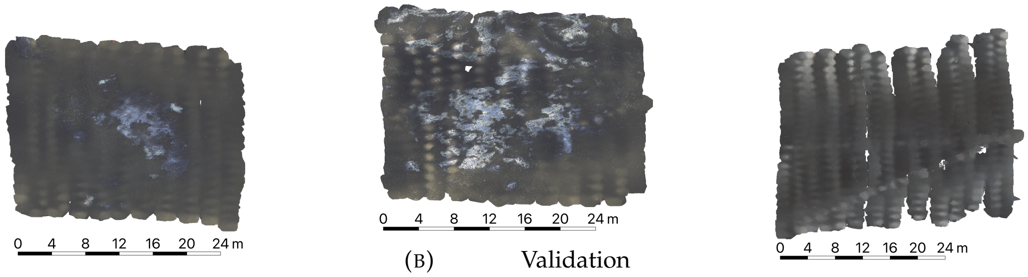


FIGURE 4.2: Mosaic of the South Hydrates Ridge from the Adaptive Robotics 2018 Expedition from the High Altitude Vehicle



(A) Train Dataset

(B) Validation Dataset

(C) Test Dataset

FIGURE 4.3: Mosaics of the South Hydrates Ridge from the Adaptive Robotics 2018 Expedition from the Low Altitude Vehicle

4.1.1 Tunasand

The Tunasand AUV was created by the University of Tokyo in 2007 [17], and is a hover-style AUV with a low speed and high manoeuvrability, resulting in high resolution images from close to the seafloor covering a relatively small spatial area. Table 4.1 shows the specifications for this AUV, and Figure 4.4 shows a photo of the vehicle.

4.1.2 High Altitude, AE2000f

The AE AUV was originally developed as a commercial vehicle, and was more recently adapted for research purposes at the University of Tokyo in 2016 [99; 4]. It is a

TABLE 4.1: Tunasand AUV Dataset Specifications

Dimensions	1.1 x 0.7 x 1.1 m
Weight in air	180 kgf
Vehicle Depth Rating	1500m
Speed	0.4 knots
Target Altitude	2.2 m
Endurance	8 hours
Batteries	Li-ion rechargeable 3.8 kWh
Diving and surfacing methods	2 x drop weights, 5 or 10 kg
Camera Name	Unagi Visual Mapping
Opening Angle	55.96° x 47.82°
Resolution	2464 × 2056
Port	Dome
Camera Depth Rating	6000m
Labelled Images	522
Training Images	207
Validation Images	123
Test Images	192

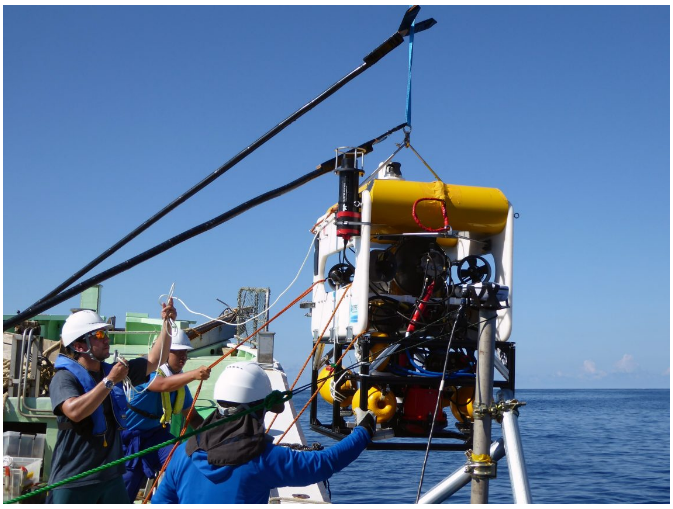


FIGURE 4.4: A photograph of the Tunasand AUV being deployed in the field.

gliding AUV with a higher speed and lower manoeuvrability than Tunasand, resulting in lower resolution images but covering a much larger spatial area in a similar amount of time. Table 4.2 shows the specifications for this AUV, and Figure 4.5 shows a photo of the vehicle.

4.1.3 Classes

Many classes of mega-fauna were identified in the Falkor 2018 expedition dataset, of which a subset of the larger classes present in both the Tunasand and AE2000f images were selected for use in the presented experiments. This subset is presented below,

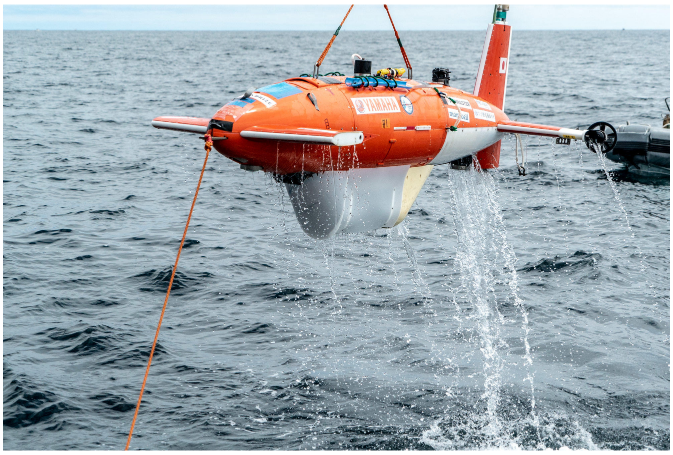
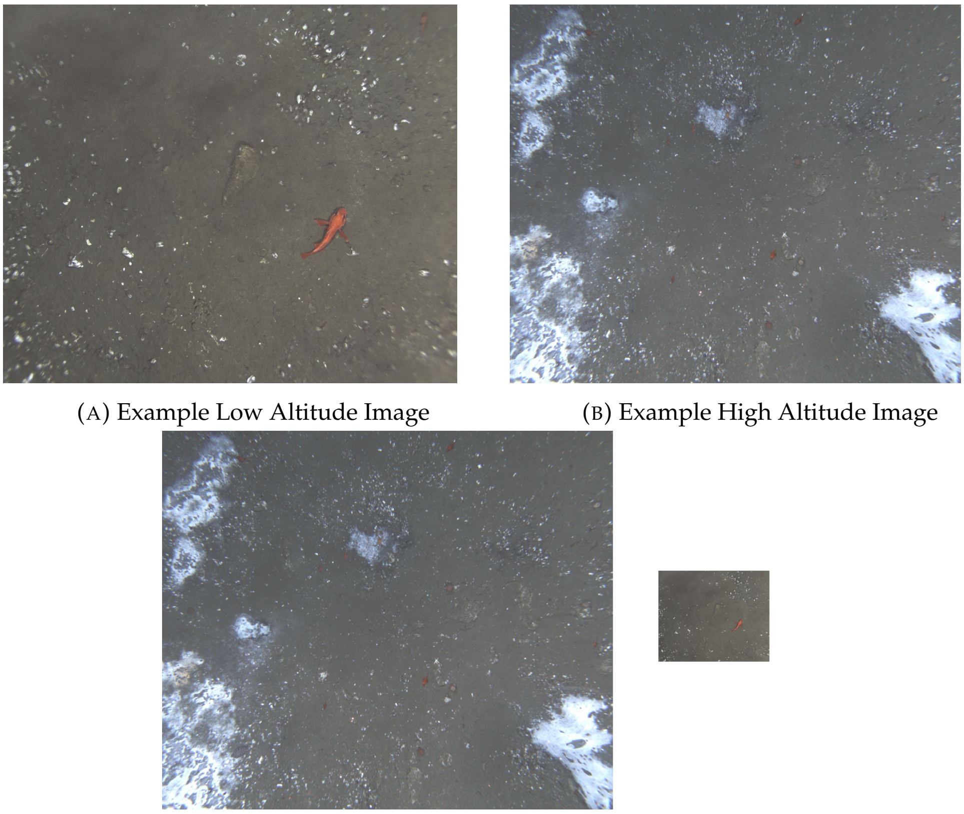


FIGURE 4.5: A photograph of the AE2000f AUV being deployed in the field.



(A) Example Low Altitude Image

(B) Example High Altitude Image

(C) Relative Difference in Spatial Scale

FIGURE 4.6: Example images from each dataset and their relative difference in spatial scale

TABLE 4.2: AE2000 AUV Dataset Specifications

Dimensions	3.0 x 1.3 x 0.9 m
Weight in air	370 kgf
Vehicle Depth Rating	2000m
Speed	1.7 knots
Target Altitude	8 m
Endurance	8 hours
Batteries	Li-ion rechargeable 3.3 kWh
Diving and surfacing methods	2 x drop weights, 5 or 8 kg
Camera Name	SeaXerocks 3 Visual Mapping
Opening Angle	60.41° x 52.12°
Resolution	1280 × 1024
Port	Flat
Camera Depth Rating	3000m
Labelled Images	60
Training Images	20
Validation Images	20
Test Images	20

showing the population counts in each dataset in Figure 4.8, and example images of each class in Figure 4.7. The population counts for these classes are incredibly unbalanced, with very few examples of the Hagfish class, and many examples of the Rockfish class, with varying numbers in-between. While this is unusual for curated machine learning datasets, it is representative of marine imagery data sets where the abundance of different morphotypes varies massively. For the purposes of this study, the class imbalance will not be counteracted in any way, with the most abundant class having the highest impact on the performance scores, and likely being the most accurately identified class. With the aim of making biomass estimates summed across all classes, the most abundant class will have the highest impact on the results, and therefore being prioritised is not an issue. It is advised that when using machine learning for another application that requires all classes to be prioritised equally, for example when inaccuracy on a rarer class has a large adverse effect, that class balancing techniques are used such as augmenting examples of rarer classes to artificially create balance, or weighted impacts on loss scores.

All five of the presented classes were used in training, however for results analysis purposes, Sole and Sea Stars have been excluded. The Sole class contains two quite distinct morphotypes with relatively few examples, causing unreliable performance on this class. Sea Stars, while relatively abundant in these datasets, are close in physical appearance to another class that was excluded from the labelling efforts, the Brittle Star. This class is highly abundant in the area, however it is hard to label reliably in part due to the small size of many of the individuals. In many cases, Brittle Stars are wrongly identified as being Sea Stars, causing unreliable performance on this class.

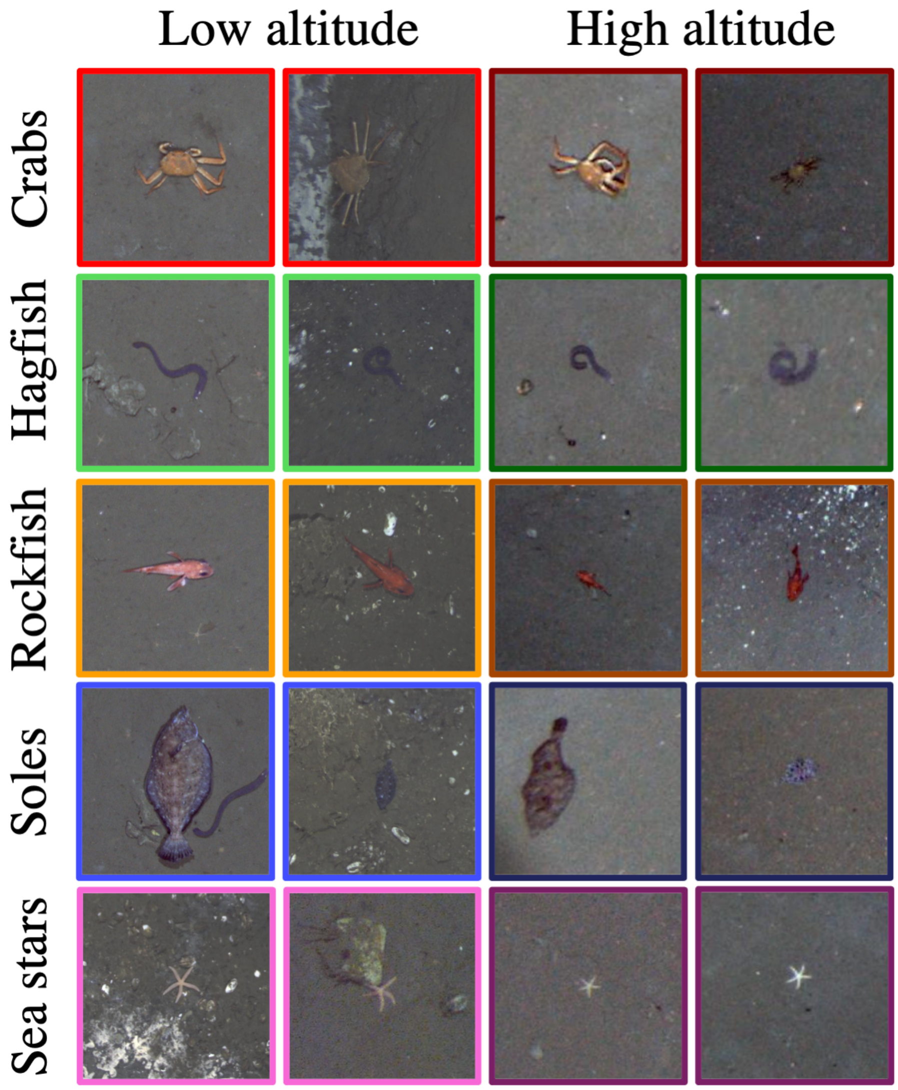


FIGURE 4.7: Examples of each class labelled and used in the training data. The colours used for each class will be used throughout this thesis, and any example labelled images will use these colours to indicate which class the label is for. Labels coloured black are classes which were labelled but which were not included in this thesis.

Some of the classes included in this study are of environmental and commercial importance, with the snow crab being a commercially fished species with concerns around its over-fishing and population sustainability, [100] and Roberts et al. found Rockfish to be a generally commercial fish in cold water coral reefs in the Pacific Ocean [101].

4.1.4 Example Images

In Figure 4.9, both a low altitude image, $\mathcal{L}$, and a high altitude image, $\mathcal{H}$ are shown and a random selection of augmentations are applied to them. These examples firstly show the difference in size of objects appearing in the $\mathcal{L}$ dataset compared to those in

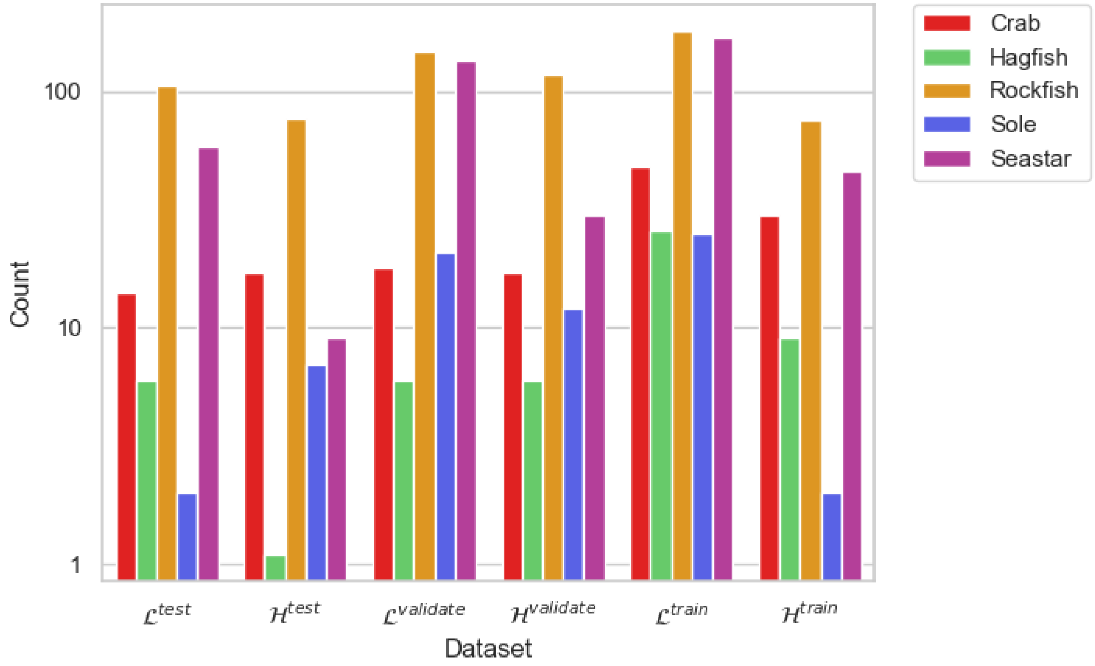
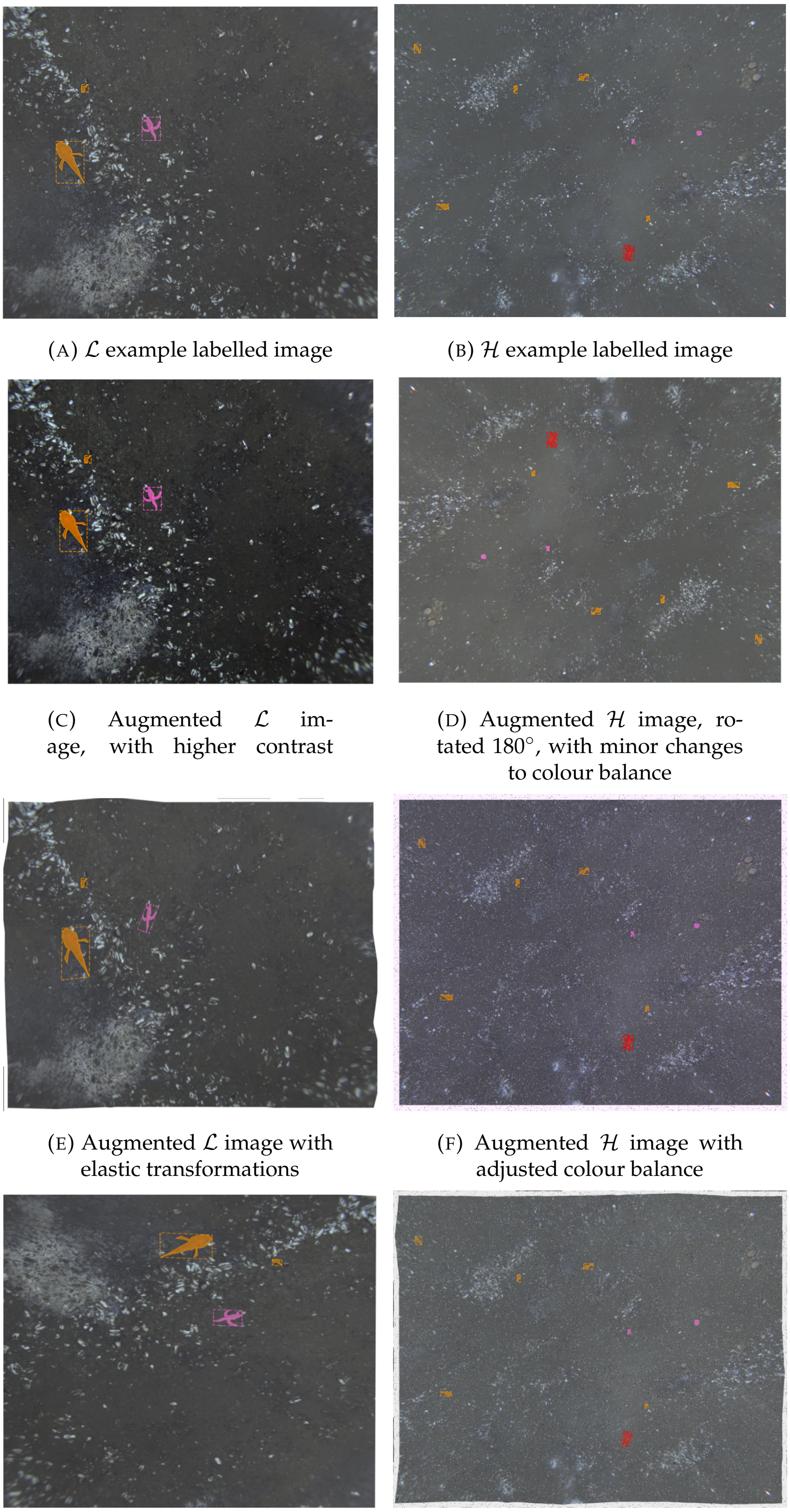


FIGURE 4.8: Class count per dataset in each of the test, validate, and train datasets for both low ($\mathcal{L}$) and high ($\mathcal{H}$) altitude.

the $\mathcal{H}$ dataset, with those in the $\mathcal{L}$ dataset appearing larger and in higher resolution. The number of individuals in each image is relatively representative, with the $\mathcal{H}$ images covering a much larger spatial scale than the $\mathcal{L}$ images, resulting in more individuals present per image. We see the effects of some of the augmentations applied for this study when applied to the full image, it is worth noting that the augmentations are applied to a random cropping of the image in practice to maintain spatial scale when acquiring a 1024x1024 patch for training the neural network. Augmentations that effect the orientation of the image, the colour balance, the contrast, and elastic distortions are all shown in the figure for both low and high altitude datasets, $\mathcal{L}$ and $\mathcal{H}$.

Figure 4.10 shows two example labelled images from the low altitude and high altitude datasets, $\mathcal{L}$ and $\mathcal{H}$, and the effect of distortion correction on these images. Figure 4.11 shows examples zoomed in on individuals in the images show the differences in effect on the appearance of individuals. A much more visible effect is seen on individuals towards the corners of $\mathcal{H}$ images due to the flat port of the camera. There is a much larger lens distortion effect from flat port cameras than from dome port cameras [102].



(A) $\mathcal{L}$ example labelled image

(B) $\mathcal{H}$ example labelled image

(C) Augmented $\mathcal{L}$ image, with higher contrast

(D) Augmented $\mathcal{H}$ image, rotated 180°, with minor changes to colour balance

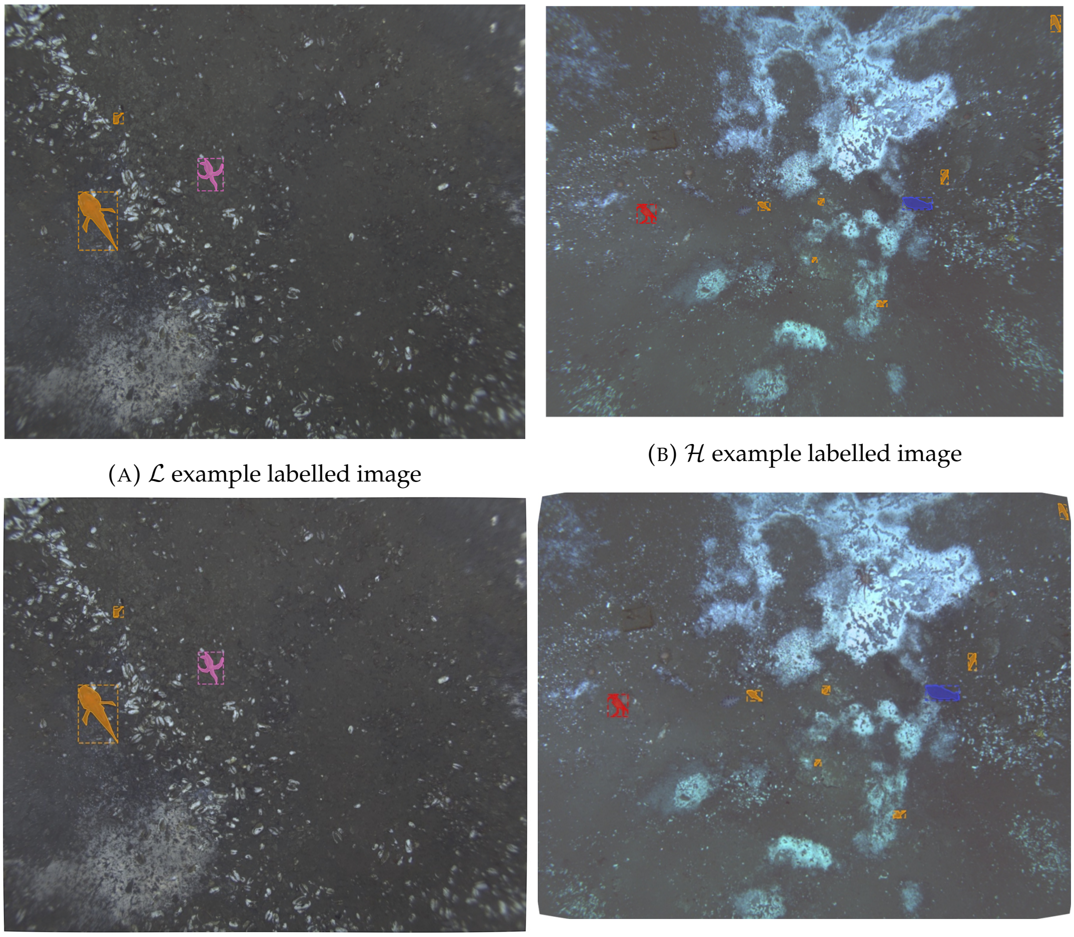
(E) Augmented $\mathcal{L}$ image with elastic transformations

(F) Augmented $\mathcal{H}$ image with adjusted colour balance

(G) Augmented $\mathcal{L}$ image with 90° rotation and stretching

(H) Augmented $\mathcal{H}$ image with elastic transformations

FIGURE 4.9: Examples of labelled images, and those images under different random augmentations



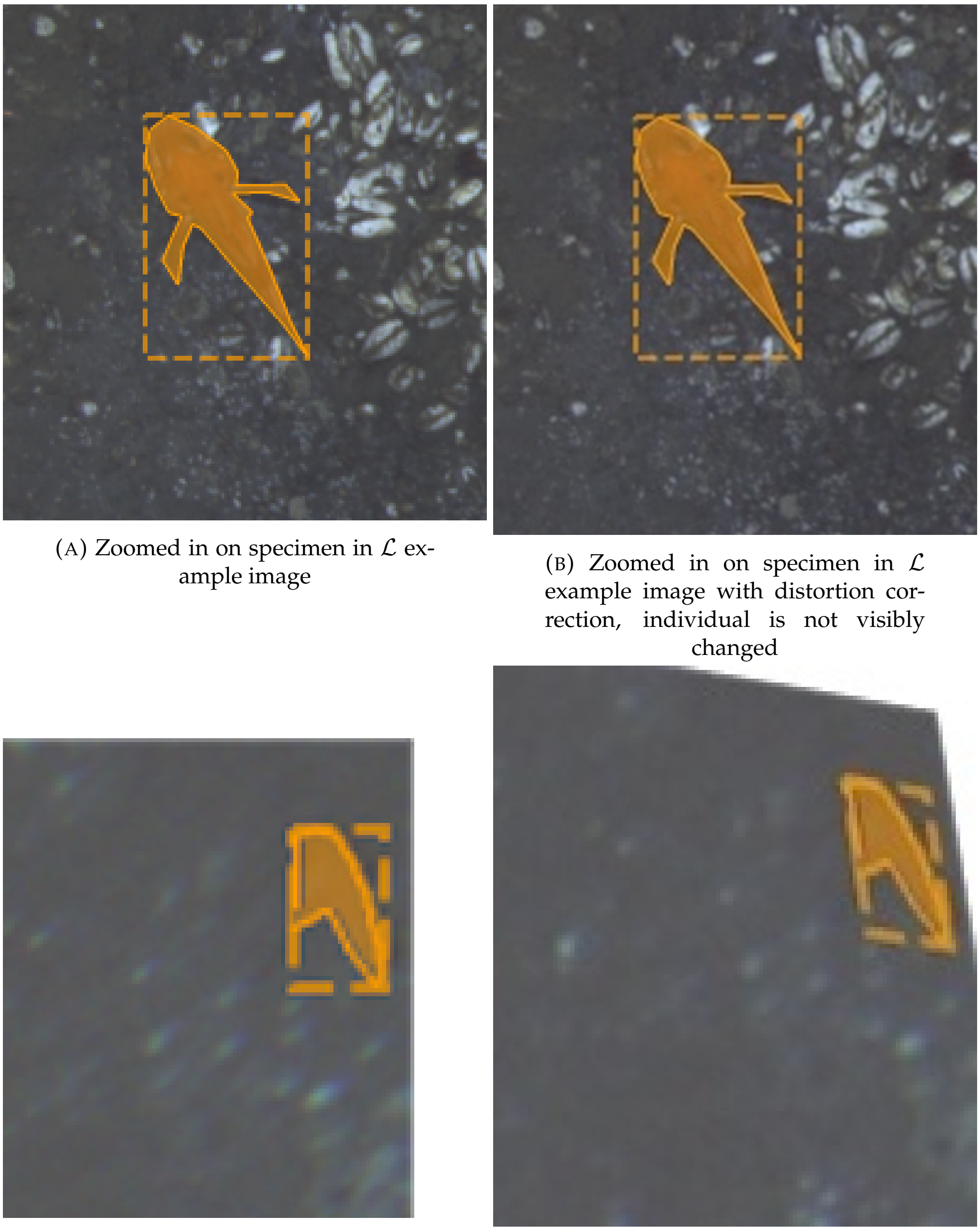
(A) $\mathcal{L}$ example labelled image

(B) $\mathcal{H}$ example labelled image

(C) $\mathcal{L}$ example corrected for distortion, hardly visible due to the dome lens of the camera

(D) $\mathcal{H}$ example corrected for distortion, more visible effect due to the flat lens of the camera

FIGURE 4.10: Example images and their labels from the low altitude dataset, $\mathcal{L}$, and the high altitude dataset, $\mathcal{H}$, and the effect of lens distortion correction on each image. There is minimal difference for the $\mathcal{L}$ image, and a large difference for the $\mathcal{H}$ image due to the dome port of the $\mathcal{L}$ camera and the flat port of the $\mathcal{H}$ camera.



(A) Zoomed in on specimen in $\mathcal{L}$ example image

(B) Zoomed in on specimen in $\mathcal{L}$ example image with distortion correction, individual is not visibly changed

(C) Zoomed in on specimens in $\mathcal{H}$ example image

(D) Zoomed in on specimens in $\mathcal{H}$ example image with distortion correction, causing visible changes in the shape of individuals

FIGURE 4.11: Zoomed in examples of an $\mathcal{L}$ image and a $\mathcal{H}$ image, showing a small difference in appearance for the individual in the $\mathcal{L}$ image and a much more obvious difference in shape for the individual in the $\mathcal{H}$ image.

Chapter 5

Effects of augmentation and normalisation on learning

The results presented in this chapter show the impact of the previously described augmentation and normalisation methods on learning and performance on images taken in different, but nearby, spatial regions by the same vehicle. This is referred to as intra-vehicle transferability within this thesis. These experiments provide an insight into each combination of normalisation and augmentation techniques' impact on performance of the Mask R-CNN neural network architecture. The datasets used are, as described in previous chapters, taken by two AUVs, one at a low altitude and one at a relatively higher altitude. These datasets were collected on the same expedition with the same lighting and camera setups, on different days and with possibly differing environmental conditions.

