Automating Laser Point Detection in Antarctic Benthic Imagery to aid Biodiversity Monitoring

A REPORT PRESENTED
BY
Jasmine Ng

Department
Department of Physics

Collaborating Institution
British Antarctic Survey

Degree
MPhil Data Intensive Science

Supervision
Dr. Cameron Trotter

Trinity Hall
University of Cambridge
7th July 2024

Abstract

This thesis replicates the DELPHI ("DEtection of Laser Points in Huge image collections using Iterative learning") method and the evaluation framework introduced by Schoening et al. for detecting laser points in benthic imagery. The method combines k-means color clustering, spatial filtering based on laser geometry, and morphological operations to predict laser points. It was evaluated on a benthic image dataset captured by a towed camera and annotated by the British Antarctic Survey (BAS). The dataset includes two subsets: t1 (hard substrate) and t2 (soft substrate). t2 images consistently performed better than t1, likely due to t1 containing noisier backgrounds and often darker images. Given the small size of the dataset, evaluation was supported by Monte Carlo and k-fold cross-validation. These approaches revealed a similar performance trend to Schoening et al.'s study, where increasing the number of training images improved detection performance. However, stable performance in this study was only reached after using 19 to 22 training images, compared to 13 in Schoening et al.'s study. Combining t1 and t2 into a joint training set improved the classification performance on t2 compared to training on t2 alone. Future work could explore replacing the current basic color thresholding and morphological denoising methods with more robust machine learning approaches, in order to better model the variation in color, texture, and lighting found in benthic imagery.

List of Figures and Tables

List of Figures

List of Tables

Contents

Total LaTeX Word Count: 6888/7000

1 Introduction

1.1 Background on Benthic Research

Benthic research is the study of seafloor environnments and the organisms that habituate these environments. This field of research has become an increasingly important field in recent years due to the growing pressures of climate change. Current hot topics include coral reef conservation, benthic ecosystems of seamounts, and deep-sea mining of manganese nodules to fuel green technologies [1, 2, 3]. These efforts are essential for shaping policy in conservation through data-driven recommendations. For example, the 2023 Intergovernmental Panel on Climate Change (IPCC) report references benthic research on coral reefs, kelp forests, and seafloor-dwelling fauna, highlighting their vulnerability to climate change [4]. Hence, these benthic research publications are important for providing essential guidance in developing strategies and adaptation plans for future marine conservation efforts.

Many benthic biodiversity and exploration studies rely on image-based surveys [5, 6], where RGB (Red, Green, Blue) cameras collect true colour images in the visible light spectrum [7]. These surveys are often conducted during cruises expedition where cameras, attached to the boat or vessel, is submerged in the water and towed behind.

1.2 Motivation for Automating Benthic Image Analysis

As the volume of benthic imagery continues to grow, researchers face two major challenges. The first is the sheer quantity of data to process. For example, in 2025, BenthicNet compiled a global collection of benthic imagery from academic and federal institutions worldwide, curating 1.3 million images representing diverse seafloor environments captured using various imaging systems [8]. Long-established platforms like PANGAEA also host a wide range of openly accessible benthic image datasets from published research [9]. However, most of these images are unlabelled. Manual annotation, such as identifying the species present, is time-consuming and labor intensive, which can significantly delay the production of research outputs that inform policy. The second challenge lies in the complexity of the imagery. Varying environmental conditions, such as suspended sediments during image capture, can degrade image quality and make interpretation more difficult [10]. This further complicates the already challenging task of manual classification. Studies have shown that taxonomic experts are often prone to error during manual species labelling and may arrive at different conclusions when analysing the same data [11].

To mitigate these issues, there is growing interest in automating image analysis in marine science [12]. While machine learning holds promise, existing methods often fall short in accuracy or generalizability, and improvements are needed throughout the pipeline [13].

1.3 Laser Point Detection in Benthic Imagery

One key component of automating this pipeline is image calibration, which ensures consistent spatial resolution across datasets. Environmental factors such as camera depth, seabed slope, and water conditions can cause substantial variation in image scale. This inconsistency must be corrected for tasks that rely on precise spatial measurements, such as quantifying the size of sessile organisms [1]. A common approach to address this is laser calibration, where cameras are fitted with lasers that project three fixed laser points (LPs) onto the seafloor at known distances. By identifying LPs in an image, pixel distances can be converted into real-world units, such as centimeters. However, because LPs are often small red dots that are often difficult to detect manually, the process is tedious and error-prone, making automated LP detection a promising direction forward.

Automatic detection of laser points in benthic imagery presents significant challenges. For instance, red light, which has a longer wavelength, is rapidly attenuated in water compared to blue and green light, making red laser points difficult to detect [14]. Additionally, high water turbidity caused by suspended particles and organic matter leads to light scattering and image blurring, further complicating LP detection [15].

1.4 The DELPHI Method

A key advancement in automating the detection of LPs is the DELPHI method ("DEtection of Laser Points in Huge image collections using Iterative learning"), proposed by Schoening et al. [16]. DELPHI is a web-based tool that enables semi-automated LP detection. It allows users to manually annotate a small number of LPs, which the model then uses to learn relevant colour and spatial features. The trained model can subsequently detect LPs across large sets of unlabelled images. Experimental results from Schoening et al. demonstrated that with as few as 13 annotated samples, the method achieved an F1-score exceeding 80% [16].

1.5 Thesis Objectives

This thesis aims to implement the DELPHI method for the British Antarctic Survey (BAS), which collects large volumes of benthic imagery to monitor biodiversity in Antarctic seafloor environments. During initial investigation, it was found that the original DELPHI web tool is no longer available. Furthermore, the source code is written in C++ and is only accessible upon request from the original authors. As a result, BAS requires support in reimplementing DELPHI in a more accessible and researcher-friendly format, such as a Python package. This thesis addresses that need by reimplementing the DELPHI method as a Python package.

The first objective of this thesis is to reproduce the DELPHI method architecture and detection pipeline in Python, as described in Section 3. The second objective is to evaluate the implementation using the same performance metrics as Schoening et al.'s paper, specifically precision, recall, and F1 score, in order to assess reproducibility on datasets provided by BAS [16]. The aim is to determine whether comparable detection performance can be achieved. In addition to replicating the baseline, this thesis investigates several extensions to improve performance. These include applying cross-validation to assess the robustness of the results, optimizing colour clustering and morphological filtering parameters to enhance detection accuracy, and testing the model on combined transects to evaluate its generalizability across varying environmental and substrate conditions.

2 Dataset Overview

From this point onward, the paper by Schoening et al. will be referred to as the DELPHI paper, and the algorithm will be referred to as the DELPHI method [16].

2.1 Selection of the Dataset

The original datasets (T1 from the Clarion-Clipperton Zone and T2 from the Arctic) used by the DELPHI paper were not available for reuse. Upon contacting the authors, it was confirmed that the datasets could not be shared due to non-compliance with FAIR principles (Findable, Accessible, Interoperable, Reusable).

As an alternative, suitable datasets were sourced in collaboration with the BAS. Two benthic transects were selected from expedition PS118, conducted by the Alfred Wegener Institute between February and April 2019, specifically from cruises 6-9 [17] and 69-1 [18]. These datasets were previously used in a BAS study on predator–prey interactions, in which 39 images from cruise 6-9 and 61 images from cruise 69-1 were manually annotated with LP labels [19]. This combined set of 100 LP-labelled images forms the basis of the current study.

Cruises 6-9 and 69-1 gathered data from two sites along the western Antarctic Peninsula (Figure 1). Both the Arctic and Antarctic are known for their high turbidity, defined as the haziness in the water caused by sediments scattering light. This phenomenon is a direct result of surface melting and glacial runoff from the ice [20].

All images were acquired using the Ocean Floor Observation and Bathymetry System (OFOBS), a towed platform equipped with a high-resolution camera and three fixed LPs spaced 50 cm apart. Each image captures approximately $4\,\mathrm{m}^2$ of the seafloor at a resolution of 3840×5760 pixels. This setup differs from that used in the DELPHI paper, which employed the Ocean Floor Observation System (OFOS), a

similar towed camera platform. OFOBS is an enhanced version of OFOS, featuring additional capabilities such as acoustic sensors and synchronised bathymetric mapping [21]. The images collected using both OFOS and OFOBS are similar in photographic characteristics, which ensures comparable results to be evaluated between the DELPHI paper and this thesis.

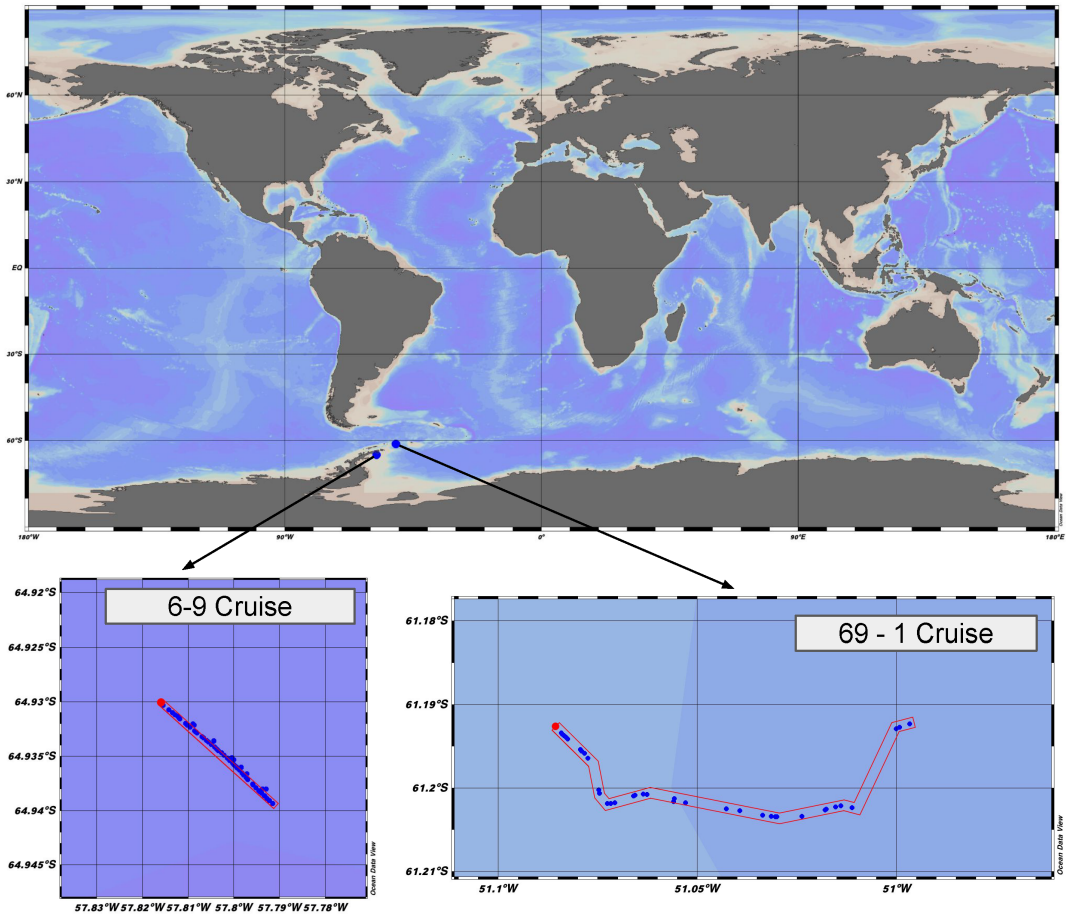
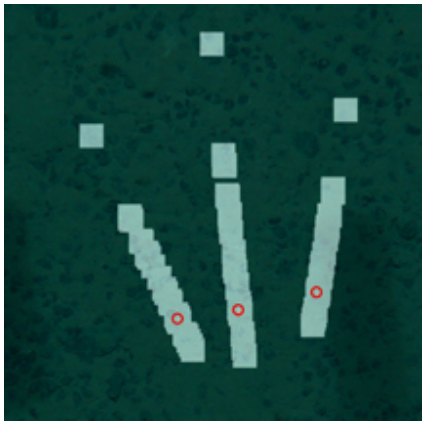


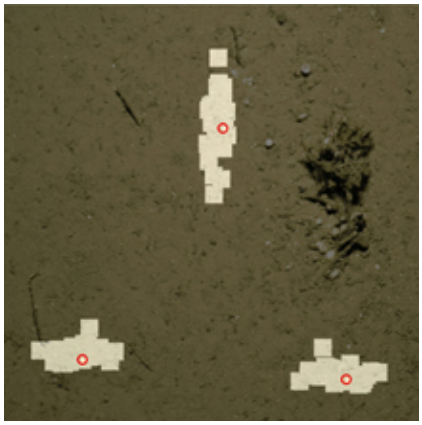
Figure 1: Cruises 6-9 and 69-1 data transects on a map

Images from Cruise 69-1 depict a hard substrate composed primarily of rocky material, as shown in Figure 2c. This terrain is visually and structurally more similar to the Clarion Clipperton Zone, where polymetallic nodules create a textured, high-contrast seafloor as seen in Figure 2a. In this thesis, data from the cruise 69-1 transect is referred to as t1 (lowercase) to distinguish it from the corresponding dataset in the DELPHI paper, which is denoted T1 (uppercase).

In contrast, images from Cruise 6-9 show a softer, clay-based substrate, as illustrated in Figure 2d. This environment more closely resembles the Arctic dataset T2 in Figure 2b. In this thesis, data from the cruise 6-9 transect is referred to as t2 (lowercase) to distinguish it from the corresponding dataset in the DELPHI paper, which is denoted T2 (uppercase). The similarity in substrate type and visual appearance was a key motivation for selecting these transects, with the aim of replicating the original data conditions as closely as possible.

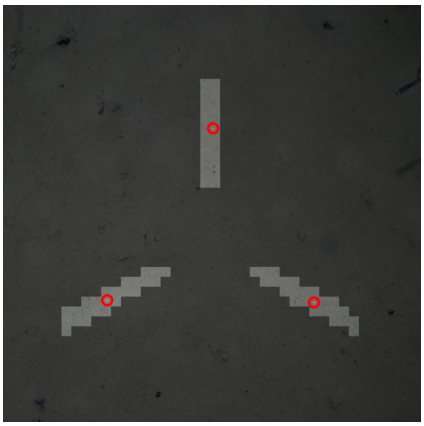


(a) Example image from the T1 transect in the Clarion Clipperton Zone, showing a seafloor substrate with polymetallic nodules, as presented in the DELPHI paper



(b) Example image from the T2 transect in the Arctic, showing a clay-based substrate, as presented in the DELPHI paper

(c) Example image with ID 0381 from t1 from Cruise 69-1 as part of the BAS dataset.



(d) Example image with ID 0027 from t2, from Cruise 6-9 as part of the BAS dataset.

Figure 2: Example images from the datasets used in this thesis and in the DELPHI paper. Panels (a) and (b) show T1 and T2 from the DELPHI paper. Panels (c) and (d) show the corresponding t1 and t2 transects from the BAS dataset. Translucent squares indicate the typical locations of LPs, with automatically detected LPs circled in red. Higher-resolution versions of the original images were unavailable due to lack of access to the original datasets.

Another motivation for selecting t1 and t2 as this thesis's experimental data was to capture the variation in LP colour distributions observed in the DELPHI paper's T1 and T2 datasets. t1 and t2 of the BAS data shown in Figure 3 also exhibit contrasting average RGB values, with t2 displaying higher intensity across all colour channels. Thus, the BAS t1 and t2 data provides a valuable test case for evaluating the adaptability of the DELPHI detection method under diverse imaging conditions.

