Analyzing Ocean Boundary Phenomena in Echograms: A Deep Learning Approach

by

Femina Bharatkumar Senjaliya

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF APPLIED SCIENCE

in the Department of Electrical and Computer Engineering

Analyzing Ocean Boundary Phenomena in Echograms: A Deep Learning Approach

by

Femina Bharatkumar Senjaliya
B.Eng., Gujarat Technological University (Ahmedabad, India), 2021

Supervisory Committee

Dr. Alexandra Branzan Albu, Supervisor
(Department of Electrical and Computer Engineering)

Dr. Tunai Porto Marques, Departmental Member
(Department of Electrical and Computer Engineering)

ABSTRACT

This research puts emphasis on the fundamental part of marine monitoring as an instrument to study how the oceans influence the global climate, biodiversity and ecological systems under the condition of the Arctic region. Utilizing underwater active acoustic surveys conducted with moored multi-frequency echosounders as our source gives us the opportunity to reflect on the complexity of ocean settings. We propose a deep-learning approach to automate the identification of sea surface boundaries and near-surface phenomena in echograms to assist oceanographers who currently rely heavily on the time-consuming manual analyses. The identification of boundaries at the surface and the occurrence of bubble phenomena are vital to those who investigate marine environments. These factors greatly affect the complex interactions between organisms. We propose a two-step, end-to-end, deep learning approach where the first step uses an image classification framework to categorize echograms based on surface conditions and is followed by the second step where we employ semantic segmentation frameworks that help to delineate sea surface and near-surface bubbles within the water column. This segmentation in the second step is equipped with a type-specific model that has been proven to outperform a single global segmentation model. Furthermore, our methodology incorporates innovative learning strategies, including a tailored boundary loss function, to enhance model performance. Through comprehensive testing with a range of image classification and semantic segmentation architectures, we identify the most effective models for Arctic echogram analysis. Our proposed deep learning pipeline showcases noteworthy capabilities in accurately characterizing and analyzing marine acoustic data.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

Funding from NSERC Canada and ASL Environmental Sciences, partially supported by the Digital Research Alliance of Canada and through the Alliance-Mitacs Grants program, made this work possible.

I was fortunate to spend time with some truly remarkable people on this incredible tour. I am really grateful to Dr. Alexandra Branzan Albu for all of her advice and support as I go forward in both my personal and academic lives. Her compassion, wisdom, and direction have greatly impacted both my studies and personal development. A particular thank you to our lab's research associate, Dr. Mélissa Côté, for her counsel, mentoring, and direction on this project. Simply said, without her assistance, this effort would not have been feasible.

I would like to thank Dr. Stéphane Gauthier for the data insights and Andrea Niemi for the data. I'm also thankful to Dr. Kaan Ersahin, Dr. Steve Pearce, Dr. Julek Chawarski, Keath Borg, and Amanda Dash for their great guidance and assistance.

A special thanks to my friends and colleagues at the computer vision lab at the University of Victoria.

I am incredibly appreciative of my family, whose love and support have enabled me to accomplish this great achievement. My friends Shalini Shah and Meet Diwan have been a continual source of love and support for me on this journey.

Chapter 1

Introduction

1.1 Oceanographic Research and Echograms

Oceans play a significant role in biodiversity and ecological systems. Through active acoustic surveys, the underwater environment can be studied. The information collected from these acoustic surveys is visualized in the form of echograms that give accurate acoustic maps of sea ecosystems, thus describing complex ecological dynamics underwater. Understanding the role of the ocean in shaping climate, biodiversity, and ecological dynamics could be done through echogram analysis.

Proper identification and analysis of marine events is necessary for comprehensive oceanographic studies, and this requires an understanding of acoustic reflections. The measurement of acoustic backscatter inside the water column yields acoustic maps. The traditional method for gathering this backscatter data is by using multi-frequency echosounders (such as the Acoustic Zooplankton and Fish Profiler (AZFP) from ASL Environmental Sciences [12]), which emit a series of acoustic pulses (pings) at different frequencies and listen for echoes from potential targets to generate visualizations of the water column in a minimally invasive manner. In order to analyze the sea surface, sea floor, and other regions of the water column, they can be fixed to the sea floor and pointed upward, or they can be fastened to vehicles and platforms pointing downwards to collect reflected acoustic echoes. This technique relies on the principle that objects and boundaries in the water, like biology, bubble layers, seabeds, and sea surfaces, reflect sound differently because of their acoustic characteristics, as explained by Lurton et al. [13]. Biological phenomena such as the presence of zooplankton and fish schools significantly influence the pattern and intensity observed in acoustic backscatter on echograms. Moreover, the surface border of

the ocean and subsurface bubble formations are examples of physical events that impact the interpretation of acoustic backscatter in echograms. These events influence the nature of the scattered sound waves due to their unique acoustic impedance in comparison to the surrounding water.

The data from moored multi-frequency single-beam echosounders are typically visualized as echograms. As Vohra et al. discuss in their study [14], echograms are presented as sets of single-frequency 2D images. These images capture different pings over time, with the x-axis representing temporal units for a given frequency and the y-axis depicting the depth or range of the instrument in the water column. The intensity of each pixel in the echograms is color-coded to reflect the amplitude of the echoes at that particular frequency and time, typically calculated from the volume backscattering strength (S_v), providing a visual map of the underwater scatterers.

Transforming S_v values, which in our case normally range from approximately -125 to 0 dB, into the vivid colors of the "Jet" color map improves the presentation of echograms significantly. The 125 kHz echogram sample provided in Figure 1.1 serves as an example of this. The red band at the top displays the air-water interface in vivid hues, while the lower boundary of the red band represents the sea surface boundary. Large fluctuations in chroma and brightness, which are essential for differentiating between various underwater elements, are particularly effectively displayed by the whole visual spectrum of the 'Jet' color map in this context. Biological entities, as shown in Figure 1.1, appear as sporadic cyan-blue and occasionally yellow patches and spots throughout the water column, their presence is emphasized against the darker background. The contrasting hues emphasize the near-surface air bubbles' proximity to the sea surface boundary, which are shown as prominent dark yellow clusters. The color map is a useful tool for evaluating acoustic data related to maritime environments because of its gradient, which clearly delineates these elements from each other. More aquatic organisms are indicated by denser light blue groupings. This echogram provides a visual reference for the details of interpreting acoustic data as well as the makeup of the marine environment.

Oceanographers and biologists traditionally analyze echograms using manual or semi-automatic methods with software like Echoview [15], focusing on the statistics of organism aggregations [16, 14]. These methods are time-consuming and error-prone, underscoring the need for more efficient, automatic or semi-automatic analysis techniques. Such advancements are vital not just for species abundance but also for identifying marine features like surface boundaries and bubble phenomena. The shift towards automatic processing aims to improve the accuracy and efficiency of interpreting complex marine ecosystem

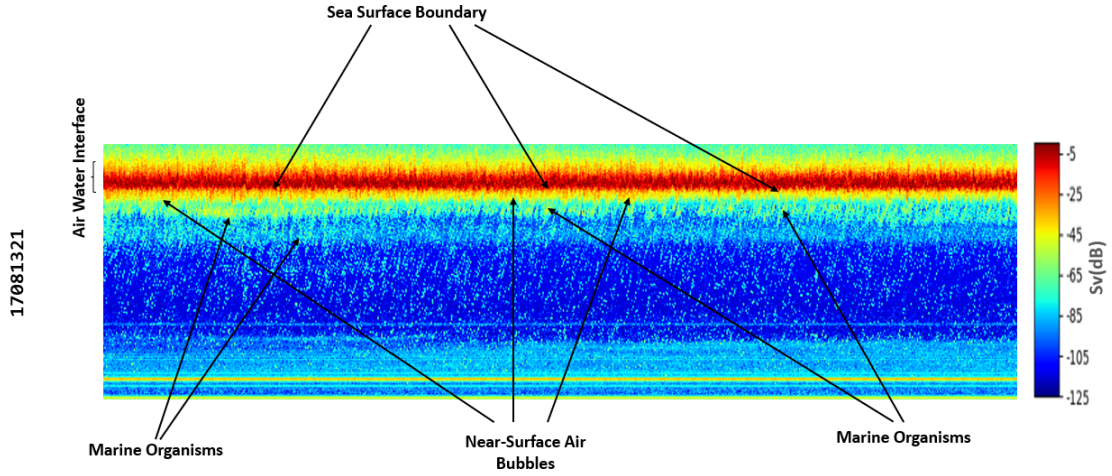

Figure 1.1: *The 125kHz echogram from August 13, 2017, 9-10 pm, under windy, ice-free conditions, uses the 'Jet' colormap to depict marine features. The bright red top band represents the air-water interface, with the sea surface boundary just below. Marine life is shown in blue and cyan blue clusters throughout the water column, contrasting with the background. Near-surface air bubbles appear as yellowish clusters near the boundary. This echogram encodes S_v values (-125 to 0 dB), visually differentiating between various marine elements and aiding in environmental analysis.*

features.

1.2 Deep Learning-Based Techniques in Marine Ecosystem Analysis

In particular, the interpretation and application of echograms have undergone a radical shift as a result of the development of deep learning (DL) in marine environment studies. Significant progress in our understanding of marine ecosystems has been made thanks to its enhanced capacity for analyzing acoustic data.

Research by Marques et al. [17] and Rezvanifar et al. [18] demonstrate its efficacy in semantic segmentation of marine species and automatic fish school detection, respectively. Using developments such as the "Echofilter" by Lowe et al. [19], the technology can automate the analysis of echosounder data. Brautaset et al. have demonstrated how deep convolutional neural networks (CNN) can be used to improve acoustic signal processing for underwater analysis [8]. As compared to conventional techniques of analysis, tools such as CNN-based methodologies for bottom correction labeling and ECOPAMPA, investigated by Sarr et al. [20], show how effective deep learning is in marine acoustics.

In addition, these advanced methodologies allow for gaining detailed and accurate insights into ocean settings as showcased by studies that include semantic segmentation, instance segmentation, and tasks involving multifrequency echograms [21, 8, 14, 22, 17, 23].

We present a systematical review in the chapter 2 on efforts aimed at improving marine ecosystem assessment.

1.3 Thesis Objectives and Significance

1.3.1 Objectives

This research highlights the significance of studying physical phenomena in marine environments, such as bubbles [24] and turbulence [25], which have received attention in prior studies. We present a method to refine the analysis of near-surface ocean phenomena, which includes bubbles and boundary types, through multi-frequency echograms. Traditional methods, heavily manual and informed by direct observations [26], are slow and prone to bias [27, 28]. To overcome these issues, our method incorporates deep learning techniques for automatic annotation. This advancement offers a heightened precision and speed of ocean boundaries and events analysis that are hindered by manual procedures.

