Image Analysis: Automatic Detection of Fish Schools in Acoustic Data by Two Methods

Presented by
Jiao LIU
Master's in Statistics
University of Strasbourg

Supervised by
Claire Saraux (CR, IPHC)
Lilia Guillet (PhD candidate, IPHC)
Céline Meillier (McU, ICUBE)
Benoît Naegel (Pr, ICUBE)

Dates
February 12, 2024 - August 9, 2024

Abstract

Understanding the distribution of fish schools and the influence of environmental factors on seabird foraging behavior is critical, particularly for seabirds such as penguins, which are constrained by physiological limitations and the need to frequently return to their colonies. These constraints impact their foraging range and success. To investigate these dynamics, we analyzed acoustic images collected from September 2023 to January 2024 off Phillip Island using two detection methods: a double threshold method and a deep learning-based approach. Our analysis of 5,606 acoustic images revealed that the two methods yielded significantly different results in terms of fish school intensity, size, and depth, etc. And the detection methods are dependent on both the sea depth levels and the timing of fish school detection by the echosounder. When combined, these methods provided improved predictive accuracy for fish school detection. This combined approach enhances our understanding of fish distribution patterns, which is vital for optimizing seabird foraging strategies and ensuring sustainable management of marine resources amidst environmental changes and fisheries exploitation.

Contents

1 Introduction

Seabirds, such as penguins, are central place foragers, meaning they travel between their colonies on land for breeding or molting and the sea, where they primarily feed on forage fish [7]. Due to their physiological limitations, such as resistance to pressure and lung capacity, penguins cannot dive too deep. Additionally, they are constrained by the need to frequently return to the colony to feed their chicks, which prevents them from foraging too far or for extended periods. Consequently, the distribution of fish schools, both horizontally and vertically, significantly influences their foraging success. To gain a deeper understanding of how environmental factors impact fish distribution and, at the same time, how factors like fish position relative to the colony, school depth, and fish density, etc., influence seabird foraging behavior, it is crucial to map fish distribution in three dimensions [3]. Understanding these dynamics is particularly vital in the face of climate change and the intense exploitation of forage fish by global fisheries, emphasizing the need for sustainable management practices that protect both marine ecosystems and predator species like seabirds.

To address these research needs, my internship supervised by Claire Saraux and Lilia Guillet at the Institut Pluridisciplinaire Hubert Curien (IPHC) in Strasbourg, France, from February 12, 2024, to August 9, 2024, aims to develop a method for automatically detecting fish schools in acoustic echograms. The IPHC is a joint research unit under the supervision of CNRS and the University of Strasbourg (UMR7178), known for its successful multidisciplinary research. Teams from diverse scientific backgrounds, including ecology, physiology and ethology, chemistry, and subatomic physics, collaborate on high-level programs using advanced scientific instrumentation. Additionally, my co-supervision by Céline Meillier and Benoît Naegel from the ICUBE laboratory, specializing in computer science, imaging, and remote sensing, ensures a comprehensive approach to this research. The collaboration between IPHC and ICUBE leverages the strengths of both institutes, providing an ideal environment to develop this method.

The acoustic data used in this project are collected from a continuous survey conducted over several months off Phillip Island, Australia, using a sailing drone equipped with an echosounder. The sailing drone follows a specific trajectory around the colony of little penguins (*Eudyptula minor*) to collect acoustic data from the water column below the drone. From these acoustic data, information such as position (longitude, latitude), quantity of fish, and fish depth can be extracted.

This internship consists of two main stages: data preprocessing and school extraction using both thresholding and deep learning methods. The results obtained from the two methods will be compared. Furthermore, we aim to extract additional information from the acoustic image, such as as shape, density, or size to later classify them and help identifying fish species [17].

2 Related Research

In the domain of acoustic image object detection, researchers have explored various methodologies aiming for accurate and efficient detection. These methodologies fall into three main categories: segmentation-based techniques, machine learning-based techniques, and deep learning-based techniques. Each approach offers distinct advantages and has been applied in various research studies to address specific challenges in fish school object detection.

Segmentation-based techniques involve partitioning an image into meaningful segments or regions based on certain criteria. Two common approaches in this category are region-based segmentation techniques and edge-based segmentation techniques. Region-based segmentation techniques divide an image into regions sharing similar characteristics, such as color or intensity. These techniques often employ image filtering and dilation processes to enhance object delineation, thereby enabling precise object detection and tracking in acoustic images. Representative methods in this category include thresholding, region growing, region splitting, and region merging, which determine the edges within objects. On the other hand, edge-based segmentation techniques focus on detecting edges or boundaries within an image, relying on edge detection algorithms to highlight object boundaries. This approach is particularly useful for detecting objects with distinct boundaries, such as marine organisms or underwater structures [13].

Machine learning-based techniques involve training algorithms to learn patterns from labeled training data and make predictions or decisions based on new, unseen data. These techniques often require manual feature engineering, where relevant features are extracted from the data and provided as input to the model. Common machine learning algorithms used for object detection include support vector machines (SVMs),

decision trees, and random forests. For example, Robotham et al. employed support vector machines (SVMs) in their work [15], while Fallon et al. utilized random forests in their study on the classification of Southern Ocean krill and icefish echoes [6].

Deep learning techniques involve training neural networks with multiple layers to learn hierarchical representations of echo data. These models can automatically learn relevant features from raw data, eliminating the need for manual feature engineering. Examples include convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which have been widely used for image classification, object detection, and segmentation tasks. For instance, in the work by Brautaset et al. [4], CNNs were used for sandeel detection in multifrequency echosounder data. Similarly, Marques et al. [11] detected pelagic species from echograms using a deep learning framework based on instance segmentation.

In summary, segmentation-based techniques are simple to implement and computationally efficient, while machine learning and deep learning techniques offer higher accuracy and flexibility for complex tasks with sufficient resources. In this internship, we aim to apply both segmentation-based and deep learning-based methods to our acoustic data to evaluate their respective performance. Notably, due to the absence of annotated data, we will directly apply the Segment Anything model for the deep learning component.

3 Data Preprocessing

3.1 Introduction

3.1.1 Echosounder

Echosounders operate by emitting bursts of electrical energy at specific frequencies, typically ranging from 38 kHz to 420 kHz. These bursts are transmitted through transducers, which convert electrical energy into acoustic energy propagating through the water. The transducer projects sound in a directional beam towards the bottom, with the beam's width inversely proportional to the frequency of the sound. Consequently, higher frequency echosounders have narrower beam widths, enabling finer resolution but limited range, while lower frequency echosounders offer longer range but lower resolution. As the transmitted pulse of sound propagates through the water, it encounters various targets such as fish or the seabed. Every surface that has a different density than water will reflect the sound: fish (with their swim bladders), air bubbles from waves, and the bottom. These targets reflect or scatter the pulse, with some energy returning towards the transducer as backscattered sound or echo. The time when the echo is received determines the distance of the target from the transducer. Then, the depth of the fish is calculated by multiplying the velocity of sound wave travel in the sea by the time and dividing by 2, that is :

$$\text{Distance} = \text{Velocity} \times \text{Time}/2$$

The received signal is then amplified and displayed on an echogram, traditionally shown with dark traces on a white background, though modern displays often use color for better visual contrast [18]. See the echosounder system in Figure 1.

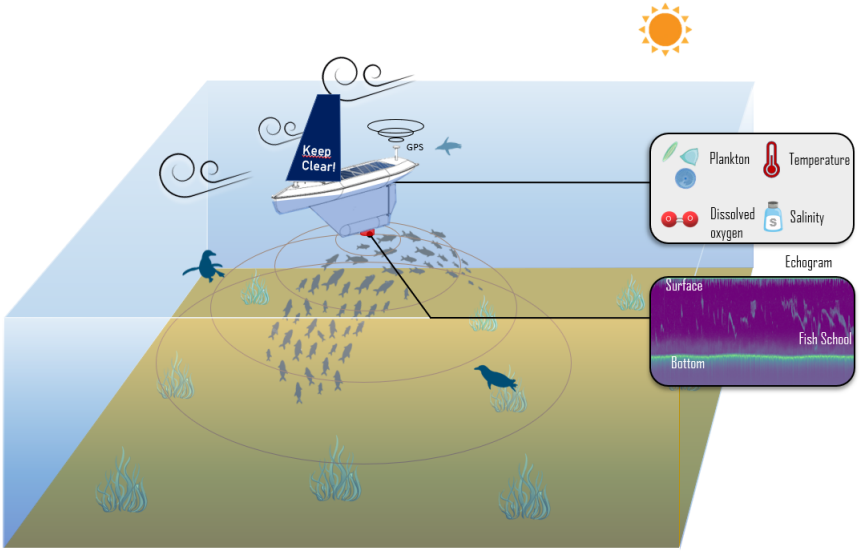

Figure 1: Illustration of the echosounder system

The echosounder used here is the WBAT Echosounder, with the following parameters:

- CW for continuous wave
- 200kHz: FM (frequency modulated) range from 185 to 255kHz
- Pulse duration: 2048µs
- Power: 150W
- Provides acoustic data via EK80 echograms

We will use images of acoustic echograms as input, provided by a signal processing library. The data has already been processed by our collaborators to preprocess and clean the raw acoustic data collected by drones. They detect the seabed from the signal and remove the repeated parts where the signal bounces between the bottom of the drone and the seabed. This preprocessing includes useful functionalities such as calibration, wave detection, and seafloor smoothing[17]. The following figure (See Figure 2) illustrates the differences before and after this processing.

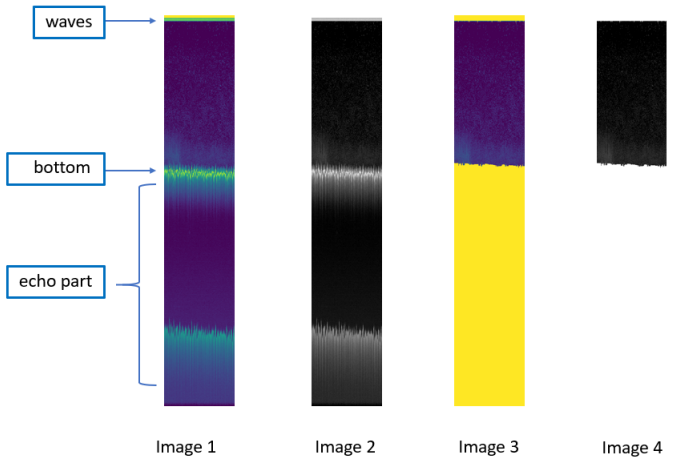

Figure 2: Images 1 (in color) and 2 (in grayscale) show the data before processing, while images 3 (in color) and 4 (in grayscale) show the data after processing.

After basic interpolation, ensuring comprehensive data coverage within each 2-second interval (matching the echosounder's ping interval), we obtained 5606 acoustic image outputs as shown in Figure 2. Each image comprises a .npy file containing original echo data in matrix format, alongside a .csv file containing key metadata such as sampling time, fish depth and quantity, drone location (longitude, latitude), velocity, and other relevant parameters.

The .npy files store matrix data similar to grayscale images. By plotting a .npy file, we visualize the acoustic image where the background appears black, and detected objects like fish or plankton are represented by gray or white pixels. In the visualization, darker shades correspond to smaller matrix values, while brighter shades, closer to white, correspond to larger matrix values.

See in Figure 3, the vertical axis represents depth, and the horizontal axis represents time. Each pixel horizontally corresponds to a two-second time interval and vertically a 10-centimeter distance. All images corresponded to 6-minute periods of echosounder activation, i.e. 180 pixels.

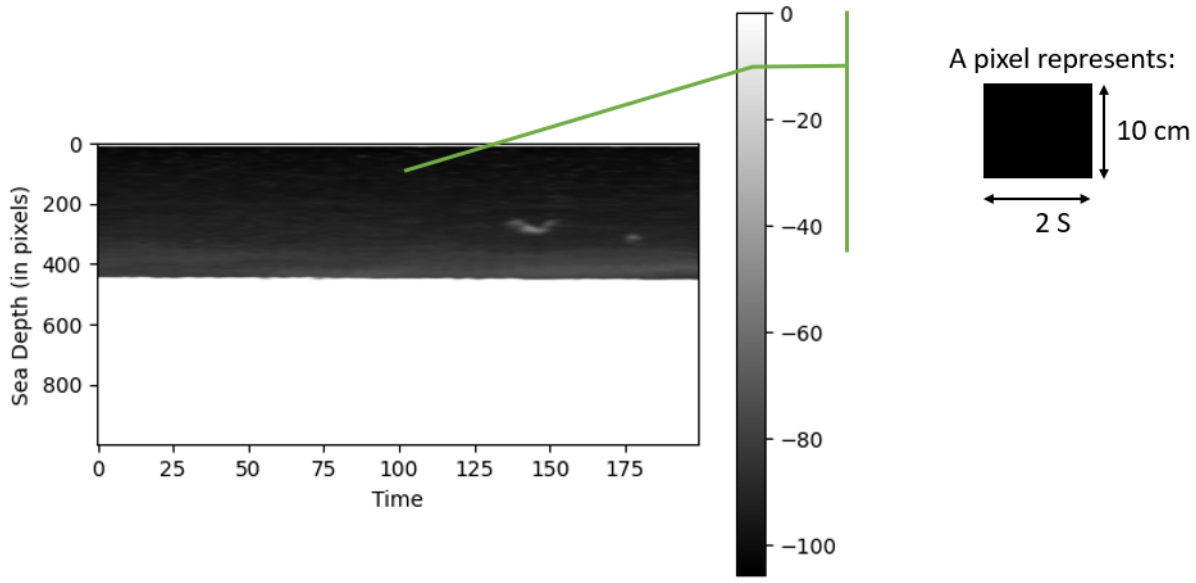

Figure 3: Example of one acoustic image displaying 2 fish schools: one around 28m depth and the other around 31m depth. As the drone moves the seabed depth may vary resulting in a non-flat seabed depth in the image. The values in the color bar represent the echo strength and that it's in decibels.

3.1.2 Sampling method: a sailing drone collecting acoustic data

While acoustic surveys are common in fisheries science, what distinguishes this project is the use of an echosounder mounted on an automated sailing drone. This setup allows for continuous data collection and provides access to surface and coastal waters, which is challenging for larger oceanographic vessels. However, this approach introduces specific challenges for data collection. Notably, the drone lacks motorization and relies solely on wind and currents for navigation. As it follows predefined transects within a set radius, its variable speed influences the detection and characterization of fish schools.

Additionally, the sensors are powered exclusively by solar panels, leading to sequential activation. For instance, the echosounder operates for 6 minutes every 10 minutes, with the remaining time allocated to positioning and other environmental sensors (such as temperature). These operational conditions significantly impact data collection efficiency and require careful consideration during data interpretation and analysis.

Data were collected from September 2023 to January 2024 along four transects (Figure 4).

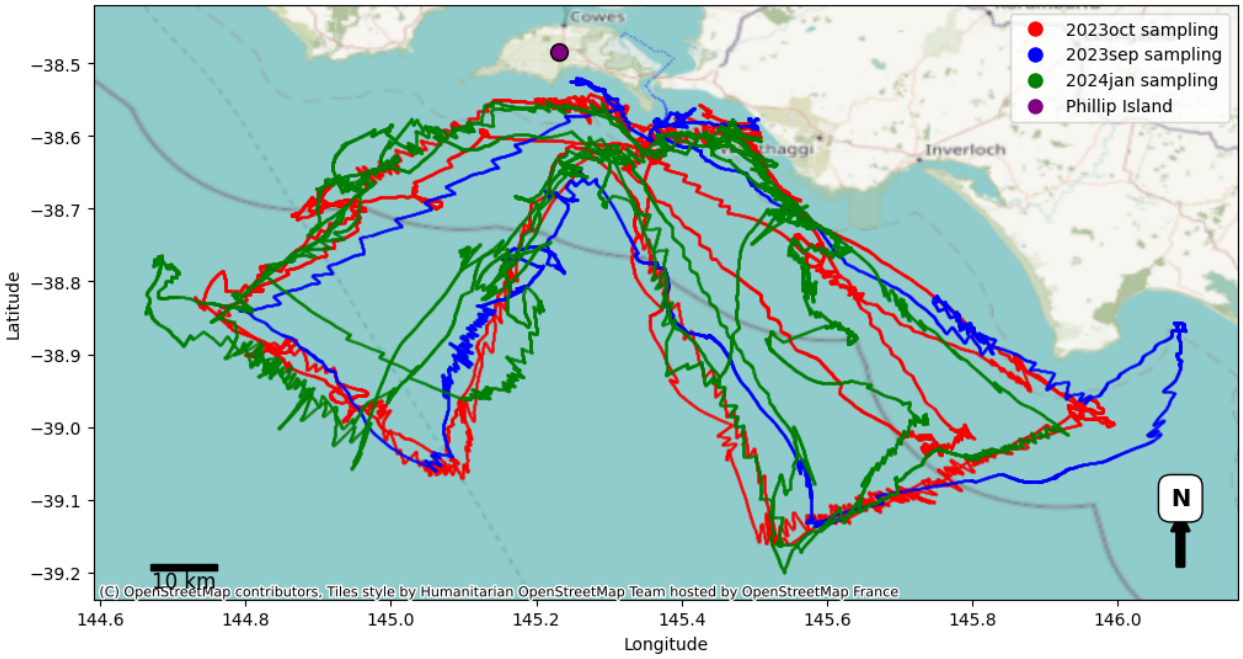

Figure 4: Map showing the trajectory path of the sampling process from September 2023 to January 2024. The blue line represents the sampling trajectory for September 2023, the red line for October 2023, and the green line for January 2024. Phillip Island is marked on the map with a large purple point.