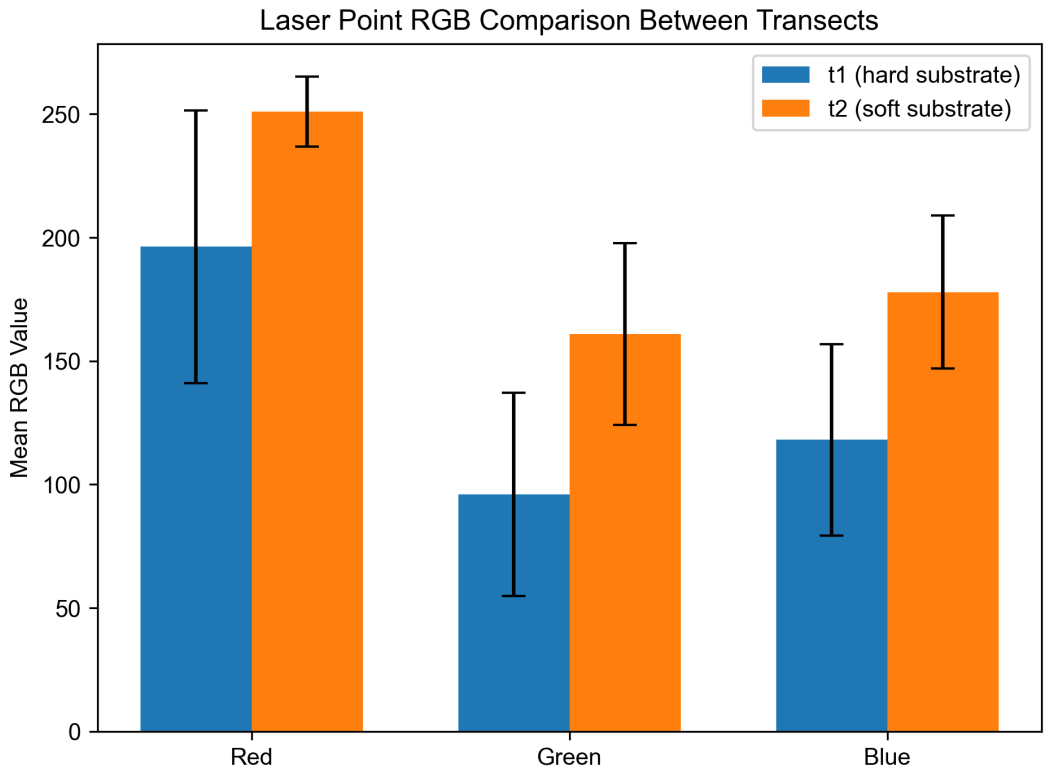
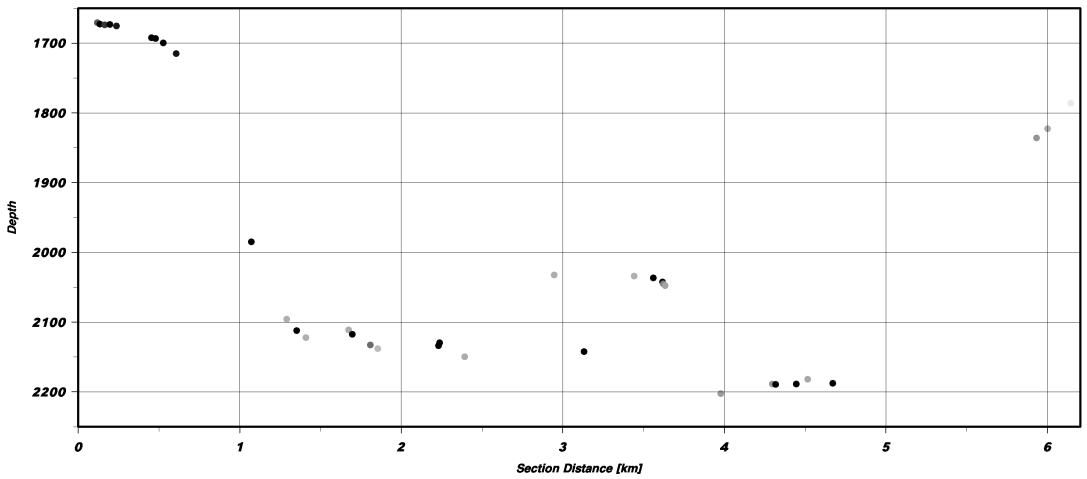
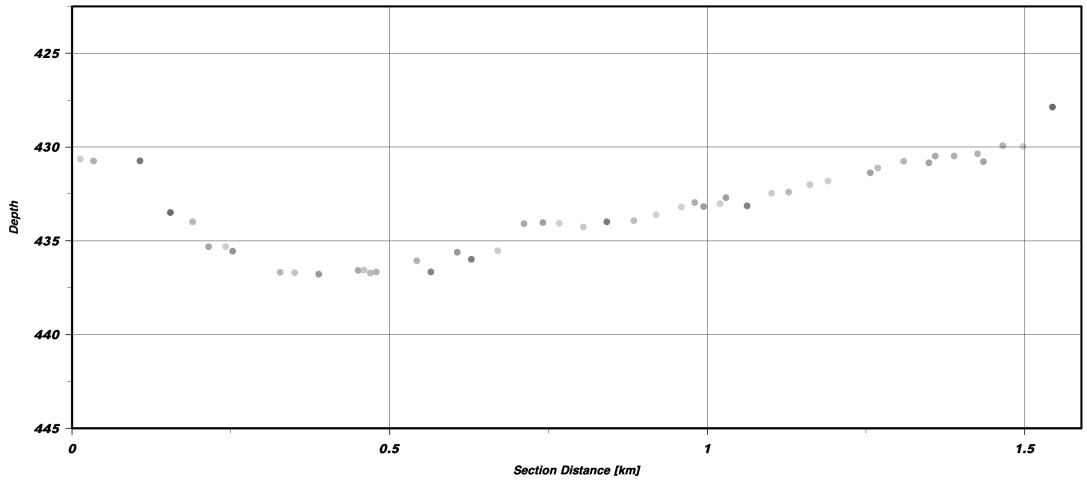


Figure 3: Average RGB values for t1 and t2 transects

The darker appearance of images captured in the t1 transect is likely associated with the depth at which the transect was conducted. As shown in Figures 4a and 4b, t1 covers a broader range from 1650 to 2200 metres, situated in the deeper and darker aphotic zone, whereas t2 spans a narrow depth range between 425 and 440 metres, placing it within the twilight zone. It is important to note that depth here refers to the absolute depth below the ocean surface and does not reflect the distance between the camera and the seafloor, which remains relatively constant across both transects.



(a) t1 depth profile (1650–2200 m, aphotic zone).



(b) t2 depth profile (425–440 m, twilight zone).

Figure 4: Depth profiles for the t1 and t2 transects. These reflect absolute depth below sea level, not the camera-to-seafloor distance.

2.2 Limitations of the dataset

However, there are two key limitations associated with the selected dataset. Firstly, both t1 and t2 use the same isosceles triangle LP configuration, as shown in Figures 2c and 2d. In contrast, T1 and T2 from the DELPHI paper (Figures 2a and 2b) feature differing LP spatial layouts. This limitation reduces the ability to assess the detection method's robustness to variation in LP geometry. Nevertheless, it is important to note that most Antarctic benthic research currently uses the isosceles triangle configuration, so this limitation may have limited practical impact.

Secondly, the BAS dataset is relatively small, with t1 and t2 making up to a total of 100 annotated images. This may lead to unreliable model performance when using standard evaluation metrics. In comparison, the DELPHI study used approximately 1200 test images, providing a more statistically robust evaluation. It is also important to note that most Antarctic benthic image datasets are small, as manual annotation is time-consuming. As such, this limitation cannot be addressed within the scope of this project. However, it will be mitigated in the extended methodology (Section 3.5) through the use of Monte Carlo cross-validation and k-fold cross-validation to produce more reliable performance estimates and quantify variability.

2.3 Preprocessing of the Dataset

Preprocessing of the dataset is required to ensure the correct functioning of the current methodology. While future versions of the package may reduce these requirements, this preliminary version assumes that input data adheres to a specific structure.

1. Image and Annotation Format
Each image must be provided in .jpg format, accompanied by a .json annotation file containing metadata about the LP bounding boxes. These annotation files should follow the format generated by the LabelMe tool [1]. Images and annotations must be stored in separate folders with corresponding filenames to enable correct pairing during training and evaluation.

2. Three annotated LPs per image
For spatial layout analysis to function correctly, each image must contain exactly three labeled LPs. This is essential for the algorithm to learn consistent geometric patterns. Any image with fewer or more than three annotations is excluded from the dataset to maintain the validity of the method.

3. Uniform image dimensions
All images must have identical dimensions. Spatial layout learning is sensitive to image size, and inconsistent dimensions can disrupt the detection process. While resizing, padding, or cropping could theoretically address size mismatches, these operations distort the spatial relationships between LPs and significantly degrade performance. Therefore, it is strongly recommended to ensure that all images are of the same resolution, typically guaranteed if captured using the same camera system under consistent settings.

4. Unique image identifiers
Each image and its corresponding annotation file must have a unique identifier, typically a four-digit numerical ID appended to the filename. For example, a valid filename might be `TIMER_2019_03_06_at_06_21_00_IMG_0144.JPG`. Duplicate IDs can lead to conflicts during data loading or training. During preprocessing of the BAS dataset, several filenames were modified to resolve such conflicts and ensure uniqueness across the dataset.

3 Methodology

3.1 Overview of the DELPHI method

This thesis will implement DELPHI as a package, which is a semi-automated laser point detection method that learns from a small set of manually annotated images. A high-level overview of the DELPHI learning and detection process is shown in Figure 5, with details in the training and detection phase in Section 3.2 and Section 3.3.

[1] `https://github.com/wkentaro/labelme`

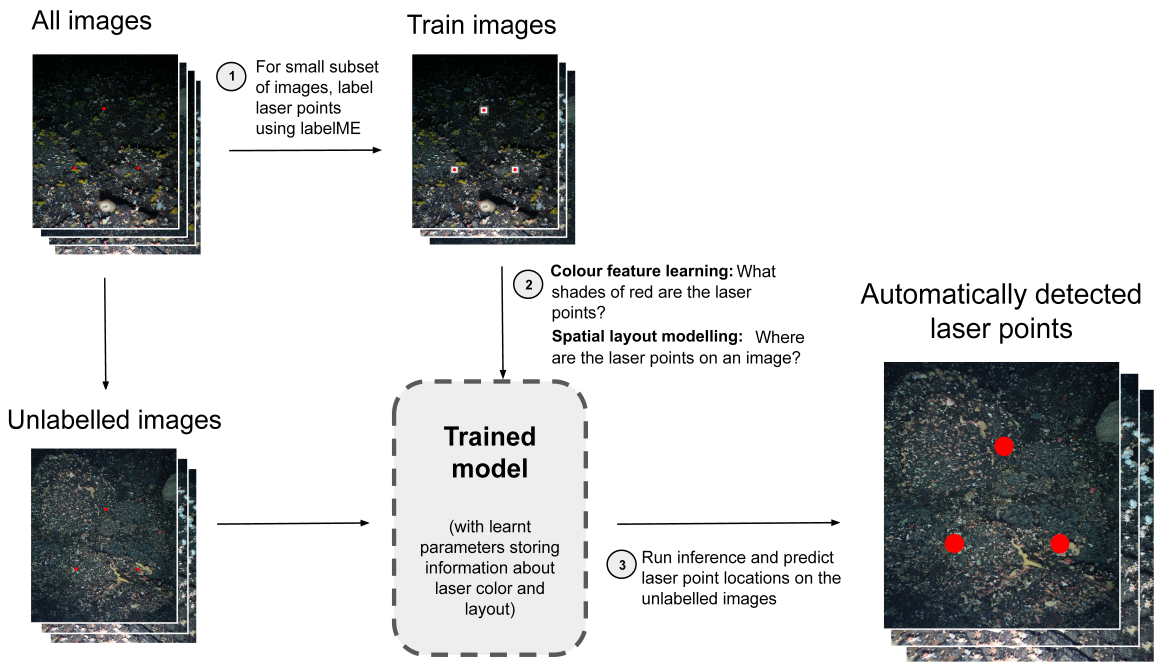


Figure 5: High-level overview of the DELPHI training and testing process

Using the LabelMe tool, researchers label laser points in a few training images. DELPHI then learns typical colour features and spatial layouts from these annotations and stores the parameters in a model. This model is used to detect laser points in the remaining unlabelled images.

3.2 Training Step

The training process consists of two components: colour feature learning and spatial layout modelling, both of which are described in mathematical detail in the methodology section of the DELPHI paper.

3.2.1 Colour feature learning

Colour feature learning aims to distinguish laser point colours from the background. This is achieved using two predefined radii, δ_1 and δ_2, centred on annotated laser point positions, as depicted in Figure 6. Pixels within δ_1 are labelled as LP colours, while pixels near the perimeter of δ_2 are treated as background colours. These two sets of RGB values are combined into a single dataset and used as input to a k-means clustering algorithm.

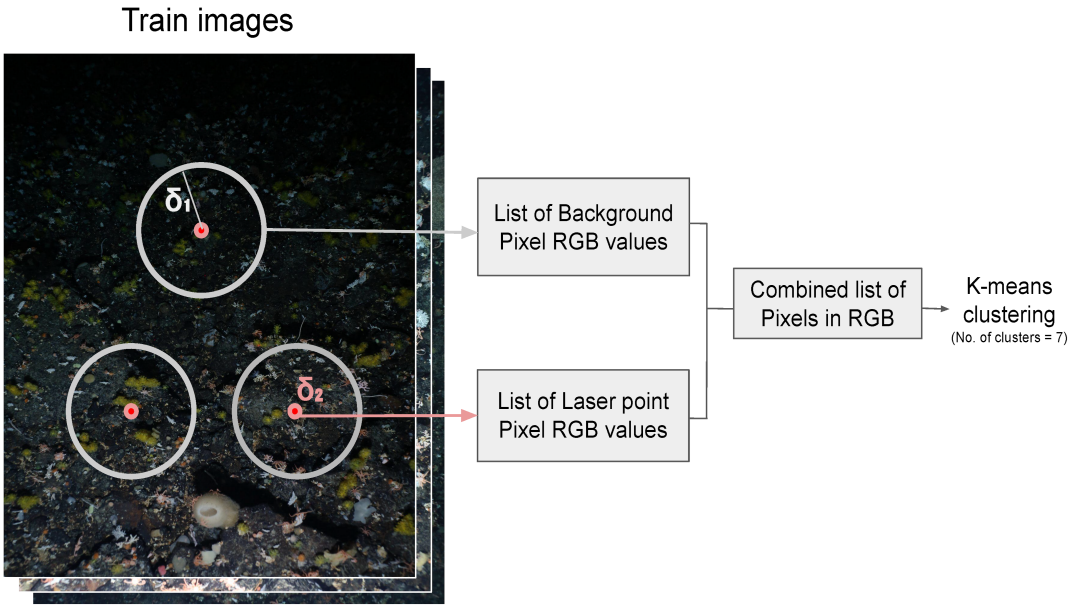


Figure 6: Demonstrating the selection of laser point and background colours using δ_1 and δ_2

K-means clustering is an unsupervised machine learning technique used to learn representative laser point colour features in a flexible, data-driven manner [22]. Unlike fixed-threshold or rule-based methods, k-means adapts to the actual range of colours present in the training data, effectively learning the specific imaging conditions embedded in the annotated laser points. The algorithm begins with a user-defined number of clusters, set to seven as in the DELPHI paper, and iteratively assigns pixels to the nearest cluster centre based on Euclidean distance in RGB space. Cluster centres are updated until convergence. The cluster containing the highest proportion of LP-labelled pixels is selected as the representative laser point colour group. From this cluster, two parameters are extracted and stored for downstream detection.

- **Threshold:** the average Euclidean distance from each LP colour in the selected cluster to its cluster center
- **Curated laser point colours:** the RGB values of all pixels belonging to the selected cluster

3.2.2 Spatial Layout Modelling

Spatial layout modelling is used to incorporate spatial priors into the detection process by identifying regions where laser points are likely to appear, as shown in Figure 7. This is based on the assumption that the spatial distribution of laser points remains relatively consistent across similar imaging conditions.