Overall, our research focuses on analyzing near-surface phenomena, offering new software tools to oceanographers. The focus on automation addresses the pressing need for advanced processing capabilities in marine data analysis and aims to reduce the dependence on labor-intensive, error-prone manual methods.

1.3.2 The Significance of Identifying Boundaries and Bubbles

Classification of sea surface types in echogram records useful for automatic fish species detection. The correct classification enables quality species monitoring and enhances the understanding of the ecological dynamics. Consequently, it preserves biological data that could also be missing or not clear due to surface bubble-noise layer disruptions. This method also aids in resolving the issues associated with a class imbalance in echogram data, wherein background and sea surface elements are frequently more common than fish and bubble classes, which are comparatively uncommon occurrences. Sea surface type classification that is accurate and efficient makes oceanographic research more manageable and economical, freeing up scientists to concentrate more on data interpretation. The identification and classification of the sea surface are essential for tracking ice conditions and

verifying the retrieval of ice parameters in Arctic studies.

Another task of interest in echogram analysis is near-surface bubble detection, which is necessary to differentiate bubbles from biological species like juvenile salmon and krill. Because bubbles frequently resemble juvenile salmon and krill aggregations in specific echogram channels, this separation prevents errors in school detection and biomass calculations. The validity of marine biological research depends on the accurate identification of these characteristics. The impact of global warming on multiyear and first-year ice is making near-surface bubble detection in Arctic sea ice research more significant [29]. Additionally, the detection of bubble clouds in the ocean is significant because of their substantial acoustic target strength and their role in air-sea interaction processes [24].

1.3.3 Contributions

The work conducted in this research presents novel techniques and methodologies that improve the precision and effectiveness of analysing near-surface physical phenomena. Highlighted below are the key contributions of this research:

1. **Two-step classification and segmentation process:** We present an efficient two-step classification and segmentation process. In the first step, a classification model categorizes every echogram into one of six surface types. In the second step, a dedicated segmentation model for each surface type trained on class-specific data is employed to segment the echogram into bubbles, surface, and background. This two-step process turns out to be more effective than one single global segmentation model.

2. **Optimized model for sea surface type classification:** We employed an image classification model to classify echograms based on surface types, comparing its performance against many popular DL image classification models. The model selected was identified as best performing for our data.

3. **Proposed methodology for sea surface and near-surface bubbles detection:** We explored and compared many image segmentation models to find the sea surface boundary and near-surface bubbles on a pixel-level basis and identified the most effective model. In the process of learning echogram segmentation, we advocate three strategies.

In addition to the contributions highlighted above, the research provides new ways of data interpretation and analysis in biological and physical conclusions about marine ecosystems thereby enabling specialists to draw precise findings. Moreover, the process of first

classifying and then segmenting images has broader implications beyond marine research. Similar methods may be helpful in areas like medical imaging where initial classification could detect possible anomalies while subsequent anomaly-specific segmentations can map out exact pathological areas for diagnosis. In agricultural monitoring for instance, using this approach we could first classify different crop types or health statuses from aerial imagery and then engage more detailed segmentation for targeted intervention. Furthermore, urban planning could use this method to classify land use types in large-scale satellite images before segmenting specific areas for developmental analysis. These broader applications highlight the versatility and potential of the two-step process developed in this thesis, suggesting its utility in a range of other application domains requiring nuanced and accurate data interpretation.

1.4 Publications

The contributions of this study are published in the paper titled "***Deep Learning-Based Identification of Arctic Ocean Boundaries and Near-Surface Phenomena in Underwater Echograms***" in the Perception Beyond the Visible Spectrum Workshop (IEEE PBVS) at the 2024 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Chapter 2

Literature Review

The first section of the review, Section 2.1, explores the fundamentals of echogram analysis research and provides a background for the techniques and difficulties that are specific to this discipline. In Section 2.2, we turn our attention to machine learning (ML) and DL techniques used to analyze near-surface events and ocean boundaries. Finally, Section 2.3 offers a detailed of the approaches used in the following chapters.

2.1 Methods for Analyzing Echograms

Echogram analysis in marine ecosystem studies started from analog methods and evolved to include ML and DL approaches. In the beginning, the field used analog tape recordings and basic hardware, dated back to the 1980s, such as mainframe computers with software written in Fortran that needed data input via reel-to-reel tapes and punch cards for software feeding into the computer [30]. Subsequently, a shift towards digital systems occurred [30]. Historically, manual or semi-automated techniques using software like Echoview [15] have been used to analyze echograms, enabling marine researchers to identify and extract different features and information from echogram data. The three primary categories of these features are bathymetric (connected to the location of organisms in the water column), morphometric (pertaining to attributes like the height, width, form, or perimeter of organism aggregations), and energetic (relating to backscatter signal properties) [31]. Auxiliary external data such as moon phases, time of day, or year may also be taken into consideration. Multifrequency techniques take advantage of backscatter variations. An extensive analysis of multifrequency target classification techniques are based on ML as well as conventional techniques [32]. For detection tasks in echograms, DL techniques have garnered atten-

tion recently [17]. An overview of classical multifrequency approaches, traditional ML methods, and recent DL methods in echogram analysis are provided in this section.

2.1.1 Multifrequency Methods

The fisheries acoustics community has been using multifrequency acoustic data since the late 1970s, especially for zooplankton identification [33]. Throughout the 2000s, multifrequency approaches to backscatter categorization became more common due to advancements in apparatus, techniques, and software enabling multifrequency acoustic measurement using numerous narrow-band echosounders [34]. The "conventional" multifrequency approaches often rely on the frequency response, either relative, differential, or combined, of various species. The relative frequency response idea $r(f)$ was first introduced by Korneliussen and Ona [33]. It compares the volume backscattering coefficient S_v at a given frequency to a reference frequency, usually 38 kHz. Focusing on the frequency-dependent backscattering properties, this characterization approach is especially useful in mixed layers of fish and small scattering creatures such as Euphausiids. Additionally, in an extension of work mentioned in [33], the authors presented development of the Large Scale Survey System (LSSS) software in [35], a system for analyzing and displaying multi-frequency acoustic data that improves performance during large-scale operations in their previous works. Through the use of multi-frequency information for species discrimination, this method facilitates the efficient post-processing of sonar and echosounder data, enabling manual echogram interpretation, classification, and echo integration for biomass calculation.

One significant contribution has been the creation of software such as Echoview [15], Echoview provides a stable and adaptable framework for single and multi-frequency data analysis. Multi-frequency echosounder data visualization and post-processing have benefited greatly from its focus on real-time processing and the generation of virtual echograms [36]. With its advanced image processing tools, arithmetic operations, and logical bitmap and data combinations, Echoview is an incredible tool in the field that enables advanced analytical approaches [37]. It is the first extensively used commercial software package for this kind of application, supporting a wide range of acoustic data from different manufacturers of echosounders, including Kaijo, BioSonics, and Simrad. Enhancing research capabilities is its capacity to generate virtual echograms via mathematical operations over several channels.

2.1.2 Machine Learning Methods for Echogram Interpretation

In the analysis of echograms, traditional machine learning classifiers have been widely used, especially for problems involving fish species identification and acoustic target characterization. A variety of classifiers have been used to classify acoustic data via handcrafted features, including decision trees, random forests, minimum distance classifiers, gradient boost classifiers, general gaussian mixture models, support vector machines, and various types of shallow artificial neural networks [17]. Gauthier et al. [38], employed a classification decision tree to differentiate between various mesopelagic groups according to aggregate morphological features that were taken from echograms using programs such as Echoview. Proud et al.'s [39] work used random forests (RF) to classify schools of silver cyprinid Rastrineobola argentea (dagaa) in Lake Victoria with exceptional classification accuracy. Fallon et al. [40] also developed a novel random forest technique to distinguish weak acoustic-scattering organisms such as krill, mixed groundfish echoes, and mackerel icefish. LeFeuvre et al. [41] used a mahalanobis distance classifier to separate Atlantic cod and capelin species in high-resolution echograms in an effort to improve species differentiation in underwater sound-based surveys carried out in nearshore waters. Charef et al. [42] studied fish school classification by artificial neural networks (ANN) and discriminant function analysis (DFA). Minelli et al. [43] developed a semi-supervised ML system based on multibeam echosounder data, utilizing gradient boost classifiers (GBC) to distinguish fish schools from other targets, including gas bubbles, with high accuracy. Woillez et al. [44] employed general gaussian mixture models (GGMM) to classify multifrequency acoustic backscatter, demonstrating the ability to identify walleye pollock, krill, and other major classes occurring in the upper water column. ANN approaches in echogram analysis [45] and studies focusing on ambient noise data in the open ocean [46] to understand underwater physics highlight the expansive potential of these technologies in marine research. Traditional methods like Digital Image Processing (DIP) for image segmentation, morphological operations, edge detection, image representation, Binary Large Object (BLOB) analysis, and classification are integrated with ANNs. According to Robotham et al. [47], support vector machines (SVM) show potential in distinguishing species against a background when compared to ANNs in categorizing small pelagic fish species.

2.1.3 Deep Learning Methods

A systematic survey by Yassir et al. [48] highlights the benefits of DL methods over conventional ML techniques for automating fish species echo categorization based on acoustic

data. It illustrates the potential of DL models to improve the effectiveness and precision of acoustic fish species identification by showing how they can produce more accurate results even with a small amount of annotated data. By providing advanced tools that automatically learn and extract pertinent characteristics for more precise species detection, this shift towards DL has further aided in the field of echogram analysis [19].

Hirama et al. [49] suggested employing a CNN architecture in an image classification-based method to distinguish between five fish species in echograms: yellowtail, salmon, cuttlefish, sardine, and juvenille tuna. They classified non-overlapping tiles on echograms using the theory that only one type of fish was present in each tile.

A hybrid approach was used in another study [18] which dealt with the identification of schools of herring using the classification of echograms. Following the traditional identification of regions of interest (ROIs) in echograms, these were recognized and classified using three widely reviewed and accepted CNN methods. DenseNet [4] was shown to attain the best overall performance.

Brautaset et al. [8] used deep convolutional neural networks for acoustic classification in the Norwegian sandeel survey's multifrequency echosounder to identify sandeel schools from schools of other species and background pixels. They faced the issue of unbalanced classes and inaccurate annotations and they overcame these issues by exposing the network to balanced mini-batches and developing sampling and preprocessing strategies. Choi et al. [50] applied a novel semi-supervised deep learning method for acoustic target classification in multi-frequency echosounder data. The method utilized the structure of the data along with the available annotations in a single CNN architecture.

One study worth highlighting by author Vohra et al. [14] is based on a semantic segmentation approach that precisely spots the existence and roughly indicates the position of discrete scatterers, such as jellyfish tracks, over other methods like object detection or instance segmentation. This approach compares different DL networks and the impacts of multi-frequency data on early vs. late fusion techniques [14].