3.2 Resizing the Image

3.2.1 Why Resize the Image?

The echosounder, which collects the data, is equipped at the bottom of the drone, and the drone moves solely on wind power without motorization. Therefore, the drone's speed is not constant [17]. Consequently, for the same time interval, when the drone travels quickly, the image compresses, covering more distance and capturing more information within the same image width. Conversely, when the drone travels slowly, the image dilates, covering less distance and resulting in less information within the same image width. This dilation and compression of object information can affect the shape of detected fish schools. As speed will influence school characteristics and possible later classification, we rescale all images to a common speed.

3.2.2 How to Resize the Image?

First, we compute the mean velocity across all our sampling data and use it as a reference to scale all other images (Figure 5). Subsequently, we compare the velocities within each image to this mean velocity to determine if stretching or shrinking is necessary. Notably, a single image may correspond to multiple velocities, as illustrated in our example (Figure 6). During this process, we adjust the image's width while maintaining a constant height, which represents the sea depth. Given that the speed of sound in seawater is approximately 1500 m/s and considering our mean sampling depth of 60 meters, we can regard this speed as constant without significantly affecting the vertical information within our images. Finally, we resize the image to arrive at the standardized spatio-temporal scale determined by the mean velocity.

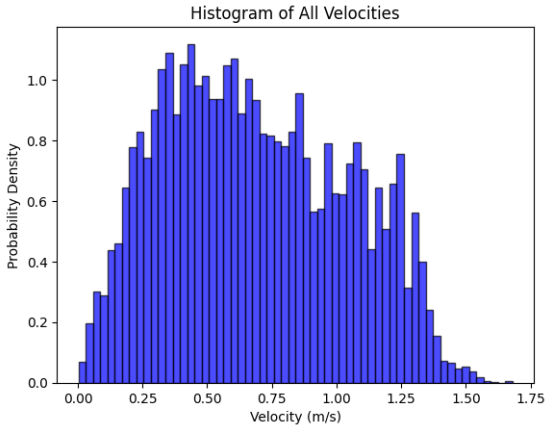

Figure 5: Histogram of all drone velocities estimated for each 2-second interval across our 5606 6-minute acoustic images. The velocities range from 0.00278 m/s to 1.68 m/s with an average of 0.685m/s.

- Step 1: Check the velocity in the image using the .csv file (as all the velocity information is saved in the .csv file). We take one acoustic image as an example. In this image, we find it contains 3 different velocity parts:

$$\begin{array}{ll} \text{Pixel interval: } 1-84 & \text{Speed } v_1 = 0.366\,\text{m/s} \\ \text{Pixel interval: } 85\text{-}175 & \text{Speed } v_2 = 0.295\,\text{m/s} \\ \text{Pixel interval: } 176\text{-}180 & \text{Speed } v_3 = 0.281\,\text{m/s} \end{array}$$

- Step 2: Resize the image. After the resizing step, the image size is compressed to 1000×86. The calculation is as follows:

$$\frac{(84\times0.366+91\times0.295+5\times0.281)}{0.685} = 86 \text{ pixels}$$

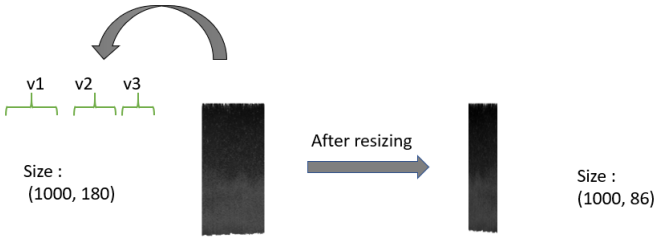

Figure 6: Resizing detail

3.3 Masking the Image

In each acoustic image, alongside the useful underwater information intended for analysis, there exists a portion that represents repeated signals. These signals occur when sound waves bounce off the seabed and return to the sailing drone, creating echoes that can repeat depending on sea depth and boat speed. The repeated parts are set to 0 in the `.npy` matrix, while the useful underwater information is represented by negative values. When comparing pixel intensities across different parts of the image, the relationship is as follows:

$$\text{Sea bottom} \approx \text{Sea surface} \approx 0 > \text{Fish schools (plankton or other noise at this level)} > \text{Rest of the sea background}$$

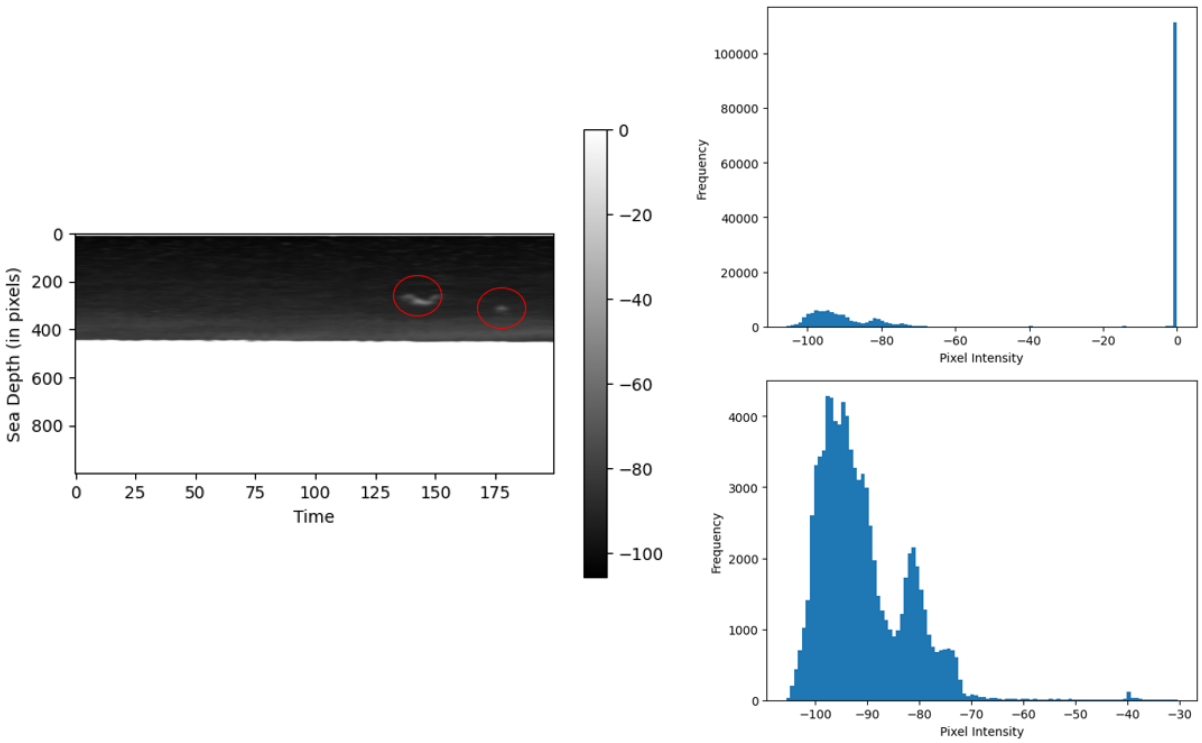

Figure 7: Left image: Acoustic image after resampling; Top right image: Pixel intensity distribution of this acoustic image; Bottom right image: Pixel intensity distribution (for pixel intensity $<$ -30) of this acoustic image.

Consider the acoustic image in the above figure as an example. Here, we can clearly distinguish two fish schools. In the histogram at the top right of Figure 7, which represents all the pixels, a large number of pixels are clustered around 0, corresponding to the sea surface, seabed, and repeated parts depicted as white in the image. These areas exhibit the strongest pixel intensities in the image. In the histogram at the bottom right of Figure 7, to better visualize the distribution of the parts of interest, we filter out dominant pixels with intensities close to 0. Now, we observe that for the darkest parts of the image (representing the sea), pixel values range around -95. There are two peaks in pixel values located between -90 and -50, corresponding to the detected fish schools.

In our approach, we utilize the `numpy.ma` module to mask our `.npy` matrix. We mask the image by selecting pixels with values greater than -30 and then extend the mask using the `binary_dilation` function from the `scikit-image` package to reduce noise, using a square structuring element to determine the neighboring regions. Due to surface waves being much more irregular than the seafloor, we use a square structuring element of size 85x85 for the first 200 lines in the acoustic images and a square structuring element of size 20x20 for the remaining lines. See Figure 8, which clearly shows that the white parts representing the surface and repeated signals have been successfully masked out, while the remaining parts exhibit a more pronounced contrast.

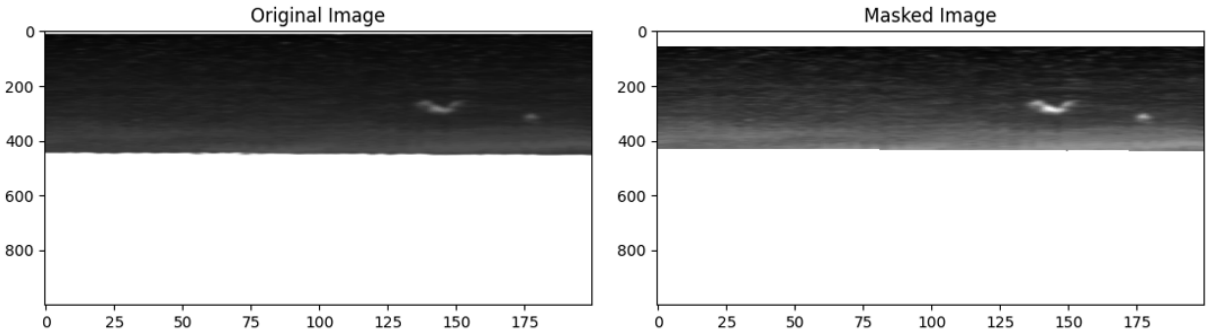

Figure 8: Left: Image before masking; Right: Image after masking

3.4 Image Correction

3.4.1 Why Do We Need Image Correction?

Even with a mask applied to the top and bottom parts, the low contrast between fish schools and the background poses challenges for segmentation. If these influences are not removed, they can adversely affect the accuracy of fish school detection. Enhancing the contrast between fish schools and the background is essential for effective segmentation. Therefore, applying corrections to the images becomes necessary.

By calculating the median pixel intensity at various depth levels for one example of our acoustic images, we obtained the following results (Figure 9):

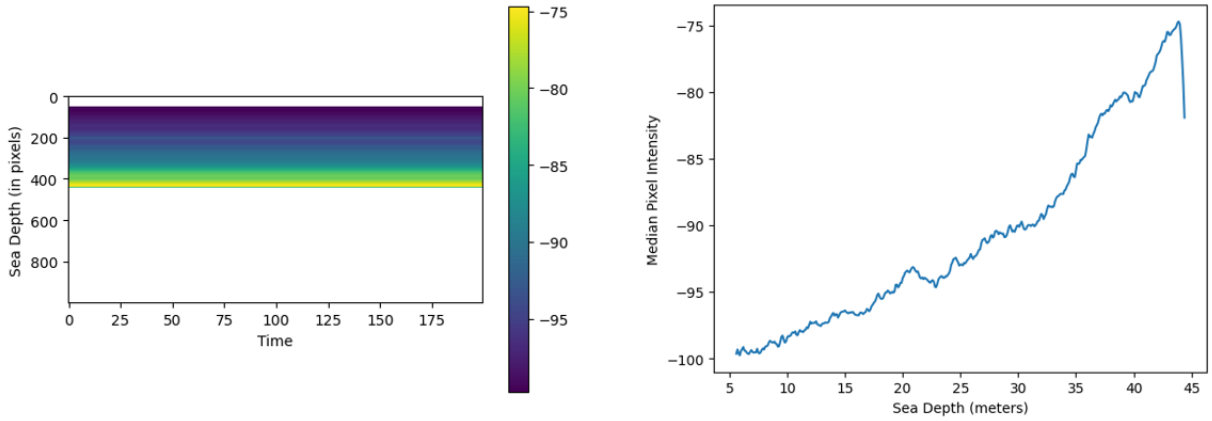

Figure 9: Overall change in median pixel intensity with respect to depth in one acoustic image

In the right figure of Figure 9, pixel intensity increases significantly with depth, possibly due to higher turbidity or plankton density. This increase in intensity complicates the effectiveness of thresholding methods for image segmentation.

3.4.2 Image Correction Method

We propose applying the following correction formula:

$$\text{corrected\_img} = 1 - \frac{\text{masked\_img}}{\text{im\_fond}}$$

Here:

- corrected_img: The corrected image.

- masked_img: The original masked image.
- im_fond: Represents the median pixel value for each row in the .npy matrix.

This correction method elegantly adjusts the image pixel intensity based on the difference between the background signal and the signal from detected fish schools.

When no fish school is present, $\frac{\text{masked\_img}}{\text{im\_fond}}$ is greater than 1 (since background intensity $>$ fish school intensity > 0). For areas containing fish schools, this term ranges between 0 and 1. By taking the complementary this term, we emphasize the fish school part correctly after correction. Below, we present an acoustic image before and after correction for comparison. In figure 10, the enhanced contrast between the fish school and the background is evident, and the background appears more homogeneous as well. Once the acoustic image is corrected, we can proceed with applying detection methods to identify fish schools.

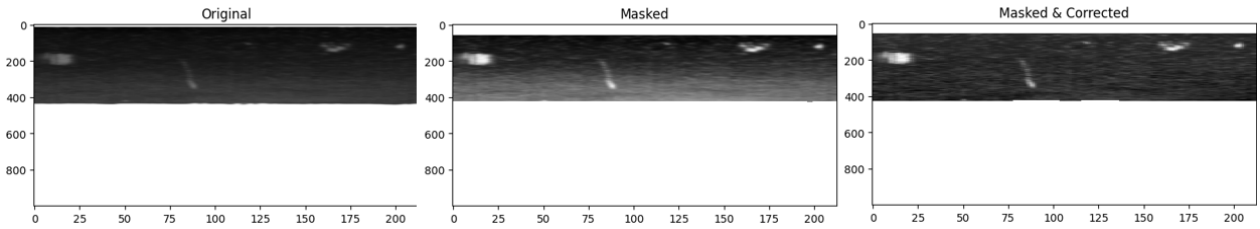

Figure 10: Effect of image correction

4 Image Segmentation Methods: Double Threshold and Segment Anything

4.1 Double Threshold

4.1.1 Introduction

Once the data are properly prepared, we will implement our first method: the double threshold segmentation technique. Thresholding is a fundamental method in image processing that transforms a grayscale image into a binary image, making it a straightforward approach for segmenting objects from a background.

In this method, our objective is twofold: first, to detect images containing fish schools, and second, for these identified images, to apply two different thresholds (hence 'double' threshold). This results in two thresholded images: a smaller image (seed image) and a larger image (mask image). Subsequently, we reconstruct the final segmented image by performing morphological reconstruction through dilation between the two images. The larger mask image defines the maximum extent achievable [1].