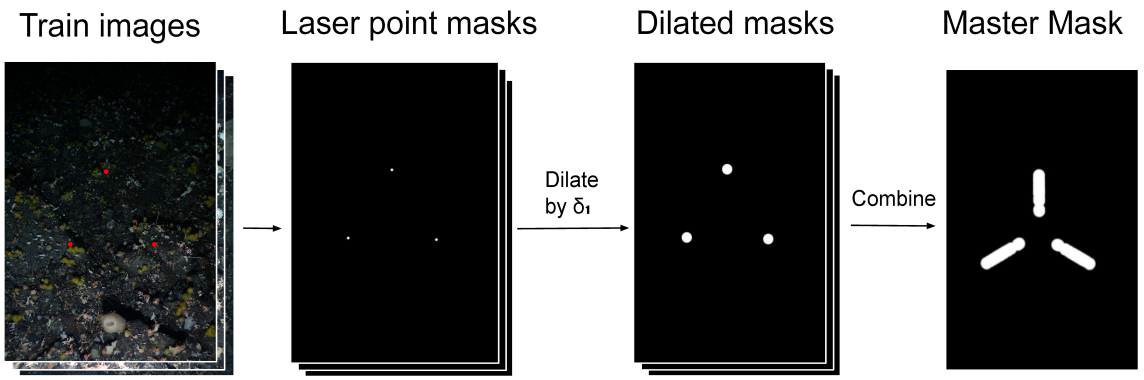


Figure 7: Demonstrating the construction of a master mask using spatial layout modelling through dilating laser points by δ_1

All annotated laser point positions in the training set are used as reference locations, and each point is expanded into a circular region by applying a fixed dilation radius δ_2. These circular regions represent plausible areas around each ground truth point that account for minor variation in position due to imaging noise or annotation uncertainty. From this, a master mask is extracted and stored for downstream detection:

- **Master Mask:** a binary image formed through dilating all annotated laser point locations by radius δ_2 and taking their union, used to restrict detection to spatially plausible regions

3.3 Detection Step

The detection process applies the trained model to unlabeled images using the stored colour features, threshold, and master mask, as seen in Figure 8. This subsection will expand on each of the steps associated in the model detection pipeline.

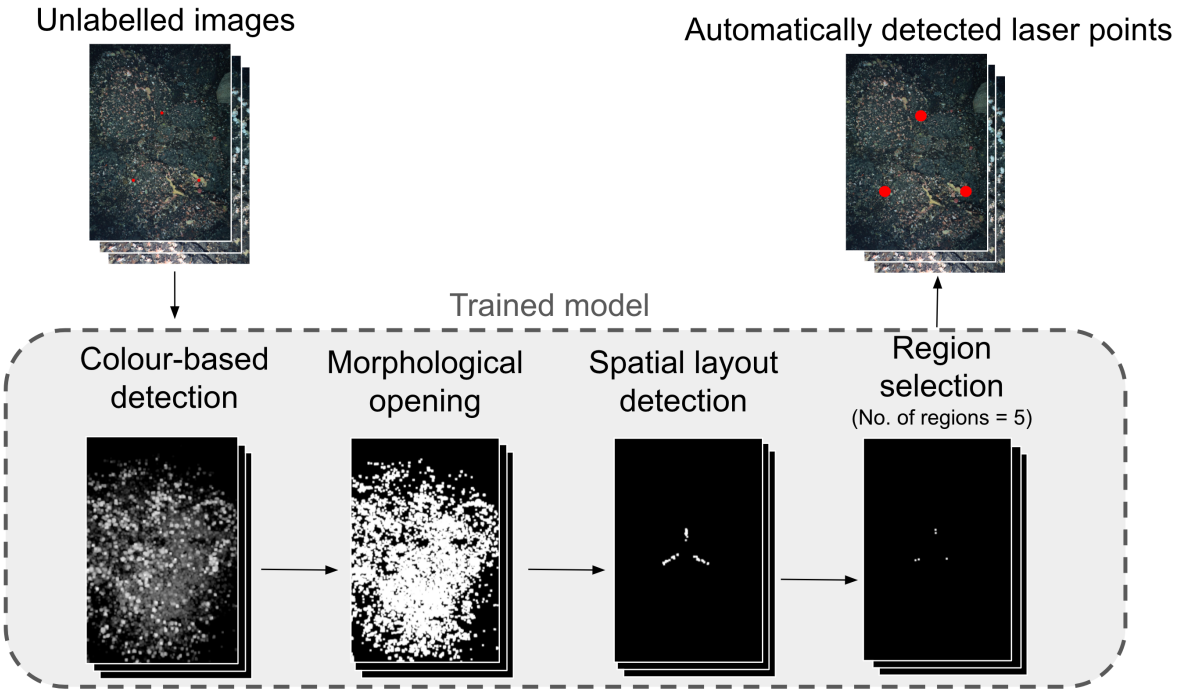


Figure 8: Demonstrating the DELPHI detection pipeline using trained model

3.3.1 Colour-based detection with threshold and curated LP colours

Candidate pixel selection evaluates each pixel by computing its minimum Euclidean distance to the curated laser point colours in RGB space. Pixels with a distance below the learned threshold are selected as candidates. These distances are further converted into grey values using the following formula:

$$\text{grey\_value} = \max\left(0, \frac{\text{threshold} - \text{min\_dist}}{\text{threshold}^2}\right)$$

This transformation assigns higher grey values to pixels that are more similar in colour to the laser point features, while suppressing values that exceed the threshold.

3.3.2 Morphological Opening for Denoising images

To suppress noise, a morphological opening operation is applied to the grayscale response image. This consists of an erosion followed by a dilation, which removes isolated pixels unlikely to represent true laser points while preserving larger, connected regions.

3.3.3 Spatial detection with Master Mask

After this, the master mask is applied to restrict detection to plausible regions where laser points were previously observed during training. While the DELPHI paper only mentioned the mask in the training phase and did not apply it during detection, correspondence with the authors and inspection of their original C++ code revealed that this mask was in fact used in the final detection pipeline. Specifically, the mask is applied through element-wise multiplication to suppress false positives in implausible spatial locations.

3.3.4 Region selection and Point Prediction

Region selection and scoring are performed by identifying connected components in the refined binary mask. Each region is assigned a confidence score by summing the grey values of all pixels it contains. As the DELPHI paper selected the top five highest-scoring regions as final candidates for further analysis, this thesis follows the same approach by retaining five regions.

To predict laser points, the centre point of each selected region was computed. All triangle combinations formed by these five centers are evaluated against reference triangle configurations derived from the training data. The triangle with the most similar spatial configuration is selected, and its three vertices define the predicted laser point positions.

3.4 DELPHI Performance Evaluation

After implementing the DELPHI package, this thesis reproduces the systematic evaluation conducted by the DELPHI paper to investigate how detection performance scales with the number of training images. The evaluation workflow was adapted to incorporate a train/test split, and varying training set sizes (e.g., 1, 2, 3, ..., 10, 13, 16, ..., 27, 25, 30 images) were tested. Each iteration followed the same three-stage pipeline described in Figure 5, with slight modifications as detailed below:

1. **Train/Test Split**
   All available images were first manually annotated to mark LP locations. A subset of these labelled images was selected as the training set, while the remainder formed the test set, with ground truth annotations reserved for evaluation. As the DELPHI paper did not specify the method for splitting, a standard 80:20 train-test ratio was used with the `train_test_split()` function from Scikit-learn[2]. When N' images are used for training, then N' images are randomly sampled from the training portion.

2. **Model Training**
   The DELPHI model was trained on the selected training subset. This included both colour feature learning and spatial layout modelling, thereby capturing the characteristic colour and spatial distribution of LPs specific to the training subset.

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html`

3. **Model Evaluation**
   The trained model was then applied to the test images to predict LP locations. The predicted LPs were evaluated against the ground truth annotations of where LPs were located.

   As the DELPHI paper did not specify a distance threshold for matching predicted and ground truth LPs, we adopted a practical assumption: the same 25-pixel radius (δ_1) used during spatial layout modelling was applied as the matching criterion. That is, a predicted LP was counted as a true positive if it lay within 25 pixels of a ground truth LP.

   Standard classification metrics were computed to quantify performance:

   - **Precision:** The proportion of correctly identified laser points from the total number of identified laser points
   - **Recall:** the proportion of correctly identified laser points from the total number of ground truth laser points
   - **F1-score:** the harmonic mean of precision and recall.

3.5 Extensions

3.5.1 Cross-validation

Cross-validation was performed using two approaches: Monte Carlo and k-fold validation. This was motivated by initial observations that performance metrics varied significantly depending on the specific test images likely due to the small test data size. Monte Carlo cross-validation mitigates this by repeatedly sampling random train-test splits 5 times and computing the average performance along with standard deviation. Additionally, 5-fold cross-validation was used, where the dataset was divided into five equal parts. Each fold was used once as the test set while the remaining four served as training data.

3.5.2 Parameter Tuning

Parameter tuning was conducted to evaluate whether adjustments to the default DELPHI settings could improve detection performance on this dataset. Two set of parameters were selected for tuning:

1. δ_1 **and** δ_2**:** δ_1 defines the background radius and determines the diameter of the dilated master mask used to restrict where laser points can be located. It controls how strictly background regions are excluded from consideration. δ_2 defines the laser point radius and determines the area used to extract pixel values for laser point colour curation. It influences the colour thresholds and curated LP colours by defining which pixels are labelled as laser points for the training step.
2. **Morphological Opening kernel size:** Morphological opening is applied to denoise test images during the detection step. However, larger kernel sizes may unintentionally remove valid laser point pixels, especially in noisy or low-contrast images. This can increase false negatives and reduce recall.

A grid search was conducted over a range of values for δ_1, δ_2, and the morphological kernel size using a held-out validation set. The goal was to identify the combination that maximised F1-score.

3.5.3 Mixture of substrates

Both substrate types (t1 and t2) are combined into a single training and testing set, allowing the algorithm to generalise across varying conditions. This was iterated though increasing training sizes to evaluate performance.

4 Results and Discussion

4.1 Results: Baseline replication

This thesis's implementation of DELPHI on the BAS dataset demonstrated increasing precision, recall and F1-score with increasing training data size from 1 to 30 images. This is consistent with findings from the DELPHI paper, as shown in Figures 9 and 10. However, the performance trends differ between the two studies. Higher detection performance was achieved on the t2 (soft substrate) dataset and lower

performance on t1 (hard substrate). By contrast, the DELPHI paper reported better performance on T1 (hard substrate) than on T2 (soft substrate). Additionally, the performance across different training sizes varied greater in the BAS dataset than in the DELPHI paper, which may be due to the small test data size of t1 and t2.

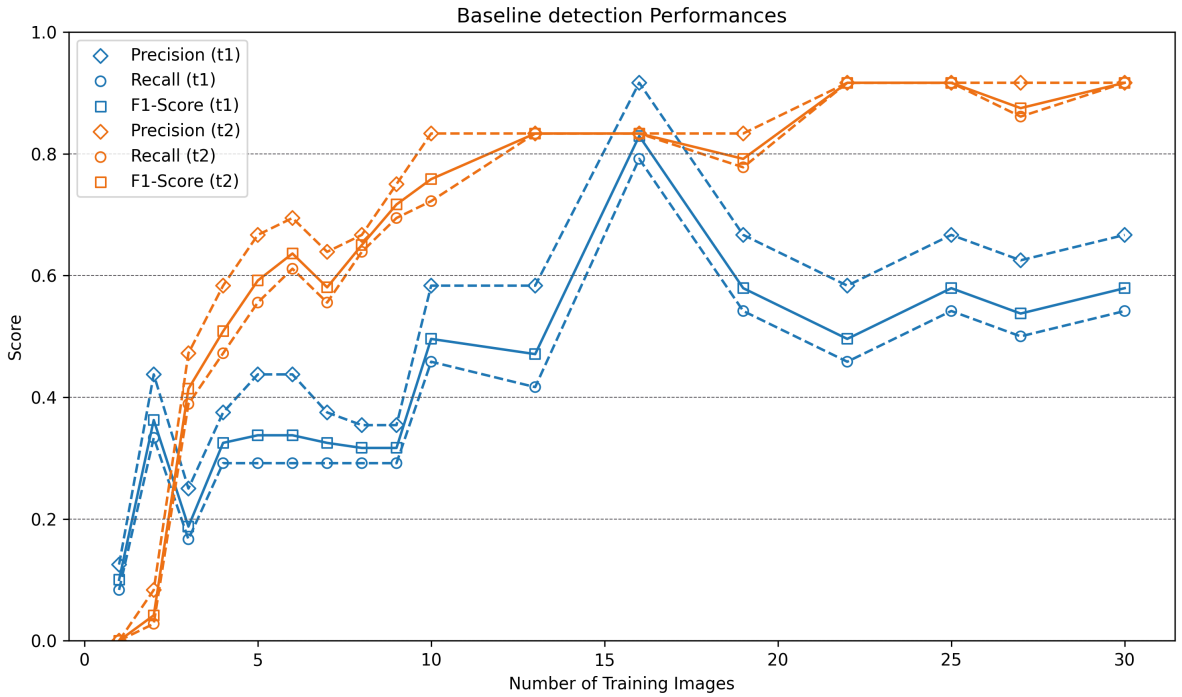


Figure 9: Precision, recall, and F1-score of DELPHI performance on the BAS dataset

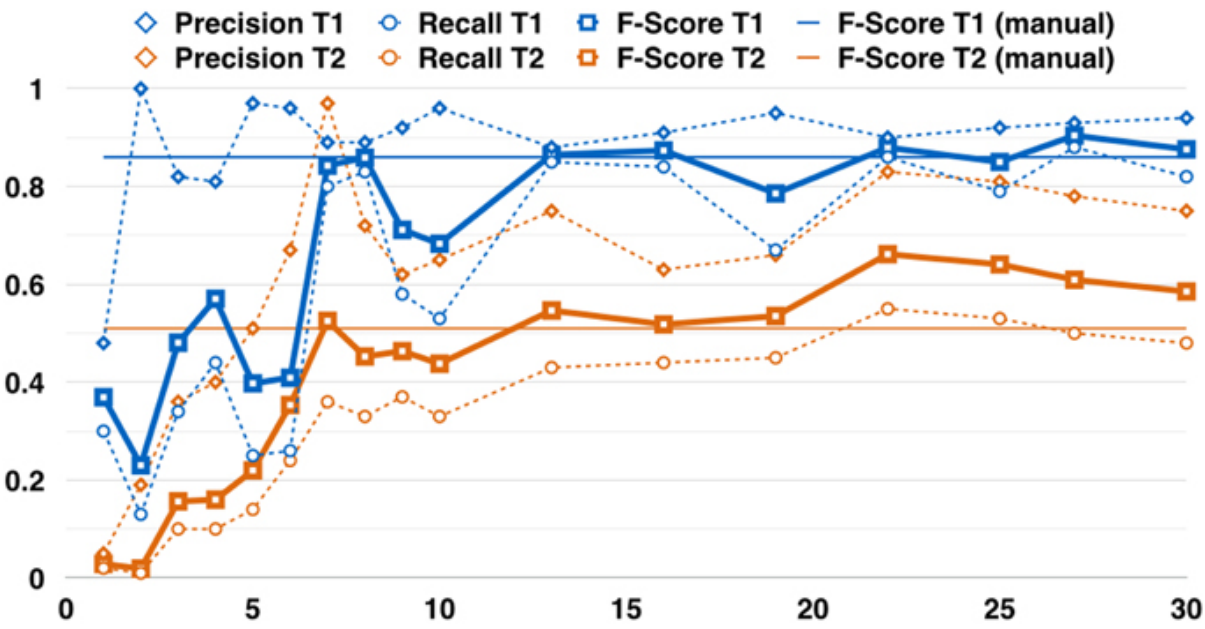


Figure 10: Performance results from the DELPHI paper

The DELPHI paper assessed performance saturation using a percentage difference metric, defined in Equation 1, where N' denotes the training size and K the number of larger training sizes larger than N'. If the absolute percentage difference falls below 8%, performance is considered to have plateaued.

$$\text{Percentage difference} = F1(N') - \frac{1}{K}\sum_{i=1}^{K} F1(N_i), \quad \text{where } N_i > 13 \tag{1}$$

According to this metric, the DELPHI paper observed a performance plateau at 8% percentage difference after 13 training images, with average F1-scores of 0.86 for T1 and 0.58 for T2, as summarised in Table 1. In contrast, the BAS dataset did not reach this plateau at the same point. At 13 training images, the percentage differences were 13% for t1 and 4% for t2, with corresponding F1-scores of 0.60 and 0.88. Performance saturation in the BAS data was instead reached at 19 images for t1 and 22 images for t2, with plateau F1-scores of 0.58 and 0.90, respectively.