Experiment notation consists of; the altitude of the dataset, signified by $\mathcal{L}$ for low altitude with a target altitude of 2.2 metres, and $\mathcal{H}$ for high altitude with a target altitude of 8 metres; whether the dataset is the train or test dataset, signified in superscript as *train*, and *test*; and the applied normalisation and augmentation techniques, signified in subscript using the codes described in Table 5.1, such as $C1.S3.E1$ for no colour correction, normalising to high altitude scale, and using elastic augmentations. For example $\mathcal{L}^{train}_{C2.S2.E0} \rightarrow \mathcal{H}^{test}_{C2.S2.E0}$ describes the experiments that were trained on low altitude data with greyworld colour correction applied, normalised to low altitude scale, and without elastic augmentations, and tested on high altitude data under the same conditions. For the independent variables not included, this indicates that the experiments have not been filtered by this variable and that all versions of this variable are included in these results.

The results are split into two sections, one for the low altitude experiments, ie. $\mathcal{L}^{train} \rightarrow \mathcal{L}^{test}$, and one for the high altitude experiments, ie. $\mathcal{H}^{train} \rightarrow \mathcal{H}^{test}$. The results for each section are shown as mAP boxplots broken down by independent variable, and a table of average mAP scores and standard deviations per combination of independent variable, to provide the case by case statistics.

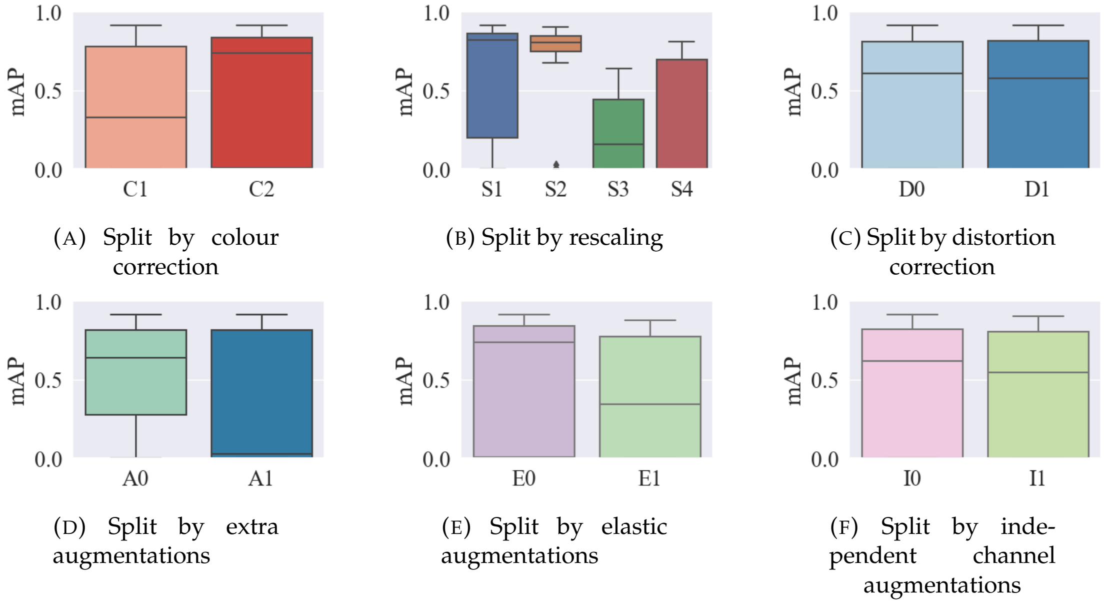
TABLE 5.1: Independent Variable Codes

	C - Colour Correction
C1	Raw images
C2	Grey world correction
	S - Scale
S1	Original scale
S2	Low altitude spatial scale
S3	High altitude spatial scale
S4	Low altitude scale with high altitude resolution
	D - Distortion Correction
D0	Without Distortion Correction
D1	With Distortion Correction
	E - Elastic Distortion Type Augmentations
E0	Without Elastic Distortion Type Augmentations
E1	With Elastic Distortion Type Augmentations
	I - Individual Channel Augmentations
I0	Without Individual Channel Augmentations
I1	With Individual Channel Augmentations
	A - Extra Augmentations
A0	Without Extra Augmentations
A1	With Extra Augmentations

5.1 Low Altitude

This section presents the results for the low altitude dataset, taken by the Tunasand AUV on the Falkor 2018 expedition. This covers every combination of independent variables for all experiments of the form $\mathcal{L}^{train} \rightarrow \mathcal{L}^{test}$. The analysis starts by looking at a breakdown of the mAP scores by independent variable, filtering those results down to the higher performing experiments, and then showing some examples of estimated segmentation masks alongside their manually annotated counterparts. A table of results for every combination of independent variables is also presented, allowing for deeper analysis of the variables interactions with one another.

Figure 5.1 shows a breakdown of the mAP scores by independent variables. As shown in Figure 5.1a, there is an improvement in performance between $\mathcal{L}_{C1}^{train} \rightarrow \mathcal{L}_{C1}^{test}$ and $\mathcal{L}_{C2}^{train} \rightarrow \mathcal{L}_{C2}^{test}$, improving from an average of 42.57% to 74.14%. This is most likely because effects such as vignetting caused by the artificial lighting in deep sea imagery are only corrected for when using C2, pixel statistics correction. Correcting for these vignetting affects provides a significant improvement in performance for this set up, the significance of which is analysed in further detail later, and implies that the variation caused by these effects is not trivially overcome by this neural network with limited training examples. It is worth noting that the Mask R-CNN architecture applies histogram stretching to the input images, so simple colour imbalance does not account for these differences in performance.



(A) Split by colour correction

(B) Split by rescaling

(C) Split by distortion correction

(D) Split by extra augmentations

(E) Split by elastic augmentations

(F) Split by independent channel augmentations

FIGURE 5.1: mAP scores for low altitude vehicle, $\mathcal{L}^{train} \rightarrow \mathcal{L}^{test}$, split by each independent variable, and aggregating over all other independent variables.

The statistical significance of these results can begin to be analysed in Table 5.3, showing the ANOVA test results for each category independently to one another. The F-value is a measure of how much the variation in the mAP can be explained by the variation in the independent variables. When the F-value is higher, it indicates that the independent variables have a greater impact on the outcome variable, the mAP. The p-value is a measure of the statistical significance of the F-value indicating whether the observed F-value is likely to have occurred by chance. A small p-value (usually less than 0.05) suggests that there is a significant relationship between the predictor variables and the outcome variable. Conversely, a large p-value suggests that there is no significant relationship. From this table we see the most significant impact comes from the rescaling method, followed by the elastic transformations and the colour correction method. Other categories have a much lower, statistically insignificant impact.

Looking at the relationships between independent variables, Table 5.4 shows the ANOVA test results for each combination of independent variables. The combination of Colour Correction and Scale, both statistically significant individually, is stronger than each variable alone, with an F-value of 29.253 and a p-value of $3.590e^{-32}$. Another combination stronger than each variable alone is Scale and Extra Augmentations, with an F-value of 31.898 and a p-value of $1.127e^{-34}$.

TABLE 5.2: mAP scores and variance for low altitude vehicle, $\mathcal{L}^{train} \rightarrow \mathcal{L}^{test}$

				F0	F0	F1	F1
				I0	I1	I0	I1
C1	S1	D0	E0	**0.87 ± 0.03**	0.85 ± 0.04	**0.88 ± 0.01**	**0.87 ± 0.01**
C1	S1	D0	E1	0.53 ± 0.38	0.80 ± 0.04	0.52 ± 0.37	0.00 ± 0.00
C1	S1	D1	E0	0.59 ± 0.39	0.55 ± 0.39	0.28 ± 0.40	0.56 ± 0.39
C1	S1	D1	E1	0.55 ± 0.39	0.00 ± 0.00	0.25 ± 0.36	0.00 ± 0.00
C1	S2	D0	E0	0.78 ± 0.01	0.27 ± 0.39	0.83 ± 0.03	0.84 ± 0.02
C1	S2	D0	E1	0.78 ± 0.04	0.71 ± 0.02	0.50 ± 0.35	0.52 ± 0.35
C1	S2	D1	E0	0.83 ± 0.01	0.53 ± 0.37	0.28 ± 0.40	0.82 ± 0.03
C1	S2	D1	E1	0.77 ± 0.03	0.50 ± 0.35	0.52 ± 0.37	0.54 ± 0.38
C1	S3	D0	E0	0.43 ± 0.05	0.11 ± 0.13	0.07 ± 0.10	0.00 ± 0.00
C1	S3	D0	E1	0.36 ± 0.04	0.28 ± 0.07	0.06 ± 0.07	0.00 ± 0.00
C1	S3	D1	E0	0.36 ± 0.09	0.43 ± 0.02	0.00 ± 0.00	0.00 ± 0.00
C1	S3	D1	E1	0.40 ± 0.03	0.05 ± 0.08	0.00 ± 0.00	0.00 ± 0.00
C1	S4	D0	E0	0.00 ± 0.00	0.22 ± 0.31	0.48 ± 0.34	0.00 ± 0.00
C1	S4	D0	E1	0.30 ± 0.23	0.11 ± 0.16	0.00 ± 0.00	0.00 ± 0.00
C1	S4	D1	E0	0.56 ± 0.18	0.24 ± 0.34	0.00 ± 0.00	0.70 ± 0.03
C1	S4	D1	E1	0.32 ± 0.23	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C2	S1	D0	E0	0.86 ± 0.03	0.54 ± 0.38	**0.88 ± 0.03**	**0.88 ± 0.02**
C2	S1	D0	E1	0.82 ± 0.02	0.84 ± 0.04	0.25 ± 0.36	0.62 ± 0.26
C2	S1	D1	E0	0.86 ± 0.02	0.82 ± 0.02	0.86 ± 0.01	**0.89 ± 0.01**
C2	S1	D1	E1	0.85 ± 0.02	0.80 ± 0.08	0.26 ± 0.37	0.78 ± 0.06
C2	S2	D0	E0	0.84 ± 0.02	0.80 ± 0.01	0.84 ± 0.02	**0.88 ± 0.02**
C2	S2	D0	E1	0.82 ± 0.02	0.80 ± 0.05	0.56 ± 0.39	0.54 ± 0.39
C2	S2	D1	E0	0.85 ± 0.02	0.55 ± 0.39	0.86 ± 0.00	0.85 ± 0.02
C2	S2	D1	E1	0.83 ± 0.03	0.81 ± 0.02	0.29 ± 0.41	0.81 ± 0.03
C2	S3	D0	E0	0.59 ± 0.03	0.55 ± 0.04	0.00 ± 0.00	0.01 ± 0.01
C2	S3	D0	E1	0.47 ± 0.03	0.53 ± 0.03	0.09 ± 0.13	0.00 ± 0.00
C2	S3	D1	E0	0.61 ± 0.02	0.51 ± 0.07	0.16 ± 0.12	0.00 ± 0.01
C2	S3	D1	E1	0.49 ± 0.06	0.37 ± 0.17	0.00 ± 0.00	0.00 ± 0.00
C2	S4	D0	E0	0.47 ± 0.33	0.01 ± 0.00	0.69 ± 0.08	0.76 ± 0.04
C2	S4	D0	E1	0.65 ± 0.04	0.00 ± 0.00	0.00 ± 0.00	0.25 ± 0.35
C2	S4	D1	E0	0.73 ± 0.04	0.47 ± 0.34	0.78 ± 0.01	0.77 ± 0.02
C2	S4	D1	E1	0.00 ± 0.00	0.41 ± 0.29	0.22 ± 0.30	0.27 ± 0.38

The only two variables that didn't have a statistically significant impact on results individually were Distortion Correction and Individual Channel Augmentations, and in combination with each other there was no clear improvement on this.

Splitting the results by rescaling method, as shown in Figure 5.1b, there is a clear increase in performance comparing $\mathcal{L}^{train}_{S1} \rightarrow \mathcal{L}^{test}_{S1}$ and $\mathcal{L}^{train}_{S2} \rightarrow \mathcal{L}^{test}_{S2}$. Looking across all the subplots in Figure 5.1, this is clearly the most impactful variable by far, and $S2$ provides the highest performance boost. Based on these findings, the recommendation this thesis presents to further marine imaging and machine learning endeavours, is to normalise by scale to allow for the most efficient use of limited training data.

TABLE 5.3: ANOVA test results for each category, $\mathcal{L}^{train} \rightarrow \mathcal{L}^{test}$

Category	F-value	p-value
Colour Correction	21.392	$5.131e^{-06}$
Distortion Correction	0.322	0.571
Scale	52.827	$1.467e^{-28}$
Elastic Transformations	19.703	$1.185e^{-05}$
Individual Channel Augmentations	1.452	0.229
Extra Augmentations	14.352	0.000176

TABLE 5.4: ANOVA test results for each combination of category, $\mathcal{L}^{train} \rightarrow \mathcal{L}^{test}$

Combined Categories	F-value	p-value
Colour Correction and Distortion Correction	7.878	$4.119e^{-05}$
Colour Correction and Scale	29.253	$3.590e^{-32}$
Colour Correction and Elastic Transformations	14.450	$6.219e^{-09}$
Colour Correction and Individual Channel Operations	7.942	$3.779e^{-05}$
Colour Correction and Extra Augmentations	12.412	$9.238e^{-08}$
Distortion Correction and Scale	23.795	$9.176e^{-27}$
Distortion Correction and Elastic Transformations	6.855	$1.654e^{-04}$
Distortion Correction and Individual Channel Operations	0.891	0.446
Distortion Correction and Extra Augmentations	4.905	$2.344e^{-03}$
Scale and Elastic Transformations	30.459	$2.538e^{-33}$
Scale and Individual Channel Operations	22.929	$7.122e^{-26}$
Scale and Extra Augmentations	31.898	$1.127e^{-34}$
Elastic Transformations and Individual Channel Operations	7.099	$1.187e^{-04}$
Elastic Transformations and Extra Augmentations	15.255	$2.159e^{-09}$
Individual Channel Operations and Extra Augmentations	7.865	$4.195e^{-05}$

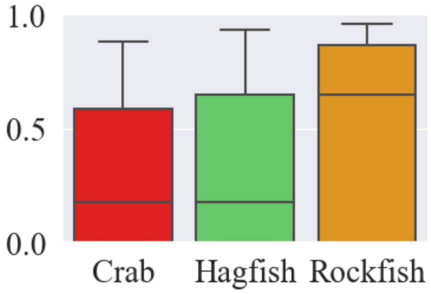
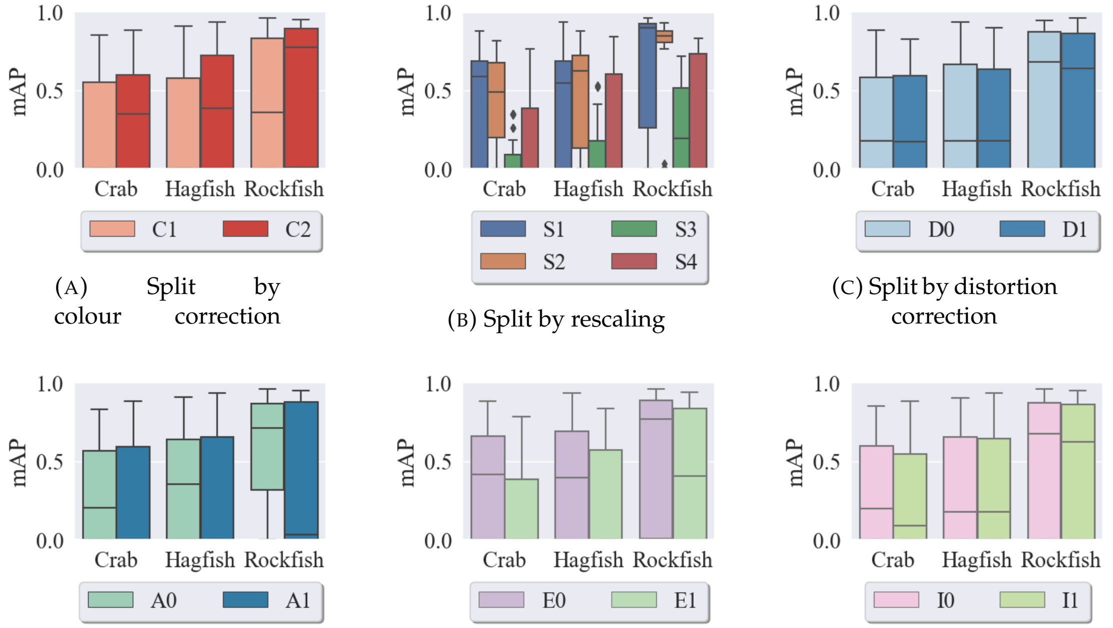


FIGURE 5.2: mAP scores for low altitude vehicle, $\mathcal{L}^{train} \rightarrow \mathcal{L}^{test}$, summary per class, aggregating over all experiments

In comparison, the performances of both $\mathcal{L}^{train}_{S3} \rightarrow \mathcal{L}^{test}_{S3}$ and $\mathcal{L}^{train}_{S4} \rightarrow \mathcal{L}^{test}_{S4}$ are very poor. This is understandable due to the reduction of information when downscaling images in this way, with less pixels to hold visual information. What is interesting, however, is the improved performance of $\mathcal{L}^{train}_{S4} \rightarrow \mathcal{L}^{test}_{S4}$ over $\mathcal{L}^{train}_{S3} \rightarrow \mathcal{L}^{test}_{S3}$ where the amount of information available before processing the images is the same, with the only difference being the spatial scale of the images mimicking that of the average low altitude image. This shows that the spatial scale of the images, and the objects within



(A) Split by colour correction

(B) Split by rescaling

(C) Split by distortion correction

(D) Split by extra augmentations

(E) Split by elastic augmentations

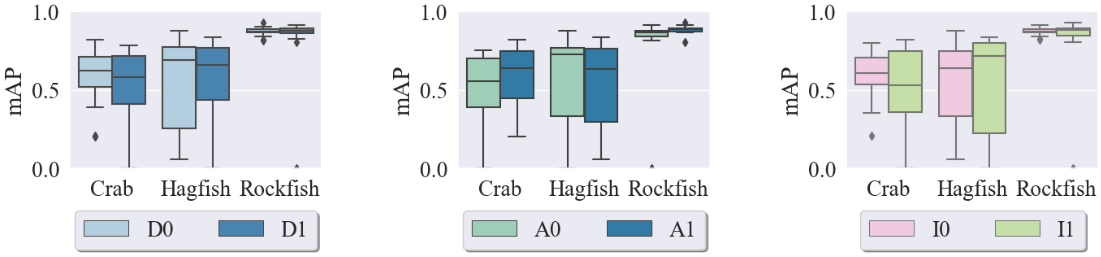
(F) Split by independent channel augmentations

FIGURE 5.3: mAP scores for each independent variable, for low altitude vehicle, $\mathcal{L}^{train} \rightarrow \mathcal{L}^{test}$, split by class and by each independent variable, and aggregating over all other independent variables.

them, have a large effect on the performance of Mask R-CNN.

Factors that impact the ability of Mask R-CNN to detect and segment objects of varying sizes includes the region proposal network anchor size parameter, which has not been investigated as part of this study, and are set as a default to the values $(32, 64, 128, 256, 512)$, giving a large enough range in size to theoretically overcome the smaller size of objects in $S3$. As it currently stands, there are few recommendations as to the preferred spatial scale of images for an object detection and segmentation system, and these findings imply that this is a severe shortcoming that may be impacting the performance of applied machine learning, especially in the marine environment. In this case, the smaller spatial scale, with objects appearing larger, performed better, however there is no evidence to conclude what spatial scale is optimal for this set up, or how that information can be transferred to other datasets or other analysis methods. The recommendation of this thesis is for further research into the effect of spatial scale be carried out, especially before relying upon machine learning to automatically analyse marine imagery for scientific discoveries.

Another significant change in performance is shown in Figure 5.1e, where the elastic transformations applied caused a significant drop in performance of Mask R-CNN, with better performance for $\mathcal{L}^{train}_{E0} \rightarrow \mathcal{L}^{test}_{E0}$ than for $\mathcal{L}^{train}_{E1} \rightarrow \mathcal{L}^{test}_{E1}$. This is not conclusive for all elastic transformations, and other forms of image distortion, or other parameters for the given methods, may produce different results. One hypothesis for



(A) Split by distortion correction

(B) Split by extra augmentations

(C) Split by independent channel augmentations

FIGURE 5.4: mAP scores for each independent variable for low altitude vehicle, $\mathcal{L}^{train}_{C2.E0.S2} \rightarrow \mathcal{L}^{test}_{C2.E0.S2}$, summary per class

further investigation is that the chosen elastic distortion parameters were too harsh. Another is that individual methods investigated have differing effects, and may not all follow the over all pattern of decreasing performance. This study combined three different types of elastic distortion into this one category, and the negative impact of any one of them cannot be differentiated from the others. The three techniques investigated were ElasticTransformation, PiecewiseAffine, and PerspectiveTransform, from the imgaug python library. All three of these techniques were chosen to simulate the changing of terrain or vehicle perspective on a scene, and while this specific combination greatly hindered the performance of Mask R-CNN, there is still an interest in future research into simulating such an effect, and selecting the techniques and specific parameters that provide the most gain to a system such as Mask R-CNN.

Figure 5.2 shows the breakdown of the mAP scores by class across all $\mathcal{L} \rightarrow \mathcal{L}$ experiments, with Figure 5.3 showing the more in depth breakdown by class and by independent variable, showing a higher performance on Rockfish, followed by Hagfish, very closely followed by Crabs. Rockfish performing the highest out of these three classes is as expected, with Rockfish not only being the most abundant class of the three, but also being far more contrasting with the background substrate than Hagfish, and a much simpler shape to segment than Crabs. Due to this much higher performance, Rockfish are the most viable class for a neural network architecture trained on this labelled dataset to then be relied on for scientific study. The other two classes provide important insight into low-shot machine learning - the training of machine learning systems on very few training examples per class. The findings in Figure 5.3 do not show a large variance in the effects of different independent variables on different classes, but one finding of note is the performance of $S2$ for the Rockfish class, resulting in a high performance with a very small amount of variance. Scale normalisation having such a strong impact on this class in particular may be worth further investigation.

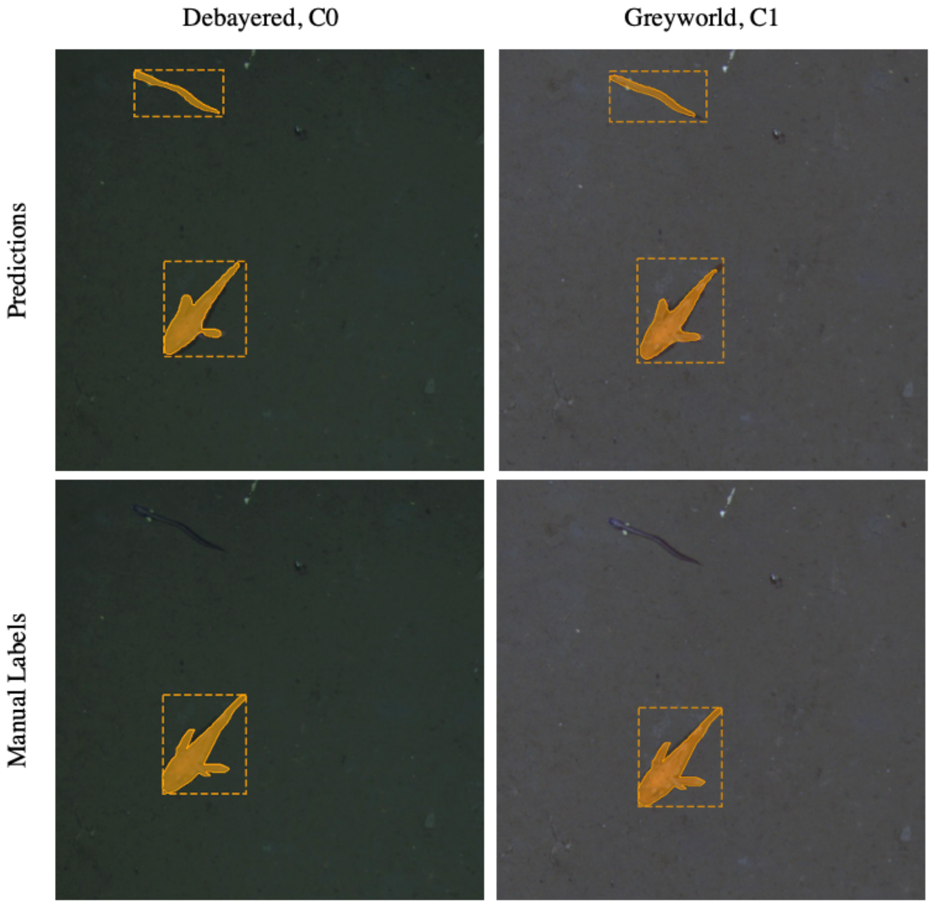


FIGURE 5.5: Example predictions and their respective manually generated labels for two instances of Mask RCNN, one trained on $\mathcal{L}^{train}_{C0.S2.D0.E0.I0.A0}$, and one on $\mathcal{L}^{train}_{C1.S2.D0.E0.I0.A0}$ images. Both of these instances make the same mistake of labelling a fish of a different morphotype as a Rockfish

Given the clear improvement in performance when using greyworld colour correction, $C2$, when not applying elastic augmentations, $E0$, and when rescaling to the average low altitude spatial scale, $S2$, Figure 5.4 shows the mAP scores for $\mathcal{L}^{train}_{C2.E0.S2} \rightarrow \mathcal{L}^{test}_{C2.E0.S2}$ split by class for closer analysis of the other independent variables. This figure shows that distortion correction has a small impact on performance, with a small decrease in performance across all classes. The effect of distortion correction being so small for the low altitude dataset is as expected due to the dome port camera used on this vehicle providing a very small amount of lens distortion in the images.

Extra augmentations improve performance for the Crab and Rockfish classes, but decrease performance for the Hagfish class. Both the Crab and Rockfish classes have a high contrast with the background substrates, with their bright orange colours, while the Hagfish is much darker. Whether this is the cause of such a difference in effect is beyond the scope of this thesis, but this is one possible explanation worth further consideration.

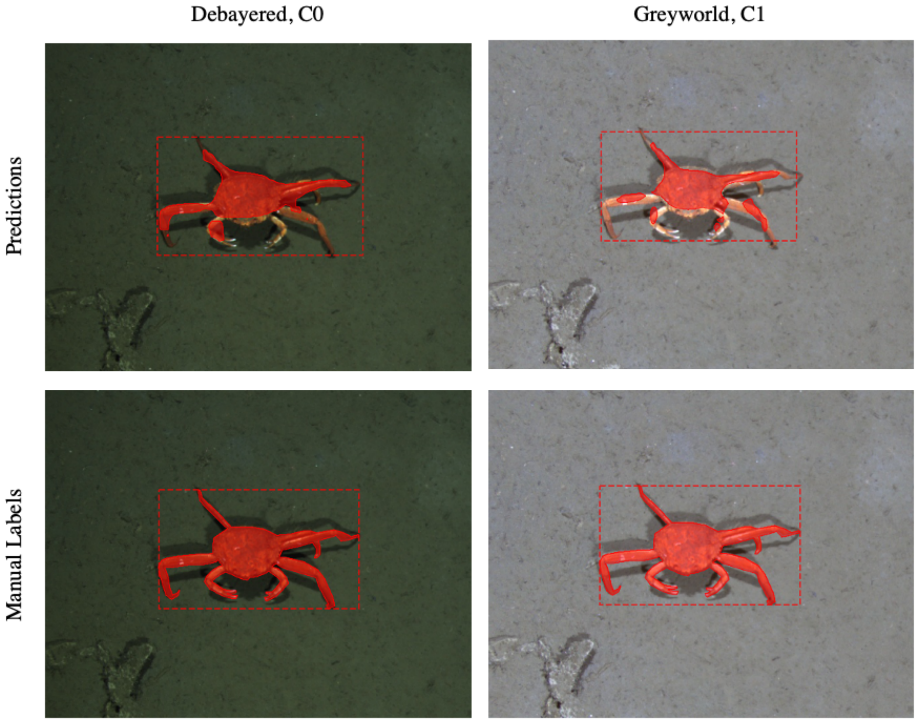


FIGURE 5.6: Example predictions and their respective manually generated labels for two instances of Mask RCNN, one trained on $\mathcal{L}^{train}_{C0.S2.D0.E0.I0.A0}$, and one on $\mathcal{L}^{train}_{C1.S2.D0.E0.I0.A0}$ images. Both of these instances successfully identify an instance of a crab, but fail to fully segment the complex shape. Both of these attempted segmentations however have a large enough IOU score to count as successful detection and segmentation.

Finally, independent channel augmentations harm the performance on the Crab class, and mildly harm performance for the Rockfish class, but provide an improvement for the Hagfish class. Again, the difference in colour and contrast between the Hagfish and the other two classes may contribute to this difference in effect. Independent colour channel augmentations provide shifts in the colour balance of the images, which will effect the appearance of the brightly coloured Rockfish and Crabs much more than the darker Hagfish class, so this hypothesis is likely but remains unconfirmed.