To delve deeper into this method, the critical step involves determining the two distinct thresholds. Different combinations of thresholds yield varying seed and mask images, thereby influencing the final segmentation outcome. Selecting the appropriate thresholds is pivotal for generating optimal seed and mask images.

Firstly, we will provide a brief introduction to the morphological dilation process in scikit-image, which forms the basis of morphological reconstruction. Secondly, we will explain the reconstruction process and its operational principles. Thirdly, we will outline the assumptions underlying the double threshold method. Lastly, we will discuss potential errors associated with this method and strategies for selecting the pair of thresholds (`thresh_max` and `thresh_min`).

4.1.2 Thresholding Method

Thresholding algorithms implemented in scikit-image can be separated into two categories:

- Histogram-based: These algorithms use the histogram of the pixels' intensity and make certain assumptions about its properties (e.g., bimodality).

- Local: These algorithms process a pixel based only on its neighboring pixels. They often require more computation time [1].

The simplest thresholding methods replace each pixel in an image with a black pixel if the image intensity $I_{i,j}$ is less than a fixed value called the threshold T, or with a white pixel if the pixel intensity is greater than that threshold [1].

Here is a simple example of thresholding (Figure 11):

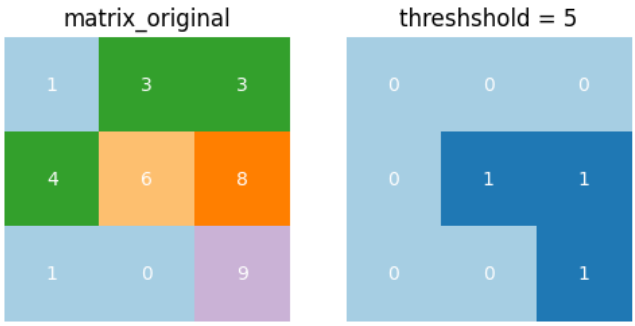

Figure 11: Left: Original matrix; Right: Matrix after thresholding with threshold = 5

The double thresholding method used in this internship is based on histogram-based thresholding, as shown in this example. Each threshold is compared with all the pixel values in the image to generate a thresholded image. Double thresholding produces one seed image and one mask image [1].

4.1.3 Image Processing: Dilation

When performing a morphological reconstruction operator in scikit-image, it can be considered a series of morphological dilations. Morphological dilation sets the value of a pixel to the maximum value of all pixels within a local neighborhood centered around it. The local neighborhood is determined by a structuring element, such as a square, rectangle, disk, or any other element you define. Dilation enlarges bright regions and shrinks dark regions, enhancing the features of the image [1].

For example, we apply a 3x3 square as a footprint to determine the neighbors in this area and compare each pixel in the original matrix with its neighbors. Each pixel is reattributed with the maximum pixel value in its local neighborhood centered around it. This process expands the area with pixel values equal to 1, as shown in the dilated matrix (Figure 12). The green positions in the matrix are the new pixels dilated to 1.

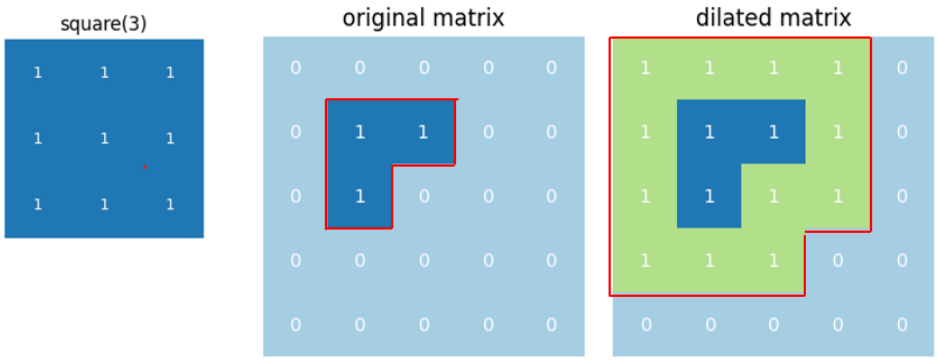

Figure 12: Image Processing : Dilation

4.1.4 Image processing : Reconstruction

From the seed image, we expand the pixels through a series of dilations constrained by the mask image provided. This process may involve multiple dilations. Reconstruction can be thought of as a way to isolate the connected regions of an image. For dilation, reconstruction connects regions marked by local maxima in the seed image; neighboring pixels less than or equal to those seeds are connected to the seeded region [1].

To clearly visualize the process, we introduce the following symbols:

- F: Represents the seed image
- G: Represents the mask image
- B: Structural element (used to determine the local neighbors)
- D: Dilation
- $\oplus$: Dilation operator
- $\cap$: Intersection between two matrices (here we use binary matrices, with values of 0 or 1)

Using these symbols, we can express one geodesic dilation (considered as a single dilation from the seed image F to the mask image G by the structural element B during the reconstruction of grayscale images) as follows. The difference from simple dilation is the constraint imposed by G, which is the maximum extent allowed. Therefore, an additional intersection with the mask image is performed after the dilation:

$$D_G^1(F) = (F \oplus B) \cap G$$

This means that we first perform the dilation using the structural element B on our seed image F, and then intersect this output with the mask image G. If we perform n dilations, it can be expressed as:

$$D_G^n(F) = D_G^1(D_G^{n-1}(F)) \text{ with } D_G^0(F) = F$$

These relations can be used to express any number of geodesic dilations in morphological reconstruction. Here is an example of one geodesic dilation [10] (Figure 13):

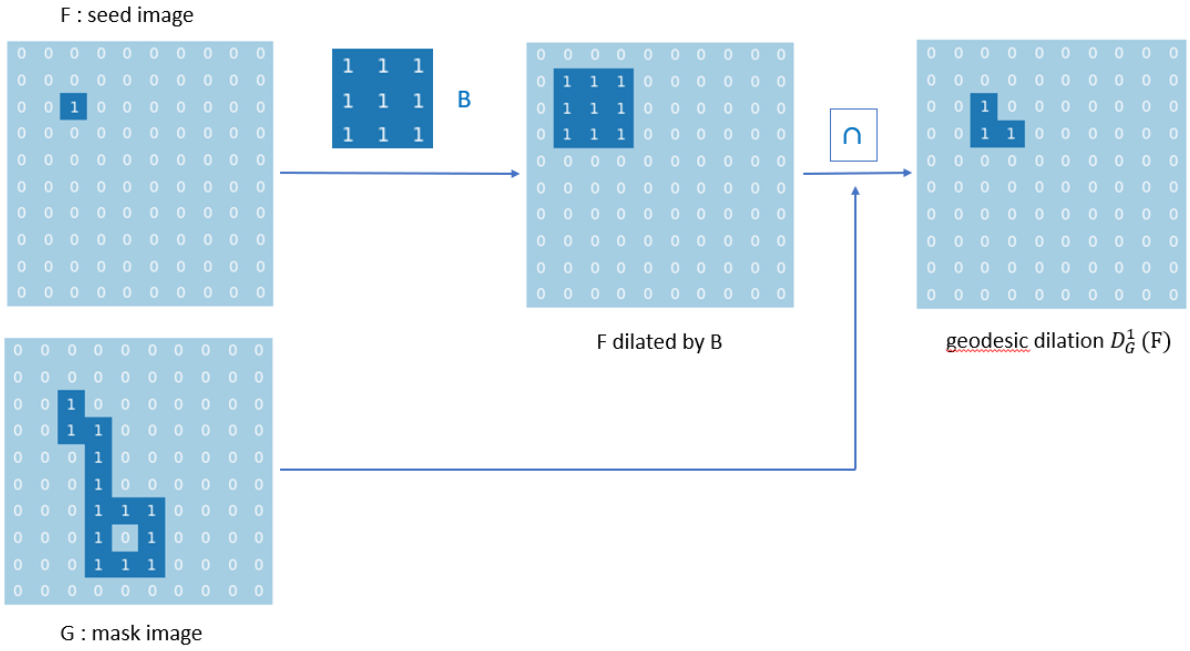

Figure 13: Illustration of geodesic dilation

To visualize the effect of the reconstruction, we apply the double threshold (the two thresholds found by our algorithme in the next part) to the corresponding corrected acoustic image, see in Figure 17.

4.1.5 Assumptions and Limitations of Double Threshold

In this section, we adopt the following variables and concepts from the referenced literature:

- I: A grayscale image where pixel intensities range between g_0 and g_1 (e.g., $g_0 = 0$ and $g_1 = 255$).
- $I(p)$: The intensity value of pixel p in image I.
- U: A set of intensity values representing the object of interest in the image.
- $S_U(I)$: The thresholded image where pixels in I with intensity values in U are set to 1 (foreground), and others are set to 0 (background). This can be written as:

$$S_U(I)(p) = \begin{cases} 1 & \text{if } I(p) \in U \\ 0 & \text{if } I(p) \notin U \end{cases}$$

- U_e: A narrow thresholding interval corresponding to the seed image, with few false positives but some false negatives.
- U_l: A wide thresholding interval corresponding to the mask image, with few false negatives but some false positives.
- $S_{Ue}(I)$: The thresholded image using the interval U_e.
- $S_{Ul}(I)$: The thresholded image using the interval U_l [16].

Generally, false positives (extracted pixels that are not part of the objects of interest) and false negatives (pixels that are part of the objects of interest but are not extracted) coexist in the thresholded image. However, we can make the following two assumptions:

1. With a narrow thresholding interval U_e (corresponding to the seed image), there will be few false positives in $S_{Ue}(I)$, but a certain number of false negatives.
2. With a wide thresholding interval U_l (corresponding to the mask image), there will be very few false negatives in $S_{Ul}(I)$, but a certain number of false positives.

Thus, $S_{Ue}(I)$ consists essentially of a part of the objects of interest that are extracted, while $S_{Ul}(I) \backslash S_{Ue}(I)$ consists mainly of the parts of the objects of interest missed by the narrow thresholding $S_{Ue}(I)$ (the false negatives), as well as the foreign structures and noise introduced by the wide thresholding $S_{Ul}(I)$ (the false positives). We can make two additional assumptions:

3. The false negatives of $S_{Ue}(I)$ (the objects of interest in the area of $S_{Ul}(I) \backslash S_{Ue}(I)$) are generally adjacent to $S_{Ue}(I)$.
4. The false positives of $S_{Ul}(I)$ (the regions of $S_{Ul}(I) \backslash S_{Ue}(I)$ that are foreign to the objects of interest) are generally not adjacent to $S_{Ue}(I)$.

We then take the union of the connected components of $S_{Ul}(I)$ that have a non-empty intersection with $S_{Ue}(I)$, which should help us identify the objects of interest. The operation of taking the union of the connected components of a figure that have a non-empty intersection with a certain part of the figure is well-known and has been the subject of several algorithms, notably queue-based algorithms. This describes the process of reconstruction from the narrow image to the wide image [16].

Despite the advantages, the result of double thresholding is generally better than that of a single thresholding, some errors will persist because our four assumptions can fail:

1. $S_{Ue}(I)$ may have false positives.
2. $S_{Ul}(I)$ may have false negatives.
3. False negatives in $S_{Ue}(I)$ recovered in $S_{Ul}(I)$ may not be connected to $S_{Ue}(I)$.
4. False positives in $S_{Ul}(I)$ may be connected to $S_{Ue}(I)$.

Therefore, it is reasonable to expect that we might commit several errors, such as not being able to reconstruct all the objects of interest or leaving some noise in our reconstructed images [16].

4.1.6 Finding the threshold : threshold_max

As explained previously, we need to determine two thresholds: one to obtain a seed image or a narrow image, and another to obtain a mask image or a larger image. We denote these as follows:

- `thresh_max`: The first threshold, used to determine the seed image.
- `thresh_min`: The second threshold, used to determine the mask image.

Given their purposes, the relationship between them is:

$$\texttt{thresh\_min} < \texttt{thresh\_max}$$

To determine the thresholds, we first need to understand the distribution of pixel intensities in our corrected images. By analyzing this distribution across all corrected images, we observed that the smallest pixel intensity is around -0.7, and the largest is around 0.65. For example, the distribution of intensities in one corrected image, which we have shown multiple times (e.g., in Figure 10), looks as follows (Figure 14):

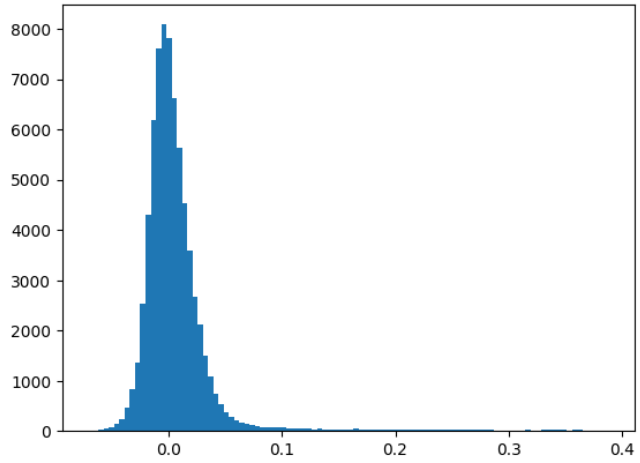

Figure 14: Distribution of pixel intensities

To find the optimal thresholds, we vary them across the entire possible range with a step size of 0.01. For example, if the maximum threshold (`thresh_max`) is 0.30, we evaluate all possible values for the minimum threshold (`thresh_min`) within the range [0, 0.30] in increments of 0.01. We start from 0 because a threshold less than 0 would retain a lot of noise in the thresholded image, resulting in more noise in our fish school extraction. After testing several corrected images (10 images with clear fish schools), we found that when `thresh_max` is within the range of approximately [0.10, 0.24] and `thresh_min` is within the range of approximately [0, 0.10], we obtain a reasonably accurate reconstructed image.

Next, we set `thresh_min` to a reasonable value, such as 0.04, and vary `thresh_max` to observe the resulting reconstructed images. There are two important considerations here. First, as `thresh_max` changes, the number of detected fish schools may vary. In this process, we aim to find an interval where changing `thresh_max` yields the same number of fish schools as observed in the acoustic image. This interval represents a good or convergent range for `thresh_max` in our detection. Second, due to varying sea conditions and sampling times, our acoustic images often differ slightly in pixel intensities. Therefore, we need to find the suitable pair of thresholds for each image. Using the same acoustic image shown in Figure 10 as an example:

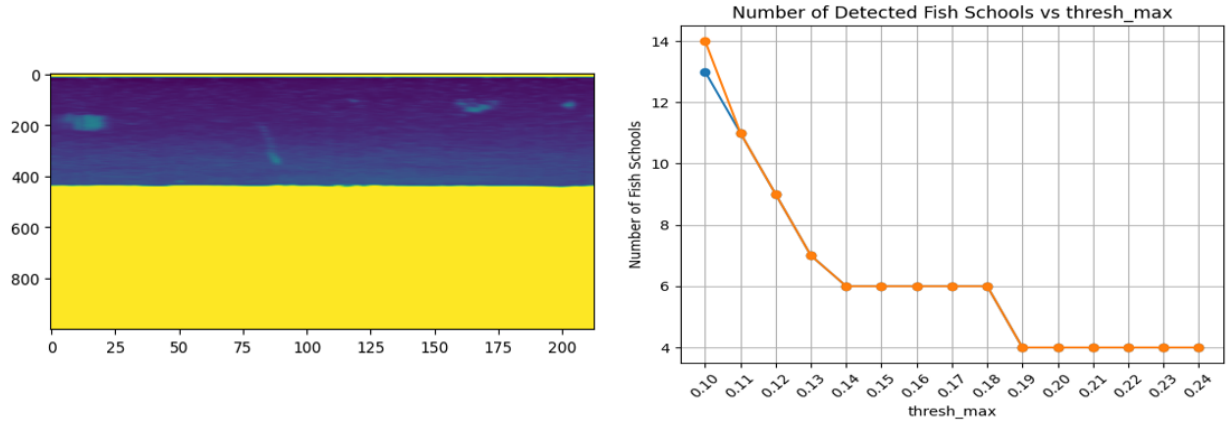

Figure 15: Left: Original resized acoustic image. Right: Number of fish schools detected from this image as `thresh_max` changes. The orange line represents the total number of fish schools, while the blue line represents fish schools larger than 20 pixels.