Where traditional ML methods utilize expert knowledge and handcrafted features, DL excels in its ability to autonomously learn and identify pertinent features. DL methods such as image classification, object detection, instance segmentation, and semantic segmentation for echogram analysis, and the comparison of these approaches against traditional ML methods, highlight the sophistication and potential of DL technologies [17].

Hybrid methods that use deep learning architectures such as U-Net [51] architecture to distinguish salmon and herring semantically in multi-frequency echograms are another advancement. This approach combines DL with heuristic techniques, classifying marine

species into distinguishable groups after pixel-level classification using U-Net. The performance of the DL models is further improved by adding environmental context, such as sun elevation angle and water depth [52].

Motivated by these noteworthy developments we investigate image segmentation and classification methods to address our task of interest.

2.2 Ocean Boundaries & Near-Surface Phenomena: Deep Learning Insights

A critical but little-studied issue in the field of echogram analysis for marine ecosystem research is the accurate classification of sea surface types and the identification of near-surface bubbles. This section delves into the body of literature that exists around these points.

Sandy et al. [53] focused on automating the detection and characterization of sea ice and surface waves using acoustic data collected from the upper 30 meters of the water column in the northeast Chukchi Sea, employs a self-organizing map (SOM) machine learning algorithm to differentiate between open water and sea ice cover in the acoustic dataset. Specifically, the SOM method produces an accurate distinction between polar oceans and ice by checking 15-minute ensembles of data. Thus, the statistical properties of surface wave height envelopes and ice drafts can be traced without defined thresholds. This approach makes the detection of key patterns more effective within large-volume acoustic datasets. Furthermore, the use of Acoustic Zooplankton Fish Profiler (AZFP) data, alongside conductivity, temperature, depth, and sea level pressure measurements, enhances the analysis. By integrating time-averaged data and statistical metrics such as skewness and kurtosis into the SOM, the algorithm is configured to identify distinct patterns within the AZFP data. Sandy automates the acoustic data processing and addresses the problem of labor-intensive and time-consuming analyses, thus increases the use of the acoustic data throughout the whole year and provides new information about the mechanisms underlying the Arctic sea ice seasonal variability and the surface wave generation in the Chukchi Sea region.

The study by Slonimer et al. [23] offers new insights into bubble detection and identification because of the 125 kHz strong presence of bubbles. This peculiarity contributes to making bubbles easily distinguishable from juvenile salmon. In this study, herring schools emitted gas bubbles which are specifically annotated as bubbles, while cases which cannot

accurately be classified as so, are filtered out as “unknown” and merged into the background class. This approach allows for the echogram to be segmented into five distinct classes: surface, background, schools of herring, juvenile salmon aggregations, and bubbles. This study reveals feasibility of using U-Net neural networks, to distinguish bubbles of the near-surface from fish like juvenile salmon [23].

Deep learning is used in the Echofilter model [19] to overcome post-processing difficulties in tidal energy streams, particularly in zone determination of entrained air, where Echoview approaches fall short. Echofilter automates the construction of boundaries and the designation of data regions of interest such as water column affected by tidal disturbance. This reduces manual labor and facilitates data standardization. Because of the echofilter model’s ability, obtaining predictions for boundary points and data periods is made simple, saving time and providing a helpful solution to previously difficult jobs of detecting fish schools and identifying physical phenomena in echograms. Three primary processes comprise core-data filtering: removing air head effects, removing erroneous data points below the sea floor, and taking into account all of the data points located on the sea floor’s surface. The model performs a great job at improving the resolution of the entrain-air lines, aligning to the training data, and accommodating the differences in depth-line and mask annotations. Because of this automation, hydroacoustic data automation with Echofilter is much advanced compared to Echoview, and it also offers a faster and more efficient analysis as well as proof of higher concurrence to human annotations.

To summarize, the literature on sea surface type classification and near–surface bubble detection in echograms is still nascent; however, the relevant studies demonstrate the capability of DL methods in transforming the way we perform echogram analysis.

2.3 Deep Learning for Computer Vision

Since DL approaches have become widely used, computer vision has undergone tremendous change. Neural network frameworks with remarkable improvements in image classification, object detection, semantic segmentation, and image restoration have emerged and include AlexNet [54], VGGNet [55], GoogLeNet [56], ResNet [5], DenseNet [4], and UNET [51]. These approaches demonstrate how DL in CV applications advances our understanding of images, as described in [57].

Image classification becomes a key task in the context of DL for CV, taking advantage of CNN architectures’ powerful capabilities. A standard for performance in CV tasks is set by the convolutional, pooling, and fully connected layers that make up the architectural

core of CNNs, which enable robust feature extraction and classification [58]. By using CNN-driven semantic segmentation in particular, DL has significantly advanced the field of image segmentation concurrently.

This section reviews key concepts and works in DL for tasks that are relevant to our two-step approach, organized into sub-sections: convolutional neural networks, image classification, and semantic segmentation.

2.3.1 Convolutional Neural Networks

Neural networks are composed of interconnected neurons, forming the basic units of computation. By adding hidden layers, neural networks become more complicated and give the model depth, hence the term ”deep learning” [59]. The network can decipher complex data patterns by activating particular subsets of input information thanks to these hidden layers. The model's learning capabilities are greatly strengthened by this improved pattern identification, especially in situations where there are several output categories. Turning now to CNNs, these models are particularly good at processing image data because they divide images into smaller parts, or tiles, and then perform convolution operations on those segments. With small kernels, usually 3x3 or 7x7 pixels, the goal is to extract and activate particular features from these segments while balancing the computational demand and feature extraction efficiency [59]. CNNs are hierarchical architectures that recognizes basic patterns in the first layers and combine them into more intricate representations in the subsequent layers is how CNNs set themselves apart. In order to achieve reliable feature detection using CNNs, sizable datasets is required [60]. These datasets are frequently enhanced by data augmentation techniques like noise addition, rotation, or mirroring. The following sections, 2.3.2 and 2.3.3, expand on the fundamental ideas of neural networks and CNNs by exploring the particular fields of DL-based image classification and semantic segmentation.

2.3.2 Deep Learning-based Image Classification

An image may be categorized as a ”animal,” ”vehicle,” or ”landscape” in image classification, for example, depending on what's captured in it.

GoogLeNet [56], detailed by Szegedy et al., introduced the inception module, a novel concept that allowed the network to process images at various scales simultaneously. This architecture's parallel convolutional paths enabled it to capture a wide range of image features without the exponential increase in computational cost typically associated with

deeper networks. GoogleNet's success in robust feature representation for image classification has become the benchmark for network design, which focused on both depth and width but without causing the slowing down or oversize of networks.

Following GoogLeNet, the development of ResNet-101 [2] by He et al. introduced a novel approach that overcomes the vanishing gradient problem which is a significant hurdle in the training of DL networks. The architecture has skip connections that were able to allow gradients to bypass layers. ResNet-101 can effectively learn from additional layers without performance degradation. This not only facilitated more complex feature learning but also improved image classification performance, showcasing the potential of depth in neural network design.

The development of Darknet-53 [5] was within the framework of YOLOv3 [5], a popular object detection framework, by Redmon and Farhadi. What made this architecture stand apart was its focus on optimizing the trade-off balance between speed and accuracy. Darknet-53 deals with real-time object detection.

DenseNet-201 [4], proposed by Huang et al., further advanced network architecture through its dense connectivity pattern. Unlike its predecessors, it connected each layer to all other layers directly to guarantee the maximum information flow through the network. This was a major engineering breakthrough that considerably limited the model's complexity because of greatly increasing feature reuse, thus generating noticeable efficacy gains and higher classification accuracy. Densenet-201 demonstrated the power of connections in stimulating the networks to become more cooperative and feature-rich.

The evolution of the inception architecture culminated in Inception-v3 [61] by Szegedy et al., which brought several optimizations to further refine the model. Factorized convolutions and label smoothing techniques reduced the computational demand and improved the model's generalization capabilities. Inception-v3's adaptations made it a valuable model among extremely scalable DL models. A wide range of image classification tasks could be solved successfully through this model, it turned out to be a cornerstone for the use of efficient and powerful DL models for CV.

2.3.3 Deep Learning-based Semantic Segmentation

Semantic segmentation is the process of assigning labels to each pixel in an image, classifying them into categories such as "human", "car", "tree", or "sky". Unlike image classification where the entire image is assigned a single label, semantic segmentation assigns one label to each pixel in the image [62]. Among the best examples of semantic segmentation

architectures are DeepLabv3 [11], Mask R-CNN [63], UNet [51], Attention UNet [64], and UNet++ [65], each contributing significantly to the field's advancement.

In DeepLabv3 [11], Chen and et al. employed atrous convolutions and atrous spatial pyramid pooling (ASPP) to gradually increase the size of the receptive field, thus, reducing the loss of spatial information. The employment of atrous convolution allows the network to effectively capture multi-scale contextual information without increasing the number of parameters. ASPP addition allows the model to deal with the scale changes and thus demonstrates its efficacy at classifying objects at different sizes.

Mask R-CNN [63], an extension of the Faster R-CNN by He et al. [63] and his collaborators, represents a significant leap by seamlessly amalgamating object detection and high-precision segmentation. This is achieved by applying a mask to each Region of Interest (RoI), effectively merging object detection and segmentation into a cohesive architecture. This fusion results in exceptional performance across both object detection and segmentation domains.

The UNet [51] network has a symmetric encoder-decoder structure initially conceived for biomedical image segmentation by Ronneberger et al and is very efficient in depicting contextual information for accurate segmentation. This architectural novelty has been the starting point and development for modern versions such as Attention UNet [51] and UNet++ [65].

Attention UNet [66] is based on the original UNet structure, but adds an attention mechanism. This mechanism acts to direct the model's focus towards meaningful image features that offer more value instead of the areas of less relevance. Consequently, Attention UNet has proved to be promising to deliver accurate segmentation even for images that have complex backgrounds (non-target areas) or that have subtle differences in large-scale structures.

UNet++ [65] is an improved version of UNet [51] that adds nested, dense skip pathways. This architectural innovation narrows the semantic gap between encoder and decoder features which results in a more efficient feature propagation and reuse. UNet++ excels in the segmentation task because it produces more accurate results for the entire image.

Chapter 3

Dataset and Annotations

In our thesis, we use Cape Bathurst Arctic Sea Surface Acoustics (CBASSA) dataset, which includes raw data collected by Fisheries and Oceans Canada. Consisting of 15 months of one-hour multi-frequency echograms, gathered from August 2017 to October 2018, the deployment utilizes moored upward-looking autonomous echosounders located near Cape Bathurst in the Northwest Territories of Canada. The area around Cape Bathurst, with its distinct Arctic geography and rich marine ecosystem, provides an ideal location for such acoustic studies. The CBASSA dataset, therefore, is a resource for understanding the marine environment of this region, offering critical insights into the habitat of various marine species and contributing significantly to Arctic environmental research.