In Figure 5.5 the inference results from two trained instances of Mask RCNN are shown along with the manually generated labels for a given image. The two instances in question were trained on $\mathcal{L}^{train}_{C0.S2.D0.E0.I0.A0}$ and $\mathcal{L}^{train}_{C1.S2.D0.E0.I0.A0}$ respectively, with the only difference in training images for these networks being the colour correction method. The example image shown contains two individuals, one Rockfish and one individual from another class not included in this study. This class was not abundant enough in the data to be included, but as is shown in these examples can provide some confusion. Both of these instances of Mask RCNN mistook this individual as an instance of a Rockfish. The quality of the segmentation, both of the correctly identified

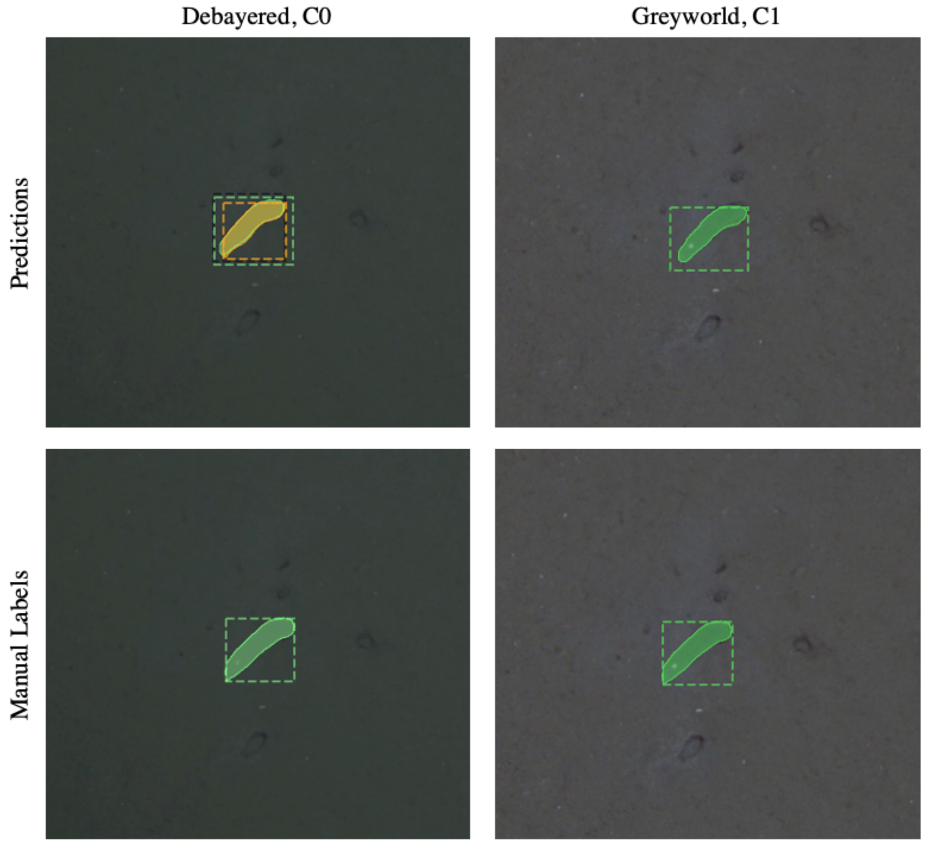


FIGURE 5.7: Example predictions and their respective manually generated labels for two instances of Mask RCNN, one trained on $\mathcal{L}^{train}_{C0.S2.D0.E0.I0.A0}$, and one on $\mathcal{L}^{train}_{C1.S2.D0.E0.I0.A0}$ images. The instance trained on $C0$ struggles to successfully identify a Hagfish, identifying one but also identifying an overlapping Rockfish, where the instance trained on $C1$ successfully identifies the Hagfish with no other complications.

Rockfish, and the incorrectly identified other morphotype, are very good with well defined boundaries and an accurate shape and size. This example demonstrates the impact of early decisions in this study, particularly the decisions made on which classes to include in the labelled training data. Including rare classes that the network has a poor chance at correctly identifying may negatively impact its performance on other more abundant classes, and it is for this reason that fewer, but more abundant classes, were selected to be labelled and included in these experiments. However, this example also shows that by not training the network on this rarer class, it can be misclassified as one of the included classes and can bring down the network's accuracy. A human labelling these images, having only learnt the classes from the labelled training data, would easily identify that the other fish is not a Rockfish, but instead belongs to a previously unseen class. This kind of reasoning is more complex than the simple object detection and segmentation that Mask R-CNN is architecturally designed to do. Perhaps a more complex neural network architecture that can identify objects, recognise uncertainty in the classification stage, and classify them as not

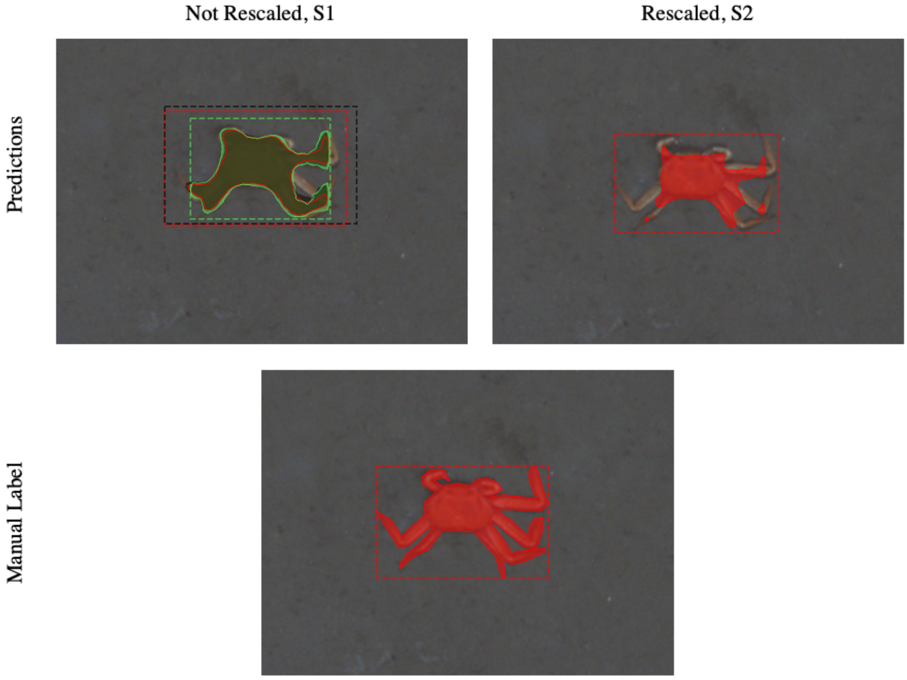


FIGURE 5.8: Example predictions and their respective manually generated labels for two instances of Mask RCNN, one trained on $\mathcal{L}^{train}_{C1.S1.D0.E0.I0.A0}$, and one on $\mathcal{L}^{train}_{C1.S2.D0.E0.I0.A0}$ images. This example contains a single instance of the Crab fish class.

belonging to the classes it has learnt from would be better suited to this task, and to real world machine learning applications in general. In this example there are instances of this unincluded class in the training data, though very few, but it is also highly probable that there will be instances of unseen classes when applying machine learning algorithms to marine imagery data, especially when applying it at scale. Handling these cases where a previously unseen class is present is an interesting question for further investigation.

Figure 5.6 again shows the output predictions from two instances of Mask R-CNN trained on $\mathcal{L}^{train}_{C0.S2.D0.E0.I0.A0}$ and $\mathcal{L}^{train}_{C1.S2.D0.E0.I0.A0}$ respectively. In this example there is a single instance of a Crab. Both instances of Mask R-CNN have correctly identified this instance of this class, and have made a relatively good attempt to segment this complex shape. The legs of the Crab class are among the most complex shapes present in the labelled datasets used in this study. The IOU score, used to determine if a segmentation counts as successful or not, for both of these cases would be high enough to count as a successful detection and segmentation, as the area covered in the predicted segment is similar enough to that of the manually generated labels. This example raises the question of what impact this less accurate segmentation may have on biomass estimates made from these predictions. Previously used biomass estimate methods use the width of a crab's carapace to calculate an estimate, and as the

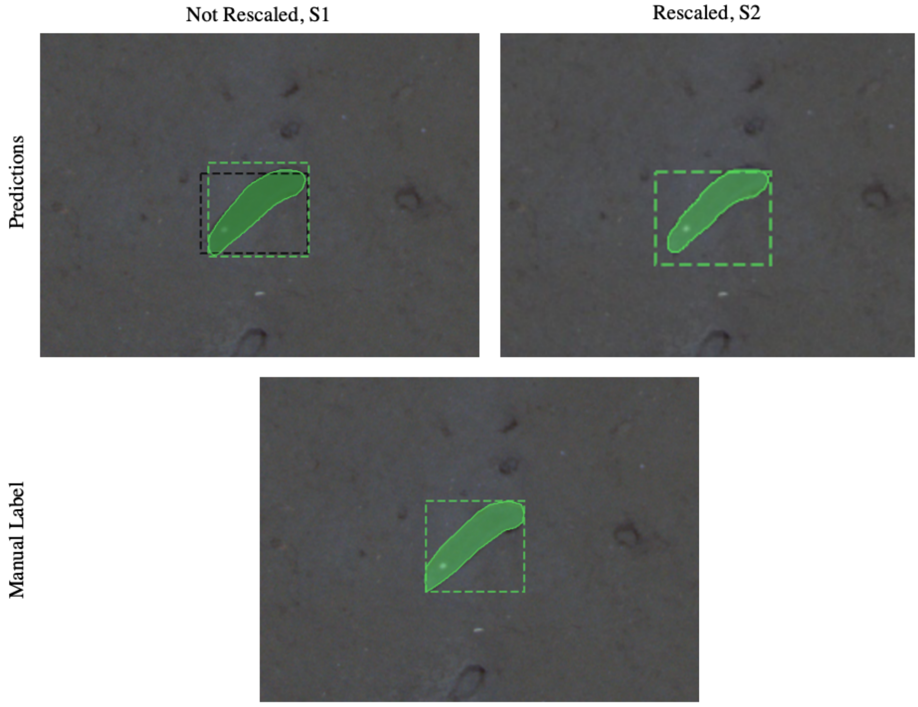


FIGURE 5.9: Example predictions and their respective manually generated labels for two instances of Mask RCNN, one trained on $\mathcal{L}^{train}_{C1.S1.D0.E0.I0.A0}$, and one on $\mathcal{L}^{train}_{C1.S2.D0.E0.I0.A0}$ images. This example contains a single instance of the Hagfish fish class.

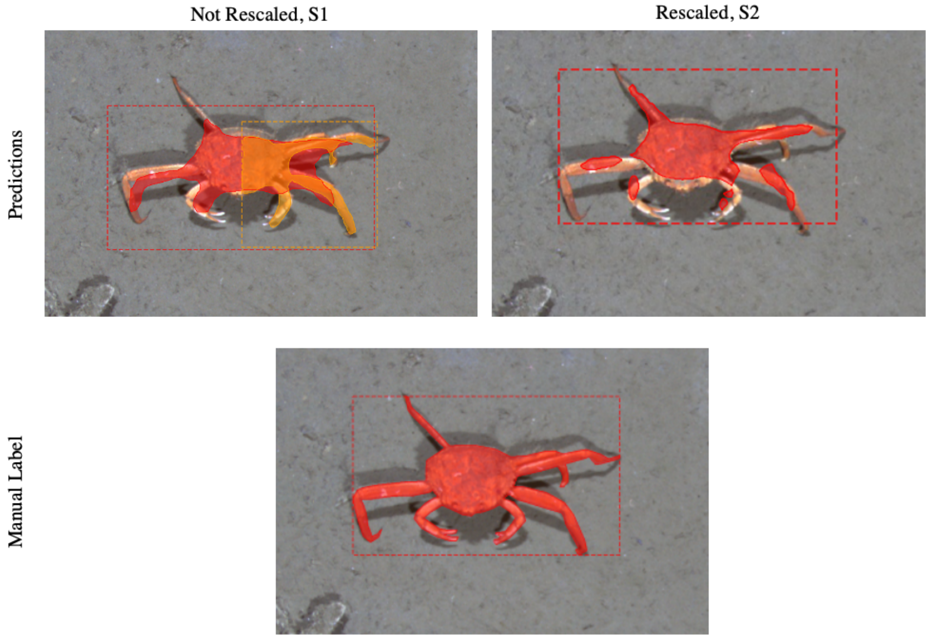


FIGURE 5.10: Example predictions and their respective manually generated labels for two instances of Mask RCNN, one trained on $\mathcal{L}^{train}_{C1.S1.D0.E0.I0.A0}$, and one on $\mathcal{L}^{train}_{C1.S2.D0.E0.I0.A0}$ images. This example contains a single instance of the Crab fish class.

majority of the crab's weight is held in the carapace perhaps disregarding the legs is a more accurate approach. Whether Mask R-CNN would adapt well to being trained on labels of just the carapace of crabs, and being expected to disregard the legs, is currently unknown.

Figure 5.7 again shows example predictions from two instances of Mask R-CNN, and is the final comparison between examples trained on $\mathcal{L}^{train}_{C0.S2.D0.E0.I0.A0}$ and $\mathcal{L}^{train}_{C1.S2.D0.E0.I0.A0}$. This example shows the rarer class from this study, the Hagfish. This class poses multiple levels of complexity, both with it's rarity in the labelled datasets - in training, validation, and testing - but also with its behaviours, with the tendency to tie itself into knots and other complex shapes, and to burrow under the sediment. The Hagfish in this example is partly submerged under the sediment, with only its head visible in the images. In this example, the network trained on $C1$ images correctly identifies the Hagfish and produces a highly accurate segmentation. On the other hand, the instance trained on $C0$ correctly identifies and segments the Hagfish, but also wrongly detects an overlapping Rockfish.

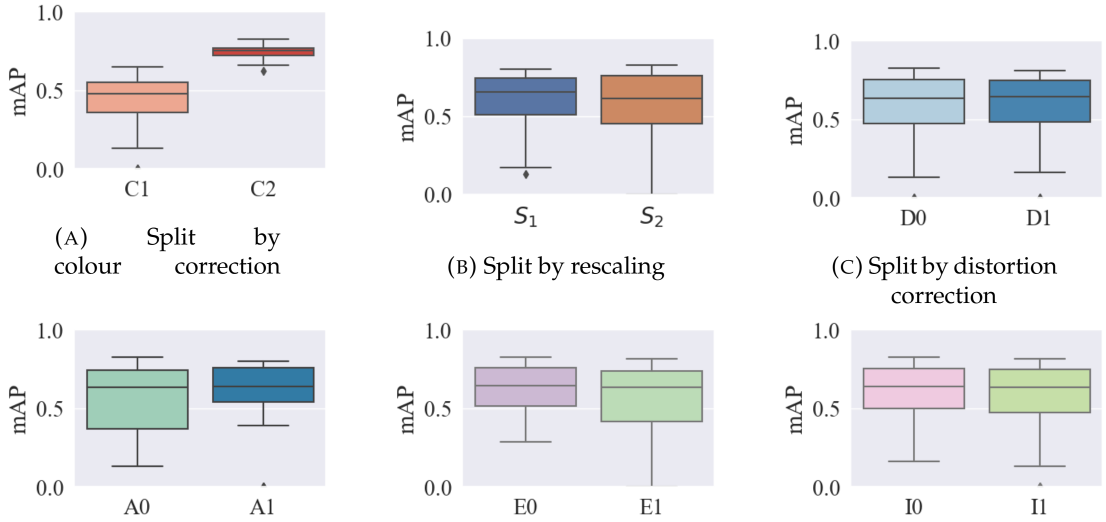
Figure 5.8 shows the results from two instances of Mask R-CNN trained on $\mathcal{L}^{train}_{C1.S1.D0.E0.I0.A0}$ and $\mathcal{L}^{train}_{C1.S2.D0.E0.I0.A0}$ respectively, with the difference in training datasets being the rescaling method, $S1$ against $S2$. This example shows a single instance of the Crab class. Similarly to the earlier example of the Crab class, the instance trained on $S2$ correctly identifies the class but struggles with segmenting the complex shape of the crab's legs. The instance of Mask R-CNN trained on $S1$ images however does not perform as well, correctly classifying the Crab instance, but also detecting an overlapping Hagfish and an overlapping Sole - a class included in the training datasets but later removed from the performance analysis of this study. Incorrectly predicting overlapping classes is one of the common mistakes made by the less accurate Mask R-CNN instances, behind the most common mistake of not identifying the presence of an object at all. The cause for this mistake is unclear, but of course brings down the accuracy of the network, and would over-predict the biomass in this example by predicting the presence of more objects than there are.

Figure 5.9 shows another set of example predictions from two instances of Mask R-CNN, trained on $\mathcal{L}^{train}_{C1.S1.D0.E0.I0.A0}$ and $\mathcal{L}^{train}_{C1.S2.D0.E0.I0.A0}$ respectively, containing one instance of a Hagfish. This is the same instance of a Hagfish seen in Figure 5.7, where it has burrowed and is only partly visible in the image. In these examples the instance of Mask R-CNN trained on $S2$ images has successfully detected and segmented the Hagfish. On the other hand, the instance trained on $S1$ images has correctly identified the Hagfish and incorrectly detected an overlapping Sole.

Figure 5.10 shows the final example comparing instances of Mask R-CNN trained on $\mathcal{L}^{train}_{C1.S1.D0.E0.I0.A0}$ and $\mathcal{L}^{train}_{C1.S2.D0.E0.I0.A0}$, showing an example with a single instance of a Crab. Again, the instance trained on $S2$ successfully detects the Crab, but as with previously seen examples fails to fully segment the complex shape of the Crab's legs. The instance trained on $S1$ images similarly struggles to segment the legs, and also wrongly identifies an overlapping Rockfish over half the Crab's body. The two classes are similar in colour and contrast with the background substrate.

5.2 High Altitude

This section presents the results for the intra-vehicle transferability on high altitude datasets, taken by the AE2000f AUV at a target altitude of 8 metres. Similarly to the previous section, these results are presented as mAP boxplots split by independent variable, then filtered to high performing experiments, and a full table of every combination of independant variables is presented. Furthermore, example segmentation outputs are shown and compared with the manually generated groundtruth.



(A) Split by colour correction

(B) Split by rescaling

(C) Split by distortion correction

(D) Split by extra augmentations

(E) Split by elastic augmentations

(F) Split by independent channel augmentations

FIGURE 5.11: mAP scores for high altitude vehicle, $\mathcal{H}^{train} \rightarrow \mathcal{H}^{test}$, split by each independent variable, and aggregating over all other independent variables.

Figure 5.11 shows a breakdown of the mAP scores by independent variables for all experiments in the form $\mathcal{H}^{train} \rightarrow \mathcal{H}^{test}$, and more detailed numerical data for each combination of variables is shown in Table 5.5.

The improvement in performance between $\mathcal{H}^{train}_{C1} \rightarrow \mathcal{H}^{test}_{C1}$ and $\mathcal{H}^{train}_{C2} \rightarrow \mathcal{H}^{test}_{C2}$, shown in Figure 5.11a, is similar to that seen in the low altitude experiments, improving from an average of 44.59% to 74.40%, and a standard deviation of 13.65% to 3.72%. The reasons for this improvement are highly likely to be the same, because of the extra features that greyworld correction, C2, corrects for, such as vignetting, over basic raw

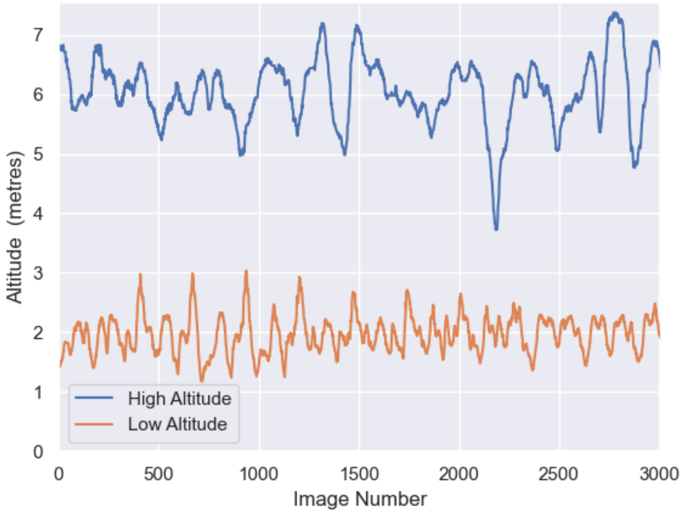


FIGURE 5.12: Altitudes recorded for each image in a subset of 3000 images from both the Low and High altitude test datasets.

images going through histogram stretching, $C1$. This method improving performance at both altitudes is not surprising, as vignetting is apparent in both datasets. It appears to be less significant for $\mathcal{H}_{C2}^{train} \rightarrow \mathcal{H}_{C2}^{test}$ than for $\mathcal{L}_{C2}^{train} \rightarrow \mathcal{L}_{C2}^{test}$, perhaps due to the different lighting set ups of each vehicle and the effect of distance off the seafloor. Visually comparing the low altitude images with the high altitude images, it appears the vignetting effect is more pronounced for low altitude images where both the camera and the lighting setup are closer to the sea floor, reaffirming this hypothesis.

The effect of rescaling, shown in Figure 5.11b, differ from the low altitude results in that rescaling to the average spatial scale for high altitude images shows no clear change in performance. There is no clear difference between $\mathcal{H}_{S2}^{train} \rightarrow \mathcal{H}_{S2}^{test}$ and $\mathcal{H}_{S1}^{train} \rightarrow \mathcal{H}_{S1}^{test}$, going from an average of 60.43% to 58.56%. This may be due to the higher altitude of the vehicle meaning small absolute fluctuations to the altitude, such as those caused by currents or terrain variation, have a lower effect on the relative distance to the seafloor than for the lower altitude vehicle. The altitude values recorded for a subset of 3000 images are shown in Figure 5.12, where although both vehicles vary in altitude throughout data collection, the relative change in the low altitude dataset is more pronounced.

The results for elastic augmentations are similar to their low altitude counterparts where $\mathcal{H}_{E0}^{train} \rightarrow \mathcal{H}_{E0}^{test}$ clearly out performs $\mathcal{H}_{E1}^{train} \rightarrow \mathcal{H}_{E1}^{test}$. This is unsurprising, but was not a guaranteed result as elastic transformations simulating changes to terrain largely depend on spatial scale to determine what kind of effect is being simulated, small scale distortions simulating variation in texture, and large scale distortions simulating variation in slope, or the appearance of hills or dips in the terrain.

TABLE 5.5: mAP scores and variance for high altitude vehicle, $\mathcal{H}^{train} \rightarrow \mathcal{H}^{test}$

				F0	F0	F1	F1
				I0	I1	I0	I1
C1	S1	D0	E0	0.45 ± 0.09	0.51 ± 0.12	0.57 ± 0.03	0.57 ± 0.02
C1	S1	D0	E1	0.41 ± 0.11	0.22 ± 0.09	0.60 ± 0.05	0.49 ± 0.01
C1	S1	D1	E0	0.44 ± 0.06	0.43 ± 0.12	0.57 ± 0.07	0.54 ± 0.02
C1	S1	D1	E1	0.33 ± 0.02	0.32 ± 0.09	0.54 ± 0.06	0.53 ± 0.04
C1	S2	D0	E0	0.39 ± 0.05	0.42 ± 0.08	0.52 ± 0.05	0.53 ± 0.06
C1	S2	D0	E1	0.29 ± 0.10	0.33 ± 0.07	0.47 ± 0.07	0.35 ± 0.25
C1	S2	D1	E0	0.49 ± 0.07	0.44 ± 0.07	0.53 ± 0.07	0.52 ± 0.04
C1	S2	D1	E1	0.27 ± 0.08	0.32 ± 0.10	0.54 ± 0.00	0.32 ± 0.23
C2	S1	D0	E0	0.74 ± 0.03	0.75 ± 0.01	0.75 ± 0.00	**0.77 ± 0.02**
C2	S1	D0	E1	0.72 ± 0.02	0.71 ± 0.02	0.75 ± 0.02	0.73 ± 0.02
C2	S1	D1	E0	0.75 ± 0.02	0.70 ± 0.04	0.75 ± 0.02	**0.77 ± 0.01**
C2	S1	D1	E1	0.70 ± 0.03	0.73 ± 0.03	**0.77 ± 0.04**	0.72 ± 0.03
C2	S2	D0	E0	**0.78 ± 0.04**	0.75 ± 0.03	0.76 ± 0.02	0.75 ± 0.03
C2	S2	D0	E1	0.73 ± 0.03	**0.79 ± 0.02**	0.72 ± 0.07	0.73 ± 0.01
C2	S2	D1	E0	**0.77 ± 0.05**	0.76 ± 0.01	0.74 ± 0.03	**0.78 ± 0.01**
C2	S2	D1	E1	0.75 ± 0.02	0.72 ± 0.02	0.74 ± 0.06	0.72 ± 0.05

TABLE 5.6: ANOVA test results for each category, $\mathcal{H}^{train} \rightarrow \mathcal{H}^{test}$

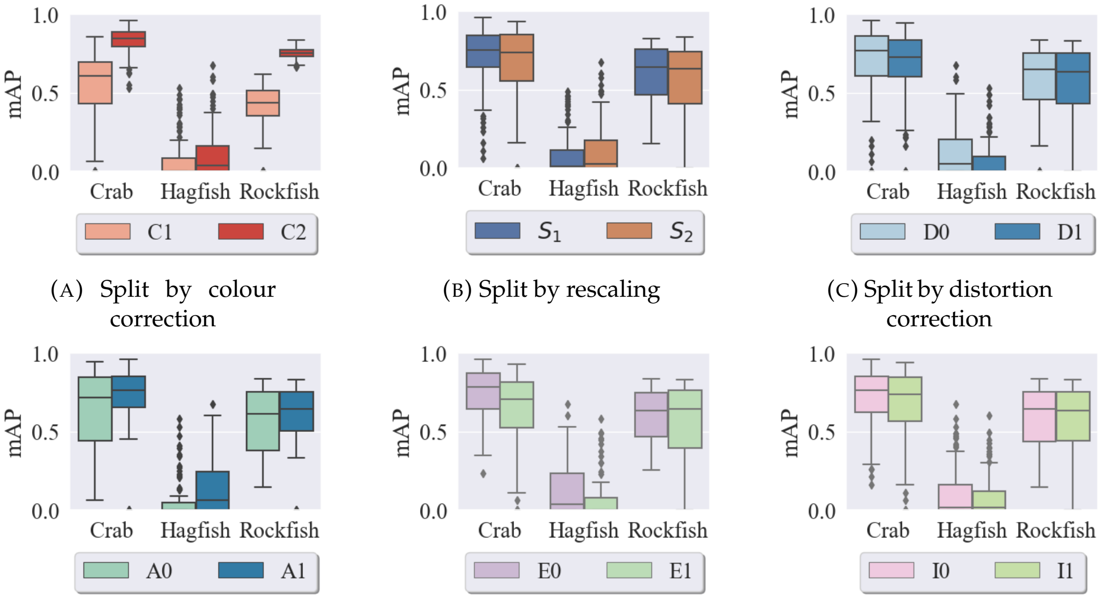
Category	**F-value**	**p-value**
Colour Correction	421.557	$4.091e^{-50}$
Distortion Correction	0.003	0.958
Scale	0.516	0.474
Elastic Transformations	5.546	0.020
Individual Channel Augmentations	0.516	0.474
Extra Augmentations	7.357	0.007

The statistical significance of these results is analysed with ANOVA tests, the results of which are shown in Table 5.6. Similarly to the Low Altitude results already presented, the categories with the highest significance are scale, colour correction, and elastic transformations. The ANOVA results for each combination of two categories is shown in 5.7, and unlike the Low Altitude results, some combinations of variables have stronger significance than either variable alone.

The results show that both Colour Correction had the strongest impact of all variables, with an F-value of 421.557 and a p-value of $4.091e^{-50}$. Extra Augmentations and Elastic Transformations also had statistically significant effects on the results, with over 95% significance, with F-values of 7.357 and 5.546, and p-values of 0.007 and 0.02 respectively. When combined, the effect of Colour Corrections and Extra Augmentations on the outcome variable is even stronger, with an F-value of 196.115 and a very low p-value of $1.222e^{-57}$, indicating a highly significant effect. The same can be said for Colour Correction and Elastic Transformations, with an F-value of

TABLE 5.7: ANOVA test results for each combination of category, $\mathcal{H}^{train} \rightarrow \mathcal{H}^{test}$

Combined Categories	F-value	p-value
Colour Correction and Distortion Correction	139.096	$1.688e^{-47}$
Colour Correction and Scale	145.373	$9.539e^{-49}$
Colour Correction and Elastic Transformations	169.694	$2.989e^{-53}$
Colour Correction and Individual Channel Augmentations	141.807	$4.829e^{-48}$
Colour Correction and Extra Augmentations	196.115	$1.222e^{-57}$
Distortion Correction and Scale	0.207	0.891
Distortion Correction and Elastic Transformations	1.830	0.143
Distortion Correction and Individual Channel Augmentations	0.176	0.912
Distortion Correction and Extra Augmentations	2.428	0.067
Scale and Elastic Transformations	2.065	0.106
Scale and Individual Channel Augmentations	0.345	0.793
Scale and Extra Augmentations	2.952	0.034
Elastic Transformations and Individual Channel Augmentations	2.184	0.091
Elastic Transformations and Extra Augmentations	4.564	0.004
Individual Channel Augmentations and Extra Augmentations	2.682	0.048



(A) Split by colour correction

(B) Split by rescaling

(C) Split by distortion correction

(D) Split by extra augmentations

(E) Split by elastic augmentations

(F) Split by independent channel augmentations

FIGURE 5.13: mAP scores for high altitude vehicle, $\mathcal{H}^{train} \rightarrow \mathcal{H}^{test}$, summary per class, , split by each independent variable, and aggregating over all other independent variables.

169.694 and a p-value of $2.989e^{-53}$, indicating a higher significance tthan Colour Correction alone.

As already stated, the results show that Colour Correction has a significant effect on the outcome variable, with an F-value of 421.55 and a very low p-value of $4.091e^{-50}$, indicating a highly significant effect. Distortion Correction, on the other hand, does

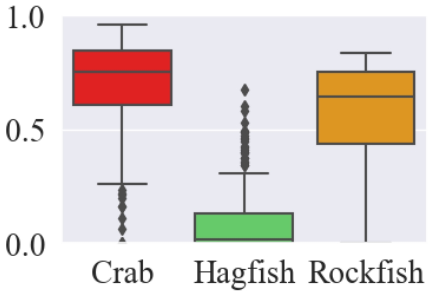
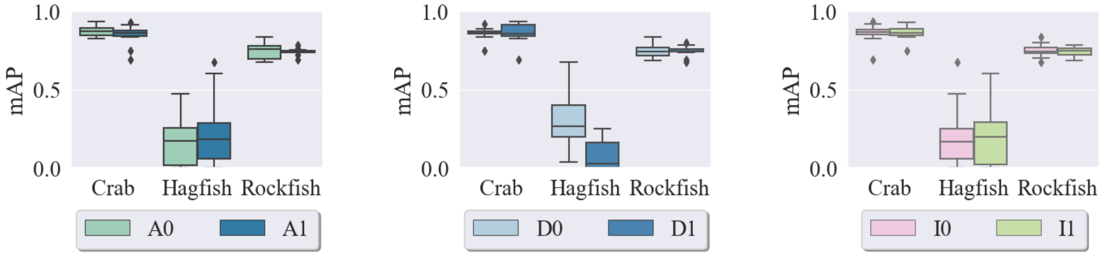


FIGURE 5.14: mAP scores for high altitude vehicle, $\mathcal{H}^{train} \rightarrow \mathcal{H}^{test}$, summary per class



(A) Split by extra augmentations

(B) Split by distortion correction

(C) Split by independent channel augmentations

FIGURE 5.15: mAP scores for high altitude vehicle, $\mathcal{H}^{train}_{C2.D0.S2} \rightarrow \mathcal{H}^{test}_{C2.D0.S2}$, summary per class

not show any significant effect, with an F-value of 0.003 and a high p-value of 0.958. When combined, Colour Correction and Distortion Correction have a significant effect on the outcome variable, with an F-value of 139.096 and a very low p-value of $1.688e^{-47}$, indicating a highly significant effect. This may be explained by the very strong effect of Colour Correction alone, but may also imply a synergy between the two variables.