Metric	**the DELPHI paper findings**		**Experiment findings**	
	T1	**T2**	**t1**	**t2**
Percentage point difference at N'=13	-8	+8	-13	-4
Average F-Score after N'=13'	0.86	0.58	0.60	0.88
Number of training images to plateau	13	13	19	22

Table 1: Comparison between the DELPHI paper and experiment findings on detection performance.

4.2 Discussion: Baseline Replication

A key distinction between this study and the DELPHI paper is the reversal in performance between hard and soft substrates. The DELPHI paper reported superior results on the hard substrate (T1), while this study observed better performance on the soft substrate (t2). This suggests that grouping images by substrate type does not guarantee consistent detection outcomes. While both T1 and t1 are rocky, and both T2 and t2 are clayey, other factors such as colour variation, surface texture, and water clarity significantly influence detection accuracy. This is further analysed below.

4.2.1 Discrepancy 1: Variable colour and texture

While both are considered hard substrate, T1 and t1 differ greatly by colour and texture. Figure 2 shows that the T1 transect from the DELPHI paper contains uniformly dark polymetallic nodules, providing strong contrast for identifying red LPs. In contrast, the t1 transect from BAS exhibits a heterogeneous, rocky substrate with a variety of benthic organisms overlaying the surface. This results in large patches of bright noise, which are likely to be falsely identified as LPs, thereby reducing detection performance.

4.2.2 Discrepancy 2: Water Turbidity

Differences in marine conditions may also have contributed to the discrepancies in detection performance between this thesis and the DELPHI paper. In the DELPHI paper, T1 was located in the Clarion Clipperton Zone of the deep Pacific, while T2 was situated in the Arctic. These environments differ significantly in turbidity. The Clarion Clipperton Zone is characterised by deep, stable waters with minimal sediment resuspension, resulting in low turbidity and improved visibility for LP detection [23]. In contrast, Arctic waters are subject to glacial melt, seasonal mixing, and high sediment loads, all of which contribute to elevated turbidity levels and degraded image quality [24].

In this thesis, the BAS transects t1 and t2 were both collected in West Antarctica, where water conditions are expected to be relatively consistent. Given this similarity in turbidity, substrate composition emerges as the primary differentiating factor between the two. The superior detection performance on t2 relative to t1, despite similar water clarity, highlights the substantial influence of substrate colour and texture on LP detection.

4.2.3 Analysing Failed Predictions

Representative examples of failed predictions are shown in Figures 11 and 12. These examples were selected because they illustrate the primary issues underlying the failed detections in t1 and t2. All remaining failed cases are included in Appendix B.

In the case of the t1 transect, two main categories of images consistently led to detection failures. The first category comprises dark images, where the main issue lies in the initial colour thresholding step. With dark images, the true LPs did not meet the trained colour threshold due to low contrast, and were consequently discarded early in the detection pipeline. This typically results in false negatives, where one or more LPs are missed, negatively impacting the recall.

The second category consists of noisy images, where many bright pixels passed the colour thresholding stage. The morphological opening step, which assumes the background noise consists of small, isolated pixels, was ineffective in these cases because the substrate contained larger, blob-like rock features that resemble LPs. As a result, this step failed to remove more complex noise patterns. While the spatial layout filtering using the master mask was effective in narrowing down candidate LPs, it occasionally selected incorrect candidates in these noisy conditions. This resulted in false positives, which in turn lowered the precision.

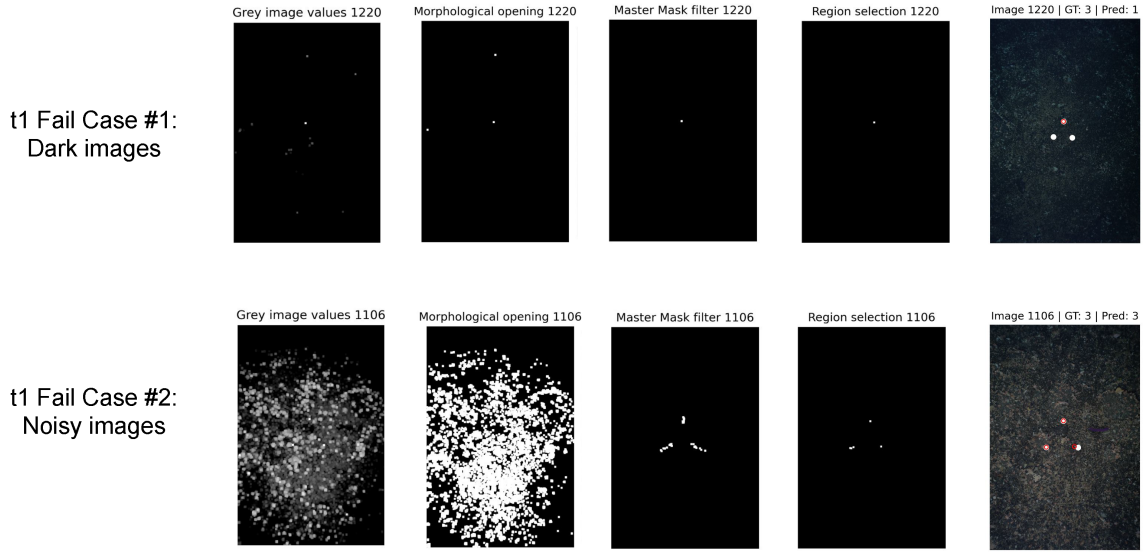


Figure 11: Detection process for two t1 examples where at least one LP was misclassified

For the t2 transect, the most common failure cases involved dark images with low contrast, likely due to the same reasons as discussed before.

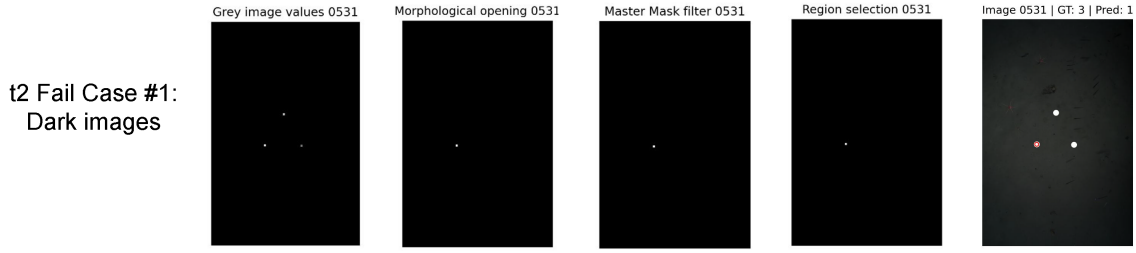


Figure 12: Detection failure in t2 where at least one LP was misclassified

These observations highlight several limitations in the baseline implementation. The colour thresholding step lacks flexibility and fails to adapt to dark images with low contrast. Morphological opening alone is not a strong enough denoising strategy to substrate-related background noise, while the spatial layout filtering using the master mask proves valuable in improving selection accuracy. Thus, the DELPHI method may not be well adapted to noisy, hard substrates.

4.2.4 Analysing Successful Predictions

Figure 13 illustrates examples that consistently achieved correct detection of all three LPs across all training sizes. Image 0282 from t1 and several from t2 performed reliably. These images are typically brighter and contain fewer large, textured features, resulting in fewer false positives and higher detection performance.

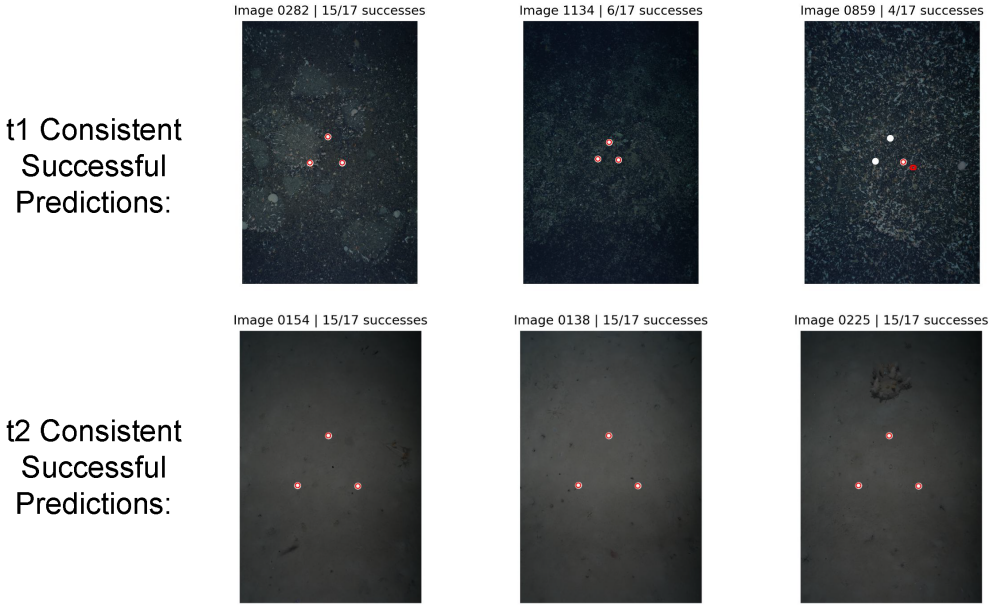


Figure 13: Examples of consistently successful detection of all three LPs across all training sizes for both t1 and t2. The number of successful detections out of the total iterations (17 different training sizes) is indicated in the title of each image.

4.2.5 Colour Clustering in K-means

Since colour clustering appears to be the primary issue, it is useful to examine how the DELPHI method grouped the training data colours to determine the colour threshold. This is illustrated in Figures 14 and 15.

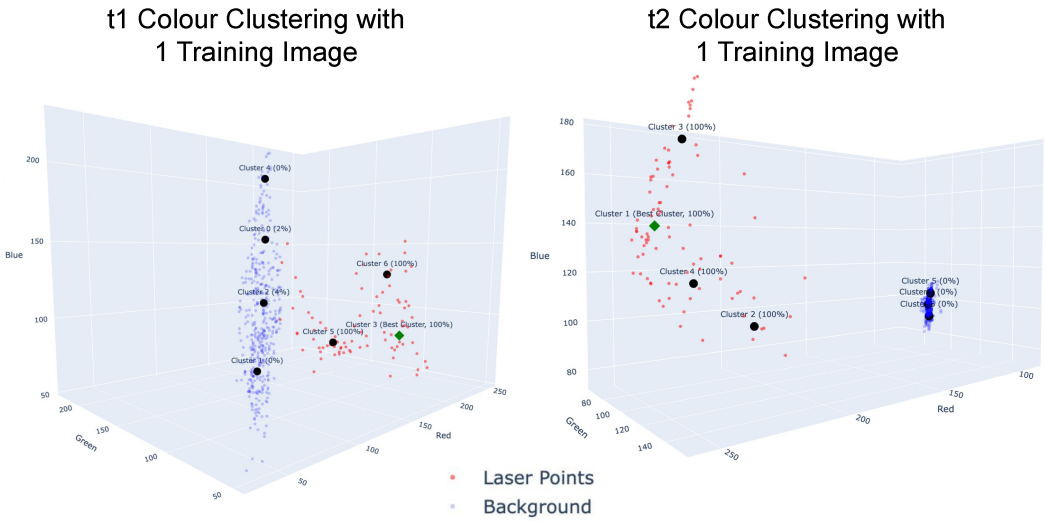


Figure 14: K-means clustering of background and LP colours using 1 training image.

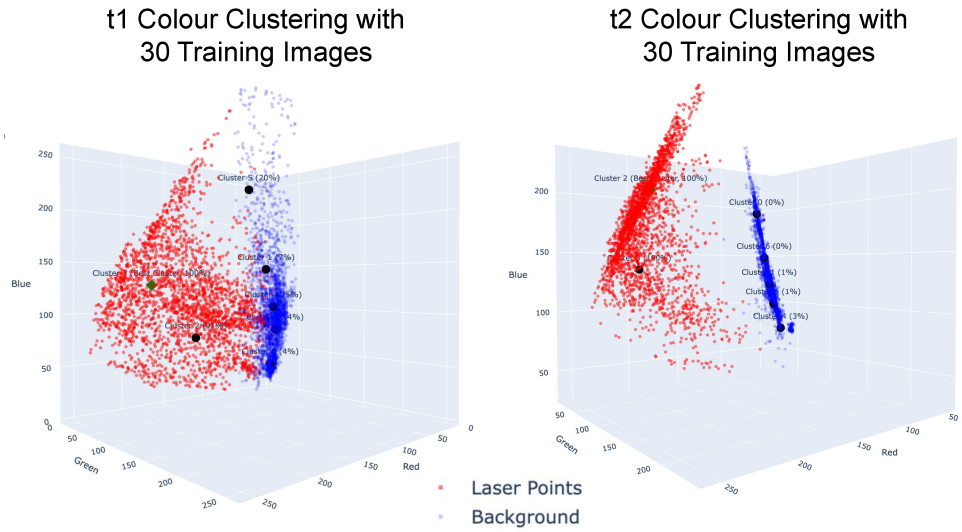


Figure 15: K-means clustering of background and LP colours using 30 training images.

At low training sizes, t1 and t2 produced different RGB clusters for the curated LPs. t1 identified high red and green with low blue, while t2 selected high red, green, and blue. With 30 training images, the clustering improved, but t2 maintained a clearer boundary between laser and background colours. In contrast, t1 exhibited overlap, suggesting that thresholds δ_1 and δ_2 might need adjustment for better discrimination.

4.2.6 Summary: Baseline Replication

These findings suggest several key takeaways and opportunities for further investigation. Firstly, while the increase in performance with larger training sets replicates the trend observed in the DELPHI paper, the trend is highly variable, likely due to the small dataset used. This limitation is addressed through Monte Carlo and k-fold cross-validation in Section 4.3. Secondly, poor colour clustering in t1 likely contributed to lower detection performance and may be improved through parameter tuning to better separate LP and background pixels. This tuning process is discussed in Section 4.4. Thirdly, substrate

composition plays a critical role by introducing varying levels of noise and texture that affect detection performance. Scoulding et al. [25] suggest that combining high and low performing images during training can enhance overall model performance on more challenging substrates. This approach is evaluated in Section 4.5, where training and testing are conducted on mixed-substrate datasets.

4.3 Extension 1: Cross validation

Monte Carlo cross-validation was used to evaluate the robustness of the detection method. As shown in Figure 16, the observed performance trends are consistent with those reported in the DELPHI paper. Variability in detection performance decreases as the training set size increases, confirming that earlier fluctuations were largely due to limited data availability. However, the t1 transect continues to exhibit large error bars, likely due to its smaller sample size of only 38 images. The presence of just a few low-quality test images, such as those with poor lighting or low contrast, can significantly impact the prediction performance results.