In this example (Figure 15), when `thresh_max` ranges from 0.19 to 0.24, the number of detected fish schools remains consistent, indicating a convergent interval for the chosen value. To minimize risk, we select the median value within this interval. Additionally, another stable interval is observed from 0.14 to 0.18. When multiple stable intervals exist, we choose the larger one. A convergent interval is considered "stable" if it remains consistent for at least four consecutive values. Therefore, for this acoustic image, we select the median value of the interval [0.19, 0.24], which is 0.215, as our final threshold_max value.

4.1.7 Finding the Second Threshold: threshold_min

Once we have determined the threshold `thresh_max`, we proceed to find `thresh_min` using a similar approach. We fix the previously determined `thresh_max` and vary `thresh_min` from 0 to `thresh_max` to observe how the number of fish schools detected changes. There are two primary aspects we consider when evaluating `thresh_min`:

Firstly, we examine whether the variation in `thresh_min` yields the desired number of fish schools, similar to what was observed in the convergent interval when determining `thresh_max`. Secondly, we assess the properties of these image regions (fish schools) using solidity from the scikit-image package (skimage.measure) to measure various properties of labeled image regions. Solidity is the ratio of pixels in the region to pixels of the convex hull image. A value of 1 signifies a solid object, while values less than 1 suggest an object with an irregular boundary or containing holes.

$$\text{solidity} = \frac{\text{area}}{\text{convex area}}$$

Among the `thresh_min` values that maintain the same number of detected fish schools as in the convergent interval, we select the one that maximizes solidity. As shown in Figure 16, by first choosing the interval that generates the correct number of fish schools, we narrowed our interval to 0-0.12. By finding the maximum solidity within this interval, we determined `thresh_min` to be 0.08. Thus, combining this with the previously found `thresh_max`, our final threshold values are 0.08 and 0.22 (0.215 rounded to two decimal places).

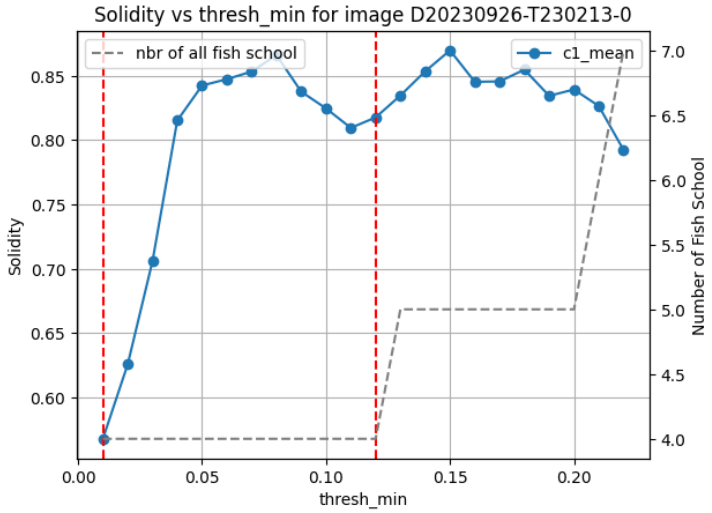

Figure 16: The gray line shows how varying `thresh_min` affects the number of detected fish schools (right y-axis), while the blue line shows the changes in solidity. The red vertical lines mark the interval during which threshold_min generates the same number of fish schools as in the convergent interval found with `thresh_max`.

Using this pair of threshold values, we can produce the final object detection image as follows (Figure 17):

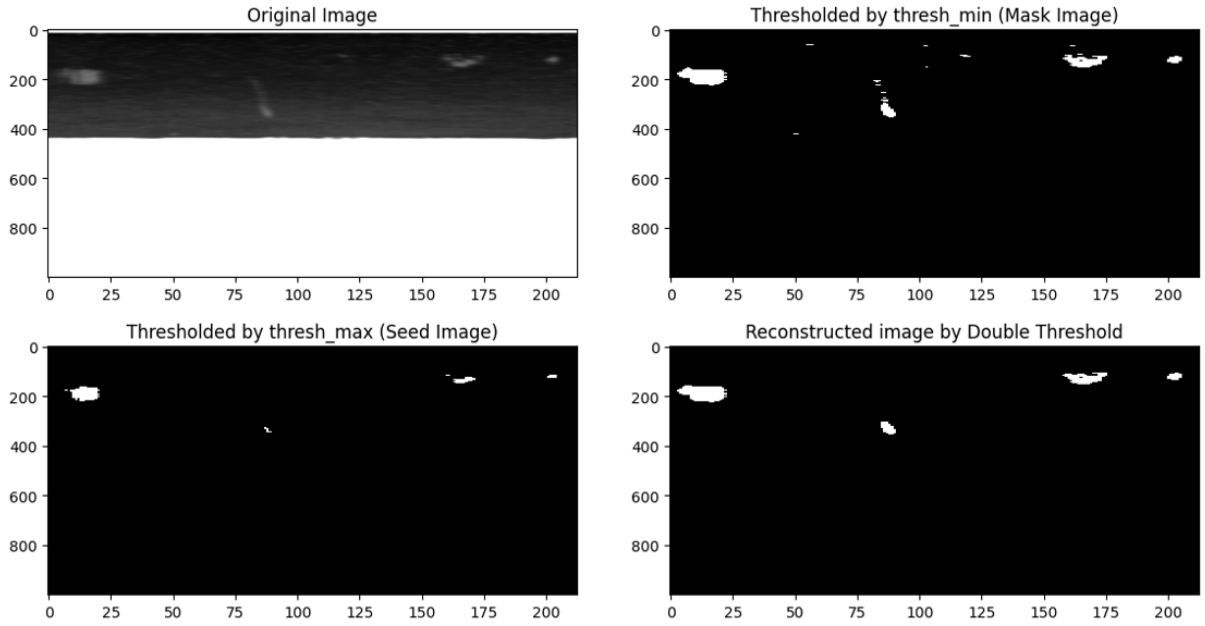

Figure 17: Double Threshold Example: Image D20230926-T230213-0

4.2 Deep learning approach

4.2.1 Introduction

Deep learning has revolutionized computer vision, particularly in the domain of object detection. Object detection involves identifying and localizing objects within images or videos, which is crucial for numerous

applications such as auto-driving, surveillance, and healthcare.

The evolution of object detection using deep learning has been profound. Early approaches, such as region-based convolutional neural networks (R-CNN) and its variants, laid the groundwork by proposing methods to generate region proposals and classify objects within those regions.

A comprehensive review provides an extensive overview of the advancements in deep learning techniques for object detection [19]. According to the review, modern approaches have shifted towards fully convolutional networks (FCNs), which perform end-to-end detection without the need for region proposal algorithms. Models like Faster R-CNN, Single Shot MultiBox Detector (SSD), and You Only Look Once (YOLO) [14] have significantly improved detection speed and accuracy, making them suitable for real-time applications.

Recent innovations have also focused on improving the efficiency of object detection models. For instance, the introduction of anchor-free detection methods and the integration of attention mechanisms have further enhanced the performance of these models.

One notable advancement is about zero-shot learning and few-shot learning capabilities in object detection models. These capabilities allow models to generalize to new classes or datasets with minimal or no additional training data, which is crucial for handling diverse and evolving environments [19].

Segment Anything (SAM), developed by Meta AI in 2023, exemplifies these advancements with its impressive zero-shot performance. SAM has shown competitiveness with or superiority over prior fully supervised methods, making it an attractive choice for tasks where extensive annotation and training data are not readily available . This capability not only saves time and labor in model development but also offers flexibility and robustness in handling various object detection challenges [9].

4.2.2 The Components of Segment Anything

The goal of Segment Anything is to build a foundational model for image segmentation. Developed by the Meta AI group, this model is promptable and pre-trained on a broad dataset, enabling powerful generalization. The success of Segment Anything hinges on three components: task, model, and data. Users can experience its robust capabilities through the official website, where they can try a demo with various images or upload their own to test its segmentation ability. Additionally, the code is open-source, making it easily accessible and applicable.

The promptable segmentation task defined by the authors is general enough to provide a powerful pre-training objective and enable a wide range of downstream applications. They adapted the concept of prompts from NLP to segmentation, where a prompt can be a set of foreground/background points, a rough box or mask, free-form text, or any information indicating what to segment in an image. The promptable segmentation task returns a valid segmentation mask for any given prompt. A "valid" mask means that even when a prompt is ambiguous and could refer to multiple objects, the output should be a reasonable mask for at least one of those objects. For example, if a point is clicked on a snail in an image (Figure 18), the mask could represent just a part of the snail, the entire snail, or even the snail and a connected frog. At least one of these masks should be reasonable.

Figure 18: SAM generates 3 valid masks from a single ambiguous point prompt (green point)[2]

The model component of Segment Anything, referred to as the Segment Anything Model (SAM), is designed to handle a variety of prompts and produce high-quality segmentation masks. SAM leverages a powerful transformer-based architecture, which has shown remarkable success in various vision tasks. By separating SAM into an image encoder and a fast prompt encoder/mask decoder, the same image embedding can be reused (and its cost amortized) with different prompts. Given an image embedding, the prompt encoder and mask decoder predict a mask from a prompt in approximately 50ms in a web browser. This architecture allows SAM to process different types of prompts effectively and produce accurate segmentation masks across a wide range of images and objects.

The data component is crucial to the success of SAM. The model is pre-trained on the largest ever segmentation dataset, consisting of over 1 billion masks across 11 million licensed and privacy-respecting images. This extensive dataset provides the model with a rich and diverse set of examples, enabling it to generalize well to new images and unseen objects. The dataset is curated to ensure high-quality annotations, which is vital for effective pretraining and downstream performance.

In summary, the components of Segment Anything—task, model, and data—work synergistically to create a robust and versatile segmentation system. The model is designed and trained to be promptable, so it can transfer zero-shot to new image distributions and tasks. Its capabilities were evaluated on numerous tasks, and its zero-shot performance was found to be impressive—often competitive with or even superior to prior fully supervised results [9]. For an example, see Figure 19 showcasing the object detection demo from the Segment Anything homepage.

In this internship, we utilize Segment Anything for object detection in acoustic images and compare its performance with our first method, double threshold.

Figure 19: Segment Anything Online Demo[2]

4.2.3 Application of Segment Anything Model (SAM)

In this part, we first installed all the necessary packages required to run the Segment Anything model, such as PyTorch and other dependencies. We then downloaded a model checkpoint, allowing us to apply the model in just a few lines to get masks from a given prompt. Each mask output includes annotations such as "segmentation," representing the mask in COCO RLE format, "bbox," denoting the bounding box in XYWH format ([x, y, w, h]), and "area," an integer indicating the mask's pixel area. Additional annotation details are stored but are not currently utilized in our analysis.

On a personal computer, generating masks for each image using SAM requires approximately 7 minutes per image. Given the total of 5606 images, this process is time-consuming, prompting us to execute it on a GPU, which reduces the processing time to about 10 hours.

After obtaining masks for all images, the next step involves selecting the appropriate masks. From the masks generated by SAM for one acoustic image, we identified several segments. Below in the Figure 20 are examples of segmentations evaluated:

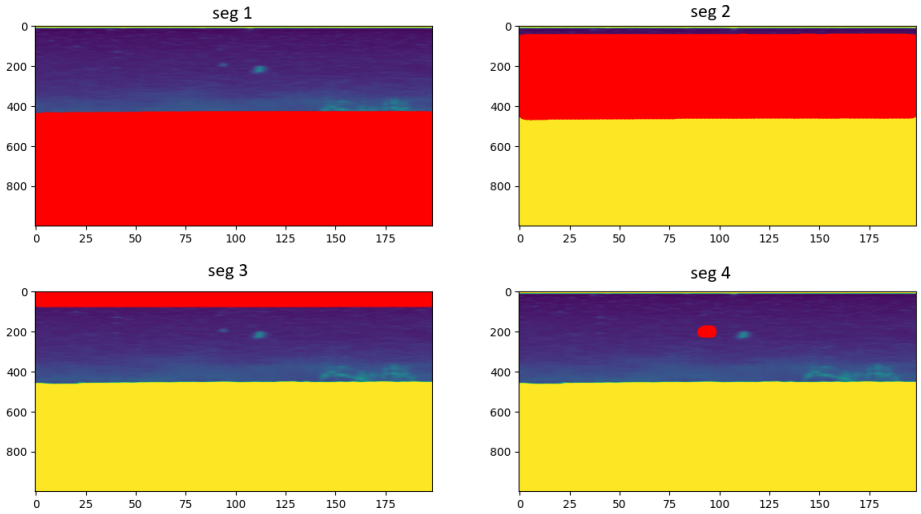

Figure 20: Segmentation is highlighted in red.

- **Segmentation 1:** This segmentation captures areas not relevant to fish schools, but the removed echo parts. Therefore, it should be disregarded. We determine its relevance by checking if it intersects with our mask matrix, the same matrix used in the masking process. Since it intersects with the mask, we disregard it.
- **Segmentation 2:** This segmentation covers almost the entire sea area without focusing on specific fish schools. Thus, it should be disregarded. We determine its relevance by checking if its area exceeds a predefined limit. Since it does, we disregard it.
- **Segmentation 3:** This segmentation captures waves rather than fish schools. It is disregarded for the same reason as Segmentation 1, due to its intersection with the mask matrix.
- **Segmentation 4:** This segmentation accurately identifies a fish school and is selected for further analysis. For small segmentation objects like this, we compare its average intensity with the average intensity in the masked image. If the intensity is higher, we choose it; otherwise, we disregard it. In this case, the intensity is higher than the average masked image intensity, so we select it.

After meticulous evaluation of SAM-generated masks, the final selected segmentations are illustrated as follows (Figure 21):

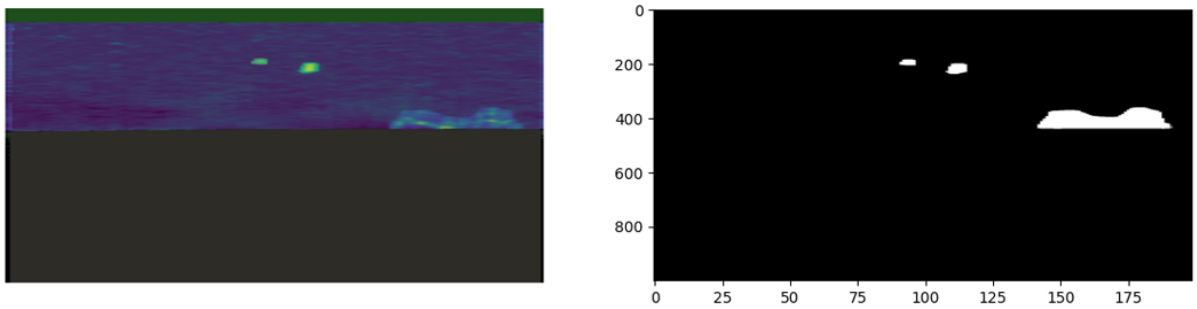

Figure 21: Left: Output by SAM showing detected segments in various colors; Right: Selected segmentations

5 Comparative Analysis of Results: Double Threshold vs. Segment Anything

5.1 General Comparison of Detection Capabilities

5.1.1 General Comparison

Out of a total of 5,606 acoustic images, the double threshold method identified 1,664 images containing fish schools. This number is based on the images where the algorithm could precisely determine the thresholds for detecting fish schools. Although more images had fish schools detected, the algorithm could not always extract the optimal threshold pairs, so only the 1,664 images with precise extraction were used. In comparison, the Segment Anything method detected 1,874 images with fish schools. Both methods identified 1,065 images in common (Figure 22).

On average, the double threshold method detected 1.95 fish schools per image, with an average fish school size of 1097 pixels. The average intensity of the fish schools was -70.50, while the average image intensity was -82.49. The mean fish depth (within one image) was 39.67 meters.

The Segment Anything method, on the other hand, detected an average of 1.80 fish schools per image, with an average fish school size of 2340 pixels. The average intensity of the fish schools was -76.79, and the average image intensity was -82.33. The mean fish depth (within one image) was 41.99 meters.