The highest performing independent variables for the high altitude experiments are clearly the greyworld colour correction, $C2$, and without elastic augmentations, $E0$. As not rescaling, $S1$, and rescaling to the average high altitude spatial scale, $S2$ have comparable performance here, $S2$ has been selected. Figure 5.15 shows the results for $\mathcal{H}^{train}_{C2,E0,S2} \rightarrow \mathcal{H}^{test}_{C2,E0,S2}$.

There is a small decrease in performance when applying extra augmentations, $E1$, for the Crab and Rockfish Classes, with a minor increase in performance for Hagfish.

The impact of distortion correction on the Crab and Rockfish classes is inclonclusive, but there is a clear decrease in performance for Hagfish. This class is the more complex shape in most cases, where the Hagfish is a flexible creature that often ties itself in knots, which may contribute to this difference from the other classes.

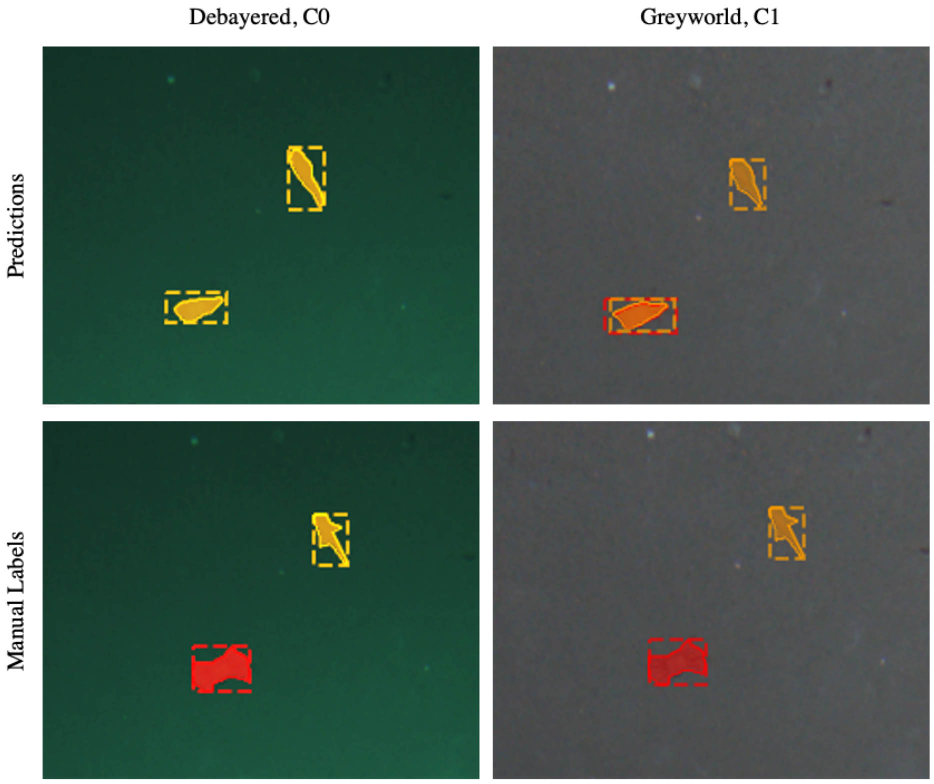


FIGURE 5.16: Example predictions and ground truth labels from two instances of Mask R-CNN trained on $\mathcal{H}^{train}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$ respectively, containing one instance of a Rockfish and one instance of a Crab. Both instances of Mask R-CNN successfully detect and segment the Rockfish, and both fail to successfully detect the Crab. Colours for labels use the legend set out in Figure 4.7.

Similarly, the impact of independent channel augmentations, $I1$, is inconclusive for both the Crab and Rockfish classes. There is a small increase in performance for Hagfish, but is not clearly significant.

Figure 5.16 shows example predictions from two instances of Mask R-CNN, and their relative manually assigned labels. These two instances of Mask R-CNN were trained on $\mathcal{H}^{train}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$ respectively, with the only difference in training data being the colour correction method applied. This example in particular shows one instance of a Rockfish and one instance of a Crab. Both of the instances of Mask R-CNN successfully detect and segment the Rockfish, which reflects the higher accuracy score we see for this class. Neither, however, successfully detected and segmented the Crab. The instance trained on C0 mistook the Crab for a Rockfish, and the instance trained on C1 detected the Crab but mistakenly also detected an overlapping Rockfish. Each of these mistakes has a different effect on automated biological analysis. The mistake of the instance trained on C0 will be referred to as mis-classification, and the mistake of the instance trained on C1 will be referred to as extra detection. The mis-classification mistake will lead to overestimating the

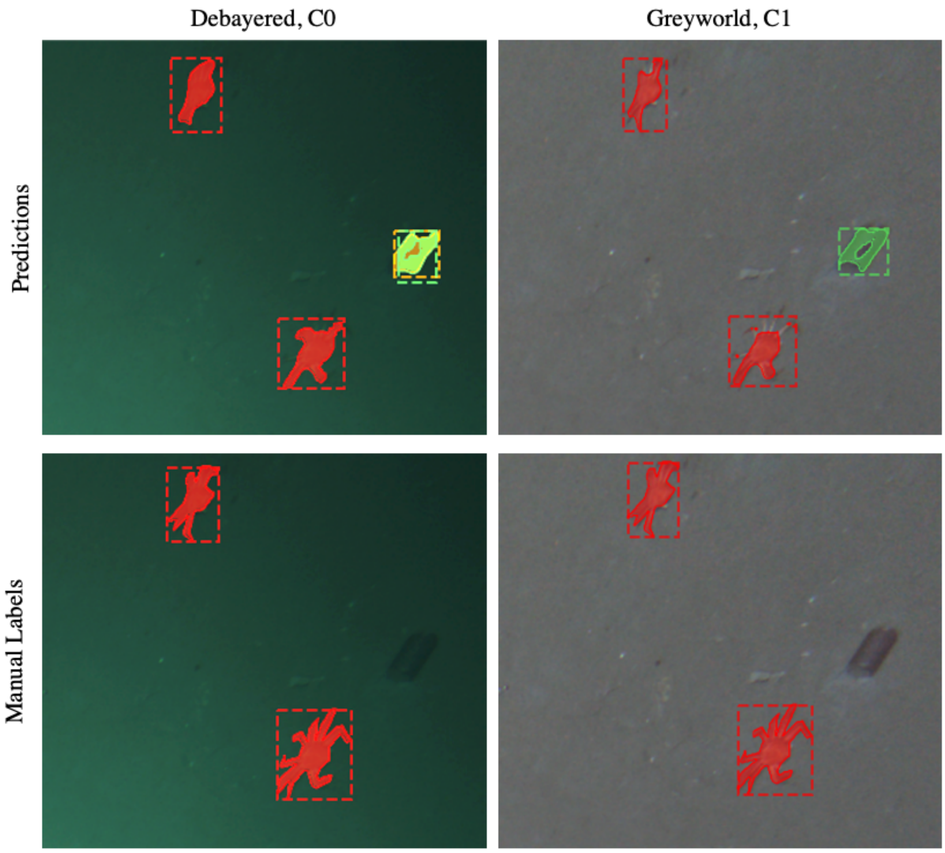


FIGURE 5.17: Example predictions and ground truth labels from two instances of Mask R-CNN trained on $\mathcal{H}^{train}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$ respectively, containing two instances of Crabs. Both instances of Mask R-CNN successfully detect and segment the Crabs, however both also mistakenly detect an object, or objects, where an artificial object appears in the image. Colours for labels use the legend set out in Figure 4.7.

population count and the biomass estimation for the Rockfish class, and underestimating the population count and the biomass estimation for the Crab class, however for aggregate statistics that don't concern individual classes, this mistake leads to the same number of organisms being counted. The extra classification mistake leads to the correct population count and biomass estimate for the Crab class, but over-estimates the population count and biomass both for the Rockfish class and for aggregate statistics. The severity of each of these mistakes depends heavily on what biological statistics are being calculated from these predictions.

Figure 5.17 also shows example predictions from two instances of Mask R-CNN, again trained on $\mathcal{H}^{train}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$ respectively, with the difference in training data being the colour correction method, $C0$ and $C1$. This example contains two instances of Crabs. Both of these instances of Mask R-CNN successfully detect and segment both of the Crabs in the image, however both struggle to segment the complex shape of the Crabs' legs. This is similar to the examples seen in the previous

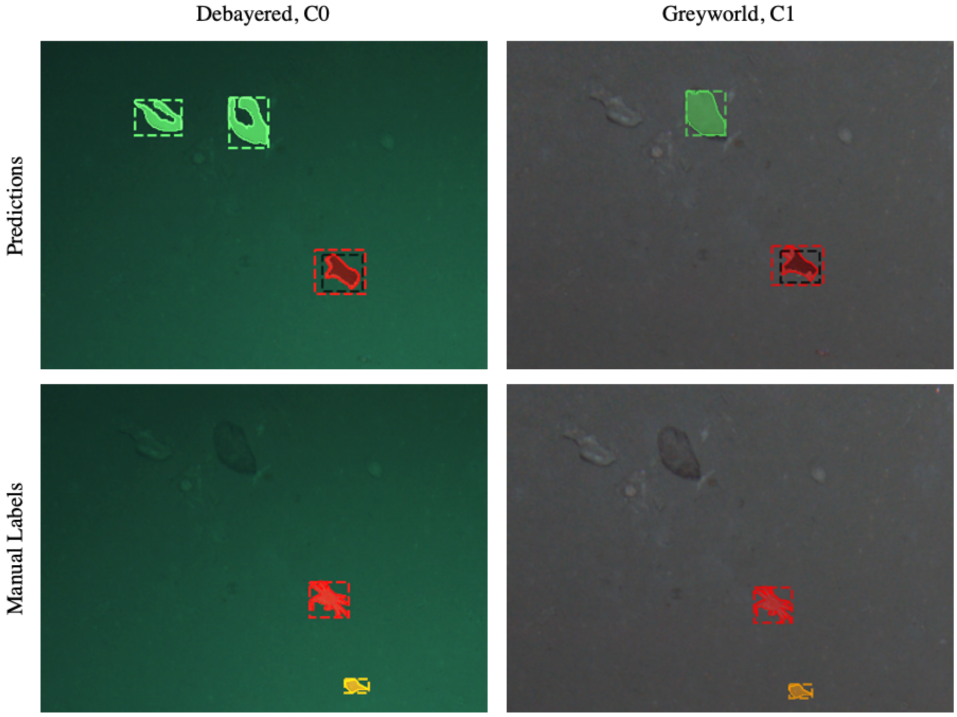


FIGURE 5.18: Example predictions and ground truth labels from two instances of Mask R-CNN trained on $\mathcal{H}^{train}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$ respectively, containing one instance of a Rockfish and one instance of a Crab. Both instances of Mask R-CNN mistakenly detect an extra object on top of the Crab, neither successfully detects the Rockfish, and they each mistakenly detect two and two Hagfish, respectively. Colours for labels use the legend set out in Figure 4.7.

section trained on low altitude imagery, showing this limitation of Mask R-CNN's segmenting abilities is cross vehicular. Both of these instances of Mask R-CNN also wrongly detected objects where an artificial object appears in the image, with the instance trained on C0 detecting both a Rockfish and a Hagfish, and the instance trained on C1 detecting just a Hagfish. These mistakes are both extra classifications, and in this case are caused by an unusual object appearing in the image that may not have been present in the training data. Similarly to the low altitude example in Figure 5.5, where a Rockfish was incorrectly detected where an organism of a class outside the training label set was visible, a more complex machine learning architecture capable of classifying objects as belonging to an unseen class may be more robust against this kind of mistake. Furthermore, as an extra classification mistake, these predictions would add to the population count and biomass estimates for the wrongly detected classes, and for aggregate statistics.

Figure 5.18 shows the third and final example outputs from two instances of Mask R-CNN trained on high altitude $\mathcal{H}^{train}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$ data respectively. This example shows one instance of a Crab and one instance of a Rockfish. Unlike the previously seen examples, neither of the instances of Mask

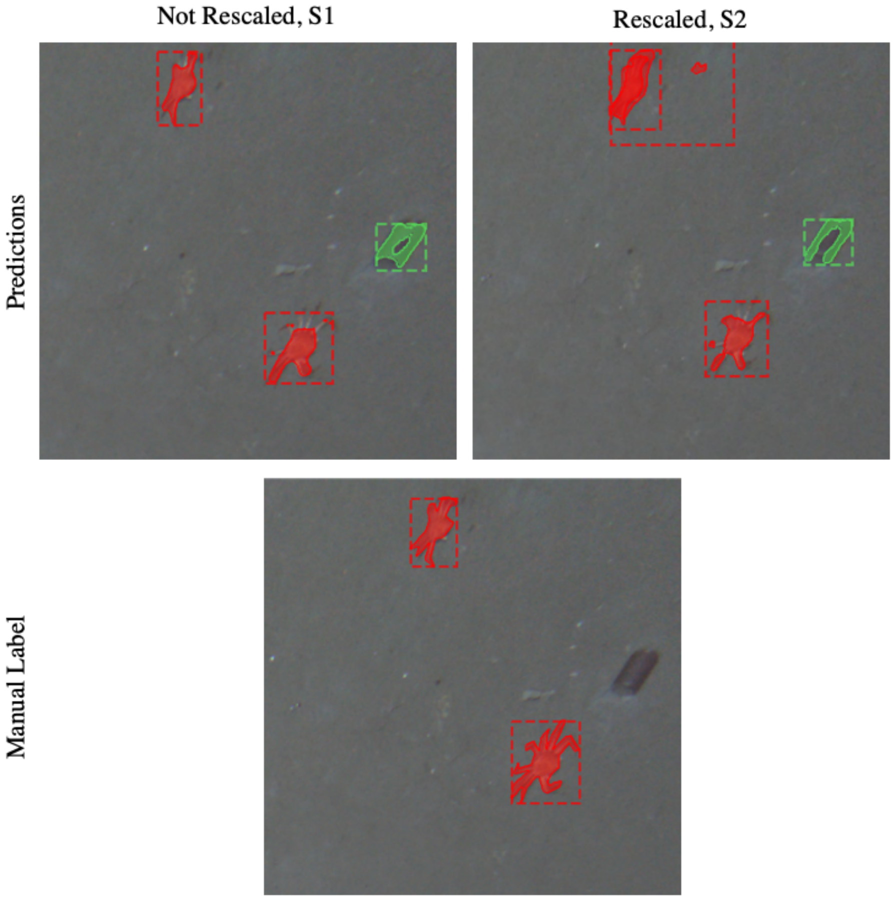


FIGURE 5.19: Example predictions and ground truth labels from two instances of Mask R-CNN trained on $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S2.D0.E0.I0.A0}$ respectively, containing two instances of Crabs. Both instances of Mask R-CNN successfully detect the Crabs, however they both also mistakenly detect a Hagfish where an artificial object appears in the image, and one mistakenly detects two Crabs where there is only one. Colours for labels use the legend set out in Figure 4.7.

R-CNN successfully detect the Rockfish. Both of the instances of Mask R-CNN detect and segment the Crab, but also detect an overlapping Sole. Furthermore, both instances also detect Hagfish, with the instance trained on $C0$ detecting two, and the instance trained on $C1$ detecting one. These mistaken detections and segmentations occur where a distinct rock appears in the image, providing the contrasting edge expected between an object and the background substrate, which likely contributed to these incorrect detections. The missed detection of the Rockfish differs from the previously seen mistakes in this section, underestimating the population count and biomass estimates for both the Rockfish class and aggregate statistics. The other mistakes in this example are all extra classifications, where the population count and biomass estimates are over-estimated for Sole, Hagfish, and aggregate statistics.

Figure 5.19 shows example predictions from two instances of Mask R-CNN trained on $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S2.D0.E0.I0.A0}$, the first example in this section comparing

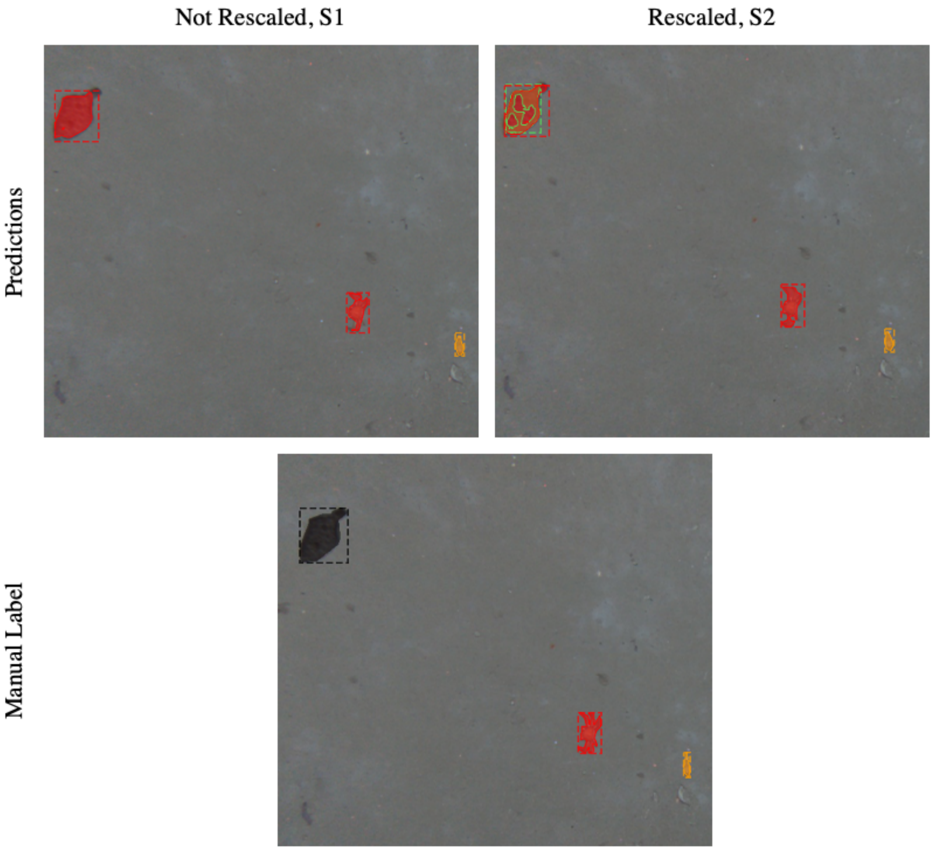


FIGURE 5.20: Example predictions and ground truth labels from two instances of Mask R-CNN trained on $\mathcal{H}^{train}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$ respectively, containing one instance of a Sole, one instance of a Rockfish, and one instance of a Crab. Both instances of Mask R-CNN successfully detect and segment the Crab and the Rockfish, however both mistakenly detect the Sole and a Rockfish, and the second instance also mistakenly detects a Hagfish overlapping the Sole. Colours for labels use the legend set out in Figure 4.7.

instances trained on differently scaled training data, $S1$ and $S2$. This example contains two instances of Crabs. Both instances of Mask R-CNN successfully detect and segment these instances, however the instance trained on scale normalised data, $S2$, wrongly detects an overlapping instance of a Crab. Both of these instances of Mask R-CNN also incorrectly detect a Hagfish where an artificial object appears in the image. This artificial object is a similar colour to many examples of Hagfish, which may have contributed to this extra detection. The duplicate detection of the Crab would add to the population count and biomass estimated of both the Crab class and the aggregate statistics, and similarly the extra detection and classification of the Hagfish would add to the population count and biomass estimates for the Hagfish class and the aggregate statistics.

Figure 5.20 shows another set of example predictions from two instances of Mask R-CNN trained on $\mathcal{H}^{train}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$ respectively, with the only

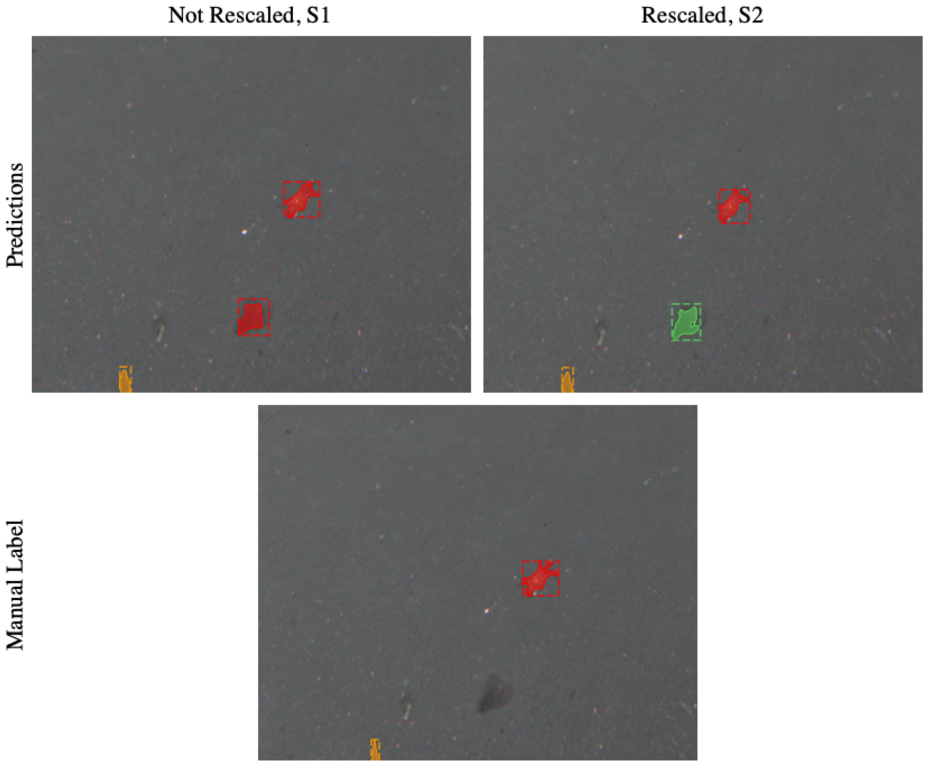


FIGURE 5.21: Example predictions and ground truth labels from two instances of Mask R-CNN trained on $\mathcal{H}^{train}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$ respectively, containing one instance of a Rockfish and one instance of a Crab. Both instances of Mask R-CNN successfully detect both the Crab and the Rockfish, however they also mistakenly detect a Crab and a Hagfish, respectively, where a dark object appears in the image. Colours for labels use the legend set out in Figure 4.7.

difference in training data being the rescaling method. This example contains one instance of a Sole, included in the training dataset but removed from the analysis for this thesis due to it's rarity in the datasets, one Crab and one Rockfish. Both of these instances of Mask R-CNN successfully detect and segment both the Crab and the Rockfish, however neither successfully detects and segments the Sole. The instance of Mask R-CNN trained on the original scale training data, $S1$, incorrectly identifies the Sole as a Crab, overestimating the population count and biomass statistics for the Crab class, and underestimating for the Sole class. The instance of Mask R-CNN trained on scale normalised data, $S2$, incorrectly identifies the Sole as an instance of a Crab, and as an overlapping instance of a Hagfish in a very unusual shape. This mistake overestimates the population count and biomass statistics for both the Crab and Hagfish classes, and overestimating aggregate statistics, while also underestimating for the Sole class.

Figure 5.21 is the final example of predicted segments from two instances of Mask R-CNN trained on $\mathcal{H}^{train}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$ respectively, and the final example over all in this section. This example shows one Crab aand one Rockfish, which both instances of Mask R-CNN successfully detect and segment. This example also contains a dark object that may be marine snow close to the camera, which the instance of Mask R-CNN trained on the original scale images, $S1$, identifies as a Crab, and the instance trained on scale normalised images, $S2$, identifies as a Hagfish. Each of these mistakes would overestimate statistics for the Crab and Hagfish classes respectively, and would overestimate aggregate statistics.

To conclude, in $\mathcal{L} \rightarrow \mathcal{L}$ experiments, the most impactful variables were colour correction and rescaling method, with the best results coming from correcting for colour and rescaling to the average low altitude spatial scale. Applying elastic augmentations and extra augmentations caused a drop in performance. For $\mathcal{H} \rightarrow \mathcal{H}$ experiments, colour correction also provided a large improvement in results, however rescaling did not. In $\mathcal{L} \rightarrow \mathcal{L}$ experiments, the best performance was on the Rockfish class, whereas for $\mathcal{H} \rightarrow \mathcal{H}$ experiments the highest performing class was Crab, possibly due to Crabs being larger in size than Rockfish, and therefore being easier to detect in the high altitude imagery than Rockfish.

Chapter 6

Transfer of learning across acquisition platforms

This chapter presents the results for transfer learning, where the training and validation datasets and the test dataset are from different vehicles. This is a valuable experiment as it provides insight into how best to apply learning from a given dataset to one collected by an entirely different vehicle, with notable differences including altitude, lighting setup, and camera opening angle. This is broken down into Low altitude to High altitude, $\mathcal{L} \rightarrow \mathcal{H}$, and High altitude to Low altitude, $\mathcal{H} \rightarrow \mathcal{L}$.

6.1 Low altitude to High altitude transfer

The general performance when transferring from low to high altitude datasets, that is training on the low altitude dataset and testing on the high altitude dataset $\mathcal{L}^{train} \rightarrow \mathcal{H}^{test}$, is very poor, with a maximum mAP score of 36%.

Figure 6.1 shows a breakdown of the mAP across the independent variables for this set of experiments. The most notable thing is the overall low performance, but other patterns of note are shown. There is a drastic increase in performance between C1 and C2, where applying pixel-statistics greyworld correction provides a clear improvement, with the majority of the $C1$ experiments achieving an mAP of 0%, identifying nothing correctly. Scale also shows to have a significant effect on performance, with $S2$, scaling to the average low altitude spatial scale, out performing other rescaling methods. Finally, there is a drop in performance when applying elastic transformations, $E1$. All three of these findings are in line with the findings of the $\mathcal{L}^{train} \rightarrow \mathcal{L}^{test}$ experiments.

The ANOVA tests for the statistical significance of these results is included here in Tables 6.2 and 6.3. The individual ANOVA scores correspond with the observed

TABLE 6.1: mAP scores and variance for low altitude vehicle, $\mathcal{L}^{train} \rightarrow \mathcal{H}^{test}$

				F0	F0	F1	F1
				I0	I1	I0	I1
C1	S1	D0	E0	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C1	S1	D0	E1	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C1	S1	D1	E0	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C1	S1	D1	E1	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C1	S2	D0	E0	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C1	S2	D0	E1	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C1	S2	D1	E0	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C1	S2	D1	E1	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C1	S3	D0	E0	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C1	S3	D0	E1	0.00 ± 0.00	0.02 ± 0.03	0.00 ± 0.00	0.00 ± 0.00
C1	S3	D1	E0	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C1	S3	D1	E1	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C1	S4	D0	E0	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C1	S4	D0	E1	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C1	S4	D1	E0	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C1	S4	D1	E1	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C2	S1	D0	E0	0.06 ± 0.01	0.02 ± 0.02	0.12 ± 0.02	0.06 ± 0.01
C2	S1	D0	E1	0.06 ± 0.02	0.06 ± 0.01	0.05 ± 0.07	0.02 ± 0.02
C2	S1	D1	E0	0.13 ± 0.03	0.08 ± 0.00	0.10 ± 0.02	0.13 ± 0.02
C2	S1	D1	E1	0.05 ± 0.01	0.07 ± 0.04	0.02 ± 0.03	0.11 ± 0.07
C2	S2	D0	E0	**0.30 ± 0.06**	**0.29 ± 0.05**	**0.28 ± 0.04**	0.26 ± 0.05
C2	S2	D0	E1	0.22 ± 0.03	0.24 ± 0.03	0.16 ± 0.11	0.09 ± 0.07
C2	S2	D1	E0	0.24 ± 0.04	0.12 ± 0.09	**0.36 ± 0.04**	**0.27 ± 0.03**
C2	S2	D1	E1	0.20 ± 0.08	0.18 ± 0.02	0.11 ± 0.16	0.24 ± 0.01
C2	S3	D0	E0	0.14 ± 0.03	0.11 ± 0.02	0.00 ± 0.00	0.00 ± 0.00
C2	S3	D0	E1	0.11 ± 0.04	0.09 ± 0.05	0.02 ± 0.02	0.00 ± 0.00
C2	S3	D1	E0	0.17 ± 0.01	0.05 ± 0.02	0.03 ± 0.03	0.00 ± 0.00
C2	S3	D1	E1	0.07 ± 0.03	0.04 ± 0.03	0.00 ± 0.00	0.00 ± 0.00
C2	S4	D0	E0	0.16 ± 0.11	0.00 ± 0.00	0.17 ± 0.05	0.22 ± 0.05
C2	S4	D0	E1	0.18 ± 0.05	0.00 ± 0.00	0.00 ± 0.00	0.09 ± 0.12
C2	S4	D1	E0	0.19 ± 0.02	0.10 ± 0.07	0.22 ± 0.05	**0.27 ± 0.05**
C2	S4	D1	E1	0.00 ± 0.00	0.09 ± 0.07	0.07 ± 0.10	0.08 ± 0.12

results above, with significance for colour correction, scale, and elastic transformations. The combined results indicate high significance for every combination of colour correction, scale, and elastic transformations with all other variables. This may imply synergy between some variables, but could also be explained by the strong significance of each of those variables alone, and due to the generally low mAP score in this set of experiments, it is not possible to draw any further conclusions.