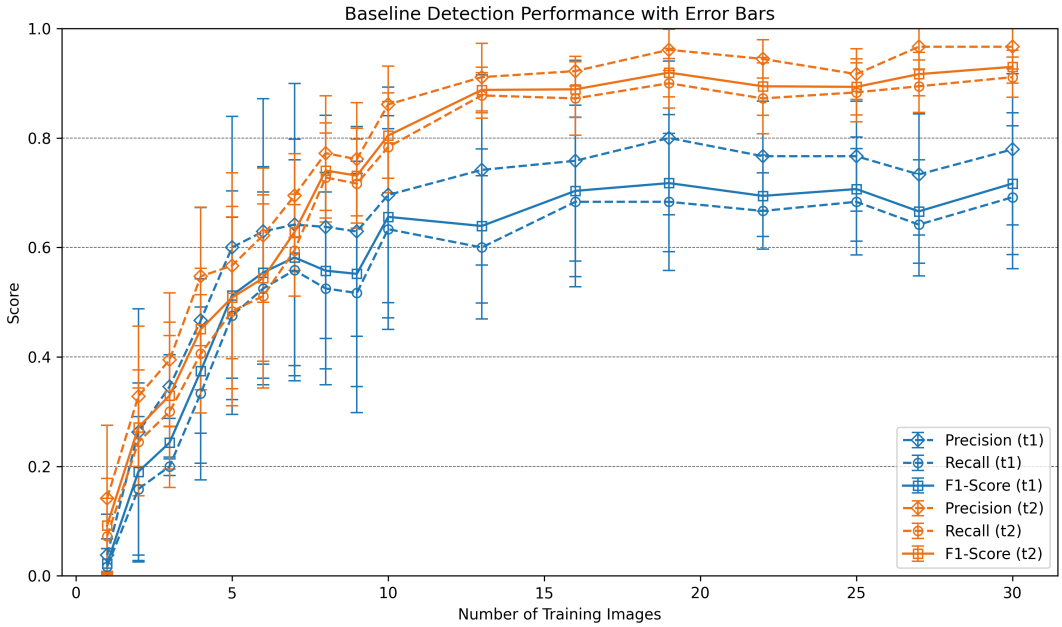


Figure 16: Baseline performance evaluated using Monte Carlo cross-validation across five random seeds

K-fold (5-fold) cross-validation across the full dataset, shown in Figure 17, demonstrate that t2 consistently outperforms t1. This provides further evidence that the DELPHI method performs more reliably on soft substrates with more homogenous backgrounds.

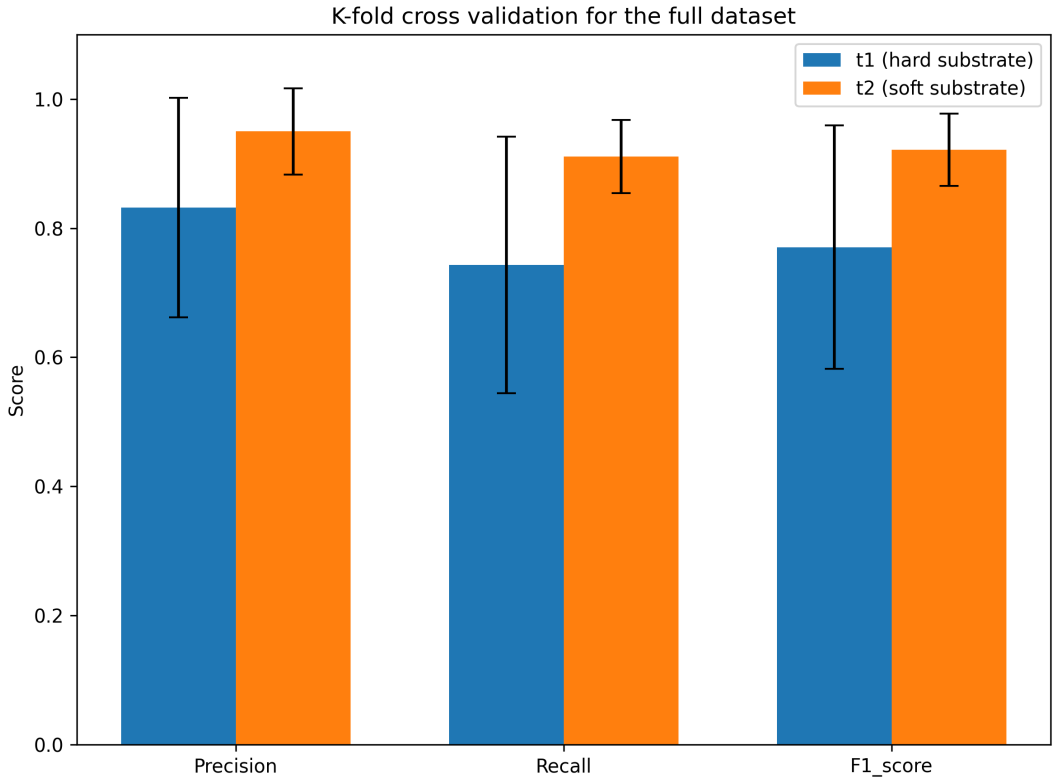


Figure 17: Five-fold cross-validation evaluation of the full dataset.

4.4 Extension 2: Parameter tuning

Tuning the morphological opening kernel reveals that increasing the kernel size leads to a decline in performance as shown in Figure 18. The DELPHI paper's default setting for a kernel size of 3 also proves effective for the BAS dataset, balancing denoising with signal retention.

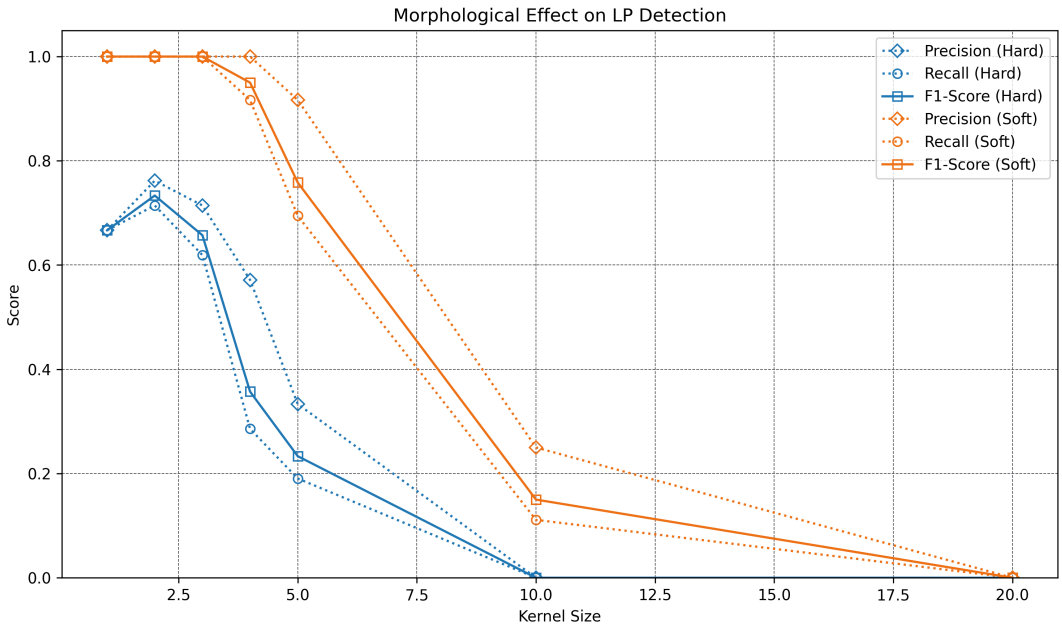


Figure 18: Detection performance with increasing morphological opening kernel size

Further tuning the colour threshold parameters δ_1 and δ_2, as shown in Figures 19 and 20, indicates

that F1-scores remain stable across a wide range of values. Significant drops in performance only occur at extreme parameter settings. This suggests that the DELPHI method is robust to moderate changes in these parameters and generalises well to new datasets, including the BAS imagery.

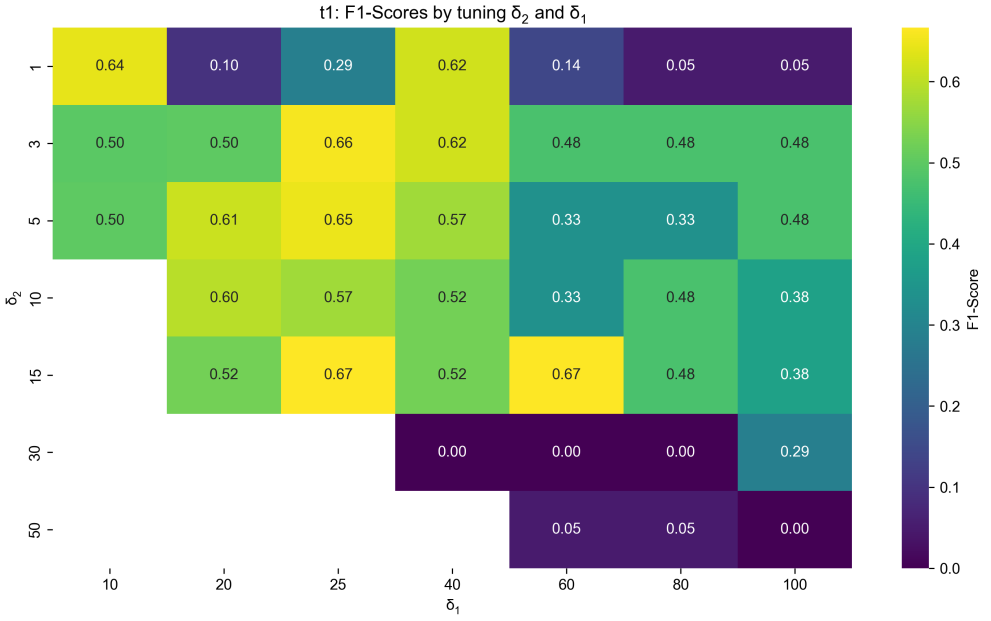


Figure 19: F1-score as a function of δ_1 and δ_2 for t1

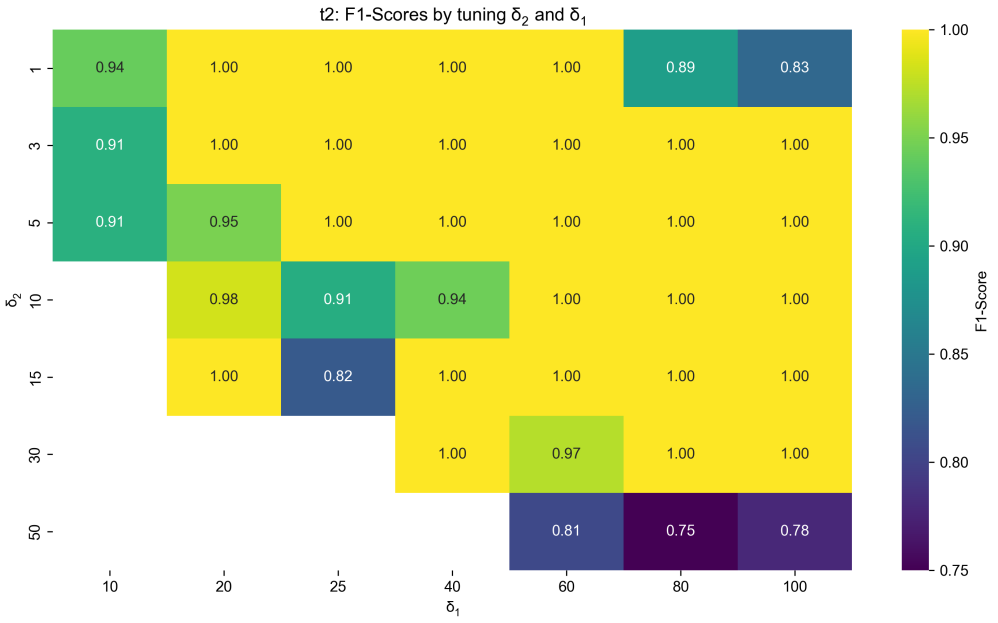


Figure 20: F1-score as a function of δ_1 and δ_2 for t2

For t1, the default parameters from the DELPHI paper ($\delta_1 = 25$, $\delta_2 = 3$) yielded the highest performance, though alternative combinations such as $\delta_1 = 60$, $\delta_2 = 15$ produced similar F1-scores. A similar pattern is observed for t2, where the default values performed comparably well to other configurations. These findings suggest that δ values do not have a strong influence on performance once reasonable thresholds are selected. As the default parameters already achieved near-optimal results, further tuning was deemed unnecessary.

4.5 Extension 3: Mixture of substrates

To evaluate generalisation across environments, the model was trained and tested using a combined dataset containing both hard (t1) and soft (t2) substrate images. Results are shown in Figure 21. The detection method performed effectively under these mixed conditions, although it required a larger number of annotated training images to reach performance saturation. This indicates that the DELPHI method can generalise across heterogeneous seabed conditions, provided sufficient training data are available.

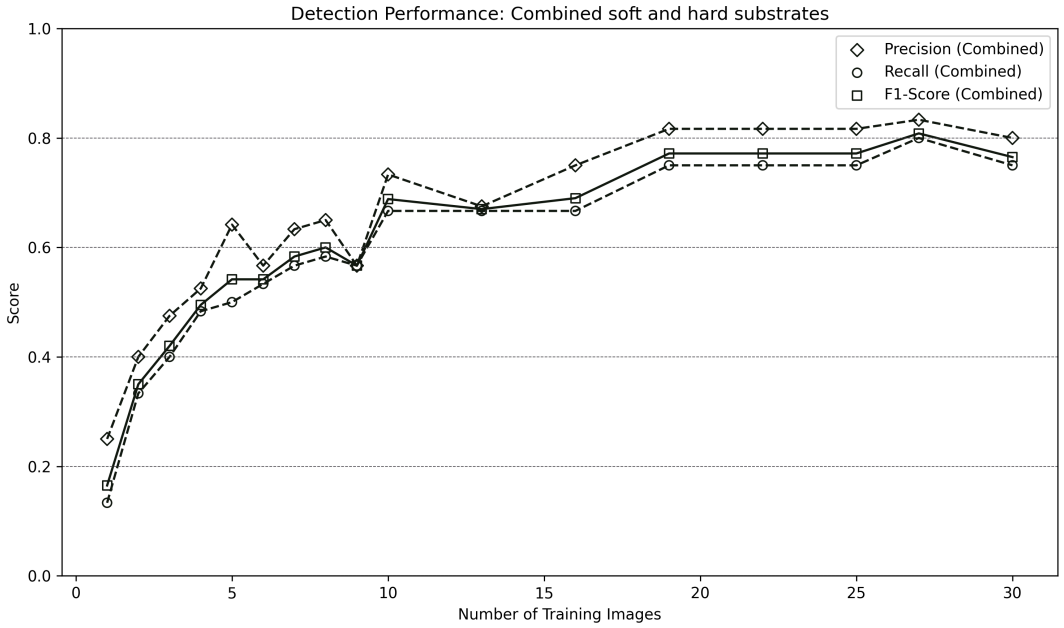


Figure 21: Detection performance on mixed substrates with increasing training data size.

These results supports prior work by Scoulding et al. [25]. This demonstrates that combining low-quality examples with those of better imaging conditions can improve model performance, as the model benefits from increased variability and exposure to diverse patterns. Further analysis of the successful and failed classifications is needed to better understand the underlying causes of these outcomes.

5 Limitations

5.1 The DELPHI paper

Data availability and transparency: One major limitation of the DELPHI paper is the lack of access to the dataset. The data was not made publicly available, making it impossible to perform a direct replication. In addition, the paper did not clearly report key details such as the number of data points used for each dataset. This omission makes it difficult to determine whether differences in performance arise from the dataset itself or from the detection method.

Incomplete method description: Several assumptions had to be made during the replication. Most notably, the paper did not specify the distance threshold used to classify a predicted LP as a true positive. This threshold is critical, as it defines how close a prediction must be to a ground truth annotation in order to be considered correct.

Another important detail omitted from the paper was the use of a spatial filter, referred to as the master mask, during the detection phase. While this mask was mentioned in the training phase, its role in testing was not described. After contacting the authors and reviewing the provided C++ implementation, it became clear that the mask was applied during detection to exclude candidate points that fell outside learned spatial regions. This omission is significant, as the mask substantially influenced final predictions by removing outliers that did not align with expected spatial layouts.

5.2 The DELPHI method

Limited generalisability: The method also assumes that exactly three LPs are always present in each image. This constraint is embedded in the detection logic, which compares triangle configurations in test images to those seen during training. While this is acceptable for certain marine imaging systems, it limits the method's general applicability to other settings where the number of LPs may vary.