	Double Threshold		Segment Anything
	only_thresh	both	only_sam
Images Detected	599	1065	809

Figure 22: Images and schools detected by the two methods

To simplify our analysis, we divided the images into the following four groups:

- **only_thresh**: Images detected only by the double threshold method (599 images, see Figure above)
- **only_sam**: Images detected only by the Segment Anything method (809 images, see Figure above)
- **both_thresh**: Images detected by both methods, using values from the double threshold method (1065 images, see Figure above)
- **both_sam**: Images detected by both methods, using values from the Segment Anything method (1065 images, see Figure above)

5.1.2 Student's T-Test for Comparing Mean School Intensities Between only_thresh and both_thresh Groups (Image Level)

We are interested in determining if there are any differences between the two methods in terms of the images or fish schools they detect. To explore this, we perform hypothesis tests on the school intensity between different groups.

We aim to test if there is a significant difference in the mean school intensities between these groups (only_thresh, only_sam, both_thresh, and both_sam). Specifically, we first compare the groups "only_thresh" and "both_thresh". This comparison allows us to assess if there is a difference in school intensities between images detected only by the double threshold method and those detected by both methods from the perspective of the double threshold.

We choose to use a t-test for this comparison because it is appropriate when comparing the means of two independent groups, assuming the data is approximately normally distributed. The t-test will help us determine if the observed difference in mean intensities between "only_thresh" and "both_thresh" is statistically significant, indicating whether this difference is likely due to the methods themselves rather than random chance.

In all our tests, we set the significance level α to 5%. All the principe of t-test used in this analysis are based on the Statistics with Python course materials by Nicolas Poulin [12].

1. **Context of the Student's t-test:**

   - X: Quantitative continuous variable representing the school intensity in the image.
   - Y: Qualitative variable with two levels (group marker, here the group only_thresh and group both_thresh).
   - We denote $X_{\text{only\_thresh}}$ and $X_{\text{both\_thresh}}$ as the variables for the two groups, respectively.
   - $\mu_{\text{only\_thresh}}$: The mean of X in the group only_thresh.
   - $\mu_{\text{both\_thresh}}$: The mean of X in the group both_thresh.
   - $\sigma_{\text{only\_thresh}}$: The standard deviation of X in the group only_thresh.
   - $\sigma_{\text{both\_thresh}}$: The standard deviation of X in the group both_thresh.
   - $n_{\text{only\_thresh}}$: The sample size for the group only_thresh, $n_{\text{only\_thresh}} = 599$.
   - $n_{\text{both\_thresh}}$: The sample size for the group both_thresh, $n_{\text{both\_thresh}} = 1065$.
   - $\overline{X}_{\text{only\_thresh}}$: The sample mean school intensities for the group only_thresh, $\overline{X}_{\text{only\_thresh}} = -70.60$.
   - $\overline{X}_{\text{both\_thresh}}$: The sample mean school intensities for the group both_thresh, $\overline{X}_{\text{both\_thresh}} = -70.45$ [12].

2. **The hypotheses are:**

   - Null hypothesis (H_0): There is no difference in mean school intensities between the "only_thresh" and "both_thresh" groups.

   $$H_0 : \mu_{\text{only\_thresh}} = \mu_{\text{both\_thresh}}$$

   - Alternative hypothesis (H_1): There is a difference in mean school intensities between the "only_thresh" and "both_thresh" groups.

   $$H_1 : \mu_{\text{only\_thresh}} \neq \mu_{\text{both\_thresh}} [12]$$

The distribution of school intensities in these two groups is shown in the following Figure 23:

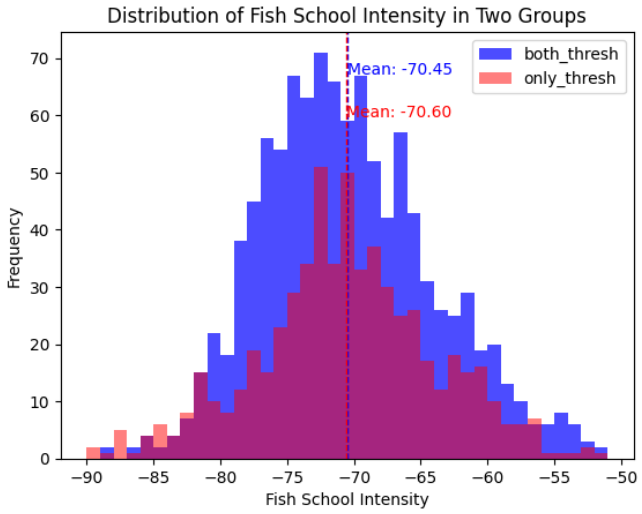

Figure 23: Distribution of mean school intensity in both_thresh group and only_thresh group

3. **Assumptions:**

- Each sample consists of independent observations.
- The two samples are independent.
- $X_{\text{only\_thresh}} \sim \mathcal{N}(\mu_{\text{only\_thresh}}, \sigma_{\text{only\_thresh}})$.
- $X_{\text{both\_thresh}} \sim \mathcal{N}(\mu_{\text{both\_thresh}}, \sigma_{\text{both\_thresh}})$.
- $\sigma_{\text{only\_thresh}}$ and $\sigma_{\text{both\_thresh}}$ are unknown.
- $\sigma_{\text{only\_thresh}} = \sigma_{\text{both\_thresh}}$ [12].

4. **Validation of the assumptions:**

   (a) **Independence:** The images are generally independent acoustic samples taken from different sampling positions or GPS coordinates in the sea.

   (b) **Normality:** Although the Shapiro-Wilk test for normality shows a low p-value, indicating deviation from normality, both groups have large number of samples each (more than 30). The distributions do not exhibit extreme skewness or heavy tails, and the QQ plots are satisfactory (Figure 24), so we assume normality.

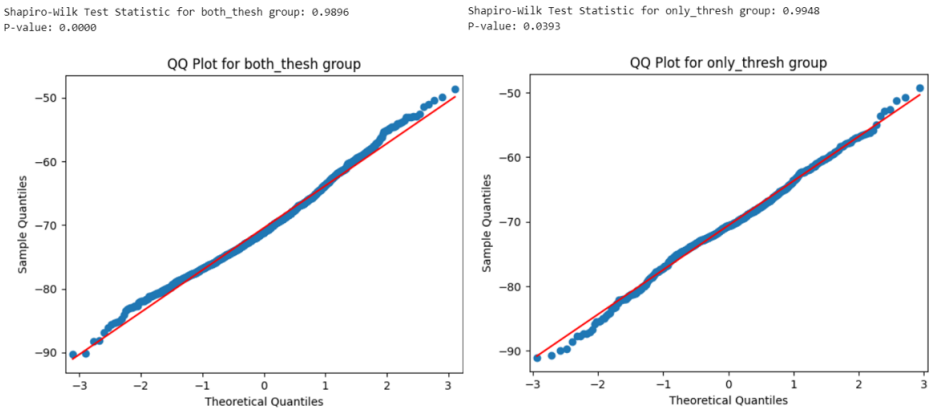

Figure 24: QQ plot for the two groups

   (c) **Equality of variances:** We use both Bartlett's Test and Levene's Test to test the equality of variances. Both tests yield p-values greater than α (Table 1), validating the assumption of equal variances.

5. **Decision Statistic:**

We use the t-statistic to compare the means of the two groups. Since we have validated the equality of variances, the t-statistic is calculated as follows:

$$t = \frac{\overline{X}_{\text{only\_thresh}} - \overline{X}_{\text{both\_thresh}}}{s_p \sqrt{\frac{1}{n_{\text{only\_thresh}}} + \frac{1}{n_{\text{both\_thresh}}}}}$$

where the pooled standard deviation s_p is given by:

$$s_p = \sqrt{\frac{(n_{\text{only\_thresh}} - 1)s^2_{\text{only\_thresh}} + (n_{\text{both\_thresh}} - 1)s^2_{\text{both\_thresh}}}{n_{\text{only\_thresh}} + n_{\text{both\_thresh}} - 2}}$$

where: $\overline{X}_{\text{only\_thresh}} = -70.60$, $\overline{X}_{\text{both\_thresh}} = -70.45$, $n_{\text{only\_thresh}} = 599$, $n_{\text{both\_thresh}} = 1065$, $s_{\text{only\_thresh}}$ represents the sample standard deviation of school intensities in the `only_thresh` group. $s_{\text{both\_thresh}}$ represents the sample standard deviation of school intensities in the `both_thresh` group.

Under H_0, the t-statistic follows the Student's t-distribution with degrees of freedom df calculated as:

$$df = n_{\text{only\_thresh}} + n_{\text{both\_thresh}} - 2 = 599 + 1065 - 2 = 1662$$

Based on the calculated t-value and the degrees of freedom, we compare the t-value to the critical value from the t-distribution table at a significance level of α. If the absolute value of the calculated t-value is greater than the critical t-value, we reject the null hypothesis. Otherwise, we fail to reject the null hypothesis [12].

6. **Student's t-test results:** Since all assumptions for the t-test are satisfied, we apply the Student's t-test to the "only_thresh" and "both_thresh" groups. The test statistic is 0.4394, and the p-value is 0.66.

	Test	**Statistic**	**p-value**	**Decision**
Normality in group *both_thresh*	Shapiro-Wilk	0.9896	0	Reject
Normality in group *only_thresh*	Shapiro-Wilk	0.9948	0.04	Reject
Equality of Variance 1	Bartlett's Test	1.3028	0.25	Accept
Equality of Variance 2	Levene's Test	0.0731	0.79	Accept
No Difference in Mean	Student's t-test	0.4394	0.66	Accept

Table 1: Summary of Tests and Decisions

7. **Conclusion:** Since the p-value is greater than our significance level α, we fail to reject the null hypothesis. Therefore, we conclude that there is no significant difference in mean school intensities between the "only_thresh" and "both_thresh" groups.

5.1.3 Welch's t-test for Comparing Mean School Intensities Between only_sam and both_sam Groups (Image Level)

Since there is no significant difference in school intensity between the two groups of images detected by the double threshold method, we want to examine the situation for the SAM method. We test the image groups "only_sam" and "both_sam". Welch's t-test is chosen because it is used for comparing mean values between two groups when the assumption of equal variances does not hold. Upon verifying the equality of variances, we found that the variances between the two groups are not equal, and the normality assuption is validated. Therefore, we use Welch's t-test to accurately compare the mean school intensities.

1. **Context of Welch's t-test:**
   - X: Quantitative continuous variable representing the school intensity in the image.
   - Y: Qualitative variable with two levels (group marker, here the group only_sam and group both_sam).
   - We denote $X_{\text{only\_sam}}$ and $X_{\text{both\_sam}}$ as the variables for the two groups, respectively.
   - $\mu_{\text{only\_sam}}$: The mean of X in the group only_sam.
   - $\mu_{\text{both\_sam}}$: The mean of X in the group both_sam.
   - $\sigma_{\text{only\_sam}}$: The standard deviation of X in the group only_sam.
   - $\sigma_{\text{both\_sam}}$: The standard deviation of X in the group both_sam.
   - $n_{\text{only\_sam}}$: The sample size for the group only_sam, $n_{\text{only\_sam}} = 809$.
   - $n_{\text{both\_sam}}$: The sample size for the group both_sam, $n_{\text{both\_sam}} = 1065$.
   - $\overline{X}_{\text{only\_sam}}$: The sample mean school intensity for the group only_sam, $\overline{X}_{\text{only\_sam}} = -79.09$.
   - $\overline{X}_{\text{both\_sam}}$: The sample mean school intensity for the group both_sam, $\overline{X}_{\text{both\_sam}} = -75.05$ [12].
2. **The hypotheses are:**

- Null hypothesis (H_0): There is no difference in mean school intensities between the "only_sam" and "both_sam" groups.

$$H_0 : \mu_{\text{only\_sam}} = \mu_{\text{both\_sam}}$$

- Alternative hypothesis (H_1): There is a difference in mean school intensities between the "only_sam" and "both_sam" groups.

$$H_1 : \mu_{\text{only\_sam}} \neq \mu_{\text{both\_sam}}$$

The distribution of school intensities in these two groups is shown in the following Figure 25:

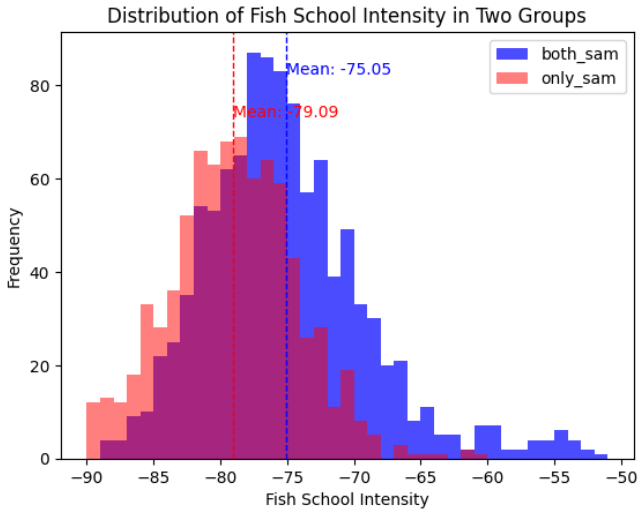

Figure 25: The distribution of mean school intensity in only_sam group and both_sam group

3. **Assumptions:**

   - Each sample consists of independent observations.
   - The two samples are independent.
   - $X_{\text{only\_sam}} \sim \mathcal{N}(\mu_{\text{only\_sam}}, \sigma_{\text{only\_sam}})$.
   - $X_{\text{both\_sam}} \sim \mathcal{N}(\mu_{\text{both\_sam}}, \sigma_{\text{both\_sam}})$.
   - $\sigma_{\text{only\_sam}}$ and $\sigma_{\text{both\_sam}}$ are unknown.
   - $\sigma_{\text{only\_sam}} \neq \sigma_{\text{both\_sam}}$ [12].

4. **Validation of the assumptions:**

   (a) **Independence:** The images are independent acoustic samples from different sampling positions or GPS coordinates in the sea. Each method (SAM and double threshold) analyzes the images independently, ensuring no interaction between them.

   (b) **Normality:** Although the Shapiro-Wilk test for normality yields a low p-value, indicating some deviation from normality, and the QQ plots do not confirm the assumption of normality (Figure 26), the Central Limit Theorem allows us to assume normality given that both groups have sample sizes well above 30 (in this case, far exceeding 30). Thus, we can reasonably assume that the distribution of the sample means approximates normality.

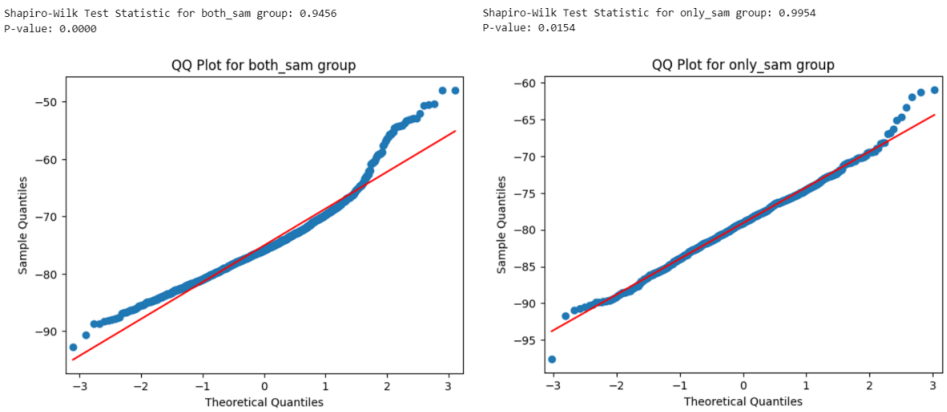

Figure 26: QQ plot for the two groups

(c) **Equality of variances:** We use both Bartlett's Test and Levene's Test to test the equality of variances. Both tests return p-values less than α (Table 2), indicating that we do not have equality of variances between these two groups. Therefore, we will use the Welch's t-test in Python to perform this test.