The training of a Convolutional Neural Network (CNN) requires the creation of a high-quality annotated dataset. The dataset, central to our study, is not merely a collection of raw acoustic data converted to echograms; they are backed up by a carefully curated set of annotated images, prepared by us with the intent to train DL models. In this chapter, we discuss the specifics of data acquisition and visualization, as well as the properties of the dataset. The chapter also aims to elucidate the meticulous process of annotation. We discuss the specific challenges encountered in annotating echogram data for the DL models and the strategies employed to overcome these hurdles.

3.1 Overview of CBASSA Dataset

As discussed in chapter 1, echograms are generated by measuring acoustic backscatter within the water column with multi-frequency echosounders (e.g. the Acoustic Zooplankton and Fish Profiler or AZFP from ASL Environmental Sciences [12]). Echosounders can

be moored to the sea floor, looking upwards, or attached to vessels looking downwards, thus allowing analysis of both the sea surface and sea floor, as well as different parts of the water column as mentioned in Section 1.1. In our case, the instruments were deployed on the seafloor, with the transducers oriented towards the sea surface with a 15° angle from the verticle axis, as illustrated in Figure 3.1. The echogram visualization techniques employed in this dataset are not only integral to its structure but are also key in capturing the unique acoustic environment of the Arctic region. The mooring embodied a bottom-mounted AZFP [12] echosounder that measured data at four frequencies (38, 125, 200, and 455 kHz) covering about 51 meters of the water column. The collected data are visualized as 201 × 712-pixel echograms, where each pixel represents approximately 25.3 cm depth resolution (height-wise) through approximately 5s of time (width-wise).

The echograms in the dataset display the standard volume backscattering strength, denoted as S_v, calculated from the raw acoustic data. S_v, which represents the sum of all acoustic responses within a volume of 1 m^3, is determined using the deployment metadata and is given by the following equation [12]:

$$S_v = EL_{max} - \frac{2.5}{a} + \frac{N}{26214a} - SL + 20\log R + 2\alpha R - 10\log\left(\frac{c\tau\Psi}{2}\right) \tag{3.1}$$

In this formulation, EL_{max} is the maximum echo level (in dB re 1μPa) that the 16-bit A/D converter can handle before reaching saturation, N the count value from the raw data, which correlate linearly with the logarithmic scale of the voltage received post-amplification, band-pass filtering, and detector processing. a the detector response's gradient (V/dB), α the seawater absorption coefficient (dB/m), R the distance from the instrument (m), SL the source level in dB re 1μPa at 1 m, c the speed of sound in the water (m/s), τ the duration of the transmitted pulse (s), and Ψ the two-way solid angle of the acoustic beam.

In order to visualize the echograms effectively, Sv values, typically ranging from around -125 to 0 dB in this case, are converted to red-green-blue (RGB) integers using a suitable colormap. The popular "jet" colormap is appealing for its full visual spectrum, showing large changes in chroma and luminance. Additionally, the "jet" colormap highlights even the smallest existing image features with high contrast, which is advantageous in our case. This is particularly beneficial because it helps distinguish between near-surface bubbles and coexisting biology, which could otherwise be difficult to detect.

Figure 3.2 depicts echograms from a bottom-mounted AZFP echosounder, illustrating

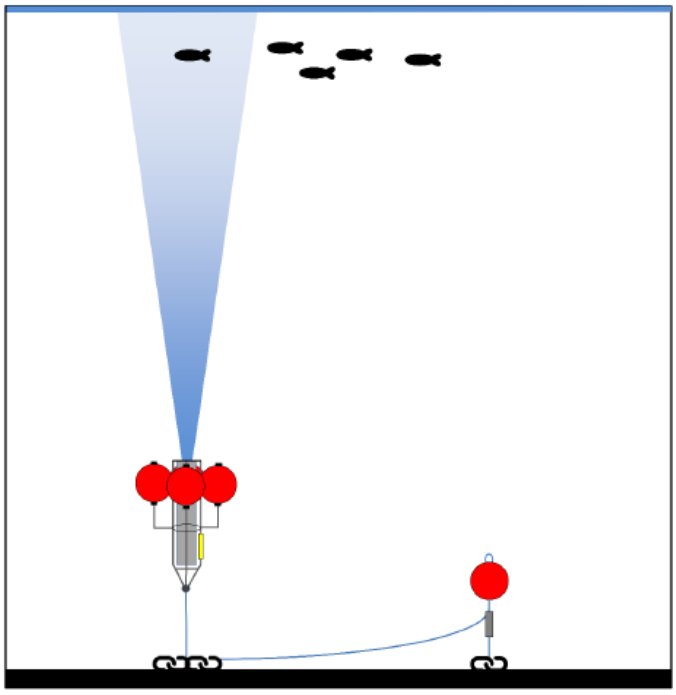

Figure 3.1: *Illustration of an upward-looking AZFP deployment: The AZFP is housed within a protective frame, stabilized by four spherical buoys to maintain vertical alignment. In our case, the echosounder is oriented with a* 15° *angle from the verticle axis. [1].*

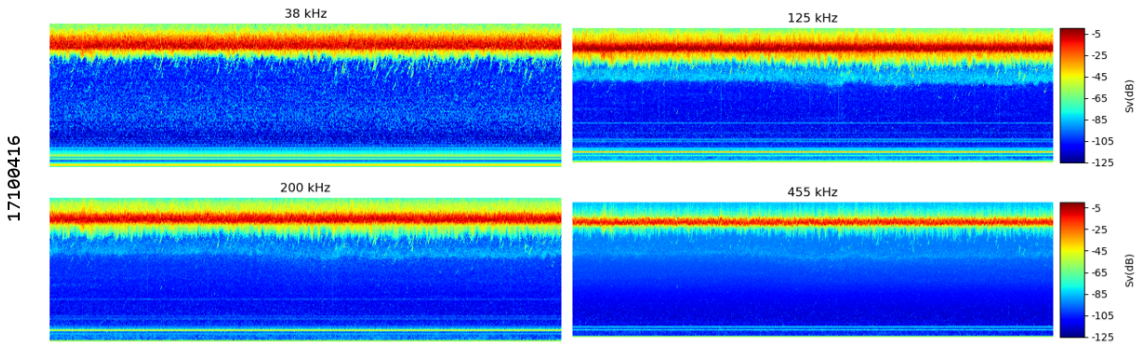

Figure 3.2: *The image shows four echograms from October 4, 2017, at 16:00 (8-digit timestamp: YYMMDDHH, with HH in the 24-hour clock format) captured at frequencies 38, 125, 200, and 455 kHz, under windy conditions. Colors range from red (strong signal) to blue (weak signal), illustrating how varying frequencies reveal unique and overlapping underwater environmental details.*

the sea surface's acoustic reflections at four different frequencies: 38 kHz, 125 kHz, 200 kHz, and 455 kHz for October 4, 2017, between 16:00 and 17:00. Each frequency in Figure 3.2 reflects the windy conditions through the pattern and intensity of surface backscatter, yet they also convey different information about the marine environment. The 38 kHz frequency offers a broader picture, less impacted by the noise and agitation at the surface, enabling clearer detection of phenomena in the deeper water column. The 125 kHz and 200 kHz frequencies, while still indicating the rough conditions at the surface, show both the surface detail and mid-water column clarity, capturing a balance of detail that is affected by wind-created bubbles and surface scattering. The 455 kHz frequency, with the highest sensitivity, shows the most detail of the near-surface disturbances but less of the deeper water column and biology. Together, these frequencies present a comprehensive acoustic snapshot, with each adding a layer of understanding to the overall picture of the marine environment under a particular condition occurring in that specific hour.

In the subsequent sections, we explore how we use the insights from the frequencies and other ancillary data to annotate the data for the classification and segmentation tasks.

3.2 CBASSA Dataset for Sea-Surface Type Classification

The dataset comprises a subset of the comprehensive CBASSA collection, annotated to facilitate image classification techniques for the sea surface classification task. With a total of 3529 labeled echograms categorized into six distinct classes, the dataset provides a robust

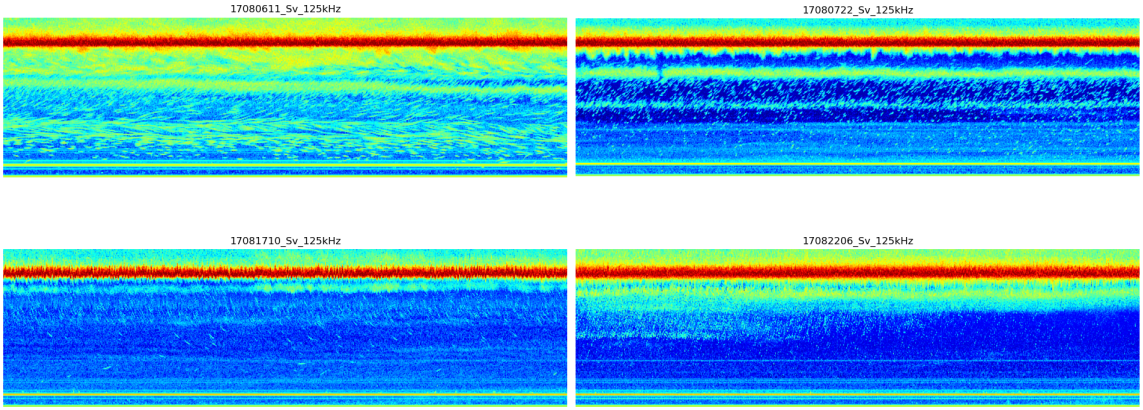

Figure 3.3: *Examples of open-water conditions as captured by 125 kHz echograms showcasing the characteristic acoustic signatures of open water: a strong, uninterrupted surface echo and a clear water column with minimal scattering. These conditions are indicative of a calm sea state.*

foundation for training and testing classification models. These classes encompass various surface conditions encountered in the Arctic environment, including open water, windy open water, ice with keels, ice without keels, slushy conditions, and mixed conditions.

3.2.1 Surface Type Identification: Annotation Cues and Criteria

Our classification task involves echograms captured at 125 kHz frequency. While echograms at other frequencies contain valuable information, utilizing the 125 kHz frequency simplifies the classification process. This decision is based on the fact that, despite all frequencies being capable of indicating surface type, the 125 kHz echograms offer a clearer representation, making it easier to discern surface conditions. Our annotation process takes into account various implicit and explicit cues for each surface type which are summarised below.