Sensitivity to image resolution: Finally, the method is sensitive to image size. Early in this project, the dataset contained images cropped to different resolutions depending on the researcher's goal. Under such conditions, the spatial layout model became distorted. This is because binary masks are sensitive to image dimensions and positioning, a property known as translation variance. Without consistent image sizes, the spatial priors do not align correctly, leading to poor performance. This represents another practical limitation of the method.

6 Future Work

The main challenges identified in this study relate to the detection of LPs in dark or highly textured images, often resulting in false negatives and false positives, respectively. As demonstrated in Section 4.2.3, LPs in dark regions may fail to meet the learned colour thresholds, while bright features in noisy backgrounds are sometimes misclassified as LPs. These errors stem from the assumption that the colour distributions and spatial configurations in the training data are consistent with those in the test images. However, variations in environmental conditions, particularly evident in the t1 dataset, can cause the model to generalize poorly, leading to incorrect classifications. This highlights the need for more robust colour correction and denoising strategies, as well as more adaptive modelling of LP features.

To address these limitations, the following sections review classical and learning-based methods drawn from current research in benthic imagery analysis, highlighting their potential for improving laser point detection.

6.1 Classical Image Analysis improvements

Colour normalisation is one way to improve LP detection under variable lighting. For instance, Beijbom et al. applied histogram stretching to benthic images to enhance contrast and reduce the impact of water turbidity in coral classification [26]. This technique improves feature separation from the background. Applying a similar approach to the DELPHI pipeline could enhance the visibility of faint red LPs in turbid or dark conditions, particularly in the t1 dataset, where low contrast caused some LPs to fall below the detection threshold. Incorporating histogram stretching as a preprocessing step could reduce false negatives rates.

Another method of colour correction proposed by Mbani et al. involves enhancing red regions by subtracting the red channel from a weighted combination of the blue and green channels [27]. This emphasises red features while suppressing background tones. Although this may increase false positives in the presence of red fauna, such as some pink starfishes present in the t2 dataset, it could still improve sensitivity in dark scenes where red LPs are otherwise too faint to detect. Combining this technique with spatial filtering using the master mask could help retain true positives while reducing false positives from flora and fauna with red tones.

To address uneven illumination, Mbani et al. also applied z-score normalisation to correct for illumination drop-off, defined as when image edges are darker than the center [27]. This lighting imbalance compresses the colour range and can obscure relevant features like LPs. A similar issue was observed in both the t1 and t2 datasets. By normalizing intensity distributions across the image, their method produced broader and more uniform histograms, which could help preserve valuable LP information in already noisy imagery.

Texture is another important factor affecting LP detection. Istenič et al. proposed a method to remove background texture in LP regions using an auxiliary image, captured either from a different camera angle or with the lasers switched off [28]. After isolating the LP region (similar to the master mask used in DELPHI), the auxiliary image is aligned with the original using normalises cross-correlation

in the Fourier domain. Once aligned, the auxiliary image is subtracted from the original image to suppress texture. This is followed by low pass filtering in the HSV color space, which may be more suitable than RGB for identifying laser points, as it explicitly encodes hue and value, which correspond to the redness and brightness of pixels.

6.2 Machine learning Improvements

Machine learning offers more adaptive strategies for handling environmental variability. Bianco et al. addressed the problem of red channel attenuation in underwater imagery by proposing a single-image dehazing algorithm [29]. Their approach leverages the known property that red light attenuates more rapidly with depth than green or blue. By comparing colour channel intensities, they estimate a coarse depth map. This is then used within a maximum a posteriori (MAP) framework to remove haze and enhance the visibility of red features. This method is particularly relevant to the t1 dataset, where steep bathymetric slopes (Figure 4a) can lead to variations in the distance between the camera and the seafloor, as well as uneven lighting conditions. These factors may have resulted in inconsistent visibility of red laser points and disrupted the effectiveness of color clustering. Incorporating a depth-aware dehazing model as a preprocessing step could normalize LP colour intensity across the dataset, improving generalization and reducing false negatives. Moreover, platforms like OFOBS provide depth metadata, which could further enhance the accuracy of this model.

Convolutional neural networks (CNNs) have also demonstrated strong performance in marine imagery tasks. Mahmood et al. applied CNNs with spatial pyramid pooling (SPP) to coral classification [30]. Their method extracts multiscale patches centered on annotated pixels and uses max pooling to retain dominant spatial features. For LP detection, a similar approach could be used to model the radial intensity profile of a laser dot, which is typically bright at the center and fades outward. This characteristic is not captured by DELPHI, which focuses only on colour and spatial layout. A CNN-based model could learn this intensity pattern and differentiate true LPs from random red noise, potentially improving both precision and recall.

Support vector machines (SVMs) also offer a promising approach for laser point classification, as they are well suited for binary classification in complex cases where simple thresholding is insufficient and the data requires transformations to become linearly separable. Rzhanov et al. developed the UVSD system using SVMs trained on pixel-level brightness and colour information to detect laser points [31]. Their model supports iterative training, allowing users to refine classifications based on error correction. This adaptive approach could enhance DELPHI by enabling model updates in response to variable conditions across image sequences. Furthermore, UVSD integrates geometric constraints derived from beam line orientation, similar to the master mask strategy used in this study. The consistent use of spatial filtering across different systems suggests it is an effective technique that should be retained in future work [28] [31].

7 Conclusion

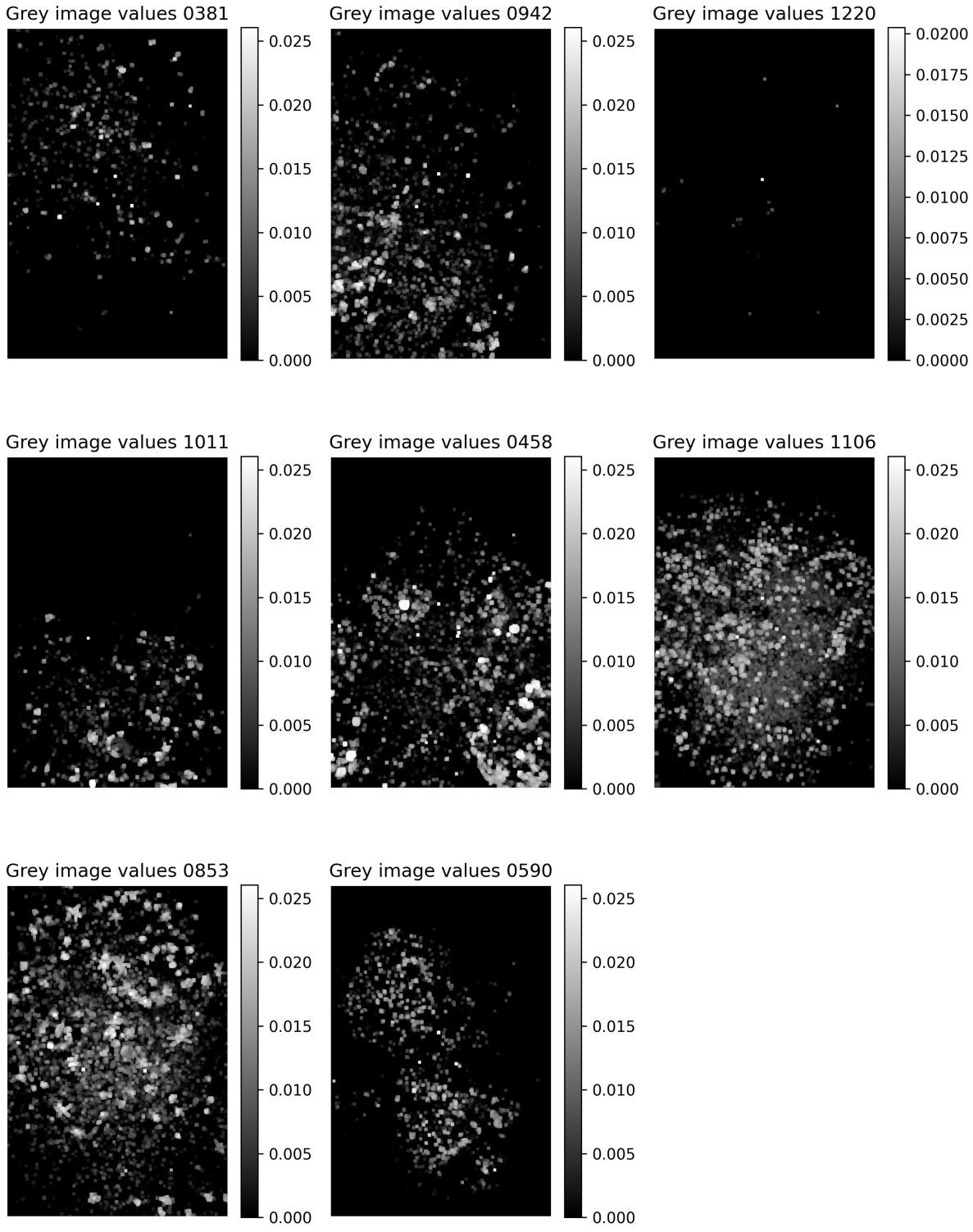
This thesis aimed to replicate the method presented by Schoening et al. by implementing DELPHI, a semi-automatic laser point detection pipeline for benthic imagery, as a Python package and evaluating its robustness on a dataset from the British Antarctic Survey. The proposed pipeline, which combines k-means color clustering, spatial filtering, and morphological operations, achieved results that were generally consistent with those reported in the DELPHI paper. However, DELPHI showed reduced performance on images with poor lighting and highly textured backgrounds, highlighting its limited flexibility in color modelling and reduced robustness to substrate noise and variation. These findings point to the need for future improvements, including classical techniques such as HSV-based color correction, histogram stretching, image and texture normalisation, and machine learning methods such as MAP, CNNs, and SVMs, all of which have been shown in the literature to enhance detection in benthic imagery. Overall, the DELPHI framework provides a solid foundation for benthic laser point detection and, with further development, could enable faster and more accurate image analysis in marine biodiversity research, ultimately supporting future conservation policy.

A Use of Auto-Generational tools

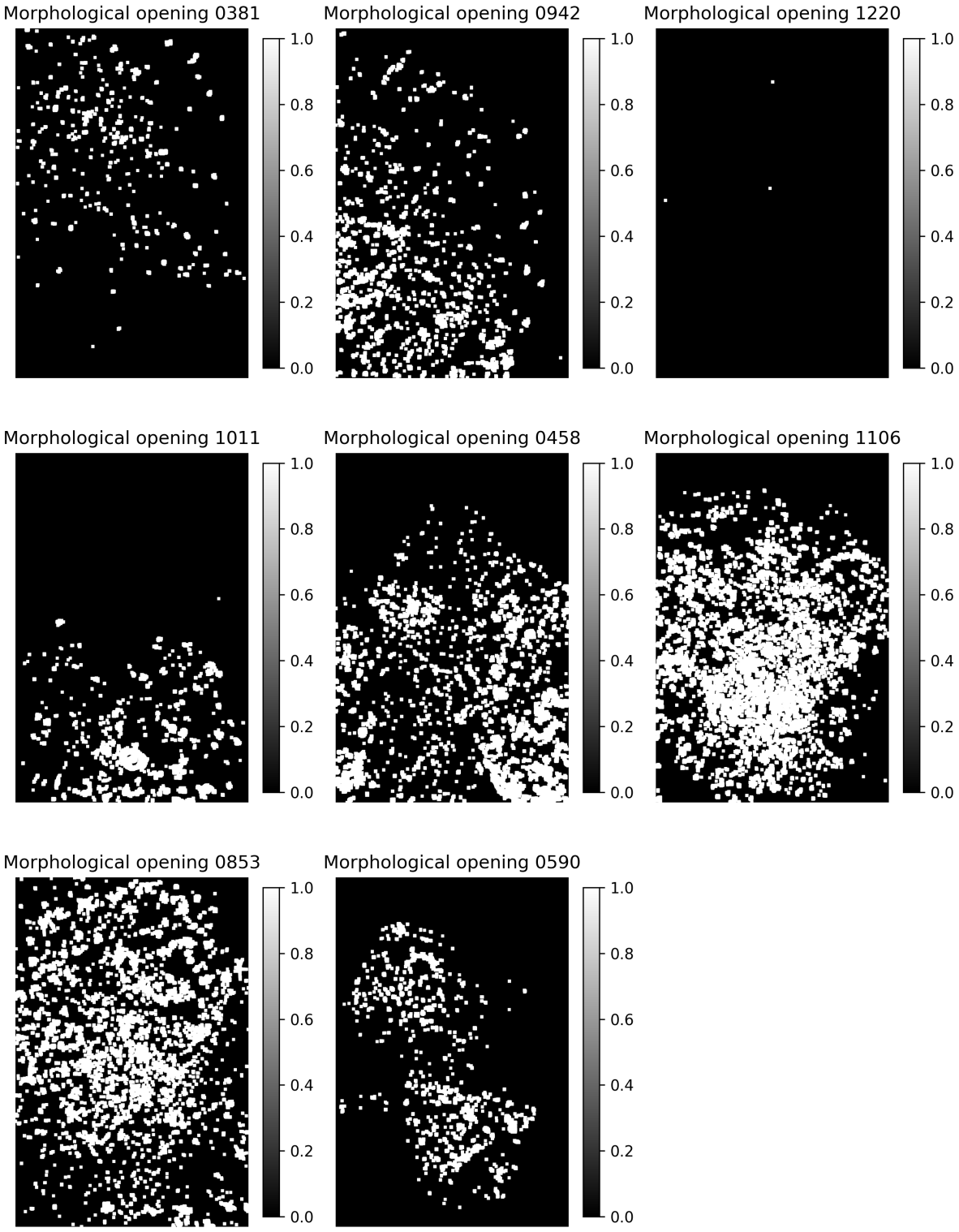
ChatGPT was used for writing and debugging code. It was also used to facilitate LaTeX formatting for figures and tables, and to perform spellchecking.