Figure 6.2 shows the breakdown of mAP scores by class, with the lowest performance for Hagfish, and the highest performance for Rockfish. This is, again, in line with the findings from the $\mathcal{L}^{train} \rightarrow \mathcal{L}^{test}$ experiments, and perhaps implies that a higher

TABLE 6.2: ANOVA test results for each category, $\mathcal{L}^{train} \to \mathcal{H}^{test}$

Category	F-value	p-value
Colour Correction Type	221.109	$8.812e^{-40}$
Distortion Correction	0.060	0.806
Rescaled	17.720	$8.730e^{-11}$
Elastic Transformations	9.630	0.002
Separate Channel Operations	0.826	0.364
Flip/Rotate	0.181	0.671

TABLE 6.3: Two-way ANOVA test results for different image augmentation techniques, $\mathcal{L}^{train} \to \mathcal{H}^{test}$

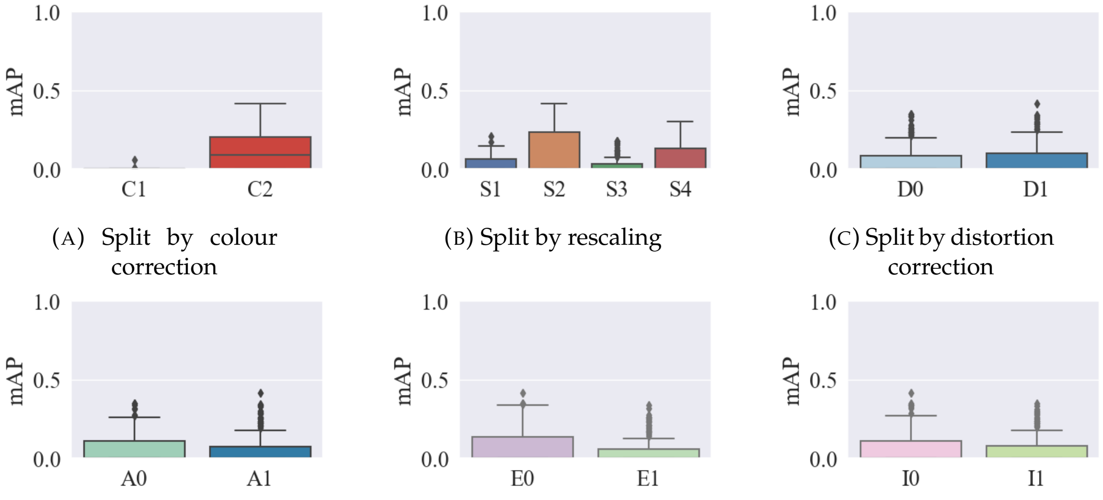
Combined Categories	F-value	p-value
Colour Correction and Distortion Correction	73.44	$1.74e^{-37}$
Colour Correction and Rescaled	85.56	$7.17e^{-74}$
Colour Correction and Elastic Transformations	90.51	$3.29e^{-44}$
Colour Correction and Separate Channel Operations	74.77	$5.00e^{-38}$
Colour Correction and Flip Rotate	73.58	$1.54e^{-37}$
Distortion Correction and Rescaled	7.76	$8.98e^{-09}$
Distortion Correction and Elastic Transformations	3.30	0.02
Distortion Correction and Separate Channel Operations	0.37	0.77
Distortion Correction and Flip Rotate	0.59	0.62
Rescaled and Elastic Transformations	9.89	$2.46e^{-11}$
Rescaled and Separate Channel Operations	7.71	$1.02e^{-08}$
Rescaled and Flip Rotate	8.93	$3.43e^{-10}$
Elastic Transformations and Separate Channel Operations	4.07	0.01
Elastic Transformations and Flip Rotate	4.10	0.01
Separate Channel Operations and Flip Rotate	0.93	0.43

number of training examples are required for higher performance on the $\mathcal{L}^{train} \to \mathcal{H}^{test}$ experiments.

6.2 High Altitude to Low Altitude

The transferability from High Altitude datasets to Low Altitude datasets proves much more promising, with performance that, while not suitable for large scale scientific analysis with any level of trust, shows potential for further work and, with additional training data and closer fine-tuning, may provide a fully scalable cross vehicle solution. These improvements are outside the scope of this thesis, and instead this section focuses on the impact of the independent variables on performance, to best inform such future work.

Figure 6.3 shows the mAP scores for $\mathcal{H}^{train} \to \mathcal{L}^{test}$. The improvement in performance when applying $C2$ over $C1$ shown here makes this effect universal across all the experiments in this thesis. The findings shown in this thesis allow for a reasonably



(A) Split by colour correction

(B) Split by rescaling

(C) Split by distortion correction

(D) Split by extra augmentations

(E) Split by elastic augmentations

(F) Split by independent channel augmentations

FIGURE 6.1: mAP scores for transfer learning from low altitude to high altitude, $\mathcal{L}^{train} \rightarrow \mathcal{H}^{test}$, split by each independent variable, and aggregating over all other independent variables.

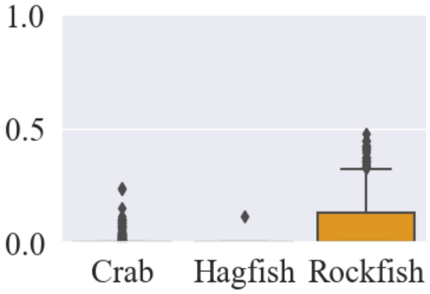


FIGURE 6.2: mAP scores for transfer learning from low altitude to high altitude, $\mathcal{L}^{train} \rightarrow \mathcal{H}^{test}$, split by class.

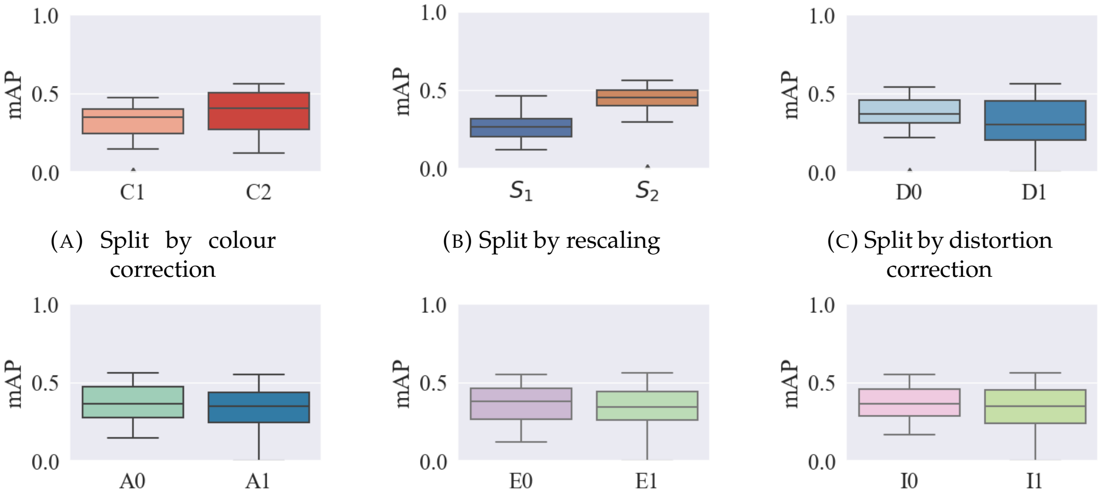
confident recommendation to apply greyworld pixel-wise statistical correction over simple histogram stretching. Any possible loss of information by this transformation is clearly made up for and more by the improvements this correction method provides, such as the removal of vignetting, the common effect in marine imagery of the illumination being brightest in the centre of the image and darker at the edges.

There is a significant improvement in performance when normalising for scale, with $\mathcal{H}^{train}_{S2} \rightarrow \mathcal{L}^{test}_{S3}$ outperforming $\mathcal{H}^{train}_{S1} \rightarrow \mathcal{L}^{test}_{S1}$, with respective average mAP scores of 44.09% and 26.52%. It is expected that scale normalisation has a large effect on inter-vehicle transferability where the two vehicles in question have very different target altitudes and the images they capture have very different typical spatial scales.

There is also a clear decrease in performance when applying $D1$ over $D0$, this is in line with the findings of $\mathcal{H}^{train} \rightarrow \mathcal{H}^{test}$. This is an unexpected finding for these experiments, as the lens distortion present in the high altitude images, $\mathcal{H}$, is much

TABLE 6.4: mAP scores and variance for low altitude vehicle, $\mathcal{H}^{train} \rightarrow \mathcal{L}^{test}$

				F0	F0	F1	F1
				I0	I1	I0	I1
C1	S1	D0	E0	0.37 ± 0.06	0.30 ± 0.06	0.34 ± 0.02	0.28 ± 0.05
C1	S1	D0	E1	0.35 ± 0.02	0.35 ± 0.03	0.27 ± 0.03	0.29 ± 0.01
C1	S1	D1	E0	0.22 ± 0.03	0.25 ± 0.01	0.22 ± 0.04	0.23 ± 0.06
C1	S1	D1	E1	0.26 ± 0.06	0.21 ± 0.02	0.23 ± 0.01	0.18 ± 0.02
C1	S3	D0	E0	0.40 ± 0.03	0.41 ± 0.01	0.39 ± 0.02	0.39 ± 0.01
C1	S3	D0	E1	0.41 ± 0.04	0.35 ± 0.02	0.40 ± 0.03	0.24 ± 0.17
C1	S3	D1	E0	0.44 ± 0.00	0.46 ± 0.01	0.40 ± 0.01	0.40 ± 0.03
C1	S3	D1	E1	0.38 ± 0.03	0.42 ± 0.05	0.42 ± 0.02	0.22 ± 0.16
C2	S1	D0	E0	0.33 ± 0.01	0.38 ± 0.02	0.31 ± 0.02	0.35 ± 0.03
C2	S1	D0	E1	0.33 ± 0.03	0.27 ± 0.01	0.29 ± 0.02	0.29 ± 0.02
C2	S1	D1	E0	0.20 ± 0.02	0.19 ± 0.04	0.19 ± 0.01	0.17 ± 0.04
C2	S1	D1	E1	0.26 ± 0.05	0.19 ± 0.01	0.24 ± 0.03	0.16 ± 0.03
C2	S3	D0	E0	$\mathbf{0.52 \pm 0.02}$	0.50 ± 0.00	0.48 ± 0.02	0.50 ± 0.03
C2	S3	D0	E1	0.49 ± 0.03	0.51 ± 0.01	0.46 ± 0.02	0.51 ± 0.01
C2	S3	D1	E0	0.49 ± 0.01	$\mathbf{0.52 \pm 0.01}$	$\mathbf{0.52 \pm 0.02}$	0.48 ± 0.03
C2	S3	D1	E1	0.51 ± 0.03	$\mathbf{0.54 \pm 0.02}$	$\mathbf{0.53 \pm 0.01}$	0.43 ± 0.02



(A) Split by colour correction

(B) Split by rescaling

(C) Split by distortion correction

(D) Split by extra augmentations

(E) Split by elastic augmentations

(F) Split by independent channel augmentations

FIGURE 6.3: mAP scores for transfer learning from high altitude to low altitude, $\mathcal{H}^{train} \rightarrow \mathcal{L}^{test}$, split by each independent variable, and aggregating over all other independent variables.

higher than that in the low altitude images, $\mathcal{L}$, leading to the expectation that correcting for this distortion would make the images more similar in appearance and would make transferability between them easier.

The statistical significance of these results is calculated with ANOVA test and shown in Tables 6.5 and 6.6. In Table 6.5 we see results backing up the observational comments above, with statistically significant effects from Scale with an F-value of

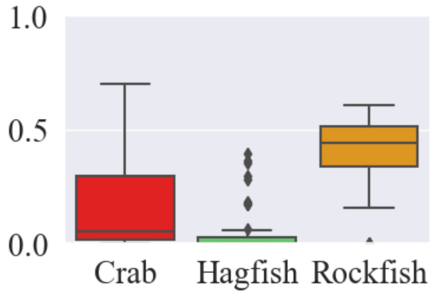


FIGURE 6.4: mAP scores for transfer learning from high altitude to low altitude, $\mathcal{H}^{train} \rightarrow \mathcal{L}^{test}$, split by class.

TABLE 6.5: One-way ANOVA test results for correction types, $\mathcal{H}^{train} \rightarrow \mathcal{L}^{test}$

Category	**F-value**	**p-value**
Colour Correction	9.86	0.00196
Distortion Correction	7.74	0.00593
Scale	225.49	$4.08e^{-34}$
Elastic transformations	1.22	0.27127
Separate Channel Operations	1.69	0.19514
Extra Augmentations	3.16	0.07714

TABLE 6.6: Two-way ANOVA test results for combined categories, $\mathcal{H}^{train} \rightarrow \mathcal{L}^{test}$

Combined Categories	**F-value**	**p-value**
Colour Correction and Distortion Correction	6.176	$5.028e^{-4}$
Colour Correction and Scale	124.579	$1.855e^{-44}$
Colour Correction and Elastic Transformations	3.878	0.010
Colour Correction and Separate Channel Operations	4.033	0.008
Colour Correction and Extra Augmentations	4.586	0.004
Distortion Correction and Scale	113.854	$4.671e^{-42}$
Distortion Correction and Elastic Transformations	3.052	0.030
Distortion Correction and Separate Channel Operations	3.217	0.024
Distortion Correction and Extra Augmentations	3.686	0.013
Scale and Elastic Transformations	76.815	$1.778e^{-32}$
Scale and Separate Channel Operations	77.070	$1.498e^{-32}$
Scale and Extra Augmentations	79.521	$2.944e^{-33}$
Elastic Transformations and Separate Channel Operations	1.427	0.236
Elastic Transformations and Extra Augmentations	1.570	0.198
Separate Channel Operations and Extra Augmentations	1.837	0.142

225.49 and a p-value of $4.08e^{-34}$, Colour Correction with an F-value of 9.86 and a p-value of 0.00196, and to a lesser extent Distortion Correction with an F-value of 7.74 and a p-value of 0.00593. As seen previously, combinations of these variables show promise of being more effective than either variable alone, with the combination with the highest impact in the two-way anova tests being Colour Correction and Scale, with an F-value of 124.579 and a p-value of $1.855e^{-44}$. This synergy may be because in combination with each other the images are better normalised than with either

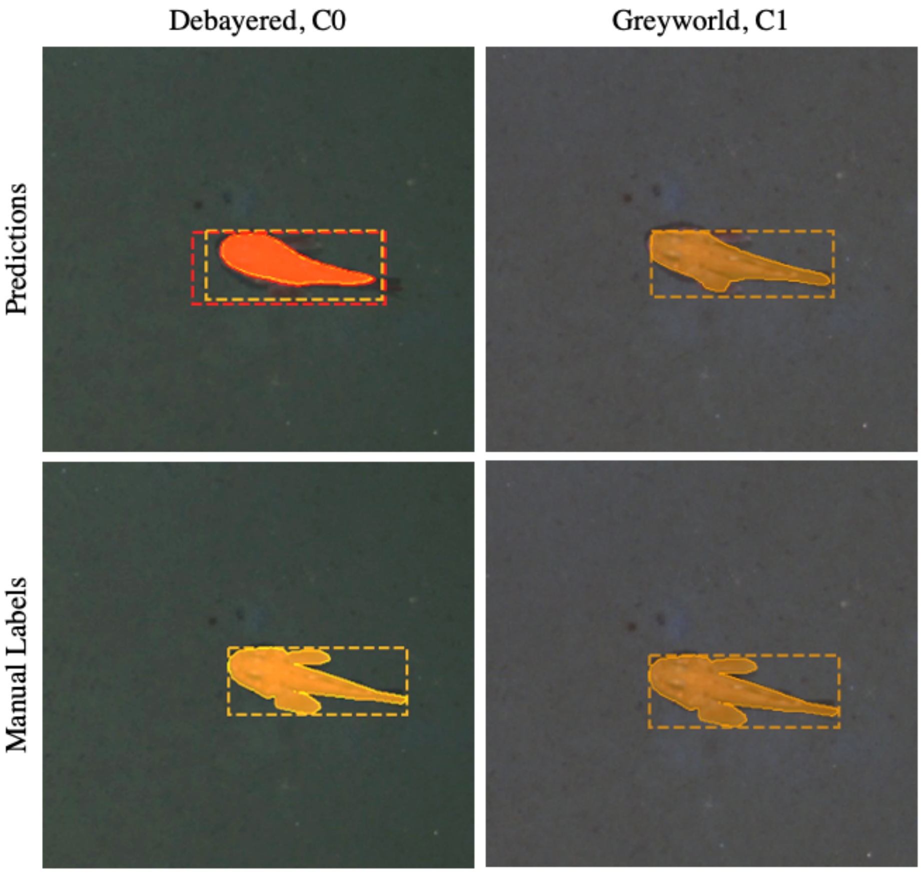


FIGURE 6.5: Example of predictions made by two instances of Mask R-CNN trained on $\mathcal{H}^{train}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$, and predicting for $\mathcal{L}^{testt}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{L}^{test}_{C1.S1.D0.E0.I0.A0}$, respectively. This example contains one instance of a Rockfish. Colours for labels use the legend set out in Figure 4.7.

correction type alone.

Figure 6.4 shows the breakdown of mAP scores by class, and again shows the lowest performance for the Hagfish class, and the highest performance for the Rockfish class. This is inline with all of the experiments in this thesis.

Figure 6.5 shows example predictions from two instances of Mask R-CNN trained on $\mathcal{H}^{train}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$ respectively, and the images being used for prediction are $\mathcal{L}^{test}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{L}^{test}_{C1.S1.D0.E0.I0.A0}$ respectively. This example shows the ability of these instances of Mask R-CNN to transfer from the high altitude images they were trained on to low altitude images taken by a different vehicle. This example contains a single instance of a Rockfish. Both of the instances of Mask R-CNN detect and segment this Rockfish, however the instance trained on C0 also wrongly detects an overlapping Crab. Neither instance produces good segmentation around the fins of the Rockfish - a feature that is likely less visible in the high altitude images they were

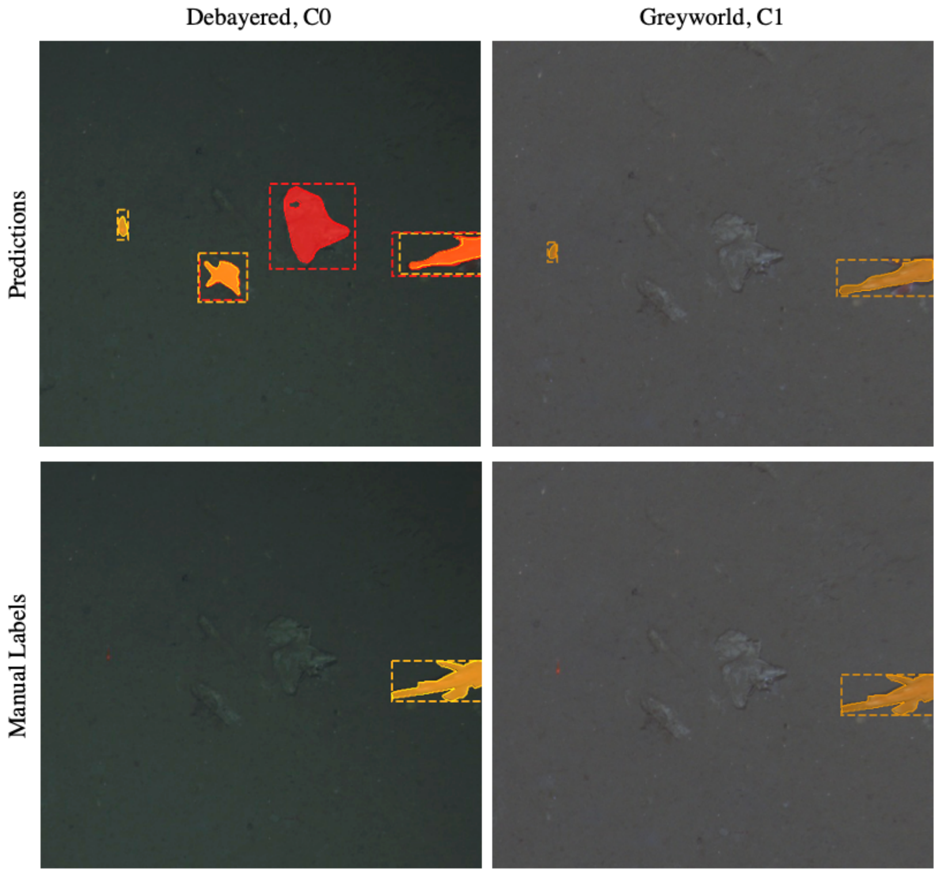


FIGURE 6.6: Example of predictions made by two instances of Mask R-CNN trained on $\mathcal{H}^{train}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$, and predicting for $\mathcal{L}^{test}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{L}^{test}_{C1.S1.D0.E0.I0.A0}$, respectively. This example contains one instance of a Rockfish, and other objects such as rocks and a shrimp that confuse these instances of Mask R-CNN. Colours for labels use the legend set out in Figure 4.7.

trained on - but the segmentation overall is good enough to count as a successful detection and segmentation.

Figure 6.6 shows exampe predictions from two instances of Mask R-CNN trained on $\mathcal{H}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{H}_{C1.S1.D0.E0.I0.A0}$ respectively. These predictions are for $\mathcal{L}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{L}_{C1.S1.D0.E0.I0.A0}$ images respectively, and show the transferability of these Mask R-CNN instances. This example contains one instance of a Rockfish, which both instances of Mask R-CNN successfully detect and segment, however the instance trained on C0 also incorrectly detects a Crab overlapping the Rockfish, and mistakes rocks and a shrimp in the image for two Crabs and two Rockfish, greatly overestimating the total population count in this example. The instance trained on $C1$ also incorrectly detects the shrimp as a Rockfish, so also overestimates the population count and biomass estimate, but not to the same degree as the instance trained on C0.

Figure 6.7 shows the final example of predicted labels from two instances of Mask R-CNN where the only difference in these instances is the colour correction method

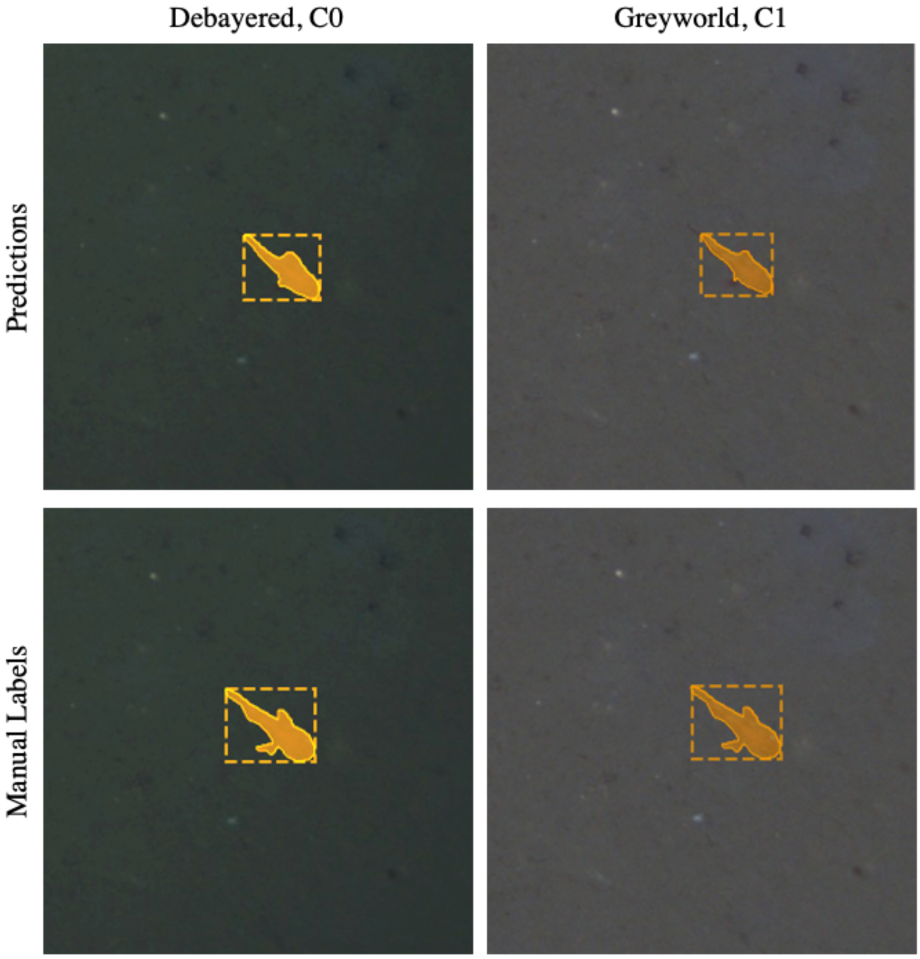


FIGURE 6.7: Example of predictions made by two instances of Mask R-CNN trained on $\mathcal{H}^{train}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$, and predicting for $\mathcal{L}^{test}_{C0.S1.D0.E0.I0.A0}$ and $\mathcal{L}^{test}_{C1.S1.D0.E0.I0.A0}$, respectively. This example contains one instance of a Rockfish that is successfully detected and segmented by both instances of Mask R-CNN. Colours for labels use the legend set out in Figure 4.7.

applied to the training data. This example contains a single instance of a Rockfish, and, unlike the earlier example in Figure 6.5, both instances of Mask R-CNN successfully detect and segment it. Neither instance fully segments the shape of the fins correctly, but correctly identifies the majority of the area of the fish.

Figure 6.8 shows example labels from two instances of Mask R-CNN trained on $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S2.D0.E0.I0.A0}$ respectively, where the difference in these two instances is the scale normalisation applied to the training images. This example contains a single instance of the Hagfish class. The instance of Mask R-CNN trained on $S1$, original scale images, incorrectly identifies the Hagfish as a Crab. The instance of Mask R-CNN trained on $S2$ images, scale normalised to the low altitude spatial scale, incorrectly identidies the Hagfish as both a Hagfish and a Crab overlapping, and fails to correctly segment the shape of the Hagfish, leaving a hole in the middle of the fish.

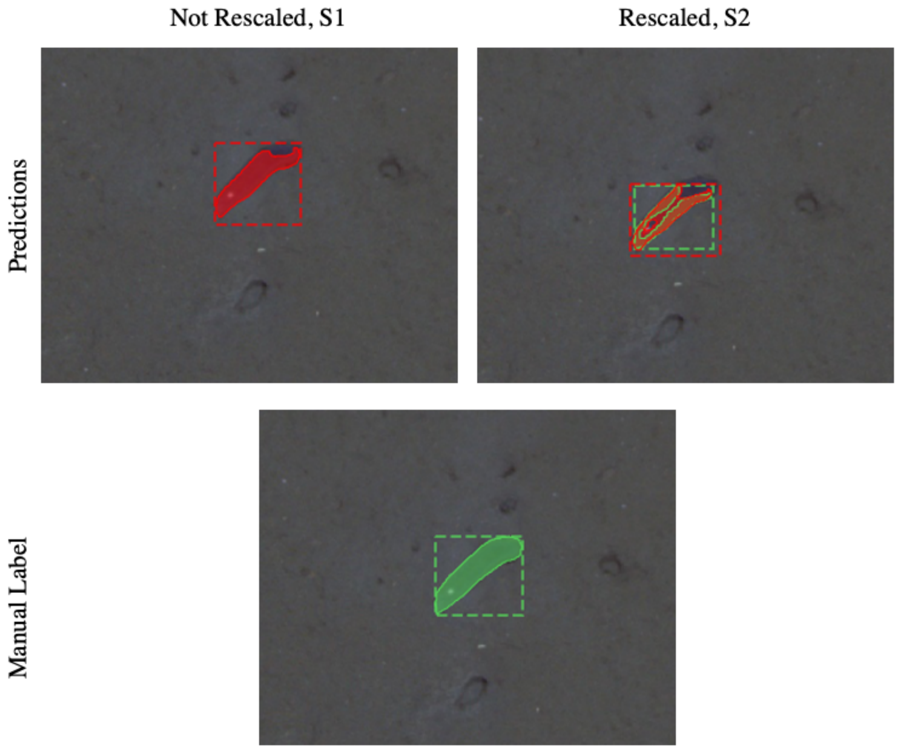


FIGURE 6.8: Example of predictions made by two instances of Mask R-CNN trained on $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S2.D0.E0.I0.A0'}$ and predicting for $\mathcal{L}^{test}_{C1.S1.D0.E0.I0.A0}$ and $\mathcal{L}^{test}_{C1.S3.D0.E0.I0.A0'}$ respectively. This example contains one instance of a Hagfish. Colours for labels use the legend set out in Figure 4.7.

The estimate made by the instance trained on $S1$ would contribute to underestimating the biomass and population count for the Hagfish class and overestimating the biomass and population count for the Crab class. The estimate made by $S2$ would correctly estimate the population count for the Hagfish but may underestimate this instance's biomass due to the incorrect segmentation result. It would also contribute to overestimating the biomass and population count of the Crab class.