Open-Water Conditions: Open-water conditions in echograms show a strong surface echo and minimal below-surface scattering, indicating calm, ice-free water. This clarity, typical in summer or when ice disperses, facilitates acoustic data analysis, as shown in Figure 3.3. Marine life, visible as yellowish-green and cyan-blue shapes, indicates a diverse underwater ecosystem. Even in winter, when ice is disrupted, these open patches can host active marine organisms.

Windy Open-Water Conditions: Windy open-water echograms, as shown in Figure 3.4, display irregular surface echoes and increased backscatter from wind-induced waves and bubbles, creating a scattered appearance deep below the surface. In these situations,

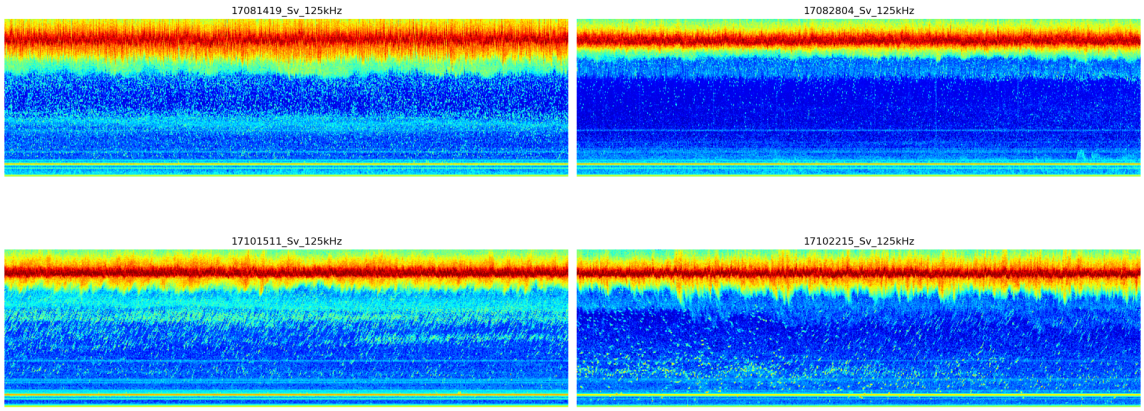

Figure 3.4: *The echograms, at 125 kHz, show windy, open-water conditions, revealing disrupted surface echoes and increased backscatter from waves and bubbles. These patterns indicate surface agitation and the effect of dynamic conditions extending into the water column.*

marine life is often detected mingling with the bubbles or along their periphery, creating a dynamic interplay that appears as cyan and yellowish hues in the echogram, signaling the presence of organisms, and highlighting underwater organisms amidst turbulence. The prevalence of deep-reaching bubbles in these echograms underscores the significant impact of wind on the water column.

Ice with Keels: In the context of the ice with keels cue, as shown in Figure 3.5, the echograms are annotated based on the smoothness of their surface echo and their strong backscatter intensities, the two features characteristic of underwater ice features known as keels. The backscatter yielded by these keels is strong and can extend up to different depths depending on the size and shape of the keels. Biomass, visualized as blue clouds against the background of dark blue water column, sometimes shows clustering behavior around these ice aggregates. When biological activity occurs, small discrete points bathed in different light colors appear.

Ice without Keels: In contrast to ice with keels, Figure 3.6 illustrates ice without keels presenting a flatter and smoother surface echo and the subdued acoustic returns from beneath the ice in echograms. This type of ice is typically older and has had time to smooth out, resulting in a more uniform echo signature. This condition is typical in winter, forming a continuous ice cover in polar regions. Near-surface biology, distinct from bubble-induced backscatter, appears as clear, strong signals in clusters. Life persists throughout the water column, visible against the dark blue background, indicating active marine ecosystems beneath the ice.

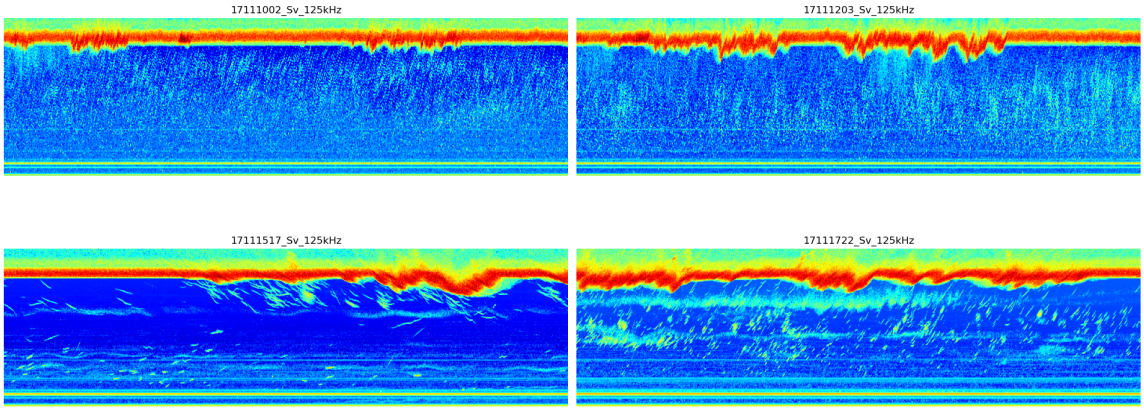

Figure 3.5: *The echograms at 125 kHz depict ice with keels, showcasing smooth surface echoes and vertical variability characteristic of keeled structures. The strong returns from the keels reveal their varied depths and shapes, highlighting the diversity of ice protrusions into the water column.*

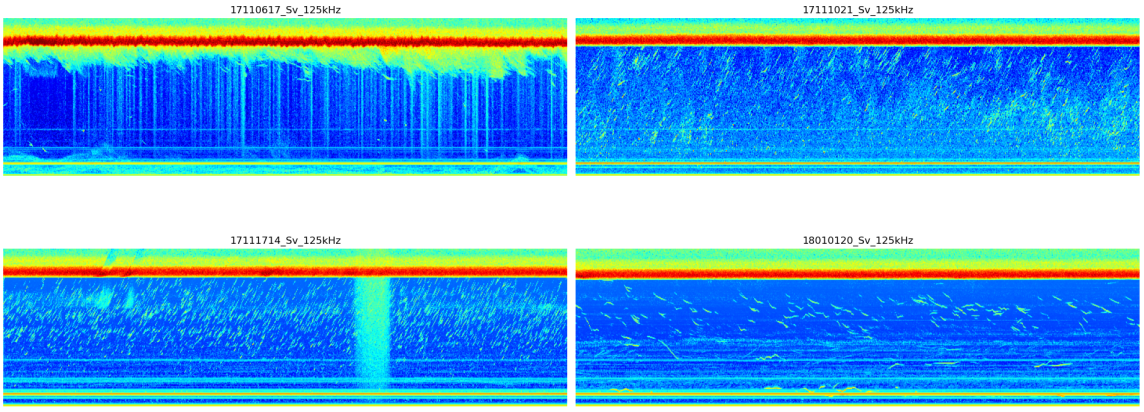

Figure 3.6: *Echograms at 125 kHz representing ice without keels. These images capture the smoother surface echo and the subdued acoustic returns from beneath the ice, characteristics of older and more uniform ice surfaces with minimal vertical relief.*

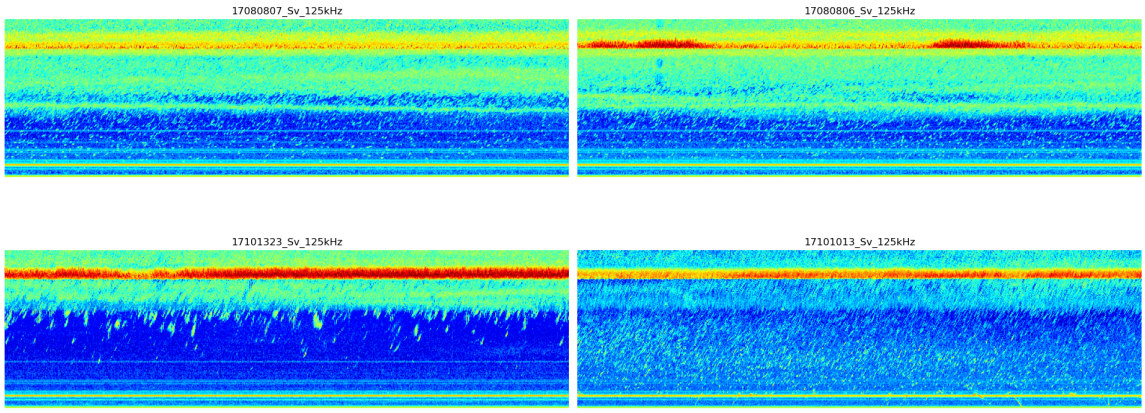

Figure 3.7: *Echograms at 125 kHz depicting slushy conditions. The images illustrate the diffused and "fuzzy" surface echoes that are characteristic of slush, which consists of a mix of ice and water with varying degrees of solidity. This slush layer scatters the acoustic signals, resulting in a less defined surface echo and a generally cluttered appearance in the upper water column.*

Slushy Conditions: Figure 3.7, provides examples of slushy conditions where the formation or melting of ice creates a semi-solid layer at the surface. The presence of slush is indicated by the diffuse, indistinct surface echoes and the cluttered acoustic returns. The slush, consisting of a mix of ice and water, scatters the acoustic signals, leading to a "smeared" or "fuzzy" appearance at the water's surface. This condition, often occurring during shoulder seasons or temperature fluctuations, results in scattered signals from the mix of ice and water. Beneath this slushy surface, marine life echoes resemble those in open water, with distinct biological activity visible throughout the water column, despite the altered surface texture.

Mixed Conditions: Mixed conditions present a combination of all the previously discussed classes in echograms as illustrated in Figure 3.8. The surface echo and scattering within the water column will be heterogeneous, reflecting the complexity of the marine environment where open water, ice, wind effects, and slush may all be present to varying degrees. This category is particularly challenging to annotate due to the diversity of acoustic signatures present. Such diverse acoustic environments are seldom observed within a single hour due to the dynamic and transient nature of these elements, making these echograms particularly valuable for the study of Arctic sea surfaces.

To enhance the accuracy of our annotations within the CBASSA dataset for the classification task, we integrated ancillary data from external sources. This supplementary data provides additional context, ensuring a more robust classification of the echograms.