B Failed Prediction Pipelines for t1 and t2

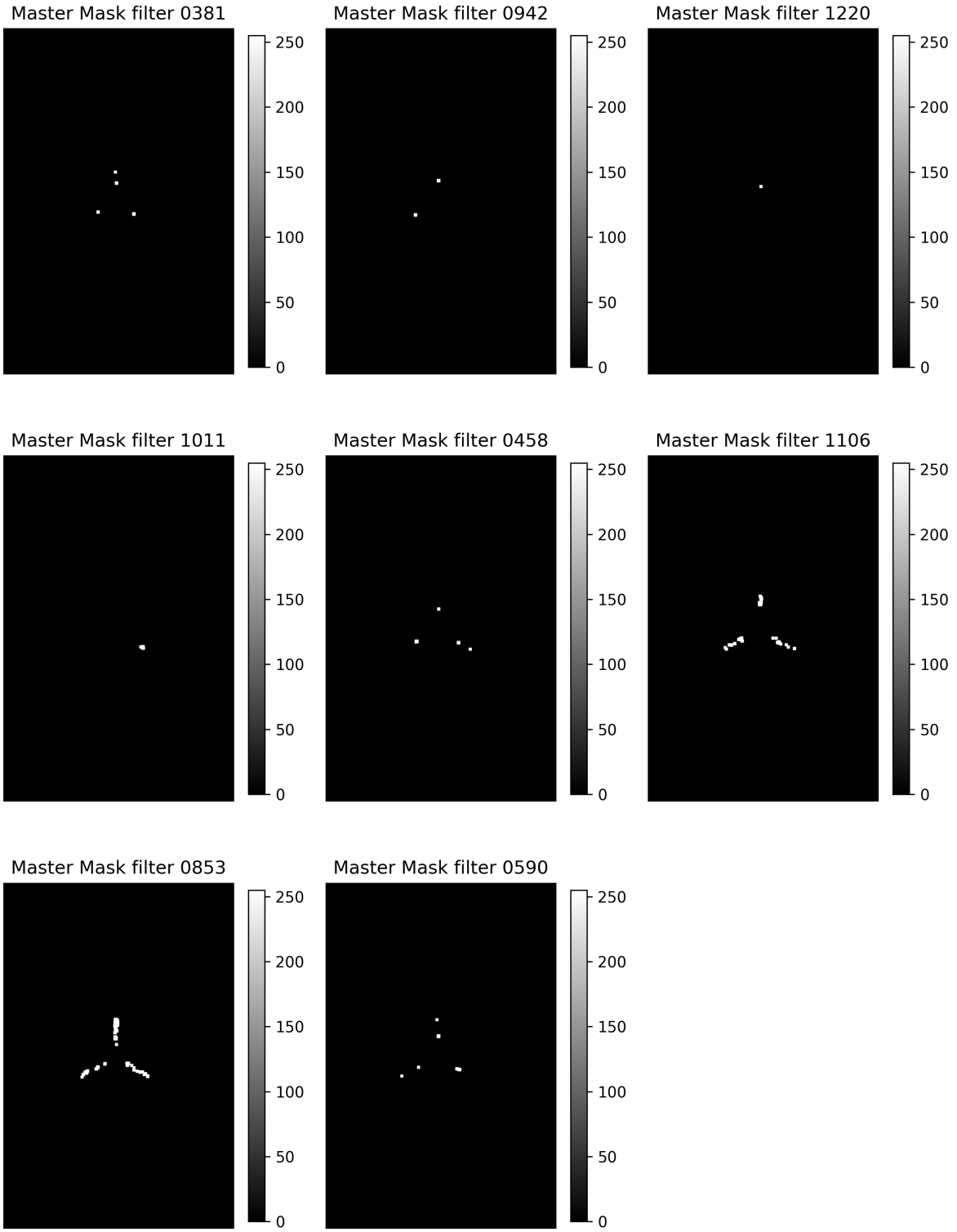
B.1 Grey value image of t1 images



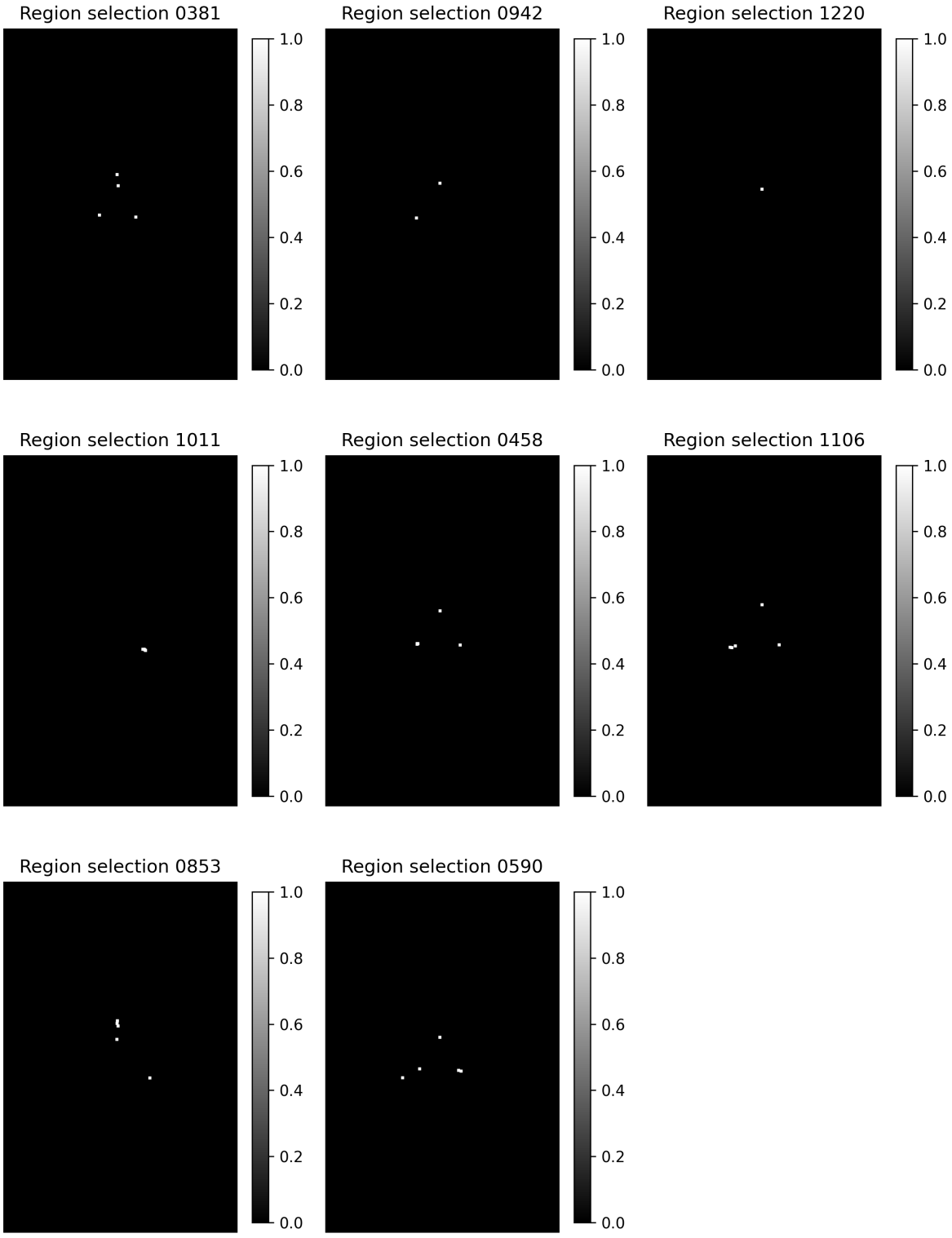
B.2 Morphological Opening of t1 images



B.3 Master Mask Filtering of t1 images



B.4 Region Selection of t1 images



B.5 Failed t1 predictions of t1 images

Image 0942 | GT: 3 | Pred: 2

Image 1220 | GT: 3 | Pred: 1

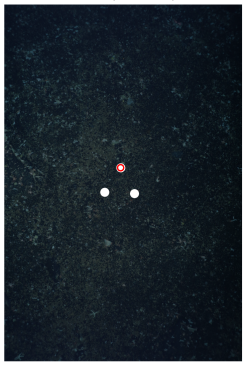


Image 1011 | GT: 3 | Pred: 3

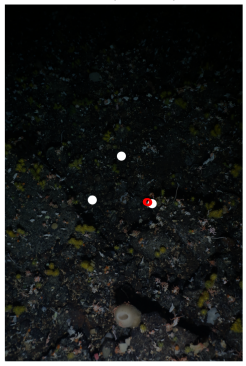
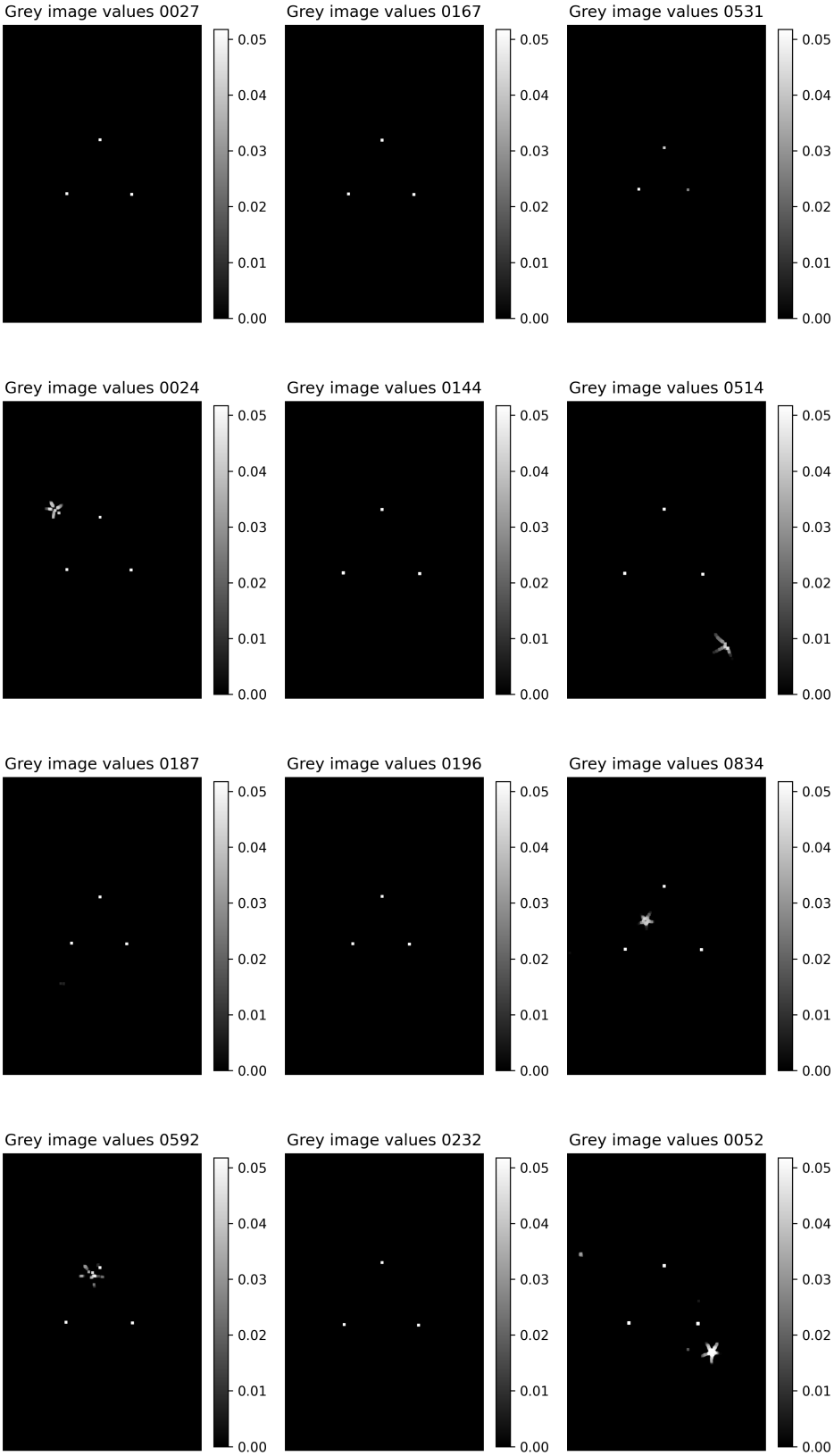