Figure 6.9 shows another example of predicted labels from two instances of Mask R-CNN where the difference between these two instances is the scale normalisation method used. This example contains a single instance of a Rockfish, which both instances of Mask R-CNN correctly identify and segment, although the instance trained on $S1$ does not correctly segment the fins of the fish. Both instances, however, also identify an extra object within the image. The instance trained on $S1$, with no scale normalisation, incorrectly identifies the shrimp in the image as a Rockfish, perhaps due to the orange colour and the similarities in shape to many Rockfish examples. Where scale is a major difference between the shrimp and the Rockfish in the collected images, an instance of Mask R-CNN trained on varying scaled images has less reliable scale information to train on, and it less likely to be capable of differentiating classes based on scale. The instance of Mask R-CNN trained on $S2$,

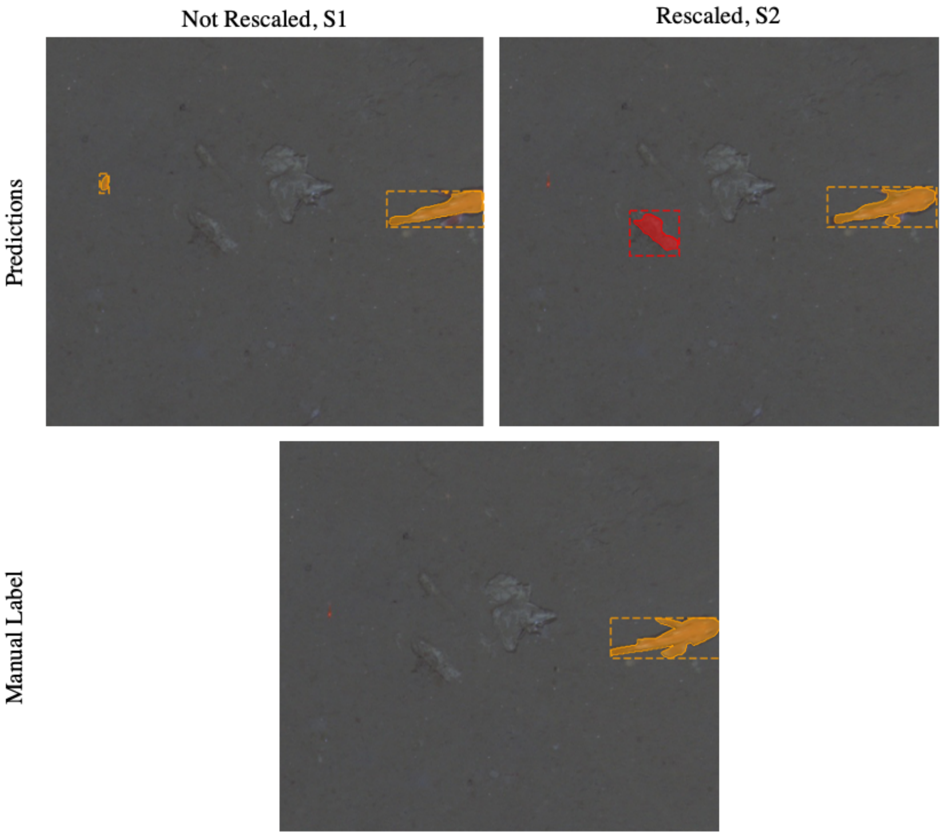


FIGURE 6.9: Example of predictions made by two instances of Mask R-CNN trained on $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S2.D0.E0.I0.A0}$, and predicting for $\mathcal{L}^{test}_{C1.S1.D0.E0.I0.A0}$ and $\mathcal{L}^{test}_{C1.S3.D0.E0.I0.A0}$, respectively. This example contains one instance of a Rockfish. Colours for labels use the legend set out in Figure 4.7.

scale normalised to the low altitude spatial scale, incorrectly identifies a rock in the image as a Crab. The prediction made by the instance $S1$ would over predict the population count and biomass estimate for the Rockfish class, and the prediction made by the instance $S2$ would correcly estimate the population count and biomass of the Rockfish class, and would overestimate for the Crab class.

Figure 6.10 shows the final example of predicted outputs from Mask R-CNN where the difference in the two instances is the rescaling method used on the training data, $S1$ being images at their original scale, and $S2$ being images normalised to the average low altitude spatial scale. In this example there are two instances of the Rockfish class. The instance of Mask R-CNN trained on $S1$ images correctly identifies both of the Rockfish, however it also identifies an overlapping instance of a Crab on one of the Rockfish, and on the other it identifies an overlapping Crab and an overlapping smaller Rockfish. The instance of Mask R-CNN trained on $S2$ also correctly identifies both of the instances of Rockfish, but also identifies one overlapping instance of a Crab on one of them. Both of these predictions lead to overestimations but to different degrees. The instance trained on $S1$ would overestimate biomass and population

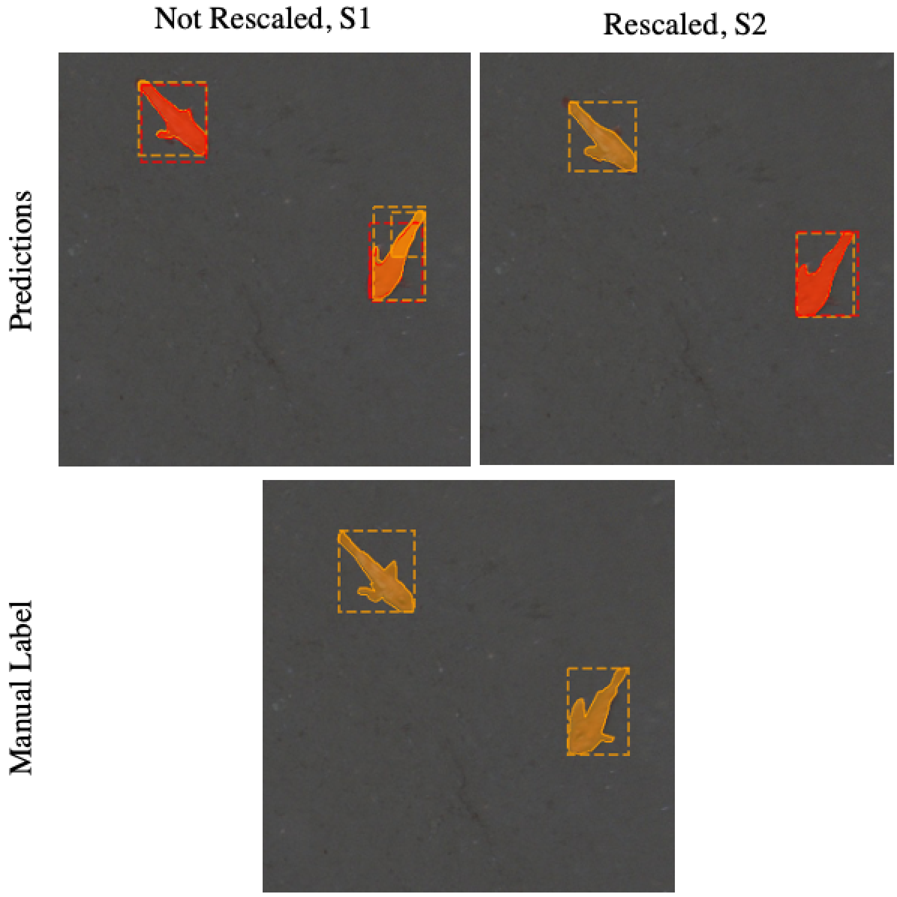


FIGURE 6.10: Example of predictions made by two instances of Mask R-CNN trained on $\mathcal{H}^{train}_{C1.S1.D0.E0.I0.A0}$ and $\mathcal{H}^{train}_{C1.S2.D0.E0.I0.A0}$, and predicting for $\mathcal{L}^{test}_{C1.S1.D0.E0.I0.A0}$ and $\mathcal{L}^{test}_{C1.S3.D0.E0.I0.A0}$, respectively. This example contains two instances of Rockfish. Colours for labels use the legend set out in Figure 4.7.

counts for both the Rockfish and Crab classes, and the instance trained on $S2$ would overestimate the biomass and population counts for the Crab class.

To conclude,$\mathcal{H} \rightarrow \mathcal{L}$ out performed $\mathcal{L} \rightarrow \mathcal{H}$, and the most significant variables were colour correction and rescaling, where applying colour correction and rescaling provided the best results. There is much further work that should be done to rule out the importance of other variables on performance, such as distortion correction and elastic augmentations, where other methods not used in this thesis may provide better results.

Chapter 7

Biomass Estimates from Object Detection and Segmentation Results

Automated image analysis is a useful tool to ease deep sea monitoring efforts, but is only useful if it can provide meaningful information about the surveyed area. For this reason, this thesis presents a novel method for estimating biomass from the output of Mask R-CNN and similar object detection and segmentation systems that have been discussed in previous chapters in this thesis. Image analysis for deep sea monitoring provides a non-invasive alternative to previous methods such as trawling. The benefit of a non-invasive method is that often the areas of most interest for surveying are protected areas where trawling either is not allowed, or is likely to disturb and destroy valuable and vulnerable habitats and ecosystems. Image analysis, however, has limitations when it comes to calculating Essential Ocean Variables (EOVs) such as biomass, where traditionally specimens from trawling would be weighed.

TABLE 7.1: Segment-Size Length Relationship, calculated from collected data for this thesis - cm^2 to cm, Length Weight Relationships, gathered from the large open source database Fishbase [5] - cm to $grams$, and derived Segment-Size Weight Relationships - cm^2 to $grams$, repeated from Table 3.2

Class Name	SLR	LWR	SWR
Crab	$2.6x^{0.38}$	$0.00036x^{2.92}$	$0.0058x^{1.11}$
Hagfish	$5.93x^{0.52}$	$0.0048x^{2.72}$	$0.61x^{1.41}$
Rockfish	$4.27x^{0.5}$	$0.012x^{3.08}$	$1.08x^{1.54}$
Soles	$2.74x^{0.54}$	$0.0074x^{3.09}$	$0.17x^{1.669}$
Seastar	$1.82x^{0.58}$	$0.00032x^{2.43}$	$0.0014x^{1.41}$

Efforts have been made to bridge this gap and estimate EOVs from images through the use of manual labelling and Length Weight Relationships (LWRs), drawing a relationship between the size of an individual and its weight, but they still suffer from a lack of scalability due to the manual analysis needed. The method developed as part of this thesis uses previously established LWRs, and newly calculated Segment Size to

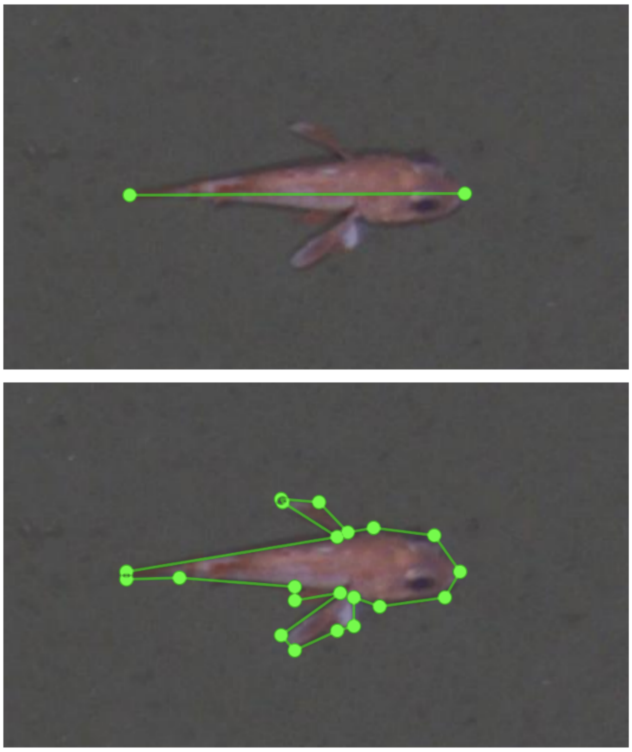


FIGURE 7.1: Example of a Rockfish labelled for length and for segment size.

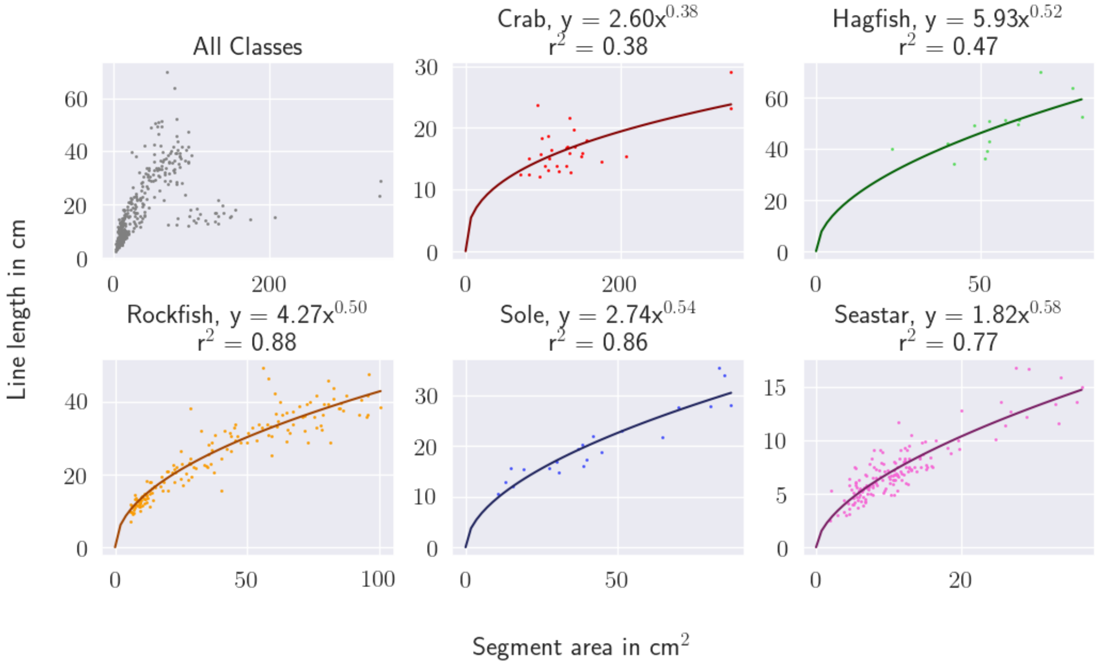


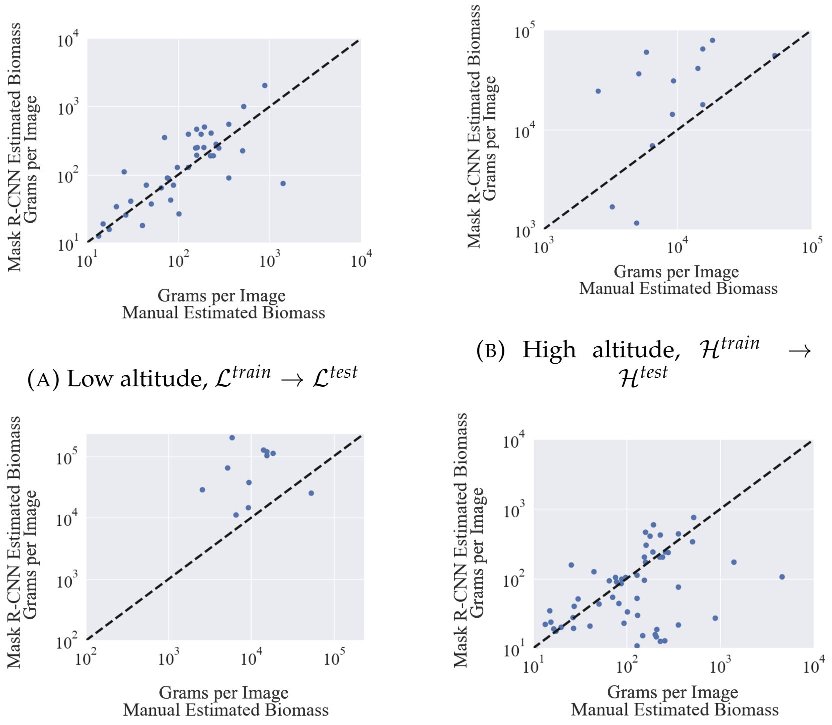
FIGURE 7.2: Estimated segment size to length relationships. The relationships are calculated using regression on measured segment sizes and line lengths, resulting in an estimated polynomial relationship. This is repeated from Figure 3.4.

Length Relationships (SSLRs), to estimate the relationship between segment sizes and weights of individuals in images. Manual labels for lengths, along with the manual labels already created for the previous experiments outlined in this thesis, were used to develop the SSLRs, as shown in Figure 7.1. The length and area values were calculated from pixel values by using the opening angle of the camera and the altitude of the vehicle off the seafloor. The polynomial relationships calculated from these labels are presented in Chapter 3, but are worth repeating here, in Table 7.1, and Figure 7.2.

This method allows for the automation of the entire process providing a fully scalable solution. Other automated systems have been investigated before which often rely on estimating the length measurements through machine learning. This form of machine learning is very domain specific, and has not been developed by the wider machine learning community. Furthermore, determining the length measurement of an object that may be curved or in an unexpected position poses a very complex problem. One benefit of the solution developed for this thesis is that it makes use of the very well developed non-domain-specific field of object detection and segmentation, reducing the complexity and increasing the reliability of the machine learning algorithms available.

Furthermore, the relationship between length and weight may not be as strong as the direct relationship between segment size and weight, forming a relationship between a two dimensional measurement and a three dimensional target metric, rather than a relationship between a one dimensional measurement and a three dimensional target metric. This novel method for estimating biomass based on segment sizes incentivises further investigation into the relationship between segment sizes when viewed from above and biomass, which may prove to be a more accurate relationship than traditional LWRs, and more easily scalable due to machine learning. For the purposes of this study, however, the combination of SSLRs and LWRs is used to estimate this relationship.

In this chapter, two individually trained instances of Mask R-CNN are used to generate biomass estimates for the test datasets, one trained on $\mathcal{L}^{train}_{C2.S2.D0.E0.I0.A0}$ and the other on $\mathcal{H}^{train}_{C2.S3.D0.E0.I0.A0}$. Each of these two instances were then used to estimate the biomass for both $\mathcal{L}^{test}_{C2.S2.D0.E0.I0.A0}$ and $\mathcal{H}^{test}_{C2.S3.D0.E0.I0.A0}$, resulting in four distinct sets of results, $\mathcal{L}^{train}_{C2.S2.D0.E0.I0.A0} \rightarrow \mathcal{L}^{test}_{C}2.S2.D0.E0.I0.A0$, $\mathcal{L}^{train}_{C2.S2.D0.E0.I0.A0} \rightarrow \mathcal{H}^{test}_{C2.S3.D0.E0.I0.A0}$, $\mathcal{H}^{train}_{C2.S3.D0.E0.I0.A0} \rightarrow \mathcal{L}^{test}_{C2.S2.D0.E0.I0.A0}$, and $\mathcal{H}^{train}_{C2.S3.D0.E0.I0.A0} \rightarrow \mathcal{H}^{test}_{C2.S3.D0.E0.I0.A0}$. The biomass estimates formed from SSLRs and LWRs combined are then compared with the estimates formed from LWRs alone, as shown in Figure 7.3. The relationships for low altitude datasets are most promising, especially for the $\mathcal{L}^{train}_{C2.S2.D0.E0.I0.A0} \rightarrow \mathcal{L}^{test}_{C2.S2.D0.E0.I0.A0}$ experiment where the accuracy of Mask R-CNN is highest, showing that improving the accuracy of the object detection and segmentation system improves the accuracy of subsequent biomass estimations.



(A) Low altitude, $\mathcal{L}^{train} \rightarrow \mathcal{L}^{test}$

(B) High altitude, $\mathcal{H}^{train} \rightarrow \mathcal{H}^{test}$

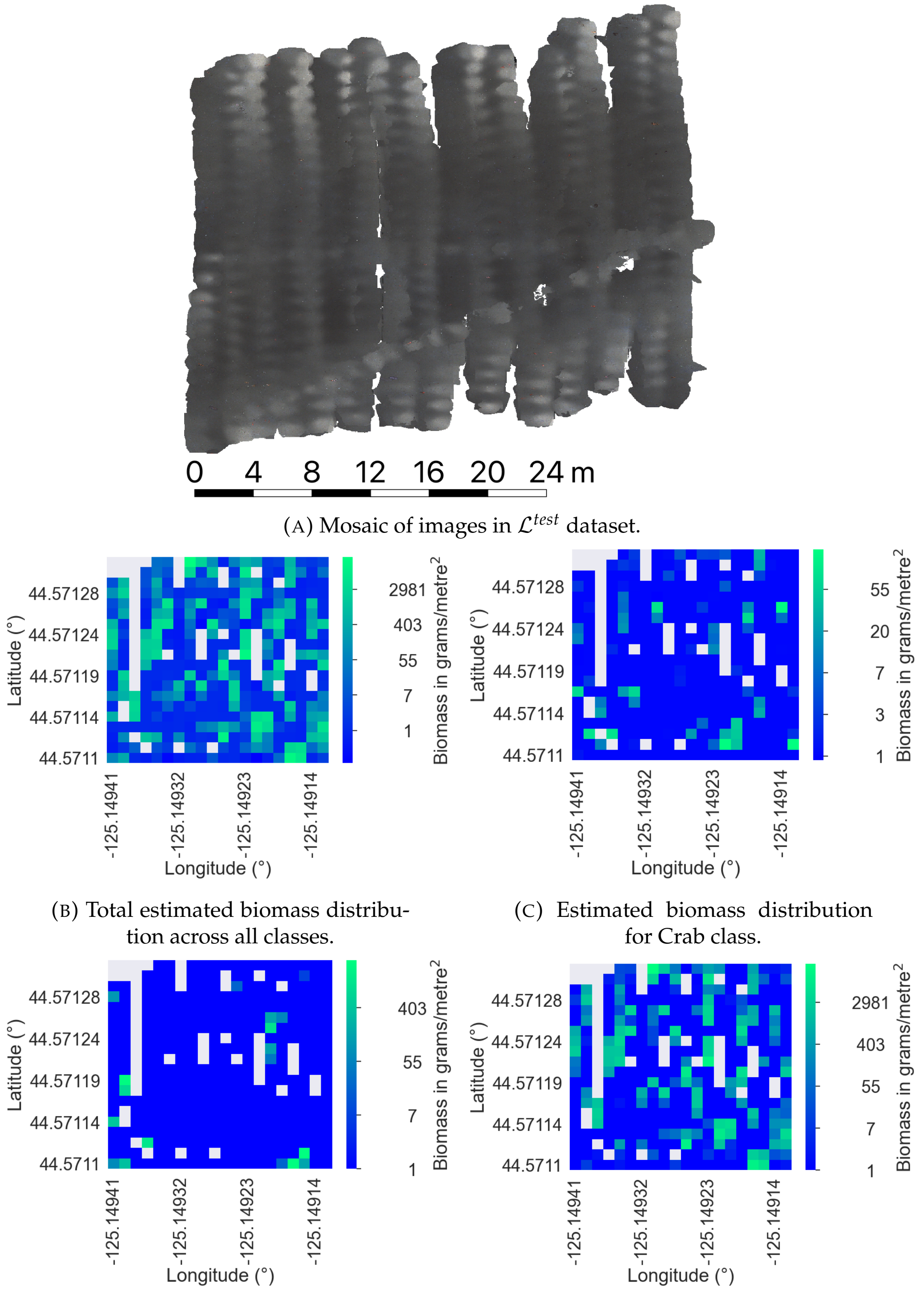
(C) Low altitude to High Altitude transfer learning, $\mathcal{L}^{train} \rightarrow \mathcal{H}^{test}$

(D) High altitude to Low Altitude transfer learning, $\mathcal{H}^{train} \rightarrow \mathcal{L}^{test}$

FIGURE 7.3: Mask R-CNN estimated biomass per image plotted against the manual label estimated biomass. The strongest correlation is shown in $\mathcal{L}^{train} \rightarrow \mathcal{L}^{test}$, where the estimated biomass from each method is most similar. In $\mathcal{L}^{train} \rightarrow \mathcal{H}^{test}$ and $\mathcal{H}^{train} \rightarrow \mathcal{H}^{test}$, there are fewer points due to the smaller number of high altitude images needed to cover the same spatial area.

The most important aspect of this approach is its scalability, with the ability to analyse any number of images with the same level of manual effort, allowing the analysis of all the images collected in a dive as opposed to the small subset often used, and allowing the analysis of multiple datasets, with the only limit being computational power and data storage capabilities. Furthermore, the most computationally expensive part of the process is the training of Mask R-CNN, compared to the less computationally intense inference stage, allowing the production of biomass distribution plots such as those shown in Figures 7.4, 7.5, and 7.6 to feasibly be done in the field, on board ships, where available computational power may be relatively low compared to on shore.

Figures 7.4 and 7.5 show example biomass distribution estimates over an entire surveyed area during a single dive of the low altitude Tunasand vehicle. It covers an area of roughly 32 by 24 metres. Figure 7.4 is an estimate made by an instance of Mask R-CNN trained on $\mathcal{L}^{train}_{C2.S2.D0.E0.I0.A0}$. Subfigure 7.4a shows the photo mosaic of the AUV dive, and subfigure 7.4b shows the total biomass distribution estimate across all three of the given classes.



(A) Mosaic of images in $\mathcal{L}^{test}$ dataset.

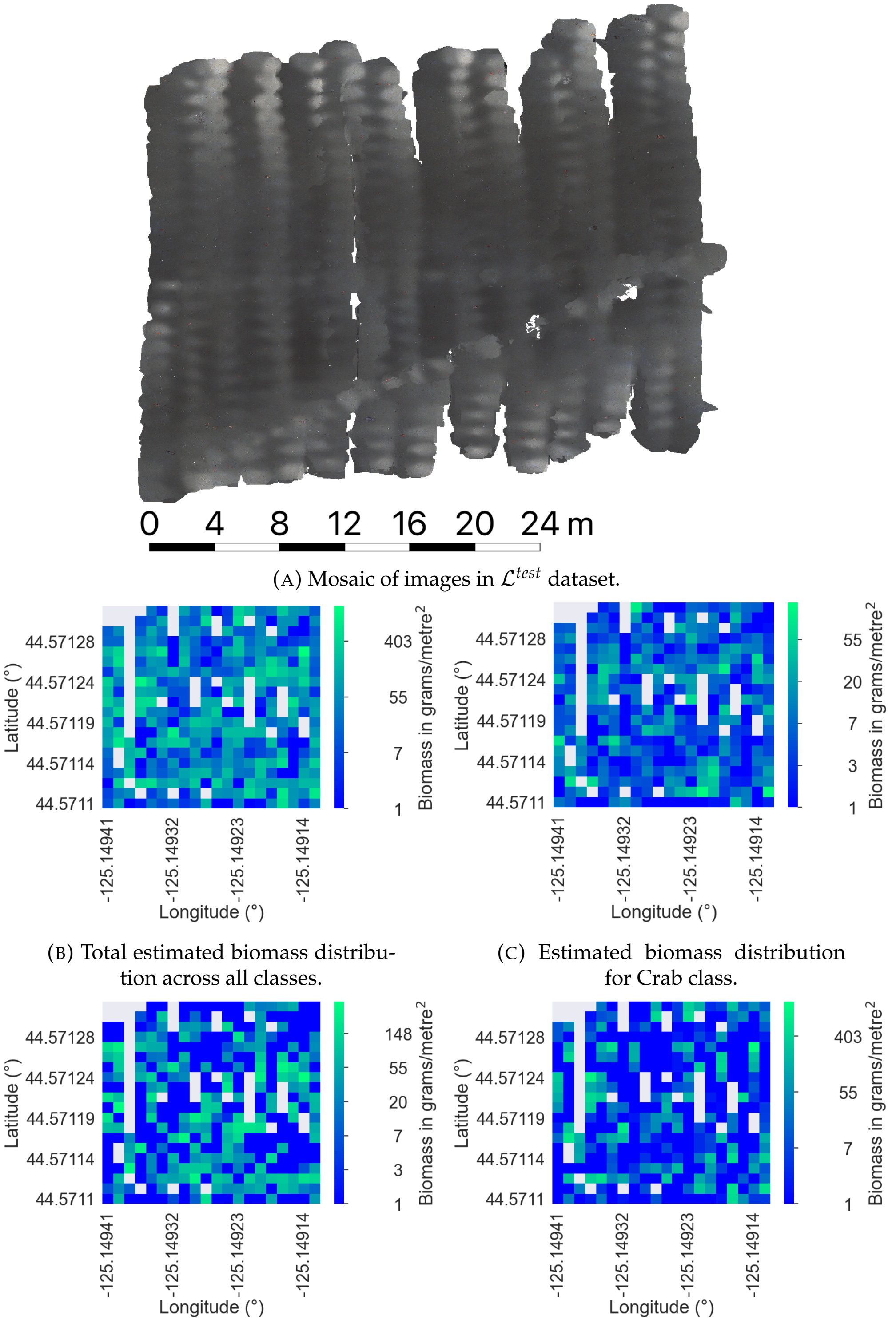
(B) Total estimated biomass distribution across all classes.

(C) Estimated biomass distribution for Crab class.

(D) Estimated biomass distribution for Hagfish class.

(E) Estimated biomass distribution for Rockfish class.

FIGURE 7.4: Biomass distributions estimate from a single instance of Mask RCNN trained on $\mathcal{L}^{train}_{C2.S2.D0.E0.I0.A0}$ and estimating on $\mathcal{L}^{test}_{C2.S2.D0.E0.I0.A0}$



(A) Mosaic of images in $\mathcal{L}^{test}$ dataset.

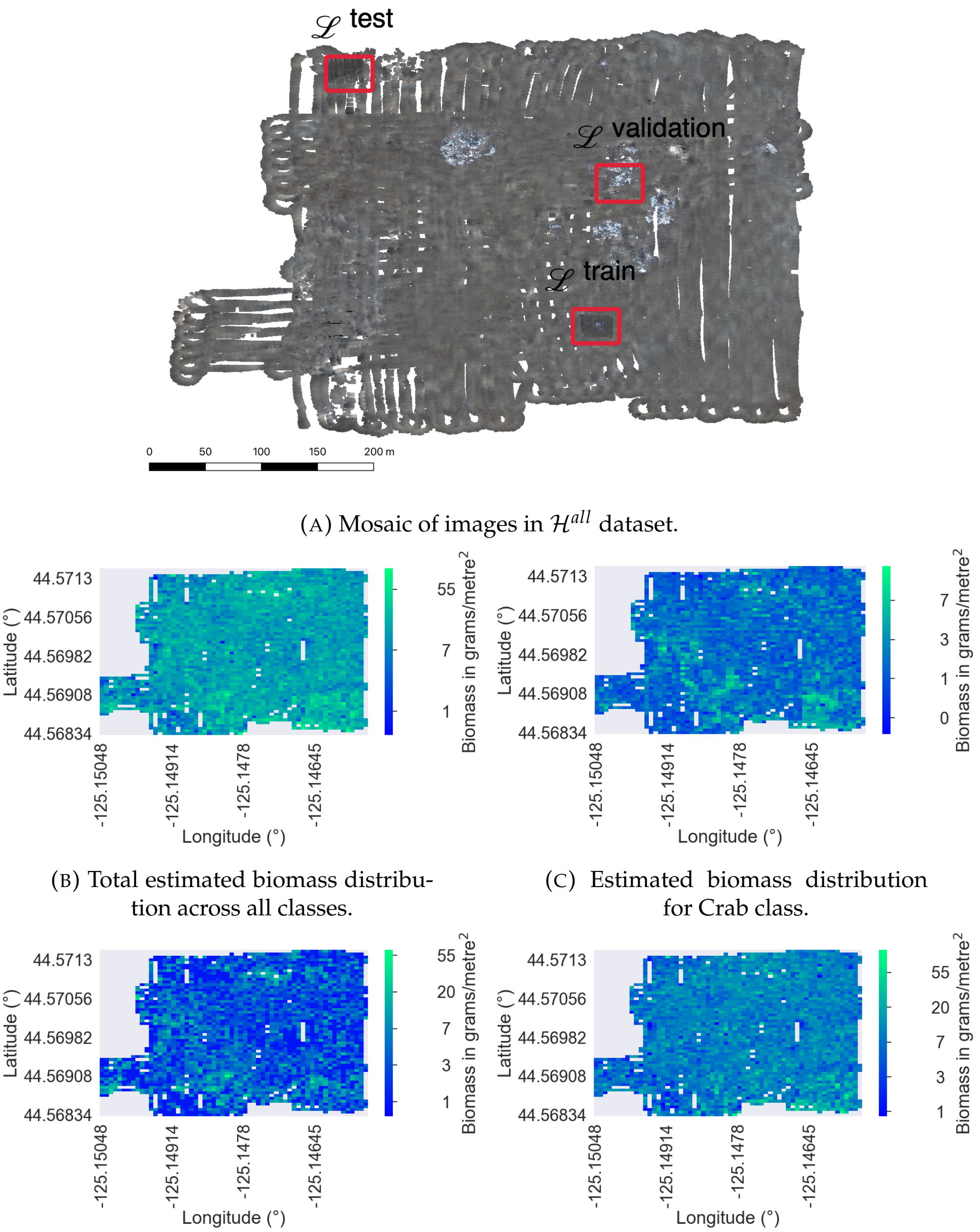
(B) Total estimated biomass distribution across all classes.