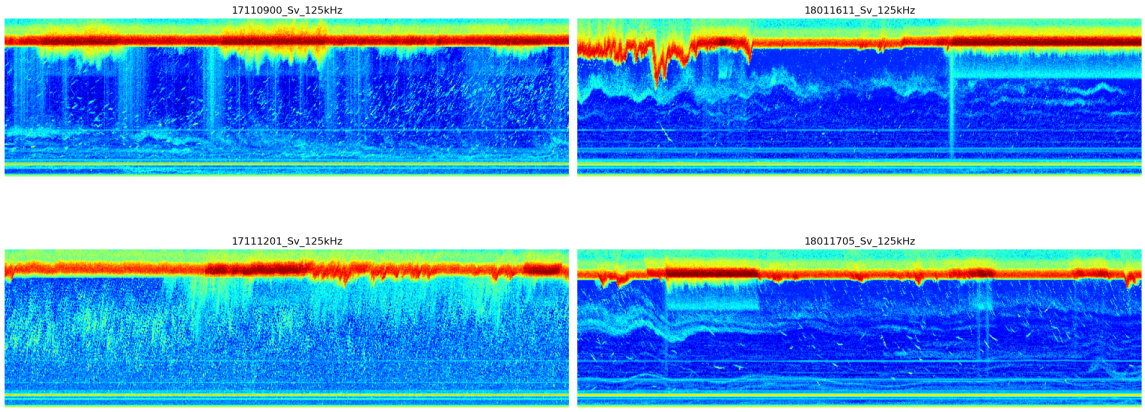

Figure 3.8: *Echograms at 125 kHz showing mixed conditions of the Arctic marine environment. These pictures present the variety of combinations, showcasing the complexity and variability of surface conditions. From left to right, the echograms show ice without keels paired with windy open water, ice with keels combined with open water, slushy conditions alongside ice, and windy open water meeting ice without keels.*

Firstly, we utilized satellite imagery from NASA Worldview [67] to observe the surface conditions at the specific coordinates of Lat: (70.576650, -127.660300) or (70°34'35.9"N 127°39'37.1"W). It is worth mentioning that data from satellites is unavailable during the polar winter during the polar midnight when the sun doesn't rise. Moreover, cloud cover can frequently obscure the view, necessitating the use of the date slider to identify clear images. When satellite data was not available or obscured, we relied more heavily on the characteristics directly observed in the echograms to inform our annotations. This ancillary data was particularly useful useful in distinguishing between surface types with ice (such as IWK, MC and INK) as opposed to slushy conditions and other open water conditions (WOW and OW). Secondly, to confirm windy conditions, we resorted to hourly report from the Cape Parry weather station, the closest to the region. These data greatly assist in confirming the presence of wind-related features on these echograms, which include increased surface backscatter and disturbed water surface, thus providing an additional proof for the annotations related to windy open-water situations.The wind data from the weather station proved essential for distinguishing between open water and windy open water conditions, which can be challenging to differentiate visually in echograms. The wind speed aided the visual classification of echograms as windy open water.

3.2.2 Graphical User Interface for Echogram Labeling

We developed a Graphical User Interface (GUI) in MATLAB to allow for the efficient creation of categories in which the images were sorted by using the cues discussed in the previous section. This GUI simplifies the annotation process by creating a user friendly platform which will enable users to classify echogram images of marine organisms. The GUI encompasses essential functionalities for a seamless annotation experience. Upon launch, as depicted in Figure 3.9, the user views a simple interface that has icons for important functions such as selecting folders of echogram images, traversing in the folder, categorizing the echograms, and saving the annotations.

The GUI enables the user to go through images in a structured way and manage the categorization, which is a critical step in building a labeled dataset. It saves the images, their corresponding names, and labels in a format that is both portable and amenable to further analysis or integration with other larger datasets.

Utilizing the above-mentioned GUI for labeling, we labeled a comprehensive dataset comprising 3,529 echograms, which are distributed across six distinct classes that reflect the various sea surface conditions encountered in the Arctic environment. The breakdown of the dataset is as follows: ice with keels (770 echograms), ice with no keels (506 echograms), mixed conditions (211 echograms), open water (646 echograms), slushy mixed (514 echograms), and windy open water (882 echograms). This classification was conducted under the supervision of experts. It is important to note that the distribution of

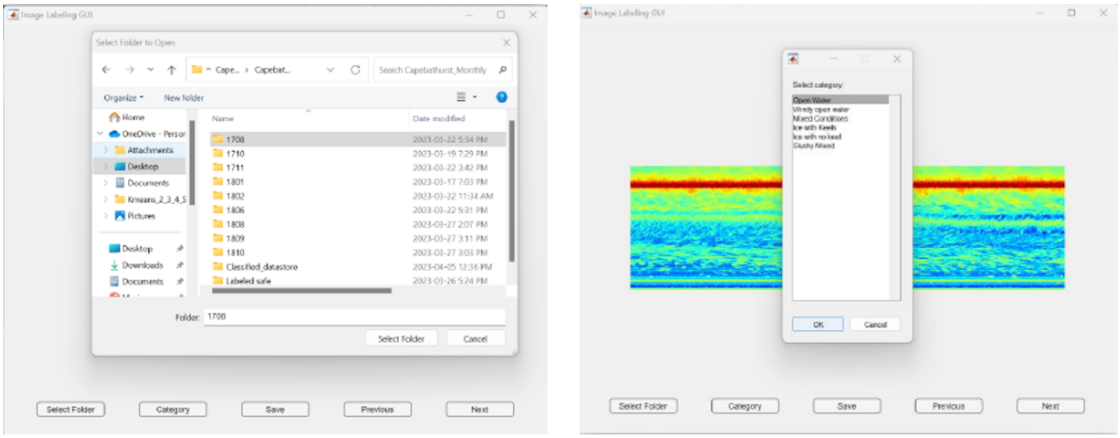

Figure 3.9: *Image Labeling Graphical User Interface for streamlining echogram classification for image annotation process*

echograms across these classes is not uniform, with the number of echograms in the mixed

conditions class being significantly lower than that in the windy open water class. This disparity in class representation mirrors the natural occurrence rates of these conditions over several months of observation. The dataset, therefore, not only provides a valuable resource for the development and testing of image classification models but also reflects the real-world variability and frequency of these marine conditions as encountered in the dataset's span, capturing the genuine ratios observed during the data collection period.

3.3 CBASSA Segmentation Dataset

The CBASSA Segmentation dataset, a subset of the CBASSA Classification dataset, is designed specifically for segmentation tasks. More specifically we aim at detecting the sea surface boundary and the bubbles near-surface. We refer to these tasks as segmentation because we aim at a precise delineation of the boundaries of the regions of interest. This dataset consists of echograms along with masks of surface, bubbles, and background segments in it. As this dataset is a subset of the dataset used for the classification task, it consists of echograms classified into six categories—open water, windy open water, ice with keels, ice without keels, slushy conditions, and mixed conditions and is organized for the development of segmentation models tailored to the Arctic marine environment.

Each echogram in the dataset corresponds to a mask that precisely marks and identify the surface and bubbles regions making the segmentation task possible. This pairing is to be used as input to the DL semantic segmentation architectures. The segmentation dataset also enhances the toolkit available to researchers and technologists working on echogram analysis.

3.3.1 Surface Phenomena Detection: Annotation Cues and Criteria

For the segmentation task, our approach leverages fused echograms that combine data from 125 kHz, 200 kHz, and 455 kHz frequencies, while excluding the 38 kHz frequency due to its lower sensitivity to small features. The reason behind this selection is based on the principle that higher frequencies, with their shorter wavelengths, are inherently more adept at detecting fine details, such as bubbles and subtle textural differences in the marine environment. Fusion is done at a pixel-level and merges multiple input images into a single, more informative composite image. This method enhances the perception of the image, either for human observers or for automated analysis systems, by integrating the strengths of individual frequencies [68]. We have adopted a fusion technique that sums up

the volume backscattering strength (Sv) values across all frequencies. The combined Sv values are translated into visual information by using the jet colormap. A systematic annotation process was applied which was informed by the distinct acoustic properties related to the Arctic seascape resulting in the precise identification of the boundaries of bubble aggregations and the sea surface to generate the ground truth masks for the segmentation task. Through careful assessment of color intensity, texture, and continuity, we were able to annotate the relevant areas on the echograms with precision. The boundary and surface bubble detection involves a detailed analysis of acoustic cues that are characteristic of these features.

The sea surface boundary and the presence of bubbles are both represented as different acoustic signals in the echograms. A consistent, powerful echo that appears as a distinct red band across the top of the image in an echogram indicates the air-water interface. The reflection of the acoustic signal at the interface between the water's surface and the air above it is indicated by this red band. Usually, the sea surface border is indicated by the lower edge of this red band, which is where the bright red starts to change to other colors. The sharp difference between the air above and the water below, shown by this echo, clearly defines the edge of the ocean.

The cues for bubble detection in echograms are the distinct visual patterns and colors. Visible yellow, amber, or lime yellow traces that begin on the sea's surface and progressively fade as they descend are the most appropriate way to identify bubbles. Visually, these bubbles resemble the above wave-like shapes. This deterioration from the surface down indicates that the density of air bubbles decreases with depth. It is important to distinguish between biology and bubbles. Bubbles are characterized by their continuity and resemblance to the undulating surface of the sea. Typically, they originate from the boundary surface. In a contrast, biological features, resembling patches or clusters, are usually displayed as light or dark cyan, light green, and light yellow, showing stronger acoustic returns. Occasionally, the biology would blend with bubble. For clarity in these visual representations, as shown in Figure 3.10, a black line is superimposed to outline the sea surface boundary, and a white line shows the boundary of the bubbles. In Figure 3.10, bubbles are notably present in the windy open water and mixed conditions categories, underscoring the influence of wave activity. Windy Open Water echograms display bubbles generated by wind-driven surface agitation, while mixed conditions reveal bubbles where calm and turbulent waters meet. The lack of bubbles in other categories implies more stable conditions without significant wave-induced air entrainment. In the CBASSA segmentation dataset, consistent with the observed patterns in the echograms shown in Figure 3.10, only the