5. **Decision Statistic:** We use Welch's t-statistic to compare the means of the two groups. As we have established that the variances are not equal, the t-statistic is calculated using the following formula:

$$t = \frac{\overline{X}_{\text{only\_sam}} - \overline{X}_{\text{both\_sam}}}{\sqrt{\frac{s^2_{\text{only\_sam}}}{n_{\text{only\_sam}}} + \frac{s^2_{\text{both\_sam}}}{n_{\text{both\_sam}}}}}$$

where: $\overline{X}_{\text{only\_sam}} = -79.09$, $\overline{X}_{\text{both\_sam}} = -75.05$, $n_{\text{only\_sam}} = 809$, $n_{\text{both\_sam}} = 1065$, $s_{\text{only\_sam}}$ represents the sample standard deviation of school intensities in the `only_sam` group. $s_{\text{both\_sam}}$ represents the sample standard deviation of school intensities in the `both_sam` group.

Under H_0, the t-statistic follows Welch's t-distribution with degrees of freedom (df) calculated using the following formula:

$$df = \frac{\left(\frac{s^2_{\text{only\_sam}}}{n_{\text{only\_sam}}} + \frac{s^2_{\text{both\_sam}}}{n_{\text{both\_sam}}}\right)^2}{\frac{\left(\frac{s^2_{\text{only\_sam}}}{n_{\text{only\_sam}}}\right)^2}{n_{\text{only\_sam}}-1} + \frac{\left(\frac{s^2_{\text{both\_sam}}}{n_{\text{both\_sam}}}\right)^2}{n_{\text{both\_sam}}-1}}$$

Based on the calculated t-value and the degrees of freedom, we compare the t-value to the critical value from the t-distribution table at a significance level of α. If the absolute value of the calculated t-value is greater than the critical t-value, we reject the null hypothesis. Otherwise, we fail to reject the null hypothesis [12].

6. **Welch's t-test results:** Since the assumption of equal variances is violated (p-value $< \alpha$, see in Table 2), we apply Welch's t-test to the "only_sam" and "both_sam" groups. The test results are as follows:

7. **Conclusion:** Since the p-value obtained from Welch's t-test is less than our significance level α, we reject the null hypothesis. Therefore, we conclude that there is a significant difference in school intensities between the "only_sam" and "both_sam" groups.

	Test	Statistic	p-value	Decision
Normality in group *both_sam*	Shapiro-Wilk	0.9896	0	Reject
Normality in group *only_sam*	Shapiro-Wilk	0.9948	0.04	Reject
Equality of Variance 1	Bartlett's Test	67.4571	2.15e-16	Reject
Equality of Variance 2	Levene's Test	24.6841	7.37e-07	Reject
No Difference in Mean	Welch's t-test	15.5298	2.98e-51	Reject

Table 2: Summary of Tests and Decisions

5.2 Depth-Based Segmentation Comparison

In this subsection, we focus on extracting detailed information about fish schools from acoustic images using two methods. Unlike the previous section, which provided average information per image (where one image may contain multiple fish schools), we now delve deeper into specific attributes such as fish school depth in the sea, distance from the seabed, and fish school intensity. The primary distinction from the previous section lies in our direct comparison of the distributions of fish schools between the two methods, rather than a general comparison based on average fish school information per image.

5.2.1 Total Distribution Comparison

1. Distribution of Fish School Intensity :

   When we compare the fish school intensity by the two methods. The standard deviation of fish school intensity for the double threshold and SAM methods are 7.08 and 6.49, respectively. The comparison is illustrated in Figure 27:

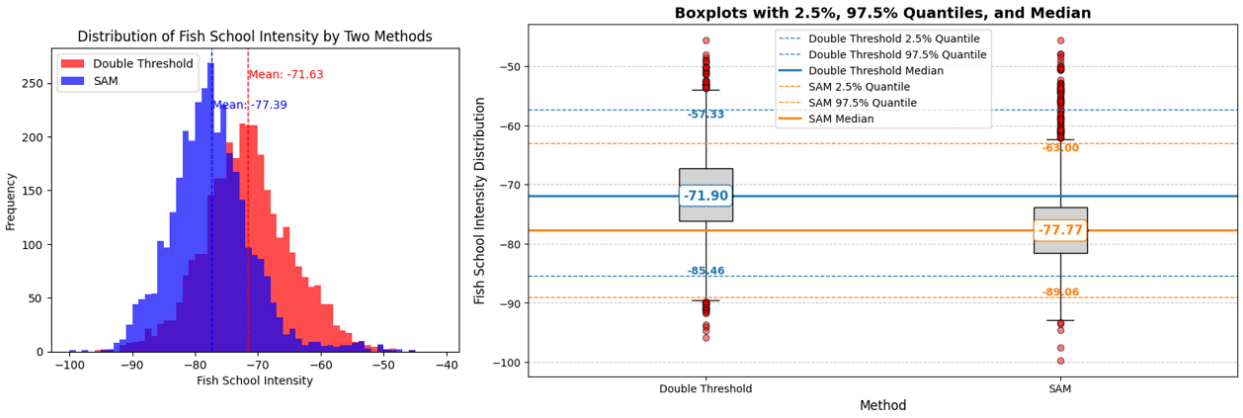

Figure 27: Distribution of Fish School Intensity by Two Methods

2. Distribution of Fish School Depth:

   In addition to fish school intensity, there are other important characteristics associated with fish schools, such as depth, size, and shape etc. We compare the distribution of fish school depth obtained using two different methods. The standard deviation of fish school depth for the double threshold and SAM methods are 19.96 and 18.92, respectively. The comparison is illustrated in Figure 28:

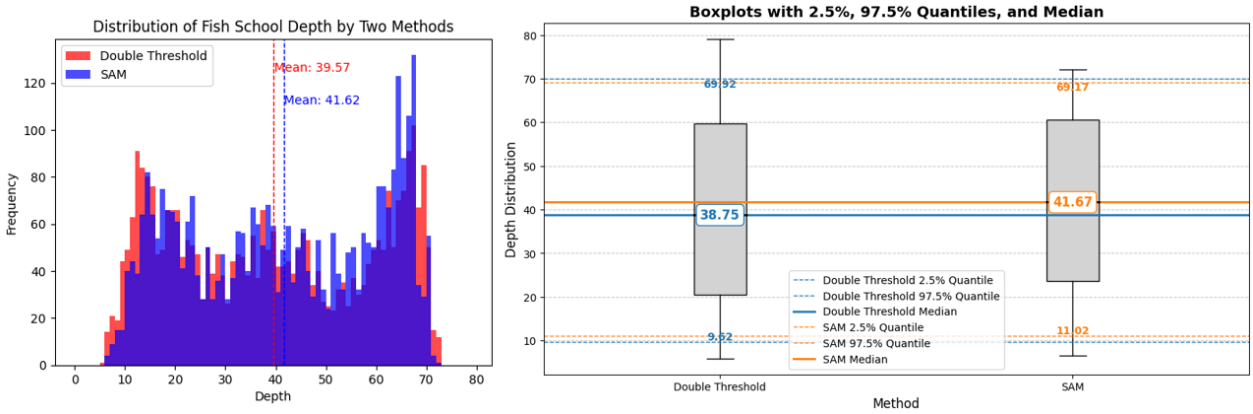

Figure 28: Distribution of Fish School Depth for Two Methods

5.2.2 Bootstrap Comparison of Fish School Intensity (Fish School Level)

The distributions of fish school intensity appear to differ, prompting us to investigate whether there is a significant difference between the intensity measured by two different methods. Initially, we consider that both methods draw from the same 5606 acoustic images. The double threshold method detects a certain number of fish schools, while the SAM method detects another set. Some fish schools may be detected by both methods, while each method also identifies unique schools. Therefore, our samples from these two groups are not entirely independent or paired.

Given this scenario, we opt to use bootstrap resampling to assess whether the mean intensity of fish schools differs between the two groups. All the principles of bootstrap resampling used in this analysis are based on the Statistical Tools S3 course materials by Davide Giraudo [8].

- Principe of Bootstrap:

  Statistical estimation involves determining the value of a statistic of interest (mean, variance, quantile, etc.) for an unknown distribution. The mathematical quantities involved are as follows:

  - An unknown distribution P. Let $X = (X_1, \ldots, X_n)$ denote the random vector containing n i.i.d. samples drawn from the distribution P.
  - A statistic of interest θ that depends on P and that we aim to estimate.
  - A statistic $T(X)$ used to estimate θ, called the estimator of θ.

  The only direct knowledge we can extract from the data is a sample, that is, a realization $x_1, \ldots, x_n$ of X. The value taken by T on the collected sample is called the estimate of θ and is denoted $\hat{\theta}$:

  $$\hat{\theta} = T(x_1, \ldots, x_n)$$

  However, this value is only an approximation of the true value of interest θ. To get an idea of the error committed (via confidence intervals) or to compare estimators, it is necessary to study the properties of T: its bias, variance, mean squared error, distribution function, etc. However, since T depends on P, which is unknown, it is impossible to access its exact distribution. Therefore, it is necessary to resort to assumptions or approximations [8].

  Advantages:

  1. Any distribution can be handled.
  2. Any estimator, regardless of its complexity, can be considered.

- Generate bootstrap samples:

  The classic procedure, introduced by Efron [5], corresponds to the one described in the previous chapter, namely that a large number B of samples are drawn according to the empirical distribution of the sample

$X = (X_1, \ldots, X_n)$. Drawing according to this distribution amounts to uniformly drawing an element from the sample. The principle of the bootstrap can thus be simply written as:

For i from 1 to B:

- Draw n times with replacement from X to obtain a bootstrap sample $X_i^* = (X_{i1}^*, \ldots, X_{in}^*)$.
- Obtain a bootstrap estimate $\hat{\theta}_i^* = T(X_i^*)$.

- **The distribution of $\hat{\theta}$**
  - Real world: we want to estimate $G(t) = \mathbb{P}(\hat{\theta} \leq t)$.
  - Bootstrap world:
    * First approximation: we approximate $G(t)$ by $G_n^*(t) = \mathbb{P}(\hat{\theta^*} \leq t)$.
    * Second approximation: we approximate $G_n^*(t)$ by the empirical distribution

$$G_{n,B}(t) = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}_{(\hat{\theta_b^*}) \leq t} \text{[8]}.$$

- **Bias estimation of $\hat{\theta}$**
  - Real world: we want to estimate $\mathbb{E}_P[\hat{\theta}] - \theta$.
  - Bootstrap world:
    * First approximation: we approximate the bias by $\mathbb{E}_{P_n}[\hat{\theta^*}] - \theta$.
    * Second approximation: we approximate the previous quantity by

$$\frac{1}{B} \sum_{b=1}^{B} \hat{\theta_b^*} - \hat{\theta}.$$

    The unbiased estimator is obtained by subtracting the bias estimate from the initial estimator. We thus obtain the estimator

$$\hat{\theta} - \left(\frac{1}{B} \sum_{b=1}^{B} \hat{\theta_b^*} - \hat{\theta} \right) = 2\hat{\theta} - \frac{1}{B} \sum_{b=1}^{B} \hat{\theta_b^*} \text{[8]}.$$

- **Variance estimation of θ_b**
  - Real world: we want to estimate $\mathbb{E}[(\mathbb{E}[\theta_b] - \mathbb{E}_P[\theta_b])^2]$.
  - Bootstrap world:
    * First approximation: we approximate by $\mathbb{E}_P[(\theta_b^* - \mathbb{E}_{P_n}[\theta_b^*])^2]$.
    * Second approximation: we approximate the previous quantity by

$$\frac{1}{B} \sum_{b=1}^{B} (\theta_b^* - \frac{1}{B} \sum_{b'=1}^{B} \hat{\theta_{b'}^*})^2 \text{[8]}.$$

- **Percentile Confidence Interval**

  One of the simplest ways to provide a confidence interval is to neglect the first approximation and thus consider that the distributions of θ_b and θ_b^* are identical. The $1 - \alpha$ confidence interval is then

$$IC_{perc}(1 - \alpha) = \left[\hat{\theta}^*_{\lceil B\alpha/2 \rceil}, \hat{\theta}^*_{\lceil B(1-\alpha)/2 \rceil} \right] \text{[8]}.$$

- **Classic Confidence Interval**

  An alternative is to consider the distribution of the error $\theta_b - \theta$, whose bootstrap realizations are $\theta_b^* - \theta$. A confidence interval for the error is then

$$\left[\hat{\theta}^*_{\lceil B\alpha/2 \rceil} - \hat{\theta}, \hat{\theta}^*_{\lceil B(1-\alpha)/2 \rceil} - \hat{\theta} \right],$$

  which gives the confidence interval

$$IC(1 - \alpha) = \left[2\hat{\theta} - \hat{\theta}^*_{\lceil B(1-\alpha)/2 \rceil}, 2\hat{\theta} - \hat{\theta}^*_{\lceil B\alpha/2 \rceil} \right] \text{[8]}.$$

5.2.3 Application of Bootstrap

In this section, we introduce the variables used in our analysis and explain the application of the bootstrap method to test the difference in means between two methods: double threshold and Segment Anything Model (SAM).

- **Context:**
  - n_{thresh}: Number of observations in the double threshold sample.
  - n_{sam}: Number of observations in the SAM sample.
  - $\bar{x}_{\text{thresh}}$: Mean fish school intensity in the double threshold sample.
  - $\bar{x}_{\text{sam}}$: Mean fish school intensity in the SAM sample.
  - boot_thresh: Bootstrap sample of mean intensities from the double threshold method.
  - boot_sam: Bootstrap sample of mean intensities from the SAM method.
  - $B = 10000$: Number of bootstrap samples drawn for the bootstrap procedure.
- **Hypothesis Test using Bootstrap:**

  The hypotheses for testing the difference in means between the two methods are as follows:
  - H_0: There is no difference in mean fish school intensity between the two methods.

    $$\mu_{\text{thresh}} = \mu_{\text{sam}}$$

    where μ_{thresh} and μ_{sam} represent the mean fish school intensities by double threshold and SAM, respectively.
  - H_1: There is a difference in mean fish school intensity between the two methods.

    $$\mu_{\text{thresh}} \neq \mu_{\text{sam}}$$
- **Test Statistic using Bootstrap:**

  The test statistic used to assess the difference in means is based on the absolute observed difference in sample means:

  $$\text{Observed Difference} = |\bar{x}_{\text{thresh}} - \bar{x}_{\text{sam}}|$$
- **Bootstrap Procedure:**

  To test the hypothesis, a bootstrap resampling procedure was conducted:
  1. **Combine Samples**: Combine the intensities from double threshold and SAM into a single dataset (`combined_intensities`).
  2. **Compute Observed Difference**:

     $$\text{Observed Difference} = |\bar{x}_{\text{thresh}} - \bar{x}_{\text{sam}}| = |\bar{x}_{\text{thresh}} - \bar{x}_{\text{sam}}| = |(-71.63) - (-77.39)| = 5.76$$
  3. **Bootstrap Sampling**: Repeat B times:
     - Sample with replacement from `combined_intensities` to create bootstrap samples: bootstrap sample for Double Threshold with sample size n_{thresh}, and bootstrap sample for SAM with sample size n_{sam} .
     - Compute the absolute difference in means for each bootstrap sample:

       $$\text{Bootstrap Difference} = |\bar{x}_{\text{boot\_thresh}} - \bar{x}_{\text{boot\_sam}}|$$
  4. **Calculate p-value**: Compute the proportion of bootstrap samples where the absolute difference in means is as extreme as or more extreme than the observed difference.

- **Results of Bootstrap:**

  The bootstrap resampling produced an observed absolute difference in means of 5.76. The 95% confidence interval for the mean difference is $(-0.35, 0.35)$.

- **Conclusion:**

  The bootstrap procedure resulted in a p-value of 0, indicating strong evidence against the null hypothesis (H_0). This suggests that there is a significant difference in mean fish school intensity between double threshold and SAM methods.