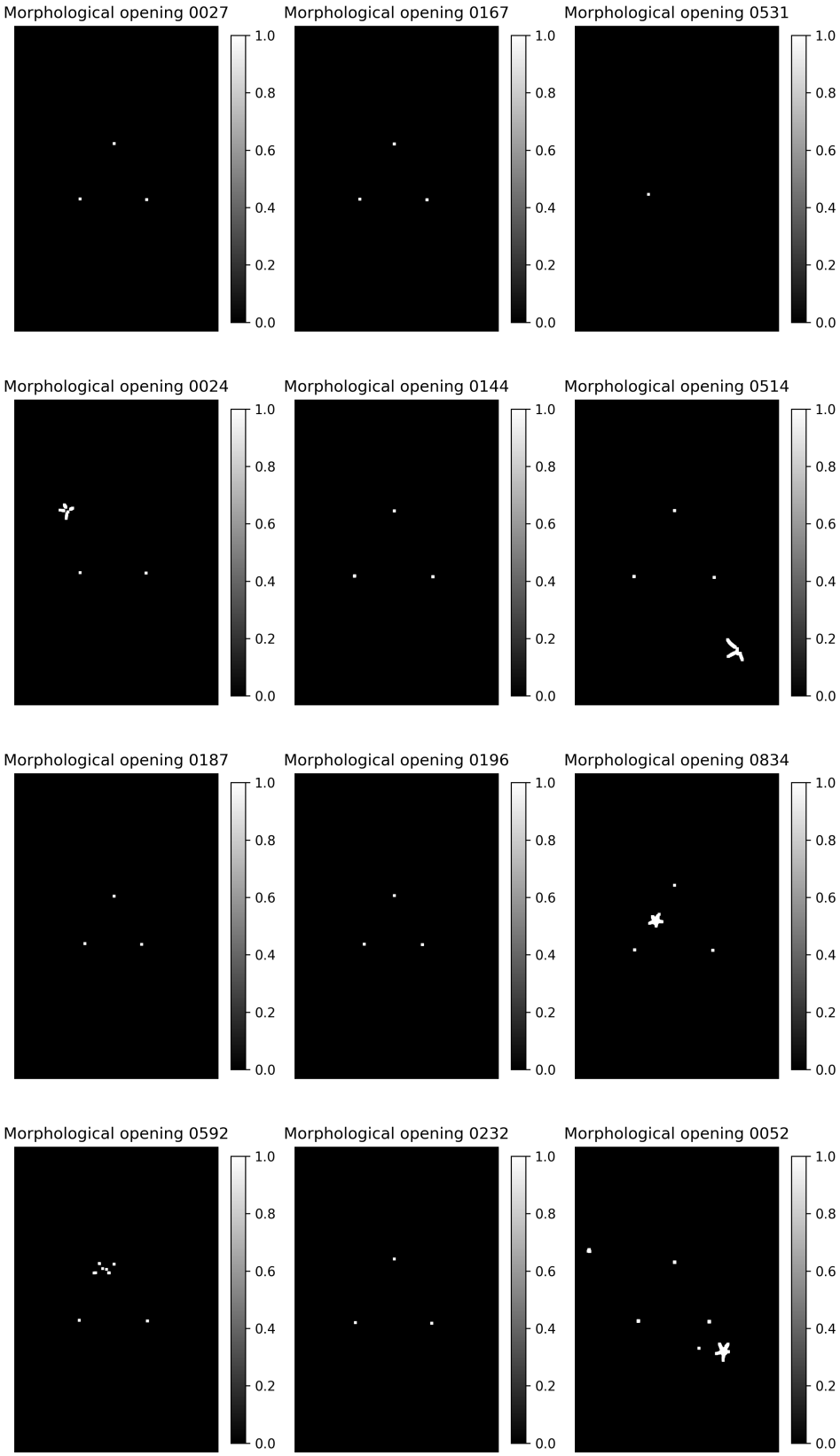
Image 1106 | GT: 3 | Pred: 3

Image 0853 | GT: 3 | Pred: 3

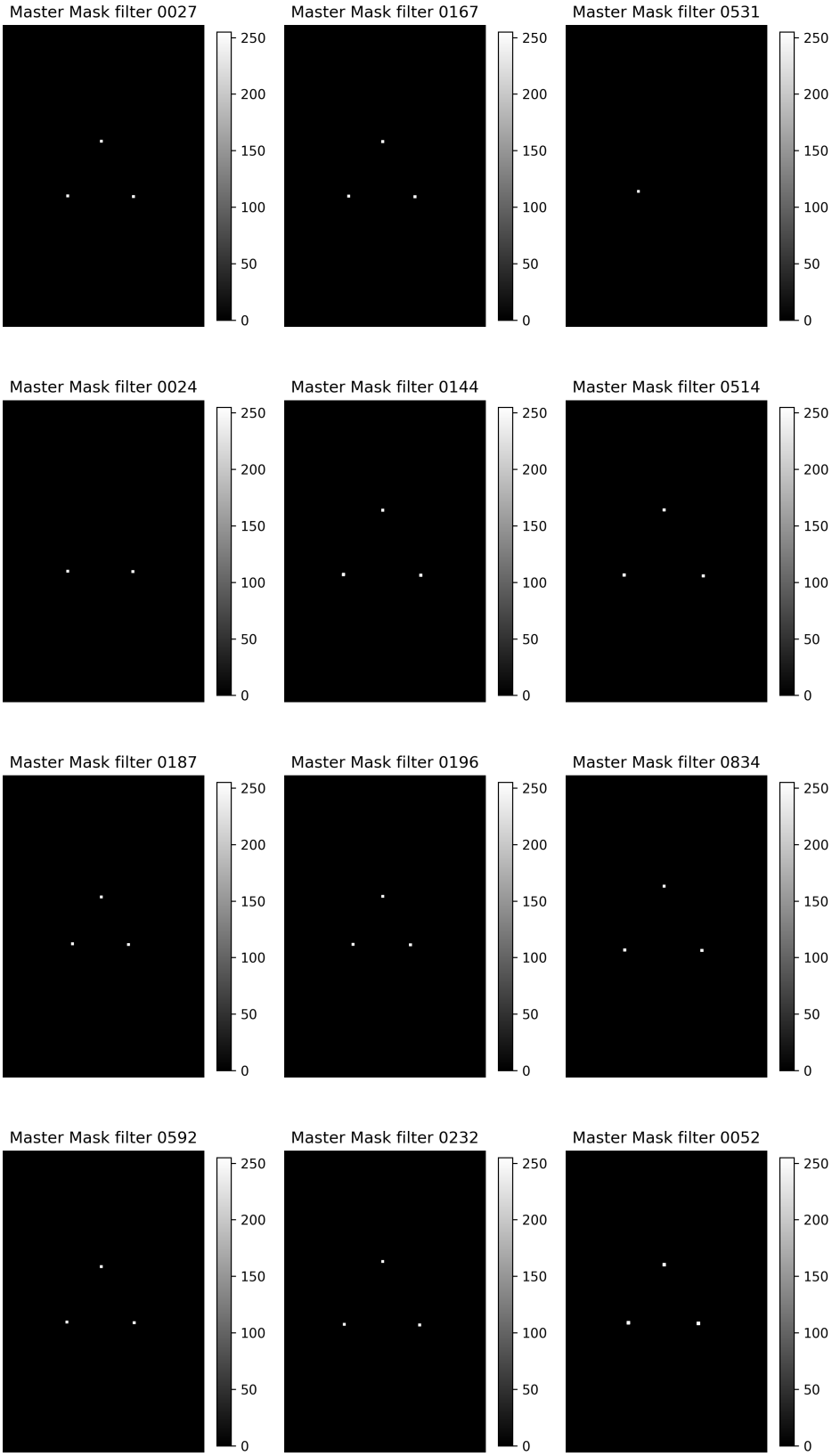
B.6 Grey value image of t2 images



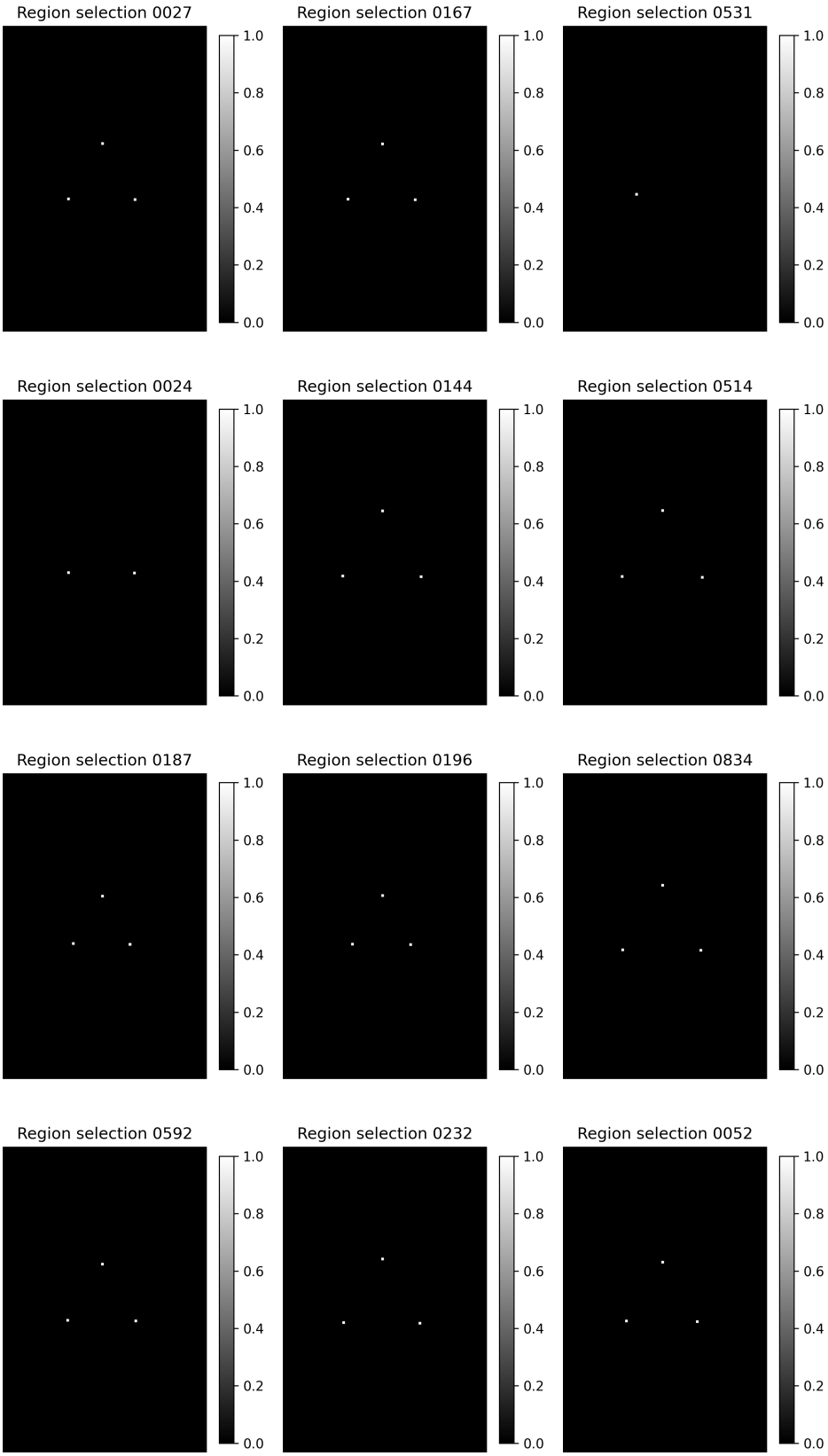
B.7 Morphological Opening of t2 images



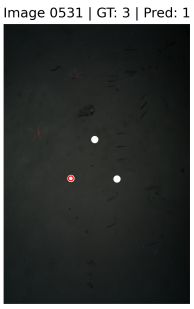
B.8 Master Mask Filtering of t2 images

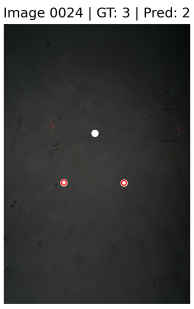


B.9 Region Selection of t2 images



B.10 Failed t2 predictions





References

[1] Ivor D. Williams, Courtney S. Couch, Oscar Beijbom, Thomas A. Oliver, Bernardo Vargas-Angel, Brett D. Schumacher, and Russell E. Brainard. Leveraging automated image analysis tools to transform our capacity to assess status and trends of coral reefs. *Frontiers in Marine Science*, Volume 6 - 2019, 2019.

[2] Alan Williams, Franziska Althaus, and Thomas Schlacher. Towed camera imagery and benthic sled catches provide different views of seamount benthic diversity. *Limnology and Oceanography: Methods*, 13, 02 2015.

[3] Daniel O. B. Jones, Maria Belen Arias, Loïc Van Audenhaege, Sabena Blackbird, Corie Boolukos, Guadalupe Bribiesca-Contreras, Jonathan T. Copley, Andrew Dale, Susan Evans, Bethany F. M. Fleming, Andrew R. Gates, Hannah Grant, Mark G. J. Hartl, Veerle A. I. Huvenne, Rachel M. Jeffreys, Pierre Josso, Lucas D. King, Erik Simon-Lledó, Tim Le Bas, Louisa Norman, Bryan O'Malley, Thomas Peacock, Tracy Shimmield, Eva C. D. Stewart, Andrew K. Sweetman, Catherine Wardell, Dmitry Aleynik, and Adrian G. Glover. Long-term impact and biological recovery in a deep-sea mining track. *Nature*, 642(8066):112–118, June 2025.

[4] M. Meredith, M. Sommerkorn, S. Cassotta, C. Derksen, A. Ekaykin, A. Hollowed, G. Kofinas, A. Mackintosh, J. Melbourne-Thomas, M. Muelbert, G. Ottersen, H. Pritchard, and E. A. G. Schuur. Polar regions. In H.-O. Pörtner, D. C. Roberts, V. Masson-Delmotte, P. Zhai, M. Tignor, E. Poloczanska, K. Mintenbeck, A. Alegría, M. Nicolai, A. Okem, J. Petzold, B. Rama, and N. M. Weyer, editors, *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*. Intergovernmental Panel on Climate Change, 2019. Accessed: 2025-07-06.

[5] V. Raoult, K. McSpadden, T.F. Gaston, J.Y.Q. Li, and J.E. Williamson. Rapid surveying of benthopelagic ecosystems with a towed mini-rov. *Marine Environmental Research*, 208:107122, 2025.

[6] Daphne Cuvelier, Martin Zurowietz, and Tim W. Nattkemper. Deep learning–assisted biodiversity assessment in deep-sea benthic megafauna communities: a case study in the context of polymetallic nodule mining. *Frontiers in Marine Science*, Volume 11 - 2024, 2024.

[7] Chuanqi Xie and Ce Yang. A review on plant high-throughput phenotyping traits using uav-based sensors. *Computers and Electronics in Agriculture*, 178:105731, 2020.

[8] S.C. Lowe, B. Misiuk, I. Xu, et al. Benthicnet: A global compilation of seafloor images for deep learning applications. *Scientific Data*, 12:230, 2025.

[9] PANGAEA. Pangaea – data publisher for earth & environmental science. `https://www.pangaea.de`, 2024. Accessed: 2025-06-29.

[10] Ketil Malde, Nils Olav Handegard, Line Eikvil, and Arnt-Børre Salberg. Machine intelligence and the data-driven future of marine science. *ICES Journal of Marine Science*, 77(4):1274–1285, 04 2019.

[11] Norman MacLeod, Mark Benfield, and Phil Culverhouse. Time to automate identification. *Nature*, 467:154–155, September 2010.

[12] Oscar Beijbom, Tali Treibitz, David I. Kline, Gal Eyal, Adi Khen, Benjamin P. Neal, Yossi Loya, B. Greg Mitchell, and David Kriegman. Improving automated annotation of benthic survey images using wide-band fluorescence. *Scientific Reports*, 6:23166, 2016.

[13] Hassan Mohamed, Kazuo Nadaoka, and Takashi Nakamura. Assessment of machine learning algorithms for automatic benthic cover monitoring and mapping using towed underwater video camera and high-resolution satellite images. *Remote Sensing*, 10(5), 2018.

[14] Min Han, Zhiyu Lyu, Tie Qiu, and Meiling Xu. A review on intelligence dehazing and color restoration for underwater images. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 50(5):1820–1832, 2020.

[15] Yong Lin Lai, Tan Fong Ang, Uzair Aslam Bhatti, Chin Soon Ku, Qi Han, and Lip Yee Por. Color correction methods for underwater image enhancement: A systematic literature review. *PLOS ONE*, 20:1–24, 03 2025.

[16] Timm Schoening, Thomas Kuhn, Melanie Bergmann, and Tim W. Nattkemper. Delphi—fast and adaptive computational laser point detection and visual footprint quantification for arbitrary underwater image collections. *Frontiers in Marine Science*, Volume 2 - 2015, 2015.

[17] Autun Purser, Laura Hehemann, Simon Dreutter, Boris Dorschel, and Axel Nordhausen. Seabed photographs taken along OFOBS profile PS118_69-1 during RV POLARSTERN cruise PS118. PANGAEA, 2020. In: Purser, A et al. (2020): OFOBS Seafloor images from the Antarctic Peninsula and Powell Basin, collected during RV POLARSTERN cruise PS118 [dataset publication series]. Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, PANGAEA, https://doi.org/10.1594/PANGAEA.911904.

[18] Autun Purser, Laura Hehemann, Simon Dreutter, Boris Dorschel, and Axel Nordhausen. Seabed photographs taken along OFOBS profile PS118_6-9 during RV POLARSTERN cruise PS118. PANGAEA, 2020. In: Purser, A et al. (2020): OFOBS Seafloor images from the Antarctic Peninsula and Powell Basin, collected during RV POLARSTERN cruise PS118 [dataset publication series]. Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, PANGAEA, https://doi.org/10.1594/PANGAEA.911904.

[19] Tasnuva Ming Khan, Huw J. Griffiths, Rowan J. Whittle, Nile P. Stephenson, Katie M. Delahooke, Autun Purser, Andrea Manica, and Emily G. Mitchell. Network analyses on photographic surveys reveal that invertebrate predators do not structure epibenthos in the deep (2000m) rocky powell basin, weddell sea, antarctica. *Frontiers in Marine Science*, 11:1408828, 2024. Open-access; published 02 Jul 2024.

[20] Jack B Pan, Michael M Gierach, Sharon Stammerjohn, Oscar Schofield, Michael P Meredith, Rik A Reynolds, Maria Vernet, F An Haumann, Annelise J Orona, and Charles E Miller. Impact of glacial meltwater on phytoplankton biomass along the western antarctic peninsula. *Communications Earth Environment*, 6(1):456, 2025.

[21] Autun Purser, Yann Marcon, Simon Dreutter, Ulrich Hoge, Burkhard Sablotny, Laura Hehemann, Johannes Lemburg, Boris Dorschel, Harald Biebow, and Antje Boetius. Ocean floor observation and bathymetry system (ofobs): A new towed camera/sonar system for deep-sea habitat surveys. *IEEE Journal of Oceanic Engineering*, 2018. Also available from AWI EPIC repository (ID 46560).

[22] Kalle Saastamoinen and Sari Penttinen. Visual seabed classification using k-means clustering, cielab colors and gabor-filters. *Procedia Computer Science*, 192:2471–2478, 2021. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 25th International Conference KES2021.

[23] Aline Bin, Jessica Smith, and Thomas Wright. Visualizing the deep–sea environment during underwater mining surveys: A systematic review. *Frontiers in Marine Science*, 9:882155, 2022.

[24] The arctic nearshore turbidity algorithm (anta) - a multi sensor turbidity algorithm for arctic nearshore environments. *Science of Remote Sensing*, 4:100036, 2021.

[25] Ben Scoulding, Kylie Maguire, and Eric C Orenstein. Evaluating automated benthic fish detection under variable conditions. *ICES Journal of Marine Science*, 79(8):2204–2216, 09 2022.

[26] Oscar Beijbom, Peter J. Edmunds, David I. Kline, B. Greg Mitchell, and David Kriegman. Automated annotation of coral reef survey images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1170–1177, 2012.

[27] Benson Mbani, Timm Schoening, Iason-Zois Gazis, Reinhard Koch, and Jens Greinert. Implementation of an automated workflow for image-based seafloor classification with examples from manganese-nodule covered seabed areas in the central pacific ocean. *Scientific Reports*, 12(1):15338, 2022.

[28] Klemen Istenič, Nuno Gracias, Aurélien Arnaubec, Javier Escartín, and Rafael Garcia. Scale accuracy evaluation of image-based 3d reconstruction strategies using laser photogrammetry. *Remote Sensing*, 11(18), 2019.

[29] Nicholas Carlevaris-Bianco, Anush Mohan, and Ryan M. Eustice. Initial results in underwater single image dehazing. In *OCEANS 2010 MTS/IEEE SEATTLE*, pages 1–8, 2010.

[30] Ammar Mahmood, Mohammed Bennamoun, Senjian An, Ferdous A. Sohel, Farid Boussaid, Renae Hovey, Gary A. Kendrick, and Robert B. Fisher. Deep image representations for coral image classification. *IEEE Journal of Oceanic Engineering*, 44(1):121–131, 2019.

[31] Yuri Rzhanov, A. Mamaenko, and Mary Yoklavich. Uvsd: software for detection of color underwater features. pages 2189 – 2192 Vol. 3, 02 2005.