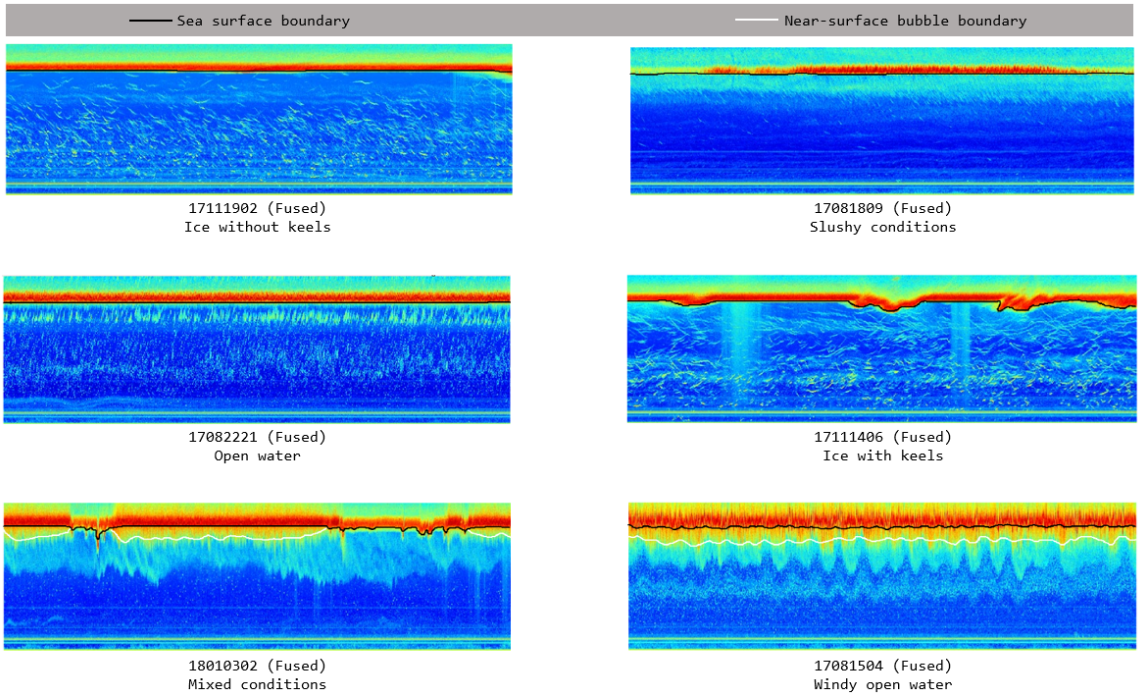

Figure 3.10: *Fused echograms show the acoustic fingerprints of the Arctic Sea surface under different conditions, emphasizing important acoustic markers for bubble detection and sea surface limits. The air-water interface is seen by the top red band in each image, where the transition is marked by a significant echo. Where the red fades into other colors, the sea surface boundary is depicted, and bubbles appear as yellow to amber tones that fade with depth. To ensure clarity, the sea surface boundary is denoted by a black line, while the near-surface bubble areas are outlined in white to distinguish them from the biological entities that are colored in cyan, green, or yellow.*

echograms categorized as windy open water and mixed conditions contained discernible bubbles. This observation aligns with the nature of these environmental conditions, where wave activity is a significant factor. The windy conditions of the open-water echograms highlight how wind develops waves that are powerful enough to encase the air and form visible bubbles. Segmentation of the data will thus be distinguishing the chaotic textures of bubbles from the other biological phenomena and water. Another illustration of windy open water echograms is shown in Figure 3.4 where the characteristic pattern of bubble formation is once again evident. Similarly, the mixed conditions class, characterized by its heterogeneous environment, shows bubbles in areas where the water surface is disturbed by varying degrees of wave action, which could be noticed in Figure 3.8. This could be due to partial wind exposure or the interface between calm zones and agitated waters, resulting in the formation of bubbles.

In other classes (see Figures 3.6, 3.7, 3.3, and 3.5 for examples), the absence of bubbles can be attributed to insufficient wave activity or surface agitation. The water surface stays steady in calm conditions, such as those found in the open water and ice classes, which reduces the likelihood of bubble formation. Bubble production and identification may also be impeded by slushy conditions, which include partially solid surface layers. In light of this, these classes have more consistent acoustic patterns than those in which bubbles are visible, a sign that the water's surface is calm or settled in this particular situation. The intricate dependence of sea surface behavior on atmospheric processes is shown in the emergence of bubbles in just two classes of the segmentation dataset.

3.3.2 Semi-automatic Annotation Process for Segmentation

We developed a semi-automatic annotation technique using a GUI to address the challenges of annotating marine echograms. The segmentation tasks for these echograms have traditionally involved manual annotations by specialists, which is a labor-intensive process prone to intra- and inter-annotator variability. Our semi-automatic approach guided by a GUI expedites this procedure, allowing more consistent pixel-level classification of echograms into near-surface bubbles and sea surface boundaries. This approach seeks to decrease the variability inherent in hand-done annotations by utilizing pixel attributes and minimizing the need for manual input, resulting in outcomes that are as consistent as feasible between users and sessions.

Boundary Detection Method: Gaussian blurring at 0° and then region-growing are the two steps in our border detection approach. Echograms are made clearer and less noisy

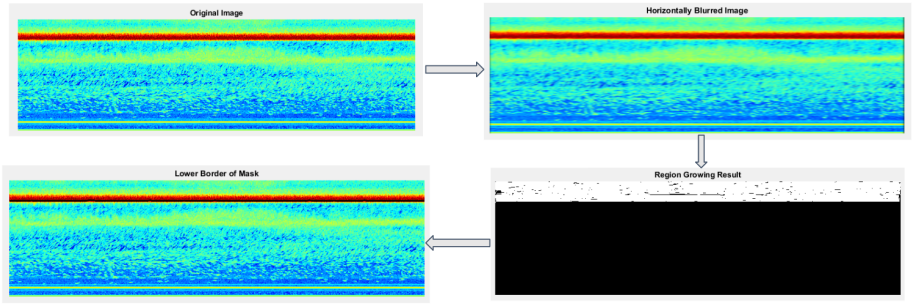

Figure 3.11: *Illustration of Boundary Detection in Echograms Using Gaussian Blurring and Region-Growing Techniques.*

by applying Gaussian blurring; clearly defined borders are made easier to draw. Following this, the region-growing algorithm is employed to accurately define borders. This algorithm starts from a seed point, which is the brightest red pixel on the air-water interface band, and iteratively adds neighboring pixels to the region that meets certain similarity criteria based on the region-growing threshold parameter [69]. The process is visually guided by a distinct line, three pixels thick, marking the progression of the region expansion. The user has to adjust two important parameters to fine-tune the annotation process: the region-growing threshold and the blur pixels for Gausian blurring parameter. They may accomplish precise and trustworthy border detection by modifying these parameters to fit the distinct qualities of every echogram. Figure 3.11 shows the intermediate results of the different steps for a specific example.

Near-Surface Bubble Detection Method: For near-surface bubble identification, first, Guassian filtering is applied on the image followed by the image conversion to grayscale and mapping to a perceptually uniform colormap (parula). Subsequently, k-means clustering is then applied on the uniform colormap image, which is followed by connected component labeling. Clarity of the visual representation is improved by the Gaussian filtering step, which helps to reduce noise in the echogram. Images in grayscale have single channel and therefore have distinct clusters as opposed to the jet colormap. The conversion of image to grayscale and then to a uniform colormap ensures that the features used for clustering are uniform in intensity resulting in consistent and distinct clusters. It is possible to discover regions with similar properties by using K-means clustering [70], which divides the image into discrete groups. Next, connected component [71] labeling is used to help accurately delineate and label bubbles suspended close to the ocean's surface as well as group pixels that belong to the same object or feature. In summary, the purpose of this method is to maximize the effectiveness of each technique in detecting near-surface bubbles in ma-

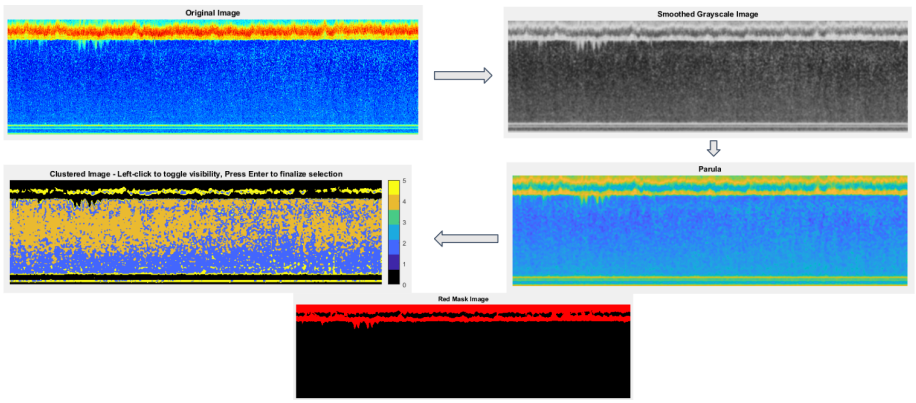

Figure 3.12: *The multi-step process for near-surface bubble detection in marine echograms. This figure illustrates the sequence starting with the original image subjected to Gaussian filtering for noise reduction, followed by a smoothed grayscale image, transitioning to parula and then k-means clustering. Finally, the red mask image highlights the connected component labeling, enabling precise bubble delineation. Users must set the number of clusters and largest blobs for effective detection.*

rine echograms by taking advantage of their respective strengths. Both the number of k for k-means clusters and the number of largest blobs to be identified are mandatory parameters that users must supply. This is determined after experimenting with the knowledge gained from preliminary tests. Figure 3.12 shows the sequential steps required in near-surface bubble detection.

Figure 3.13 shows the GUI, through which the annotation process is done in an understandable and accurate manner. The process begins by identifying the optimal surface line, which is then followed by guiding the software to detect bubble areas that accurately adhere to the annotated surface boundary. Once this is achieved, the user can repeatedly and reliably get the results that they want with very little effort for all the echogram images in the dataset. Also, the generated masks can be saved as RGB images which will prove to be quite useful especially in further analysis. In the generated masks, distinct colors represent different features: blue is used for the surface boundary, with the entire area above it filled in to create a single segment rather than just a line; red indicates bubble patches/segments (not lines), and black represents the background. It is important to note that the annotations generated by the semi-automatic method serve solely to produce ground truth masks for each echogram, specifically intended for use in the DL Segmentation Method.

A total of 1,691 echograms (out of 3,529) were annotated for the segmentation task as follows: open water (300), windy open water (299 – 264 with observable bubbles), ice with keels (298), ice without keels (299), slushy conditions (318), mixed conditions (177 –

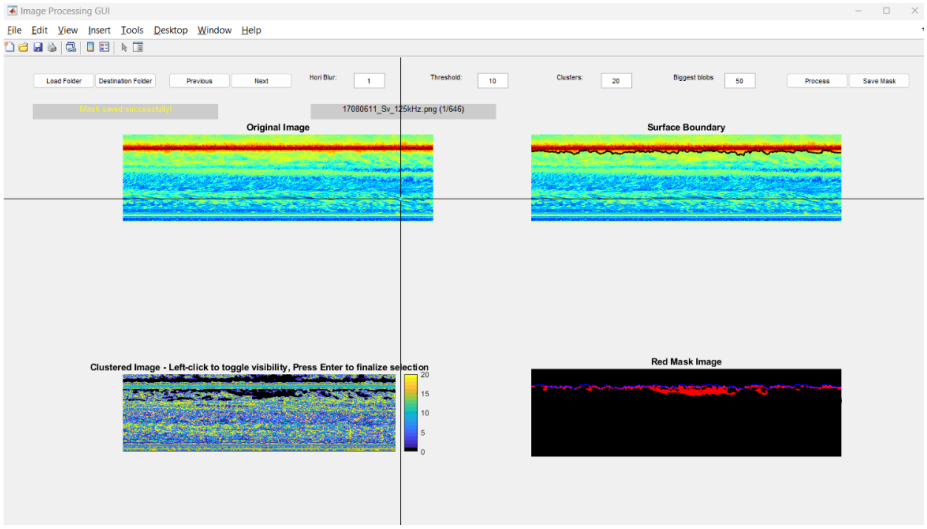

Figure 3.13: *Screenshot of the Interactive Image Processing GUI for Annotating Marine Echogram Images. This interface presents tools for users to load echogram datasets, adjust parameters like horizontal blur and threshold for segmentation, and define the number of clusters and largest bubbles for detection. The GUI displays both the original echogram and the annotated surface boundary to assist in the segmentation process. The user can save the results as RGB mask images for subsequent processing.*

50 with observable bubbles). These annotations provide an extensive dataset for additional research and model building.