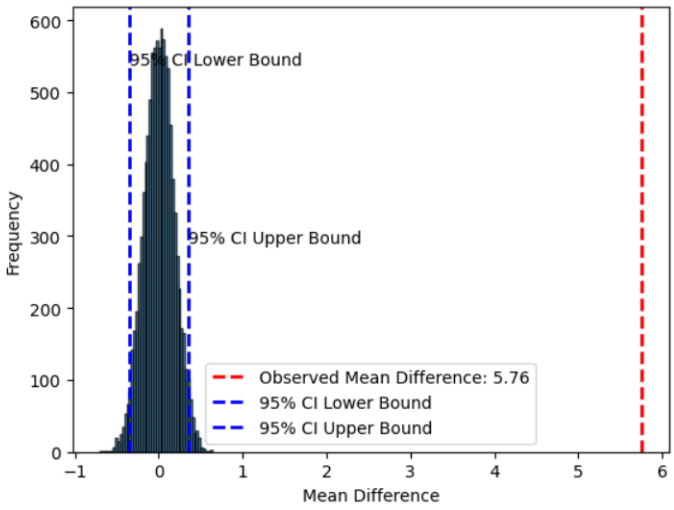

Figure 29: Bootstrap Distribution of Mean Differences of Fish School Intensity

5.2.4 Distribution Comparison by Sea Depth

When categorizing sea depth into intervals such as [0,20], [20,40], [40,60], [60,80], and [80+], the Double Threshold method detected fish schools in these respective sea depth categories: 2, 257, 574, 2405, and 14. The proportions of detected fish schools in each sea depth level are [0.06%, 7.9%, 17.65%, 73.95%, and 0.43%].

Similarly, the SAM method detected fish schools in the corresponding sea depth categories: 20, 168, 549, 2628, and 5. The proportions of detected fish schools for SAM in each sea depth level are [0.59%, 4.99%, 16.29%, 77.98%, and 0.15%]. See in figure 30.

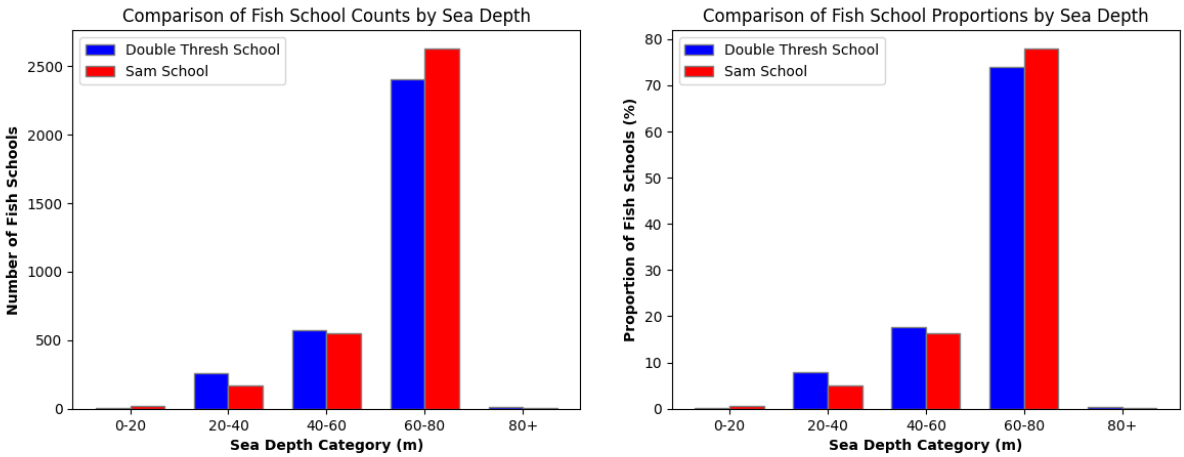

Figure 30: Visualization of fish school detections by sea depth

5.2.5 Chi-Square Test for Comparing Fish School Distributions Across Sea Depth Levels (Exclusive Detections by SAM and Double Threshold)

From Figure 30, we observe that the proportion of fish schools detected by the double threshold and by SAM does not show a significant difference. Therefore, we are interested in determining if the detection method is independant with the different sea depth levels.

As mentioned at the beginning of this chapter, both methods detected a substantial number of common images and fish schools. So to satisfy the assumptions required for the Chi-square test, we focus exclusively on the group of fish schools detected only by the double threshold method and the group of fish schools detected only by the SAM method. This ensures that each fish school is counted only once in our test. All the principe of Chi-Square used in this analysis are based on the Statistics with Python course materials by Nicolas Poulin [12].

1. **The hypotheses are:**
   - Null Hypothesis (H_0): The detection methods are independent of sea depth levels.
   - Alternative Hypothesis (H_1): The detection methods are not independent of sea depth levels.
2. **Observed Contingency Table:** The analysis is based on a contingency table that presents the cross-occurrences of two categorical variables: detection methods and sea depth levels. The observed contingency table is shown in Table 3.

	<20m	20-40m	40-60m	60-80m	>80m
Double Threshold	2	120	191	576	4
SAM	20	50	194	977	1

Table 3: Observed Contingency Table showing the number of fish schools detected by each method across different sea depth levels: <20m, 20-40m, 40-60m, 60-80m, and >80m.

3. **Assumptions:**
   - The individuals in the sample have been chosen randomly (i.e., the observations are independent).
   - The categories of the variables are mutually exclusive.
   - Cochran's Rule: At least 80% of the expected frequencies should be at least 5.
   - The sample size must be sufficiently large[12].

4. **Validation of the Assumptions:**

   (a) **Independence:** Generally, fish schools are independent; the detection of one fish school does not influence the detection of another.

   (b) **Mutually Exclusive Categories:** All fish schools in our counts are those detected exclusively by one method—either the double threshold or the SAM method—to ensure that each fish school is counted only once.

   (c) **Cochran's Rule:** Refer to Table 4. At least 80% of the expected frequencies are greater than or equal to 5.

	<20m	20-40m	40-60m	60-80m	>80m
Double Threshold	9.2	71.1	161.0	649.6	2.1
SAM	12.8	98.9	224.0	903.4	2.9

Table 4: Expected Frequency Table showing the expected frequencies of fish schools detected exclusively by the double threshold method and the SAM method across different sea depth levels.

   (d) **Sufficient Sample Size:** The sample size is sufficient, with a total of 893 samples detected by the double threshold method and 1242 samples detected by the SAM method [12].

5. **Decision Statistic:**

$$\chi^2_{\text{obs}} = \sum_{i=1}^{k} \sum_{j=1}^{c} \frac{(n_{ij} - t_{ij})^2}{t_{ij}} [12]$$

   Distribution Under H_0 (Null Hypothesis): The test statistic χ^2 follows a chi-square distribution with degrees of freedom $\nu = (k-1)(c-1)$. In this case, with $k = 2$ (the number of categories for detection methods) and $c = 5$ (the number of categories for sea depth levels), the degrees of freedom are calculated as $\nu = (2-1) \times (5-1) = 4$.

   Decision Making: The decision is based on the p-value obtained from the chi-square distribution. If the p-value is below the significance level ($\alpha = 0.05$), we reject the null hypothesis, indicating that there is a significant association between the detection method and sea depth levels. If the p-value is above the significance level, we fail to reject the null hypothesis, suggesting no significant association. [12].

6. **Chi-Square Test Results and Conclusion:** All assumptions for the chi-square test are satisfied. The test statistic calculated is 94.3889, and the p-value is 1.54×10^{-19}. Since the p-value is less than the significance level α, we reject the null hypothesis. This indicates that there is a significant association between the detection methods and the sea depth levels, meaning the detection methods are not independent of the sea depth levels.

5.3 Time-Based Segmentation Comparison

To see more about the distribution of the fish schools, we consider to divide the time when detected the fish schools by 4 different time interval, like the day time and the night time, and the time interval half un hour before the sunrise to half hour after the sunrise, and the time interval half un hour before the sunset to half hour after the sunset. And in total double threshold detect 614 fish schools in night, 108 fish schools around sunrise, 2464 fish schools in the daytime, and 66 fish schools around the sunset.in total SAM detect 747 fish schools in night, 86 fish schools around sunrise, 2435 fish schools in the daytime, and 102 fish schools around the sunset. See in figure 31.

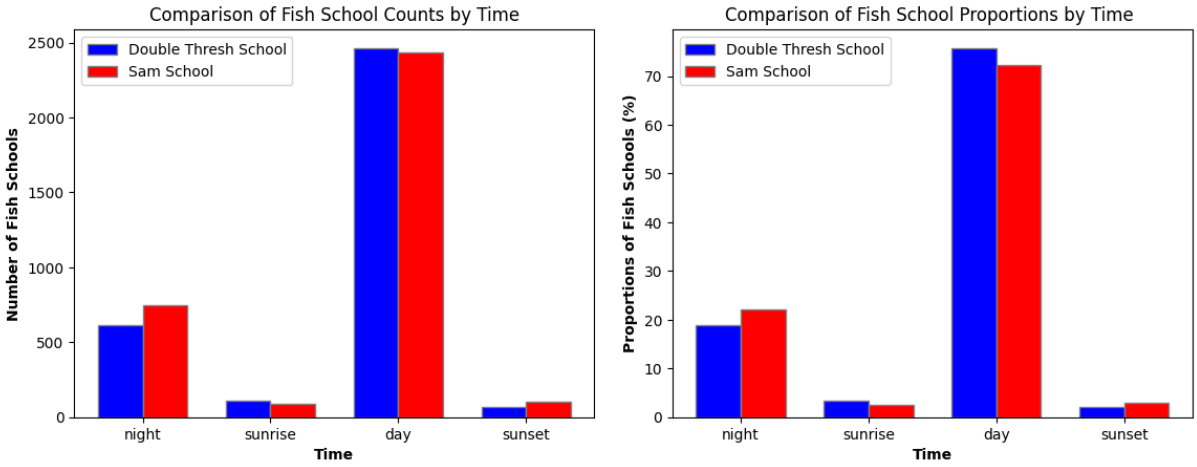

Figure 31: Distributions of Fish Schools by Different Time

We aim to determine whether the detection method is independent of the time at which fish schools are detected by the echosounder. The time categories considered are: night, sunrise, day, and sunset. We applied the chi-square test to assess the independence between these two qualitative variables, following the exact steps described in Section 5.2.5, thus, we will not provide further details here. After validating the assumptions, the chi-square test produced a p-value of 6.06×10^{-5}, which is significantly less than the significance level α. Therefore, we conclude that the detection method is dependent on the timing of fish school detection by the echosounder.

5.4 Comparison of Various Characteristics

Here, we present several important characteristics tests conducted between the groups of fish schools detected only by the double threshold method (noted as *thresh_only_school*) and those detected only by the SAM method (noted as *sam_only_school*). The goal is to identify key detection differences by comparing these relevant variables. Here are the explications of these variables:

- **sam_only_school**: Fish school intensity
- **size_school**: Fish school size by pixels
- **depth_school**: Fish school depth in the sea
- **school_seabed_distance**: The distance between the fish school and the seabed
- **dif_i_school_image**: The difference between school intensity and image intensity
- **gradient_school**: The average gradient in the fish school intensity
- **solidity**: The solidity of the fish school region
- **compactness**: Calculated by $C = \frac{4\pi \cdot \text{Area}}{\text{Perimeter}^2}$, where Area is the area of the object and Perimeter is the perimeter of the object.
- **axis_ellipse_ratio**: Calculated by $\frac{\text{region.axis\_minor\_length}}{\text{region.axis\_major\_length}}$
- **inertia_tensor_eigvals_ratio**: Calculated by $\frac{\text{region.inertia\_tensor\_eigvals[1]}}{\text{region.inertia\_tensor\_eigvals[0]}}$, where region.inertia_tensor_eigvals[1] is the smaller eigenvalue in the region.
- **perimeter_area_ratio**: Calculated by $\frac{\text{region.perimeter}}{\text{region.area}}$

Variable	thresh_only_school	sam_only_school	Statistical Test	P-value	Decision
intensity_school	-70.91	-79.32	Welch's t-test	3.66e-161	Reject
size_school	273	1612	Welch's t-test	4.72e-13	Reject
depth_school	39.19	42.28	Welch's t-test	0.000768	Reject
school_seabed_distance	20.71	21.62	Welch's t-test	0.29	Accept
dif_i_school_image	10.89	2.78	Welch's t-test	3.83e-198	Reject
gradient_school	319.49	316.34	Welch's t-test	0.72	Accept
solidity	0.84	0.92	Welch's t-test	7.42e-61	Reject
compactness	0.78	0.75	Welch's t-test	0.63	Accept
axis_ellipse_ratio	0.31	0.42	Welch's t-test	1.32e-25	Reject
width_length_ratio	0.39	0.60	Welch's t-test	1.21e-28	Reject
inertia_tensor_eigvals_ratio	0.15	0.24	Welch's t-test	4.19e-21	Reject
perimeter_area_ratio	0.65	0.35	Welch's t-test	4.40e-199	Reject

Table 5: Comparison of Various Characteristics Between Double Threshold and SAM Methods. The null hypothesis (H_0) in all tests is that the mean value in the Double Threshold group is equal to that in the SAM group, and the alternative hypothesis (H_1) is that the means are not equal. Given the large sample sizes, the Central Limit Theorem justifies the use of parametric tests, with the choice between Student's t-test and Welch's t-test depending on the equality of variances.

6 Conclusion

Detecting fish schools in acoustic images is crucial for effective fisheries management and understanding marine ecological dynamics. Accurate detection of fish schools helps in assessing their distribution and supports sustainable fisheries management. The primary challenge is to reliably identify and distinguish fish schools from noise and other artifacts present in acoustic images.

In this study, we explored two distinct methods for fish school detection: the double threshold method and a deep learning approach. Prior to applying these methods, we conducted several preprocessing steps on the acoustic images. These steps included resizing images to ensure a consistent temporal scale, masking out wave and bottom artifacts to eliminate their influence, and enhancing image contrast to improve the visibility of fish schools against the background.

For the double threshold method, we determined two thresholds for each image. These thresholds were used to create two thresholded images—a seed image and a mask image. The final image, constructed from these two thresholded images, highlighted the regions representing fish schools. This method relies heavily on accurate image preprocessing and threshold selection. Given the variability in quality and clarity of acoustic images, precise preprocessing and thresholding are crucial for accurate fish school detection.

In contrast, the deep learning approach requires extensive training of data. Due to the absence of annotated data, the Segment Anything (SAM) model was applied directly to segment the acoustic images. SAM provided automated segmentation results, from which we selected the most relevant segments representing fish schools. While SAM benefits from extensive training on diverse datasets and is highly adaptable for general object detection tasks, its performance on acoustic images was occasionally limited due to its lack of specific training on such images.

We provide a detailed comparison of the two methods based on the detection results from several acoustic images:

- **Detection Quantity and Quality:**

  The double threshold method detected 1,664 images with fish schools, whereas SAM detected 1,874 images, representing a 12.6% increase compared to the double threshold method. In terms of the number of fish schools detected, the double threshold method identified 3,252 fish schools, while SAM detected 3,370, which is 3.6% more.

  However, due to the lack of annotations, we do not have enough data to compare the prediction accuracy, precision, or sensitivity of these two methods

- **Prediction Ability with the First 999 Images**

To assess the predictive capability of the two methods, we manually inspected the first 999 acoustic images to verify the presence of fish schools. We then compared these manual inspections with the results of the two detection methods. In this comparison, we focus solely on whether the methods could detect the presence of fish schools in the images, rather than on the accuracy of fish school position extraction. It is important to note that we are using the count of images where the double threshold method detected the presence of fish schools, not the count of images where the method successfully extracted fish school positions.

		Observed	
		0	1
True	0	327	23
	1	300	349

Table 6: Double Threshold Detection

		Observed	
		0	1
True	0	299	53
	1	280	367

Table 7: SAM Detection

	0	**SAM=1 DT=0**	**DT=1 SAM=0**	**1**
0	279	48	18	5
1	190	110	92	257

Table 8: Double Threshold & SAM Detection. "0" indicates no fish school, "1" indicates with fish school.

From Table 6, we find that for the double threshold method, the false positive rate is 6% and the false negative rate is 46%. From Table 7, we find that for the SAM method, the false positive rate is 15% and the false negative rate is 43%. If we combine the results of both methods—considering a fish school as detected only if both methods agree on its presence, and as not detected if both methods do not detect it. Table 8 shows that we can significantly reduce both false positives and false negatives. In this case, the false positive rate is 1% and the false negative rate is 29%. This means that by combining the two methods, if there is no fish school, there is only a 1% chance of wrongly detecting a fish school, and if there is a fish school, there is a 29% chance of wrongly detecting that there is no fish school.

- **Statistical Comparison**

  Welch's test was used to compare the variables between the groups *thresh_only_school* and *sam_only_school*. Significant differences were observed in variables such as fish school intensity, size, and depth, indicating that the two methods differ significantly in their detection of these attributes. Additionally, the chi-square test conducted earlier showed that the detection methods are dependent on both the sea depth levels and the timing of fish school detection by the echosounder.