(C) Estimated biomass distribution for Crab class.

(D) Estimated biomass distribution for Hagfish class.

(E) Estimated biomass distribution for Rockfish class.

FIGURE 7.5: Biomass distributions estimate from a single instance of Mask RCNN trained on $\mathcal{H}^{train}_{C2.S3.D0.E0.I0.A0}$ and estimating on $\mathcal{L}^{test}_{C2.S2.D0.E0.I0.A0}$



(A) Mosaic of images in $\mathcal{H}^{all}$ dataset.

(B) Total estimated biomass distribution across all classes.

(C) Estimated biomass distribution for Crab class.

(D) Estimated biomass distribution for Hagfish class.

(E) Estimated biomass distribution for Rockfish class.

FIGURE 7.6: Biomass distributions estimate from a single instance of Mask RCNN trained on $\mathcal{H}^{train}_{C2.S3.D0.E0.I0.A0}$ and estimating on entire $\mathcal{H}$ dataset, covering entire surveyed area. Higher density clusters are visible both in the overall distribution and in each class.

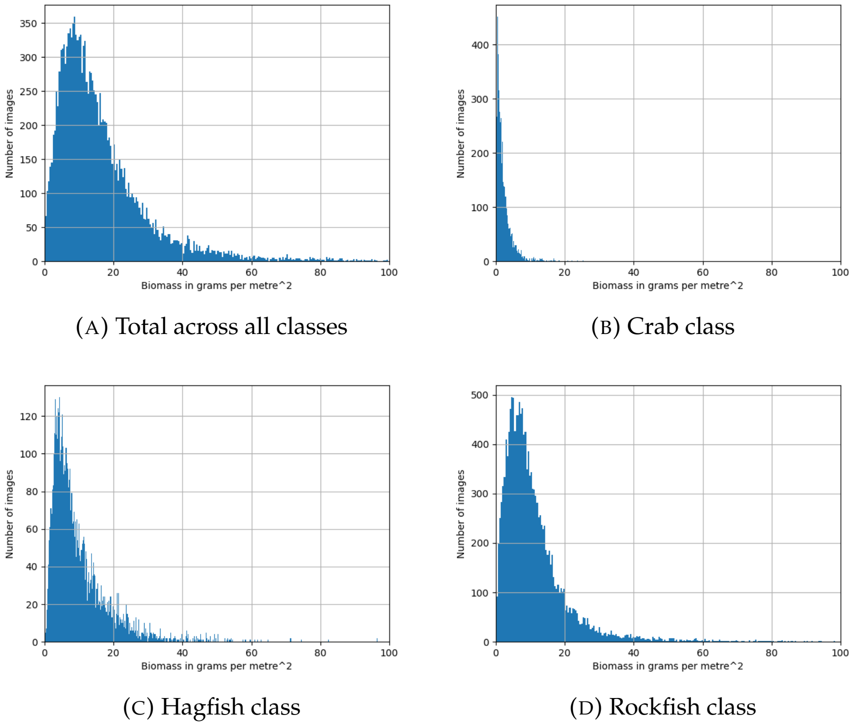
A breakdown per class is shown in subfigures 7.4c, 7.4d, and 7.4e. Both in the breadth of the distribution and in the scale of the colour map, it is clear that the Rockfish contributes the most to the overall biomass distribution, with the other classes being more localised and much smaller in terms of grams per metre^2. This is as expected where the Rockfish is by far the most abundant class in the surveyed area.

Figure 7.5 shows an estimate made by an instance of Mask R-CNN trained on high altitude data and estimating for low altitude data, $\mathcal{H}^{train}_{C2.S3.D0.E0.I0.A0} \rightarrow \mathcal{L}^{test}_{C2.S2.D0.E0.I0.A0}$, showing an higher estimate for both the Crab and the Hagfish classes, and an lower for the Rockfish class compared to the estimates made by $\mathcal{L}^{train}_{C2.S2.D0.E0.I0.A0} \rightarrow \mathcal{L}^{test+}_{C2.S2.D0.E0.I0.A0}$.

In order to demonstrate the scalability of this solution, it has been applied to all four dives made by the high altitude vehicle in the South Hydrates area in the Falkor 2018 Adaptive Robotics Expedition, covering approximately $200km^2$. The instance of Mask R-CNN in this case has been trained on $\mathcal{H}^{train}$ and is estimating for $\mathcal{H}^{all}$ The results of this are shown in Figure 7.6, with the full image mosaic in Figure 7.6a, the biomass density across all classes in Figure 7.6b, and the next three subfigures showing the breakdown by class. This shows the majority of the biomass contribution comes from Rockfish, the most abundant class, in keeping with the findings from the low altitude biomass estimations. There are areas of higher density for the Rockfish class visible in these heatmaps. The distribution of biomass is more varied for the other two classes, Hagfish and Crabs. These two classes have similar distributions with similar areas of high and low density. Findings such as this allow marine biologists to pose and perhaps even answer questions such as whether this is due to the morphotypes being comfortable in similar environments, or being drawn to the same sources of food. For expeditions concerned with a particular class of organism, set of classes, or the relationship between two classes, this sort of information from a high altitude vehicle could inform further AUV deployment plans in-situ, especially the deployment of low altitude vehicles covering much smaller areas in much higher resolution. Making more informed decisions such as this results in more valuable information collected in the same time span on a research expedition, generating more valuable scientific outputs without increasing the amount of expensive ship time required.

Another way to visualise the results is demonstrated in Figure 7.7, showing histograms of the biomass density per image, both as a total of all classes and broken down into individual classes.

This method becomes even more valuable when combined with existing information about an area, such as the map shown in Figure 7.8, developed by Yamada et al. [103] showing the habitat classification of each image in the dataset. This combination of information allows for the analysis of the relationship between biomass distributions and habitat types, in this case showing a higher density of Hagfish and Crabs on the



(A) Total across all classes

(B) Crab class

(C) Hagfish class

(D) Rockfish class

FIGURE 7.7: Histograms of biomass density per image, over all and split by class.

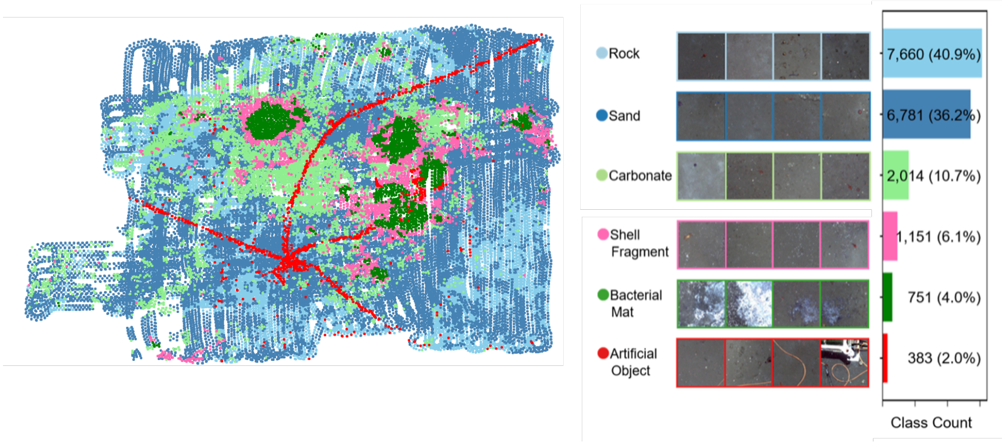


FIGURE 7.8: Substrate map of area produced by Yamada et al. [4], as an example of external datasets that can be combined with biomass density plots to form more interesting conclusions.

substrate classified as Rocky in this particular study. Again, findings such as this allow marine biologists to pose questions about this correlation and investigate possible causes.

A major difficulty in developing this method further and improving trust in this method, is the lack of validation data. The acquisition of such data is challenging, but entirely possible. Images taken from above with either a known distance to the subject and known opening angle of the camera, or with a known scale of the subject, for example with a ruler at the same distance as the subject visible in the images, and then the weights of the subjects, would need to be recorded. This is more invasive than AUV imagery alone, as to record the weight physical samples need to be taken. This data would be valuable in validating the accuracy of this method, and the collection of such a dataset is possible future work to build on this thesis.

To conclude, this chapter has introduced and demonstrated a novel method of biomass estimation through the use of segment sizes, which as demonstrated earlier in this thesis can be automatically estimated with relatively high accuracy with limited labelled training data. As such, this method is fully automated and fully scalable, providing biomass distribution estimates that make use of every single image collected with no extra manual input required to increase the number of images being utilised. With further development of direct relationships between segment size and weight or biomass, this method will become even more accurate and more effective. A possible validation dataset has been proposed. Furthermore, the computational cost of applying this method is relatively low and can be applied in-situ on board ships and, with further optimisation and improving hardware, even on AUVs themselves.

Chapter 8

Discussions

This discussions chapter delves into several key aspects related to the challenges and considerations encountered in the field of marine imagery analysis using machine learning approaches. This chapter aims to explore various topics that arise when transferring knowledge between datasets, analysing performance variations across different classes, addressing class imbalance, handling unseen classes, and the concept of an ideal universal marine imagery dataset. Each section addresses specific intricacies and complexities associated with these aspects, providing valuable insights and potential avenues for future research and improvement.

8.1 Low Transferability from Low Altitude Training Data to High Altitude Test Data

One result of interest is the incredibly low performance when transferring from low altitude to high altitude data. This disagrees with current literature on the topic, where Zurowietz et al.[87] found that transferring from Low to High altitude imagery had better results than the inverse, transferring from high altitude to low altitude data. Possible reasons for this low performance are discussed in this section.

One possible reason for the low transferability from low altitude training data to high altitude test data is the difference in altitude investigated between the two studies. The work of Zurowietz et al. suggests that transferring from low altitude to high altitude imagery yields better results than the reverse scenario. However, in this thesis, the datasets used involve high altitude data at four times the altitude of the original data. This significant difference in altitude may contribute to the observed low performance in transferring knowledge.

When Zurowietz et al. transferred from low altitude to high altitude data at two times the altitude of the original data, they might have captured a more gradual and manageable transition. On the other hand, the fourfold increase in altitude used in this thesis could result in a more drastic change in the visual appearance of the scenes. The algorithms trained on low altitude data may struggle to generalise effectively to the high altitude test data due to the substantial differences in the characteristics and features of the images.

Another potential source of discrepancy between the conflicting results lies in the datasets used. The specific characteristics and properties of the datasets can greatly influence the performance and generalisability of the trained models. It is important to consider factors such as resolution, quality, diversity, and representativeness of the data when comparing the outcomes of different studies.

In this thesis, random cropped sections of each image were utilised for training, validation, and testing. On the contrary, Zurowietz et al. employed a sampling method centered around Objects Of Interest (OOIs). This means that in their study, there is an object of interest positioned in the center of every training sample. In contrast, the experiments conducted in this thesis might have incorporated a broader range of scenes without necessarily focusing on specific objects.

The difference in sampling methods can have a significant impact on the learned representations and the subsequent transferability of knowledge. By centering the training samples on OOIs, Zurowietz et al. might have biased the training process towards object-centric features and structures. This targeted sampling strategy could enhance the model's ability to recognise and transfer object-related information. In contrast, the random cropping approach employed in this thesis might have captured a more diverse set of scene contexts, which could influence the model's generalisation capabilities differently.

These differences in experimental setups and dataset characteristics highlight the importance of careful consideration when comparing and interpreting results across different studies. Factors such as altitude difference and sampling methods can significantly affect the performance and transferability of models trained on low altitude data and tested on high altitude scenarios. Future research could investigate these factors more comprehensively to gain deeper insights into the challenges and opportunities associated with transferring knowledge between altitudes.

8.2 Variance in Performance By Class

An interesting observation in the detection and segmentation experiments is the variation in performance across different classes. Among the classes examined, Rockfish demonstrated high performance, while Crabs exhibited slightly lower performance overall. On the other hand, Hagfish displayed the lowest performance. This section delves into the factors that contribute to these performance variations.

Rockfish exhibited high performance in the transferability experiments. Several factors can explain this favourable outcome. Firstly, Rockfish are abundant in the dataset, allowing the model to encounter a substantial number of instances during training. The availability of a large number of training samples facilitates the learning process, enabling the model to capture and generalise the distinguishing characteristics of Rockfish effectively.

Furthermore, Rockfish possess a high contrast with the background. Their distinctive orange colour stands out prominently, making them visually distinguishable from the surrounding environment. This high contrast simplifies the task for the model, as it can rely on the salient visual cues to identify and classify Rockfish instances accurately.

Additionally, Rockfish exhibit a relatively simple shape compared to the other classes. The presence of fewer intricate details reduces the complexity of the segmentation task, leading to improved performance.

Crabs displayed slightly lower performance compared to Rockfish in the transferability experiments. This discrepancy can be attributed to several factors. Firstly, crabs are less abundant in the dataset, resulting in a smaller number of training instances. The limited availability of training samples may hinder the model's ability to learn and generalise the defining characteristics of crabs effectively.

Moreover, Crabs possess a more complex shape compared to Rockfish. The intricate structure of their body, including the presence of legs and pincers, adds complexity to the classification task. The increased number of distinctive features and their spatial arrangement requires the model to capture a broader range of visual cues, which can pose challenges in accurate classification.

However, despite these challenges, Crabs still exhibit a relatively high contrast with the background in terms of colour. The distinct coloration of crabs enables visual differentiation from the surrounding environment. This colour contrast provides a valuable cue for the model to distinguish crabs from the background, contributing to their reasonable performance.

Hagfish showed the lowest performance among the examined classes, primarily due to their distinct characteristics. Hagfish have long, eel-like bodies with intricate shapes, making them visually challenging to capture and recognise accurately. Their behaviours, such as knotting themselves and burying themselves, introduce additional complexity, as these actions alter their appearance and visual cues. These unique behaviours and body structure hinder the model's ability to generalise across different configurations and positions of Hagfish, resulting in reduced performance.

Furthermore, Hagfish's dark colouration adds to the difficulty of accurate classification. Their dark tones make them blend with the background, reducing the contrast and visual distinction between Hagfish and their surroundings. The lower contrast negatively impacts the model's ability to differentiate Hagfish instances accurately, contributing to the observed lower performance in the experiments.

8.3 Addressing Class Imbalance

Class imbalance is a common challenge in machine learning tasks, including marine imagery analysis. It refers to a situation where the distribution of samples across different classes is significantly uneven, leading to biased learning and potentially lower performance for minority classes. In the context of this research, the presence of class imbalance in the datasets used for the experiments raises important considerations.

In the conducted experiments, no explicit measures were taken to address class imbalance. This decision was based on several factors specific to the research objectives and dataset characteristics. Firstly, the primary focus of the investigation was to examine the effects of physics-based image normalisation and augmentation methods on the transferability of the object detection and segmentation system. The aim was to understand how these techniques could improve the overall performance and generalisation of the system across different altitudes and imaging conditions.

Additionally, the selected marine species classes exhibited varying levels of abundance within the datasets. Rockfish, for instance, showed high abundance, while Crabs had a moderate presence, and Hagfish were less abundant. The inherent imbalance in the natural distribution of these species is reflective of real-world conditions. By not explicitly addressing class imbalance, the experiments aimed to assess the system's ability to handle the inherent class distributions and reflect the real challenges faced in marine conservation scenarios.

However, it is important to acknowledge the potential impact of class imbalance on the performance of the system. The under-representation of certain classes can lead to skewed learning and a bias towards dominant classes, resulting in sub-optimal performance for minority classes. Therefore, in future experiments or applications, addressing class imbalance could be considered to ensure fair representation and improved performance for all classes.

Several strategies can be explored to mitigate class imbalance in marine imagery analysis. Oversampling techniques, such as data augmentation for underrepresented classes, can be employed to increase the number of samples and balance the class distribution. This can involve synthesising additional instances of the minority classes through techniques like image transformations or generative models.

Another approach is to leverage specialised training algorithms, such as cost-sensitive learning or class weighting, which assign higher penalties or adjust sample weights to the minority classes during the training process. These techniques can help alleviate the impact of class imbalance and encourage the model to give equal consideration to all classes.

In conclusion, while the present experiments did not explicitly address class imbalance, future studies can explore strategies to tackle this challenge. By employing techniques such as oversampling and specialised training algorithms it becomes possible to enhance the fairness and performance of the machine learning system in marine imagery analysis. Considering class imbalance is crucial for developing robust and reliable models that can effectively support marine conservation efforts.

8.4 How to Handle Unseen Classes

One of the challenges in marine imagery analysis is dealing with unseen classes, referring to classes that are not present in the training data but may appear during testing or real-world scenarios. The ability to handle unseen classes is crucial for the robustness and practicality of using machine learning models for real world applications. In this section, we discuss the difficulties encountered with unseen classes, specifically highlighting the issues encountered with the Sea Star and Brittle Star classes, and explore potential strategies to address this challenge.

In the conducted experiments, the Sea Star class was initially included in the training and testing datasets. On the other hand, although Brittle Stars were abundant in the dataset, their high abundance and unfeasible lower bound on size made manual labelling and inclusion in the tests unfeasible. This posed a problem as the model, lacking exposure to Brittle Stars during training, encountered difficulties in correctly classifying them during testing. The misclassification of Brittle Stars as Sea Stars further underscored the importance of addressing unseen classes in marine imagery analysis.

Handling unseen classes that are neither labelled nor present in the training data poses an additional level of complexity. The model's inability to recognise and classify these unseen classes can lead to erroneous outputs and limited practicality in real-world scenarios. However, instead of providing low-confidence estimates for such instances, a potential alternative approach could involve flagging these instances and highlighting them to researchers for further analysis. This can provide valuable insights into the presence and behaviour of previously unseen classes, contributing to the expansion of knowledge and understanding in marine biology and conservation.

It is worth noting that the challenges faced with handling unseen classes in marine imagery analysis highlight a weakness in classification AI. Expert human labellers possess the ability to identify and classify unseen classes based on their deep understanding and knowledge of the domain, and also higher reasoning skills. In contrast, machine learning models trained on specific datasets struggle with recognising unfamiliar classes without explicit training examples, and lack the reasoning capabilities to know that what they have seen and are struggling to identify is something new.

To address this limitation, the development of more generalised intelligence AI systems may offer potential solutions. These systems, equipped with broader knowledge and reasoning capabilities, could better handle the recognition and classification of unseen classes based on their understanding of underlying patterns and principles. By leveraging generalisation and transfer learning techniques, such AI systems can exhibit greater flexibility and adaptability to novel marine species or classes not encountered during training.

In conclusion, handling unseen classes in marine imagery analysis is a challenging task. The difficulties encountered with the Sea Star and Brittle Star classes underscore the importance of addressing this challenge. Highlighting instances of unseen classes to researchers can provide valuable insights into gaps in the given models capabilities, and perhaps in rare cases can identify entirely new morphotypes. Furthermore, the exploration of more generalised intelligence AI systems holds promise for better handling unseen classes and advancing the field of marine imagery analysis.

8.5 The Ideal Universal Marine Imagery Dataset

Creating a universal marine imagery dataset that encompasses the vast diversity of marine life and habitats presents significant challenges, if not impossibilities, due to the sheer complexity and scale of the marine ecosystem. The marine environment is a dynamic and diverse environment, with countless species, habitats, and ecological interactions. Attempting to form a single dataset that adequately represents this vastness is an ambitious task that encounters numerous obstacles.

One of the primary challenges in forming a universal marine imagery dataset lies in the incredible biological diversity present in the marine environment. From microscopic plankton to large marine mammals, the range of species inhabiting the oceans is vast and encompasses an astonishing array of morphological characteristics, behaviours, and ecological roles. Each species requires specific knowledge and expertise for accurate identification and classification, making it challenging to curate a dataset that comprehensively represents the multitude of marine organisms.

Adding to the complexity, marine habitats exhibit tremendous variability across the globe. Coral reefs, bacteria mats, kelp forests, deep-sea trenches, estuaries, and polar regions are just a few examples of the diverse ecosystems within the marine environment. Each habitat possesses distinct physical and ecological features that shape the composition and distribution of species. Collecting data across such a wide range of habitats is a logistical challenge, requiring extensive resources, specialised equipment, and expertise in various environmental conditions.

Furthermore, sampling limitations pose a significant hurdle in forming a universal marine imagery dataset. Conducting comprehensive surveys across different regions, depths, and seasons is resource-intensive and often constrained by factors such as time, budget, and accessibility. As a result, any dataset formed is inevitably influenced by the specific sampling biases and limitations of the data collection efforts. These limitations can introduce biases in the representation of species and habitats, potentially leading to incomplete or skewed datasets.

Another obstacle in creating a universal marine imagery dataset arises from the taxonomic knowledge gaps within the scientific community. Despite advancements in marine taxonomy, many species remain unidentified or undersampled. Furthermore, different taxonomic identifiers have been developed for specific research goals, resulting in a lack of universally agreed-upon labels. Additionally, taxonomic revisions and ongoing research continue to shape our understanding of marine biodiversity, making it challenging to create a dataset that accurately represents all known species.

Moreover, the marine environment is characterised by spatial and temporal variability. Factors such as water temperature, salinity, currents, and nutrient availability exhibit gradients and fluctuations across different regions and time scales. These environmental factors influence species composition, distribution, and behaviour. Capturing the full range of environmental variability within a single dataset is impractical, as it would require extensive sampling efforts covering diverse spatiotemporal scales.

Given these challenges, it is important to acknowledge the limitations of forming a universal marine imagery dataset. However, this does not diminish the value of targeted and representative datasets that focus on specific regions, habitats, or taxonomic groups. By concentrating efforts on specific areas of interest, researchers can still contribute to our understanding of localised ecosystems, uncover unique ecological interactions, and support targeted conservation initiatives.

While a comprehensive universal dataset may be unattainable, it is crucial to emphasise the importance of collaborative efforts in data sharing and integration. By sharing data across research institutions, organisations, and countries, it becomes possible to assemble larger, more diverse datasets that encompass a broader range of species, habitats, and environmental conditions. Collaboration fosters the exchange of knowledge, encourages standardised data collection methodologies, and promotes the development of more robust models and analyses.

In conclusion, the formation of a universal marine imagery dataset is an immensely complex and challenging task due to the incredible biological diversity, habitat variability, sampling limitations, taxonomic knowledge gaps, and environmental heterogeneity within the marine ecosystem. While a comprehensive dataset covering the entirety of marine life may be unfeasible, focused and representative datasets can still contribute to advancing our knowledge and conservation efforts within specific marine regions, habitats, and taxonomic groups. Collaborative approaches, data sharing, and integration are crucial for maximising the value of available data and overcoming the limitations of individual datasets.

Chapter 9

Conclusions

This thesis investigated the use of modern machine learning techniques to automate the analysis of AUV imagery. AUV surveys provide more visual imagery data than can currently be analysed by experts, and current automated methods of analysis lack the domain specific outputs needed to determine important values of interest such as biomass. This thesis presents a more automated and scalable solution that not only expands the amount of data that can be analysed, but also provides useful information in the form of biomass estimates separated by class.

9.1 Contributions

There are three major contributions from this thesis, detailed in the sections below.

9.1.1 Investigation of Intra-Vehicle Transferability

Chapter 5 presents the findings of a thorough investigation into intra-vehicle transferability of the Mask R-CNN neural network, using datasets collected at differing, but nearby, geographical locations, on the Adaptive Robotics Falkor 2018 expedition. The transferability of learning to newly seen datasets collected by the same vehicle was measured and analysed, achieving an mAP score of 89% on low altitude imagery and 79% on high altitude imagery.

The most novel aspect of this investigation is the breadth of independent variables being investigated, and the results for every subsequent combination of variables are presented. Both data normalisation and data augmentation techniques were investigated. The data normalisation techniques investigated were colour correction, scale normalisation, and distortion correction. The data augmentation techniques investigated were elastic or piece-wise transformations, independent channel augmentations, addition, multiplication, salt and pepper noise, motion blur, and linear contrast.

The key findings include that greyworld colour correction improved performance greatly for both low and high altitude datasets, improving from an average mAP of 56.6% to 74.1% for low altitude data, $\mathcal{L}^{train} \rightarrow \mathcal{L}^{test}$, and improving from an average mAP of 44.6% to 74.4% for high altitude data, $\mathcal{H}^{train} \rightarrow \mathcal{H}^{test}$. A further finding of note is that rescaling to the average low altitude spatial scale worked best for low altitude images, and rescaling had little effect on the accuracy for high altitude data. For $\mathcal{L}^{train} \rightarrow \mathcal{L}^{test}$ experiments, rescaling to the average low altitude spatial scale improved mAP scores from an average of 62.1% to 68.6%.

The other variables investigated have a less significant effect on the performance, and may need further investigating for their efficacy in different situations. In the specific set up investigated in this thesis, distortion correction harmed performance, as did elastic distortions.

9.1.2 Investigation of Inter-Vehicle Transferability

Chapter 6 presents the findings of the inter-vehicle transferability experiments for this thesis. In these experiments, the training and validation datasets were from a different vehicle to the test dataset, either transferring learning from a low altitude dataset to a high altitude dataset or vice versa, $\mathcal{L}^{train} \rightarrow \mathcal{H}^{test}$ and $\mathcal{H}^{train} \rightarrow \mathcal{L}^{test}$.

When transferring learning from a low altitude dataset to a high altitude dataset, $\mathcal{L}^{train} \rightarrow \mathcal{H}^{test}$, performance is incredibly poor. On the other hand, transferring from a high altitude dataset to a low altitude dataset, $\mathcal{H}^{train} \rightarrow \mathcal{L}^{test}$, was relatively successful, achieving as high as 54%. The independent variable with the largest impact on these results was rescaling, where rescaling to the average high altitude spatial scale improved mAAP scores from 26.5% to 44.1%. Furthermore greyworld colour correction also provides clear improvements in accuracy, improving from an average mAP score of 32.7% to 38.0%. The colour correction results are in line with the findings of the intra-vehicle results, implying that using greyworld colour correction over raw histogram stretching is applicable across many situations. The rescaling results, however, differ between $\mathcal{H}^{train} \rightarrow \mathcal{H}^{test}$ and $\mathcal{H}^{train} \rightarrow \mathcal{L}^{test}$, having little impact in the former case, and improving results drastically in the latter, showing the situational improvements scale correction offer. Scale is a major difference between the appearance of animals in high altitude imagery and low altitude imagery, and correcting for this scale variation is clearly more important when looking at inter-vehicle transfer learning than when looking at intra-vehicle learning.

Similarly to the intra-vehicle learning results, the other independent variables investigated had less significant impacts on results. Again, distortion correction reduced accuracy going against expectations. Furthermore, data augmentations all caused a small decrease in performance. In order to effectively use data augmentations more investigation is needed into optimising their parameters and identifying individual useful augmentation techniques.

9.1.3 Automated Biomass Estimation

Chapter 7 presents an automated biomass estimation method based on the use of Length Weight Relationships (LWRs) and the newly developed Segment Size to Length Relationships (SSLRs). This method allows for a fully automated process of biomass estimation from AUV imagery via Mask R-CNN or similar object detection and segmentation techniques, and the application of these relationships to form a biomass estimate for each detected individual. This thesis demonstrates the method on the datasets collected on the Adaptive Robotics FK2018 expedition on both low and high altitude datasets, and with the results broken down by class to show class specific distributions in the form of biomass density heatmaps and biomass histograms. This demonstrates the fully scalable automated method, and shows how it can be applied to large datasets containing thousands of images with no additional manual intervention required. This method will allow for the automated estimation of biomass distributions in the field, informing AUV deployment decisions in real time. Furthermore, it can be applied to existing datasets, increasing the value of collected imagery to the science community, getting more information out of expensive ship based expeditions.

9.2 Further Work

There is a wide range of possible future work that builds on and improves the findings of this thesis. There are some key directions further work may take, such as improvements to the Mask R-CNN experiments themselves, investigating the metrics by which we measure performance, and further work on improving the automated biomass estimation process. Examples of further work in each of these areas are discussed below.

9.2.1 Mask R-CNN

The training of the Mask R-CNN neural network poses many different areas for possible adjustment and investigation.

With more time and man power available, investigations with a larger labelled dataset would be possible, and would provide a clearer insight into the effects of various independent variables on performance without the interference of over-fitting, severe class imbalance, and other such effects common to small datasets.

Simply expanding the number of labelled images is not the only way the datasets in use could be enhanced. One possibility is the the addition of other vehicles to the study, seen in Zuroweitz et al.'s work [87]. In this study, a larger set of vehicles was included, including vehicles closer in altitude than the two for this study, allowing for investigation into the impact of smaller shifts in altitude. Another such possibility would be the addition of different classes. There was a clear difference in performance between different classes in this thesis, and while the data available allows us to only speculate on the reasons why, a thorough investigation into the performance on different classes could be carried out. This could look into the size of individuals in the class, their colour and contrast with their surrounding areas, their shape, and even features such as whether they are a burrowing species, or other behavioural aspects. An investigation into this class performance difference is vital if automated analysis is ever to be relied upon for scientific study and conservation decisions.

Another area for improvement in the training of Mask R-CNN is hyper-parameter finetuning. This vital aspect of machine learning is often overlooked, as it was in this thesis, due to the complex combinations of hyper-parameters and their impact on learning. Before a system such as Mask R-CNN should be used for fully automated analysis of a marine survey for scientific study, it should be finetuned to achieve the very best performance possible. Whether such finetuning of neural network parameters would change the findings of studies such as this one is mostly unknown, and the assumption made is that if gains in performance were to be made with finetuning of these hyperparameters, they would be universal and would not change the relative performance of differing independent variables. Further work can and should be done to investigate this assumption.

The way training data is provided during training can also have a large impact on performance, and there are a number of alternatives to what was used in this thesis. Repeating images containing rarer classes more often than those only containing more common classes, sometimes referred to as class balancing, is one such approach that may improve performance on the rarer classes in the training data and is worth investigating. Another approach is to repeat images that the neural network gets a lower accuracy score for, sometimes referred to as bootstrapping, and showing images it scores well on less often as it has already learnt what it needs to from those images. Both of these approaches have shown promise in other studies, and are worth investigating with marine imagery.

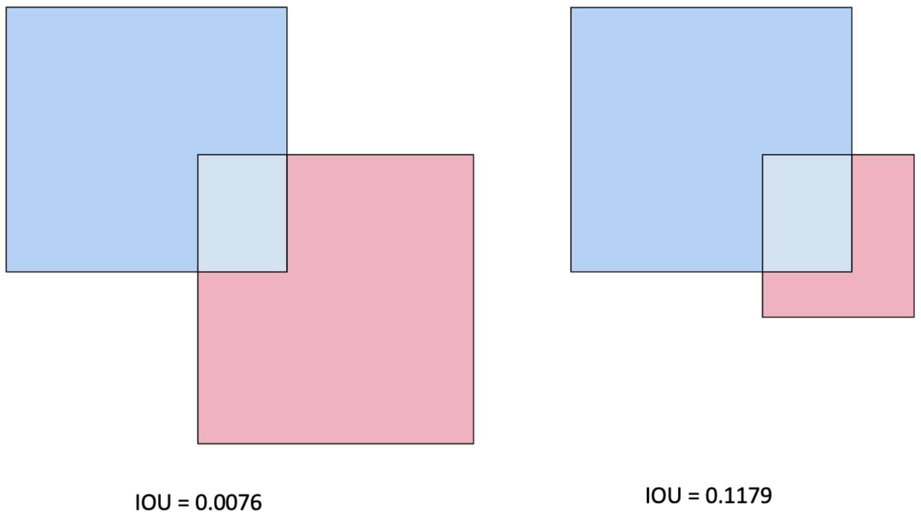


FIGURE 9.1: Example of IOU metric failing to capture context specific information. The example on the left would result in a more accurate biomass estimate, but would score lower in terms of IOU scores than the example on the right.

9.2.2 Metrics

The measurement of performance of an automated system for marine imagery is challenging, as experiments into differences between manual annotations from different experts show the level of agreement between two marine experts is often low. [104; 97]. Furthermore, these classic computer vision metrics measure the direct similarity between the two sets of labels, and do not take into account the context specific impacts of incorrect labels, such as changes this makes to population counts, size distributions, or biomass estimates. By calculating these statistics and comparing those as opposed to directly comparing labels, we can assess more accurately how well an algorithm will perform in the marine imagery context. This is demonstrated in Figure 9.1 where the example on the left gets a lower IOU score than the example on the right, but in the context of biomass estimation the example on the left would be more accurate then the example on the right, that would underestimate the biomass.

9.2.3 Biomass Estimate Improvements

LWRs were developed a long time ago for the use of fishermen and scientists, among others, to easily estimate the weight of a physical sample without the need to weigh it. These estimates were useful at the time, and continue to be used today, but they were not developed as the best way to estimate biomass from images, and, with the time and manpower required, LWRs could be super-ceded by a direct relationship between segment size and biomass. The datasets required for such a relationship to be calculated were not readily available at the time of writing this thesis, and curating such a dataset was outside the scope of the project. Creating such a dataset would involve photographing individual specimens from above with a known size per pixel, and then weighing each individual, to provide a direct comparison between visible segment size from above and weight. Such a dataset would enable a more accurate fully scalable automated biomass estimation system than the one demonstrated here, and would be a vital step towards large scale automated biomass estimation from AUV imagery for use in scientific discoveries and marine conservation efforts.

9.2.4 Large Labelled Datasets

In the application of machine learning to terrestrial imagery the image acquisition conditions are very rarely taken into account, with very large datasets collected by varying hardware in varying environmental conditions. With a large and varied enough dataset this may be possible with marine imagery too, but would require a large dataset of thousands of images to be collected and labelled in a consistent format with consistent class labels. Crowd sourcing such datasets has been successful for many different fields, including astronomy with the use of Galaxy Zoo, a project by Zooniverse, where images of space taken by a telescope are labelled by volunteers online with varying levels of expertise in the area. Another approach to gathering such a dataset may be a large scale collaboration between research institutes to provide their labelled datasets publicly with a consistent labelling schema. A dataset such as this would enable automated detection and segmentation of a larger array of classes and to a higher degree of accuracy. This would allow for the use of machine learning on AUV imagery for scientific outcomes, and would enable better investigation into the effects of normalisation and augmentation on differing classes. As previously mentioned in this thesis, the performance of Mask R-CNN varies by class, as does the effect of varying augmentation techniques, and with the three classes labelled and analysed in this thesis hypotheses can be posed, but further classes must be analysed to identify the cause of this variance. Possible causes to investigate include, but are not limited to, size, shape, colour, and behaviour such as burrowing.

References

[1] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8693 LNCS, no. PART 5, 2014, pp. 740–755. [Online]. Available: https://arxiv.org/pdf/1405.0312.pdf

[2] F. Althaus, N. Hill, R. Ferrari, L. Edwards, R. Przeslawski, C. H. Schönberg, R. Stuart-Smith, N. Barrett, G. Edgar, J. Colquhoun, M. Tran, A. Jordan, T. Rees, and K. Gowlett-Holmes, "A standardised vocabulary for identifying benthic biota and substrata from underwater imagery: The CATAMI classification scheme," *PLoS ONE*, vol. 10, no. 10, 2015. [Online]. Available: http://www.ands.org.

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. [Online]. Available: https://arxiv.org/pdf/1703.06870.pdf

[4] T. Yamada, A. Prugel-Bennett, and B. Thornton, "Learning features from georeferenced seafloor imagery with location guided autoencoders," Jan 2020. [Online]. Available: https://repository.oceanbestpractices.org/handle/11329/1520

[5] R. Froese and D. Pauly, "Fishbase," 2021. [Online]. Available: https://www.fishbase.org

[6] N. MacLeod, M. Benfield, and P. Culverhouse, "Time to automate identification," *Nature*, vol. 467, no. 7312, pp. 154–155, sep 2010. [Online]. Available: http://www.nature.com/doifinder/10.1038/467154a

[7] M. Everingham, S. M. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015. [Online]. Available: http://host.robots.ox.ac.uk/pascal/VOC/pubs/everingham15.pdf

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, 2016, pp. 3213–3223.

[9] D. Langenkämper, M. Zurowietz, T. Schoening, and T. W. Nattkemper, "BIIGLE 2.0 - browsing and annotating large marine image collections," *Frontiers in Marine Science*, vol. 4, no. MAR, mar 2017.

[10] A. Friedman, "Squidle+." [Online]. Available: https://squidle.org/

[11] K. Katija, E. Orenstein, B. Schlining, L. Lundsten, K. Barnard, G. Sainz, O. Boulais, B. Woodward, and K. C. Bell, "FathomNet: A global underwater image training set for enabling artificial intelligence in the ocean," sep 2021. [Online]. Available: https://arxiv.org/abs/2109.14646v2http://arxiv.org/abs/2109.14646

[12] K. Hayes, J. Dambacher, G. Hosack, N. Bax, P. Dunstan, E. Fulton, P. Thompson, J. Hartog, A. Hobday, R. Bradford, S. Foster, P. Hedge, D. Smith, and C. Marshall, "Identifying indicators and essential variables for marine ecosystems," *Ecological Indicators*, vol. 57, pp. 409–419, oct 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1470160X15002265

[13] M. McNutt, G. Massion, K. Raybould, J. Bellingham, and C. Paull, "MARS: a cabled observatory testbed in Monterey Bay," *EGS - AGU - EUG Joint Assembly, Abstracts from the meeting held in Nice, France, 6 - 11 April 2003,*, p. 11585, 2003. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2003EAEJA....11585M/abstract

[14] S. E. Hartman, R. S. Lampitt, K. E. Larkin, M. Pagnani, J. Campbell, T. Gkritzalis, Z.-P. Jiang, C. A. Pebody, H. A. Ruhl, A. J. Gooday *et al.*, "The porcupine abyssal plain fixed-point sustained observatory (pap-so): variations and trends from the northeast atlantic fixed-point time-series," *ICES Journal of Marine Science*, vol. 69, no. 5, pp. 776–783, 2012.

[15] L. L. Whitcomb, "Underwater robotics: out of the research laboratory and into the field," in *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 1. IEEE, 2000, pp. 709–716. [Online]. Available: http://ieeexplore.ieee.org/document/844135/

[16] M. Purcell, C. von Alt, B. Allen, T. Austin, N. Forrester, R. Goldsborough, and R. Stokey, "New capabilities of the REMUS autonomous underwater vehicle," in *OCEANS 2000 MTS/IEEE Conference and Exhibition. Conference Proceedings (Cat. No.00CH37158)*, vol. 1. IEEE, 2000, pp. 147–151. [Online]. Available: http://ieeexplore.ieee.org/document/881250/

[17] T. Nakatani, T. Ura, Y. Ito, J. Kojima, K. Tamura, T. Sakamaki, and Y. Nose, “AUV ”TUNA-SAND” and its Exploration of hydrothermal vents at Kagoshima Bay,” *Oceans 2008*, pp. 0–4, 2008.

[18] T. Ura, T. Obara, K. Nagahashi, K. Kim, Y. Oyabu, T. Sakamaki, A. Asada, H. Koyama, and M. Engineering, “Introduction to an AUV ”2D4” and its Kuroshima Knoll Survey Mission,” vol. 4000, no. m, pp. 840–845, 2003.

[19] R. B. Wynn, V. A. I. Huvenne, T. P. Le Bas, B. J. Murton, D. P. Connelly, B. J. Bett, H. A. Ruhl, K. J. Morris, J. Peakall, D. R. Parsons, E. J. Sumner, S. E. Darby, R. M. Dorrell, and J. E. Hunt, “Autonomous Underwater Vehicles (AUVs): Their past, present and future contributions to the advancement of marine geoscience,” *Marine Geology*, vol. 352, pp. 451–468, 2014. [Online]. Available: http://creativecommons.org/licenses/by/3.0/

[20] S. B. Williams, O. Pizarro, D. M. Steinberg, A. Friedman, and M. Bryson, “Reflections on a decade of autonomous underwater vehicles operations for marine survey at the Australian Centre for Field Robotics,” *Annual Reviews in Control*, 2016.

[21] S. B. Williams, O. R. Pizarro, M. V. Jakuba, C. R. Johnson, N. S. Barrett, R. C. Babcock, G. A. Kendrick, P. D. Steinberg, A. J. Heyward, P. J. Doherty, I. Mahon, M. Johnson-Roberson, D. Steinberg, and A. Friedman, “Monitoring of benthic reference sites: Using an autonomous underwater vehicle,” *IEEE Robotics and Automation Magazine*, vol. 19, no. 1, pp. 73–84, mar 2012. [Online]. Available: http://ieeexplore.ieee.org/document/6174326/

[22] M. Bewley, A. Friedman, R. Ferrari, N. Hill, R. Hovey, N. Barrett, E. M. Marzinelli, O. Pizarro, W. Figueira, L. Meyer, R. Babcock, L. Bellchambers, M. Byrne, and S. B. Williams, “Australian sea-floor survey data, with images and expert annotations,” *Scientific Data*, vol. 2, no. 1, pp. 1–13, oct 2015. [Online]. Available: https://www.nature.com/articles/sdata201557

[23] B. M. Schlining and N. J. Stout, “MBARI’s Video Annotation and Reference System,” in *OCEANS 2006*. IEEE Computer Society, jan 2006.

[24] M. Bryson, M. Johnson-Roberson, O. Pizarro, and S. B. Williams, “True color correction of autonomous underwater vehicle imagery,” *Journal of Field Robotics*, vol. 33, no. 6, p. 853–874, 2015.

[25] R. Froese, J. T. Thorson, and R. B. Reyes, “A Bayesian approach for estimating length-weight relationships in fishes,” *Journal of Applied Ichthyology*, vol. 30, no. 1, pp. 78–85, feb 2014. [Online]. Available: http://doi.wiley.com/10.1111/jai.12299

[26] F. PirahanSiah, S. N. H. S. Abdullah, and S. Sahran, “Adaptive image thresholding based on the peak signal-to-noise Ratio,” *Research Journal of Applied Sciences, Engineering and Technology*, vol. 8, no. 9, pp. 1104–1116, 2014.

[27] J. Canny, “A Computational Approach to Edge Detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986.

[28] S. X. Yu and J. Shi, “Object-specific figure-ground segregation,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2003.

[29] Ting Chuen Pong, L. G. Shapiro, and R. M. Haralick, “A facet model region growing algorithm.” *Pattern recognition and image processing. Proc. IEEE Computer Society conference, Dallas, 1981, (IEEE, New York; CH1595 8)*, pp. 279–284, 1981.

[30] A. Bala and A. K. Sharma, “Split and Merge: A Region Based Image Segmentation,” *International Journal of Emerging Research in Management and Technology*, vol. 6, no. 8, p. 306, 2018.

[31] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “Training algorithm for optimal margin classifiers,” in *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, 1992, pp. 144–152.

[32] L. Bertelli, T. Yu, D. Vu, and B. Gokturk, “Kernelized structural SVM learning for supervised object segmentation,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2011, pp. 2153–2160. [Online]. Available: http://ieeexplore.ieee.org/document/5995597/

[33] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, dec 1943. [Online]. Available: https://link.springer.com/article/10.1007/BF02478259

[34] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain,” *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[35] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning Internal Representations Error Propagation,” Tech. Rep. V, 1986. [Online]. Available: https://apps.dtic.mil/docs/citations/ADA164453

[36] D. E. Rumelhart and J. L. McClelland, *Learning Internal Representations by Error Propagation*, 1986.

[37] M. Gori, A. Tesi *et al.*, “On the problem of local minima in backpropagation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 1, pp. 76–86, 1992.

[38] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

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

[42] J. Lafferty, A. Mccallum, F. C. N. Pereira, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data Recommended Citation "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data" Conditional Random Fields: Probabilistic Models for ," pp. 282–289, 2001. [Online]. Available: http://repository.upenn.edu/cis_papershttp://portal.acm.org/citation.cfm?id=655813http://repository.upenn.edu/cis_papers/159

[43] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient Graph-Based Image Segmentation Pedro," pp. 1–26, 2015. [Online]. Available: http://people.cs.uchicago.edu/$\sim$pff/papers/seg-ijcv.pdfpapers3://publication/uuid/D1250C05-2FC7-4954-A734-E33EBBEECB95

[44] X. Haixiang, C. Wanhua, C. Wei, and G. Liyuan, "Performance evaluation of SVM in image segmentation," in *International Conference on Signal Processing Proceedings, ICSP*, 2008, pp. 1207–1210.

[45] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2014. [Online]. Available: https://people.eecs.berkeley.edu/{~}jonlong/long{_}shelhamer{_}fcn.pdf

[46] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017. [Online]. Available: https://arxiv.org/pdf/1412.7062.pdf

[47] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017. [Online]. Available: https://arxiv.org/pdf/1706.05587.pdf

[48] A. Kolesnikov and C. H. Lampert, “Seed, expand and constrain: Three principles for weakly-supervised image segmentation,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9908 LNCS, 2016, pp. 695–711.

[49] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 779–788, 2016.

[50] A. Tao, J. Barker, and S. Sarathy, “Detectnet: Deep neural network for object detection in digits,” Aug 2016. [Online]. Available: https://developer.nvidia.com/blog/detectnet-deep-neural-network-object-detection-digits/

[51] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. [Online]. Available: http://www.cs.berkeley.edu

[52] R. Girshick, “Fast R-CNN,” in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, 2015, pp. 1440–1448. [Online]. Available: https://github.com/rbgirshick/

[53] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, jun 2017. [Online]. Available: http://image-net.org/challenges/LSVRC/2015/results

[54] F. Lateef and Y. Ruichek, “Survey on semantic segmentation using deep learning techniques,” *Neurocomputing*, vol. 338, pp. 321–348, apr 2019.

[55] T. Schoening, M. Bergmann, J. Ontrup, J. Taylor, and J. Dannheim, “Semi-Automated Image Analysis for the Assessment of Megafaunal Densities at the Arctic Deep-Sea Observatory HAUSGARTEN,” *PLoS ONE*, vol. 7, no. 6, p. 38179, 2012. [Online]. Available: www.plosone.org

[56] J. Seiler, A. Williams, and N. Barrett, “Assessing size, abundance and habitat preferences of the Ocean Perch Helicolenus percoides using a AUV-borne stereo camera system,” *Fisheries Research*, vol. 129-130, pp. 64–72, 2012. [Online]. Available:
https://reader.elsevier.com/reader/sd/pii/S0165783612001956?token=9012C5A5338A6C8091C68FF1679AC3F6028AADB5AFB541C524F0BF3AF62746E006A0E4DA9BF425CB

[57] K. J. Morris, B. J. Bett, J. M. Durden, V. A. Huvenne, R. Milligan, D. O. Jones, S. McPhail, K. Robert, D. M. Bailey, and H. A. Ruhl, “A new method for ecological surveying of the abyss using autonomous underwater vehicle

photography," *Limnology and Oceanography: Methods*, vol. 12, no. NOV, pp. 795–809, 2014.

[58] N. M. Benoist, K. J. Morris, B. J. Bett, J. M. Durden, V. A. Huvenne, T. P. Le Bas, R. B. Wynn, S. J. Ware, and H. A. Ruhl, "Monitoring mosaic biotopes in a marine conservation zone by autonomous underwater vehicle," *Conservation Biology*, vol. 33, no. 5, pp. 1174–1186, oct 2019. [Online]. Available: /pmc/articles/PMC6850053//pmc/articles/PMC6850053/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6850053/

[59] B. Thornton, A. Bodenmann, O. Pizarro, S. B. Williams, A. Friedman, R. Nakajima, K. Takai, K. Motoki, T. o. Watsuji, H. Hirayama, Y. Matsui, H. Watanabe, and T. Ura, "Biometric assessment of deep-sea vent megabenthic communities using multi-resolution 3D image reconstructions," *Deep-Sea Research Part I: Oceanographic Research Papers*, 2016.

[60] K.-L. Howell, N. Piechaud, A.-L. Downie, and A. Kenny, "The distribution of deep-sea sponge aggregations in the North Atlantic and implications for their effective spatial management," *Deep Sea Research Part I: Oceanographic Research Papers*, vol. 115, pp. 309–320, sep 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0967063716300097

[61] E. Simon-Lledó, B. J. Bett, V. A. Huvenne, T. Schoening, N. M. Benoist, R. M. Jeffreys, J. M. Durden, and D. O. Jones, "Megafaunal variation in the abyssal landscape of the Clarion Clipperton Zone," *Progress in Oceanography*, vol. 170, pp. 119–133, jan 2019.

[62] N. M. Benoist, B. J. Bett, K. J. Morris, and H. A. Ruhl, "A generalised volumetric method to estimate the biomass of photographically surveyed benthic megafauna," *Progress in Oceanography*, vol. 178, p. 102188, nov 2019.

[63] J. G. Baguley, L. J. Hyde, and P. A. Montagna, "A semi-automated digital microphotographic approach to measure meiofaunal biomass," *Limnology and Oceanography: Methods*, vol. 2, no. 6, pp. 181–190, 2004.

[64] T. Blaschke, "Object based image analysis for remote sensing," pp. 2–16, 2010. [Online]. Available: www.elsevier.com/locate/isprsjprs

[65] M. Bryson, M. Johnson-Roberson, O. Pizarro, and S. Williams, "Repeatable Robotic Surveying of Marine Benthic Habitats for Monitoring Long-term Change," *Robotics Science and Systems*, pp. 3–7, 2012. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.717.6337&rep=rep1&type=pdf

[66] M. Lacharité, A. Metaxas, and P. Lawton, "Using object-based image analysis to determine seafloor fine-scale features and complexity," *Limnology and Oceanography: Methods*, 2015.

[67] O. Pizarro, P. Rigby, M. Johnson-Roberson, S. B. Williams, and J. Colquhoun, "Towards image-based marine habitat classification," in *OCEANS 2008*, 2008.

[68] A. Shihavuddin, N. Gracias, R. Garcia, A. Gleason, and B. Gintert, "Image-Based Coral Reef Classification and Thematic Mapping," *Remote Sensing*, vol. 5, no. 4, pp. 1809–1841, apr 2013. [Online]. Available: http://www.mdpi.com/2072-4292/5/4/1809/

[69] M. S. Bewley, B. Douillard, N. Nourani-Vatani, A. Friedman, O. Pizarro, and S. B. Williams, "Automated species detection: An experimental approach to kelp detection from sea-floor AUV images," in *Australasian Conference on Robotics and Automation, ACRA*, 2012. [Online]. Available: https://www.researchgate.net/publication/283257248

[70] O. Beijbom, P. J. Edmunds, D. I. Kline, B. G. Mitchell, and D. Kriegman, "Automated annotation of coral reef survey images," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012.

[71] O. Beijbom, P. J. Edmunds, C. Roelfsema, J. Smith, D. I. Kline, B. P. Neal, M. J. Dunlap, V. Moriarty, T. Y. Fan, C. J. Tan, S. Chan, T. Treibitz, A. Gamst, B. G. Mitchell, and D. Kriegman, "Towards automated annotation of benthic survey images: Variability of human experts and operational modes of automation," *PLoS ONE*, vol. 10, no. 7, p. e0130312, jul 2015. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0130312

[72] A. Mahmood, M. Bennamoun, S. An, F. Sohel, F. Boussaid, R. Hovey, G. Kendrick, and R. B. Fisher, "Automatic annotation of coral reefs using deep learning," in *OCEANS 2016 MTS/IEEE Monterey, OCE 2016*. Institute of Electrical and Electronics Engineers Inc., nov 2016.

[73] X. Sun, J. Shi, L. Liu, J. Dong, C. Plant, X. Wang, and H. Zhou, "Transferring deep knowledge for object recognition in Low-quality underwater videos," *Neurocomputing*, vol. 275, pp. 897–908, jan 2018.

[74] N. Wahidin, V. P. Siregar, B. Nababan, I. Jaya, and S. Wouthuyzen, "Object-based Image Analysis for Coral Reef Benthic Habitat Mapping with Several Classification Algorithms," *Procedia Environmental Sciences*, 2015.

[75] J.-N. Blanchet, S. Déry, J.-A. Landry, and K. Osborne, "Automated annotation of corals in natural scene images using multiple texture representations," may 2016. [Online]. Available: https://peerj.com/preprints/2026/

[76] T. Schoening, T. Kuhn, D. O. Jones, E. Simon-Lledo, and T. W. Nattkemper, "Fully automated image segmentation for benthic resource assessment of poly-metallic nodules," *Methods in Oceanography*, vol. 15-16, pp. 78–89, 2016.

[77] A. King, S. M. Bhandarkar, and B. M. Hopkinson, "Deep learning for semantic segmentation of coral reef images using multi-view information," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2019-June, 2019, pp. 1–10.

[78] K. Mizuno, K. Terayama, S. Tabeta, S. Sakamoto, Y. Matsumoto, Y. Sugimoto, T. Ogawa, K. Sugimoto, H. Fukami, M. Sakagami, M. Deki, and A. Kawakubo, "Development of an Efficient Coral-Coverage Estimation Method Using a Towed Optical Camera Array System [Speedy Sea Scanner (SSS)] and Deep-Learning-Based Segmentation: A Sea Trial at the Kujuku-Shima Islands," pp. 1386–1395, 2019. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8862868&tag=1

[79] I. Alonso, A. Cambra, A. Muñoz, T. Treibitz, and A. C. Murillo, "Coral-Segmentation: Training Dense Labeling Models with Sparse Ground Truth," Tech. Rep., 2017.

[80] I. Alonso, M. Yuval, G. Eyal, T. Treibitz, and A. C. Murillo, "CoralSeg: Learning coral segmentation from sparse annotations," *Journal of Field Robotics*, vol. 36, no. 8, pp. 1456–1477, oct 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21915

[81] D. M. Steinberg, O. Pizarro, and S. B. Williams, "Hierarchical Bayesian models for unsupervised scene understanding," *Computer Vision and Image Understanding*, vol. 131, pp. 128–144, 2015. [Online]. Available: https://ac.els-cdn.com/S1077314214001313/1-s2.0-S1077314214001313-main.pdf?{_}tid=8843d190-c488-11e7-bdd0-00000aacb362{&}acdnat=1510147815{_}631da3bfbbdc1fc2d09ba32788cc71f0

[82] E. C. Orenstein, J. M. Haag, Y. L. Gagnon, and J. S. Jaffe, "Automated classification of camouflaging cuttlefish," *Methods in Oceanography*, vol. 15-16, pp. 21–34, 2016.

[83] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, 2016, pp. 2818–2826.

[84] N. Piechaud, C. Hunt, P. F. Culverhouse, N. L. Foster, and K. L. Howell, "Automated identification of benthic epifauna with computer vision," *Marine Ecology Progress Series*, vol. 615, pp. 15–30, apr 2019.

[85] R. Mandal, R. M. Connolly, T. A. Schlacher, and B. Stantic, "Assessing fish abundance from underwater video using deep neural networks," Tech. Rep., 2018. [Online]. Available: https://pdfs.semanticscholar.org/4506/cf18906397dd8ce30fe6809d75fc97e375a4.pdf

[86] M. Zurowietz, D. Langenkämper, B. Hosking, H. A. Ruhl, and T. W. Nattkemper, "MAIA—A machine learning assisted image annotation method for environmental monitoring and exploration," *PLoS ONE*, vol. 13, no. 11, 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0207498

[87] M. Zurowietz and T. W. Nattkemper, "Unsupervised Knowledge Transfer for Object Detection in Marine Environmental Monitoring and Exploration," *IEEE Access*, vol. 8, pp. 143 558–143 568, 2020.

[88] M. Bryson, M. Johnson-Roberson, O. Pizarro, and S. Williams, "Automated registration for multi-year robotic surveys of marine benthic habitats," in *IEEE International Conference on Intelligent Robots and Systems*, 2013, pp. 3344–3349.

[89] D. Rao, M. De Deuge, N. Nourani-Vatani, B. Douillard, S. B. Williams, and O. Pizarro, "Multimodal learning for autonomous underwater vehicles from visual and bathymetric data," in *Proceedings - IEEE International Conference on Robotics and Automation*. Institute of Electrical and Electronics Engineers Inc., sep 2014, pp. 3819–3825.

[90] M. Johnson-Roberson, O. Pizarro, and S. Williams, "Saliency ranking for benthic survey using underwater images," in *11th International Conference on Control, Automation, Robotics and Vision, ICARCV 2010*. IEEE, dec 2010, pp. 459–466. [Online]. Available: http://ieeexplore.ieee.org/document/5707403/

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

[92] W. Abdulla, "Mask r-cnn for object detection and instance segmentation on keras and tensorflow," https://github.com/matterport/Mask_RCNN, 2017.

[93] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, jun 2015. [Online]. Available: http://image-net.org/challenges/LSVRC/2015/results

[94] Aleju, "Aleju/imgaug: Image augmentation for machine learning experiments." [Online]. Available: https://github.com/aleju/imgaug

[95] A. Bodenmann, B. Thornton, T. Nakataniy, and T. Ura, "3D colour reconstruction of a hydrothermally active area using an underwater robot," in *OCEANS'11 - MTS/IEEE Kona, Program Book*, 2011.

[96] J. T. Harvey, T. R. Loughlin, M. A. Perez, and D. S. Oxman, "Relationship between Fish Size and Otolith Length for 63 Species of Fishes from the Eastern North Pacific Ocean," *Fisheries Science*, no. August, p. 36, 2000.

[97] J. M. Durden, B. J. Bett, T. Schoening, K. J. Morris, T. W. Nattkemper, and H. A. Ruhl, "Comparison of image annotation data generated by multiple investigators for benthic ecology," *Marine Ecology Progress Series*, 2016.

[98] R. Garcia, R. Prados, J. Quintana, A. Tempelaar, N. Gracias, S. Rosen, H. Vågstøl, and K. Løvall, "Automatic segmentation of fish using deep learning with application to fish size measurement," *ICES Journal of Marine Science*, vol. 77, no. 4, pp. 1354–1366, oct 2020. [Online]. Available: https://academic.oup.com/icesjms/advance-article/doi/10.1093/icesjms/fsz186/5602457

[99] T. Ura, "Development timeline of the autonomous underwater vehicle in japan," *Journal of Robotics and Mechatronics*, vol. 32, no. 4, p. 713–721, 2020.

[100] L. M. Divine, F. J. Mueter, G. H. Kruse, B. A. Bluhm, S. C. Jewett, and K. Iken, "New estimates of weight-at-size, maturity-at-size, fecundity, and biomass of snow crab, Chionoecetes opilio, in the Arctic Ocean off Alaska," *Fisheries Research*, vol. 218, pp. 246–258, oct 2019.

[101] J. M. Roberts, "Cold-water coral reefs," in *Encyclopedia of Ocean Sciences*, 2019, pp. 675–687. [Online]. Available: https://www.ourplanet.com/wcmc/pdfs/Cold-waterCoralReefs.pdf

[102] M. Massot-Campos and G. Oliver-Codina, "Optical Sensors and Methods for Underwater 3D Reconstruction," *Sensors 2015, Vol. 15, Pages 31525-31557*, vol. 15, no. 12, pp. 31 525–31 557, dec 2015. [Online]. Available: https://www.mdpi.com/1424-8220/15/12/29864/htmhttps://www.mdpi.com/1424-8220/15/12/29864

[103] T. Yamada, M. Massot-Campos, A. Prugel-Bennett, O. Pizarro, S. Williams, and B. Thornton, "Guiding labelling effort for efficient learning with georeferenced images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2022.

[104] P. F. Culverhouse, R. Williams, B. Reguera, V. Herry, and S. González-Gil, "Do experts make mistakes? A comparison of human and machine identification of dinoflagellates," *Marine Ecology Progress Series*, vol. 247, pp. 17–25, feb 2003.