- **Efficiency Comparison:**

  In terms of computational efficiency, SAM took significantly longer to detect fish schools, with an average processing time of 7 minutes per image on a personal computer, compared to just 10 seconds for the double threshold method. Therefore, if computation time is a constraint, the double threshold method is highly advantageous.

- **Large Fish Schools (Layers of Fish Schools)**

  Regarding fish school size, the double threshold method detected moderate-sized fish schools, while SAM tended to detect larger fish schools, especially in cases with a large area or layer of strong signals. As shown in the first image in Figure 32, SAM detected a much larger area than the actual fish school, whereas the double threshold method provided a closer estimate of the fish school size. Conversely, in cases with a large layer of fish schools, the double threshold method sometimes failed to detect the presence of fish schools, whereas SAM could identify them, albeit often overestimating their size. SAM seems more sensitive to intensity changes across layers in images, leading to larger detections in such cases. The double threshold method, on the other hand, detected some of these images but with more accurate sizes.

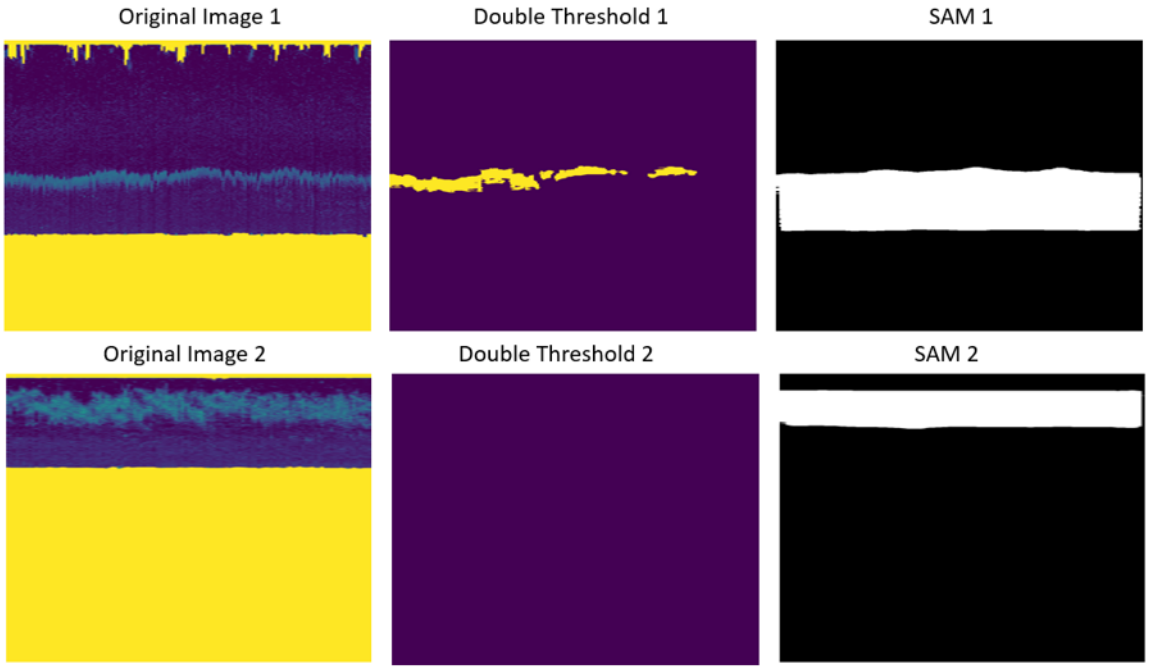

Figure 32: Detection Comparison of Large Fish Schools

- **Small or Narrow Fish Schools**

  When the fish school is small, SAM may have difficulties detecting it, as shown in Figure 33. In these two images, SAM failed to detect the fish school, while the double threshold method succeeded.

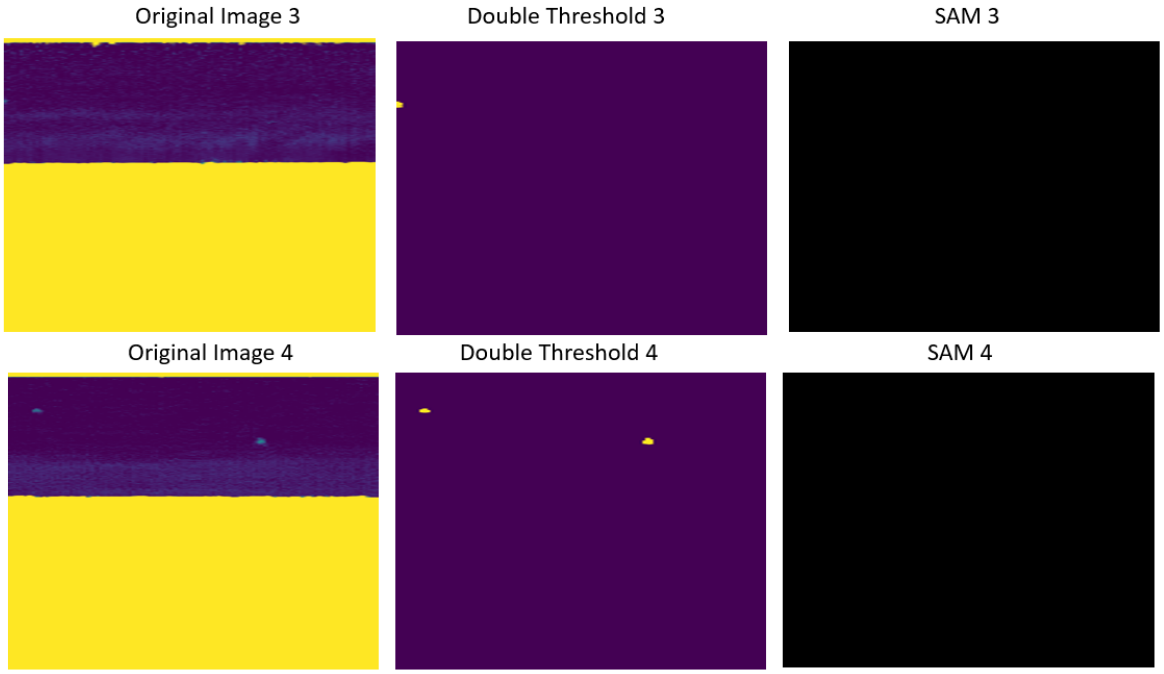

Figure 33: Detection Comparison of Small Fish Schools

- **Similar Detection**

For many images, both methods produced similar detections, though there were minor differences in size, shape, or properties when compared using image metrics such as compactness and solidity. See Figure 34.

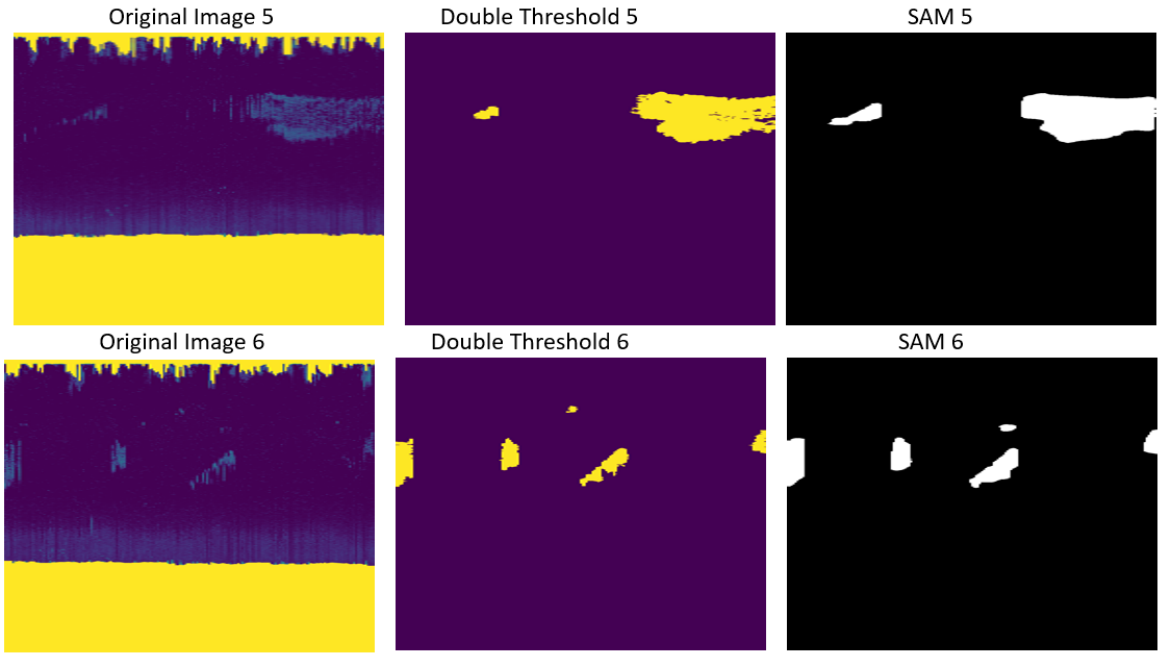

Figure 34: Detection Comparison of Similar Fish Schools

Overall, while the double threshold method requires meticulous image preprocessing and parameter tuning, it offers a straightforward and computationally efficient approach with accurate detection. On the other hand, the SAM model, known for its robust segmentation capabilities, demonstrates good accuracy but may require fine-tuning for specific applications such as acoustic image analysis. Although there are performance differences between these two methods, both have their strengths and limitations. Combining them significantly enhances the overall accuracy of fish school detection, which is a notable finding. However, even with the combined approach, the false negative rate remains somewhat high. This indicates that, despite the presence of a fish school, there is still a considerable probability of incorrectly detecting the absence of one.

To address this issue, further improvements can be made to refine our predictions and reduce the false negative rate. For instance, our analysis of corrected images revealed that fish schools close to the seafloor may be obscured by the mask used for processing. This overmasking could be a significant factor contributing to the increased false negative rate. To mitigate this problem, we can explore modifying the shape and size of the structuring element used in the mask. For example, reducing the bottom dilation size from 20 to 10 pixels or changing the shape of the structuring element from a square to other shapes such as rectangles could improve detection. Additionally, optimizing the segmentation algorithm parameters, particularly for large segmentations, may enhance prediction accuracy by refining the delineation of significant areas.

7 Acknowledgments

I would like to express my heartfelt gratitude to my supervisors for their invaluable guidance and support throughout my internship.

Firstly, I am deeply thankful to Claire Saraux for providing me with this crucial internship opportunity. The project, which involved a diverse range of techniques, has been both fascinating and highly educational. Claire's patience, meticulous scientific approach, and precise guidance have played a pivotal role in my development. Her clear vision for the project and exceptional research abilities have left a lasting impression

on me. Claire's sense of responsibility towards my progress and success was evident throughout the internship. She consistently ensured that I had the necessary support and time to complete my work and reports. Her careful planning and thoughtful consideration of my needs made a significant difference in my ability to meet deadlines and produce quality work. Claire's dedication to mentoring and commitment to high scientific standards have been a constant inspiration.

I am also profoundly grateful to Lilia Guillet for her kindness and patience, especially during the initial phase of my internship. She played a crucial role in setting up the necessary applications and explaining project intricacies. Despite my limited proficiency in French, Lilia ensured I understood everything clearly, often simplifying explanations for me. Her support extended beyond technical assistance; she guided me through project challenges, ensuring I felt confident and comfortable in my tasks. After each weekly project meeting, she patiently helped me outline tasks for the following week, keeping me on track. I feel fortunate to have been part of her team during my internship.

My sincere thanks also go to Céline Meillier, who listened to me attentively and offered effective solutions whenever I faced obstacles. Her deep involvement in our project was evident, and she consistently provided valuable suggestions. Céline's approach to image correction was very effective and a significant step forward for our double threshold method. Her problem-solving skills, blending creativity with practicality, significantly advanced my work. I have consistently been impressed by her ideas, finding them both remarkable and impactful. Her insightful ideas and constructive feedback guided the project's direction, providing continuous motivation.

I am equally grateful to Benoît Naegel, whose innovative ideas greatly contributed to our project. His elegant proposal of the double threshold method and the Segment Anything model were particularly impressive. Benoît's ability to think creatively and propose novel solutions has been truly inspiring. Though time constraints limited my exploration of the Segment Anything model, I am keenly interested in its potential applications.

I would also like to extend my thanks to IPHC (Institut pluridisciplinaire Hubert Curien) for providing an excellent working environment. Special thanks to DEPE (Ecology, physiology and ethology) where I worked during my internship. Their warm welcome and support have been deeply appreciated. Additionally, I would like to thank Lauriane Kuhn from IPHC's STI Service Technique Informatique for her invaluable informatics support in ensuring I could run the Segment Anything Model whenever needed.

Lastly, I extend my gratitude to all my professors in the Statistics Department at the University of Strasbourg. They are dedicated educators who invested considerable time in teaching and guiding us. Their high-quality courses and unwavering support for our learning and personal growth have been invaluable. Their readiness to assist whenever I had questions has been truly encouraging. The rigorous training and comprehensive curriculum provided me with a solid foundation in statistical methods, proving essential during my internship. I am proud to be pursuing a Master's degree in Statistics at the University of Strasbourg.

Thank you all for making my internship a rewarding and memorable experience. I am profoundly grateful for your guidance, support, and the knowledge you have imparted to me. This internship has marked a significant milestone in my academic journey, and I eagerly anticipate applying what I have learned in my future endeavors.

References

[1] Scikit packages. `https://scikit-image.org/docs/stable/api/api.html`. Accessed: 2024-07-23.

[2] Meta AI. Segment anything homepage. `https://segment-anything.com/`. Accessed: 2024-07-23.

[3] Charlotte Boyd, Daniel Grünbaum, George L Hunt Jr, André E Punt, Henri Weimerskirch, and Sophie Bertrand. Effects of variation in the abundance and distribution of prey on the foraging success of central place foragers. *Journal of Applied Ecology*, 54(5):1362–1372, 2017.

[4] Olav Brautaset, Anders Ueland Waldeland, Espen Johnsen, Ketil Malde, Line Eikvil, Arnt-Børre Salberg, and Nils Olav Handegard. Acoustic classification in multifrequency echosounder data using deep convolutional neural networks. *ICES Journal of Marine Science*, 77(4):1391–1400, 2020.

[5] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer, 1992.

[6] Niall G Fallon, Sophie Fielding, and Paul G Fernandes. Classification of southern ocean krill and icefish echoes using random forests. *ICES Journal of Marine Science*, 73(8):1998–2008, 2016.

[7] ORIANS GH. On the theory of central place foraging. *Analysis of ecological systems*, pages 157–177, 1979.

[8] Davide Giraudo. Statistical tools s3. Course materials, 2023. Accessed: 2024-07-23.

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[10] Andraz Krzisnik. How does geodesic dilation and erosion work? `https://epochabuse.com/geodesic-dilation-and-erosion/`. Accessed: 2024-07-23.

[11] Tunai Porto Marques, Melissa Cote, Alireza Rezvanifar, Alexandra Branzan Albu, Kaan Ersahin, Todd Mudge, and Stéphane Gauthier. Instance segmentation-based identification of pelagic species in acoustic backscatter data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4378–4387, 2021.

[12] Nicolas Poulin. Statistics with python slides. PowerPoint presentation, 2022. Accessed: 2024-07-23.

[13] R Priyadharsini and T Sree Sharmila. Object detection in underwater acoustic images using edge based segmentation method. *Procedia Computer Science*, 165:759–765, 2019.

[14] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[15] Hugo Robotham, Paul Bosch, Juan Carlos Gutiérrez-Estrada, Jorge Castillo, and Inmaculada Pulido-Calvo. Acoustic identification of small pelagic fish species in chile using support vector machines and neural networks. *Fisheries Research*, 102(1-2):115–122, 2010.

[16] Christian Ronse. Double threshold. `https://dpt-info.di.unistra.fr/~cronse/TIDOC/FEAT/`. Accessed: 2024-07-23.

[17] Claire Saraux and Lilia Guillet. Deep learning image analysis: automatic detection and classification of schools in acoustic data. 2023.

[18] John Simmonds and David N MacLennan. *Fisheries acoustics: theory and practice*. John Wiley & Sons, 2008.

[19] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.