3.3.3 Refinement of Ground Truth

Because bubbles are a quasi-random phenomenon, detecting them using acoustic signals is a challenging issue. Numerous random and unpredictable elements might affect the acoustic waves, making the process of detection and analysis more difficult. The irregular distributions of bubble sizes and locations, ambient noise that can mask or obscure bubble signals, signal attenuation as sound travels through different water conditions, and dynamic conditions that are constantly changing due to water movement and marine life movement are some of the variable sources that add to the complexity. The fact that the (lower) boundary of the entrained air penetration within the water column can be hazy, porous, and discontinuous makes the boundary placement more difficult [19].

Misclassification of the biological entities as bubbles is one of the challenges of the semi-automatic annotation method explored earlier. Hence, the k-means clustering method, an intensity-based approach, may not always be able to distinguish between high-intensity biological echoes and bubble echoes. One way to identify such cases is to track the distri-

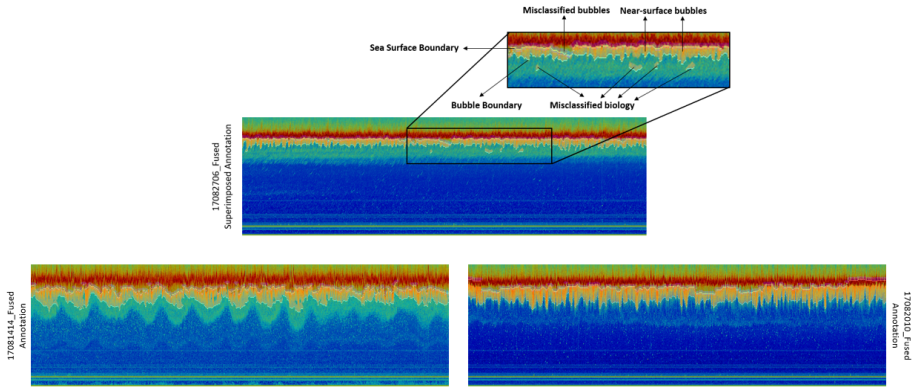

Figure 3.14: *Challenges in the semi-automatic annotation process for bubble detection in marine echograms. This zoomed-in image highlights the potential inaccuracies that can occur during the segmentation. It clearly illustrates the sea surface boundary and bubble boundary as identified by the annotation algorithm. However, it also shows areas where misclassification has occurred, with some biological features being erroneously identified as bubbles ('Misclassified biology') and some bubbles being incorrectly labeled ('Misclassified bubbles').*

bution of the detected bubbles; bubbles generally fade away on their way down. If a sudden increase in intensity is seen and it forms a distinct patch, this may be an indication that the area may not contain bubbles but rather biological entities. Although the combination of k-means clustering and connected component analysis is impactful, it might not detect the biology intertwined with the bubbles, causing misclassification. Figure 3.14 shows these annotation issues that are caused by misclassification of biological features as bubbles and areas above the surface, respectively.

Poor annotations can result in erroneous model outputs [59]. Therefore, high-quality annotations (ground truth) are essential for training successful DL models. As suggested in [14], the results of a deep learning model might potentially rectify the errors that arise from semi-automatic annotation. The model can detect bubbles more precisely and prevent entity misclassification by using its ability to distinguish patterns and learn from complex data.

The insights gleaned from a study by Singh et al. [72], have been instrumental in shaping our approach to enhancing annotation quality in DL models for image analysis. The study's exploration of ReST EM (Reward-weighted Self-Training via Expectation-Maximization) in the context of Large Language Models (LLMs) for tasks such as mathematical reasoning and code generation presents an interesting case for the efficacy of model output feedback and a voting system within a feedback loop. Drawing inspiration from

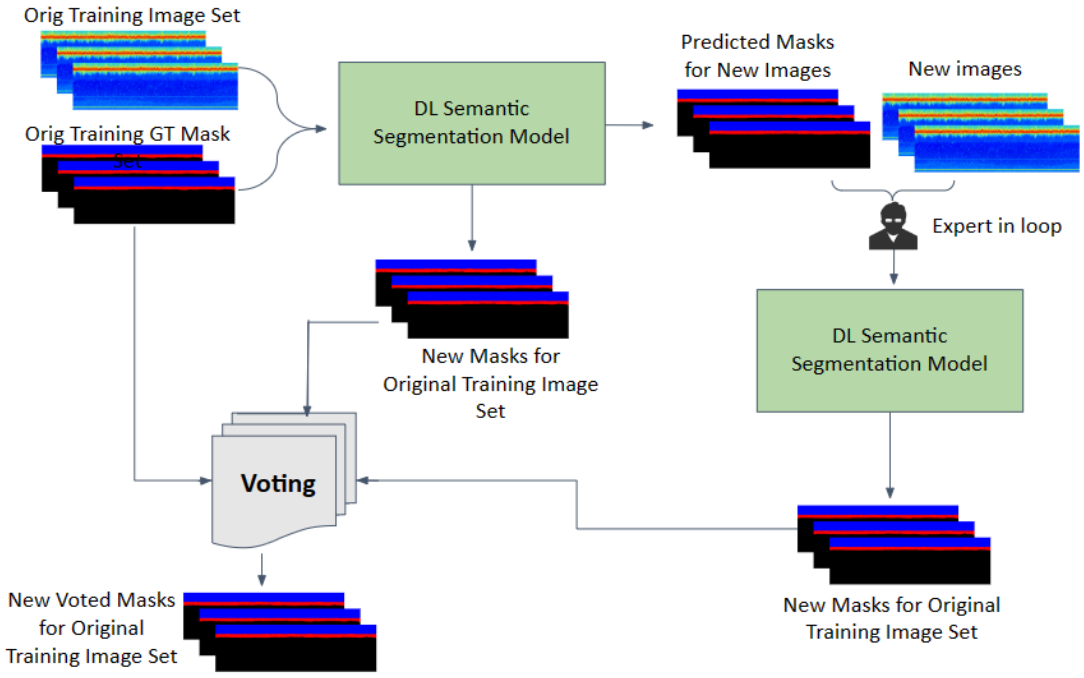

Figure 3.15: *Closed-loop annotation refinement process for DL segmentation model training. This schematic illustrates a semi-supervised loop involving a deep learning segmentation model, expert review, and a voting system to refine training masks for image segmentation, with the goal of enhancing the quality of annotations that better relate to reality.*

the paper, we recognize the parallels in the challenges faced in both language and image models, despite the differences in modalities. Just as ReST EM demonstrated significant performance gains in LLMs by efficiently generating multiple correct solutions to enhance training data, we anticipate a similar enhancement in our DL model's ability to annotate complex marine echograms. By iteratively refining annotations through a feedback loop informed by model outputs and a consensus-driven voting system, we aim to cultivate a robust dataset that mirrors the quality of expert annotations. The feedback loop allows for continuous improvement of the annotations by repeatedly integrating expert review and model output [73].

Figure 3.15 illustrates an iterative process that integrates DL models and expert input to refine annotations for image segmentation tasks. Here is how the process works:

1. Original training data: The process begins with an original set of training images paired with their corresponding masks, which indicate the target areas for segmentation (such as bubbles, area above sea surface boundary, and background in echograms).

2. DL segmentation model training: This original training data is used to train a deep

learning segmentation model. The model consists of input layers, multiple hidden dense layers, and an output layer that generates masks for new, unseen images and is discussed in detail in section 4.3. In this step, we also obtain the masks for the original training images from the model.

3. Mask generation for new Images: The trained DL model is then used to produce masks for a new set of images that were not part of the original training set.

4. Expert in the loop: Subsequent to step 3, an expert reviews the new masks generated by the DL model for the original training set, making adjustments as needed. This step ensures that the expert's knowledge is used to correct any inaccuracies in the model's output. The manual corrections involved in this step take much less time, around 30%, compared to the semi-automatic annotation process.

5. Model retraining with new images: The new images and their masks corrected by the expert are then given to the DL segmentation model as inputs. The model was then tested on the original set of images to generate new masks for the original training set.

6. Voting System: The annotation refinement process utilizes a voting system that compares across three distinct sets of annotations:

   (a) The original annotations created via GUI (step 1).

   (b) The results derived from the first application of the DL model (step 2).

   (c) The annotations produced by a DL model retrained on a new set of 300 echograms, which were initially annotated by the model trained on original data and later refined manually (step 5).

   In this label voting, the final label is determined based on the majority vote among the three annotation sets, and the most often occurring label is considered as the official annotation for that pixel. This way, the final annotations will take into account both the accuracy of the DL models and the ability to discern of the human experts.

7. New Voted Masks: The masks that receive the most votes are considered the new gold standard for the original training image set.

8. Refinement and Reiteration: These new, voted masks then replace the original training masks, and the entire process can be repeated if needed. This iterative loop

allows for continuous improvement of the DL model as it learns from the refined annotations.

It is of no consequence whether the same dataset used for training the model is fed to the model to get outputs, as the objective is to improve annotations rather than testing the model. This approach combines the scalability of the DL model with the nuanced understanding of human experts by involving expert review and a voting mechanism. Therefore, the result is a much better set of annotations which can be used for the continuous improvement of the DL segmentation model. This process was especially important for the sea surface conditions where bubbles were present, namely windy open water and mixed conditions.

Chapter 4

Methodology

The proposed sea surface boundary classification and echogram segmentation techniques for sea surface and bubble detection are covered in detail in this Chapter.

The chapter opens with an examination of the Two-step classification and segmentation process (Section 4.1), a method that initially uses a classification model to classify the echograms into one of the six sea surface conditions classes. Following the classification, the method utilizes a class-specific DL segmentation model for each sea surface condition class to segment the echogram into bubbles, surface boundary, and background classes. The chapter then describes the model utilized for the boundary classification task (Section 4.2). The segmentation task (Section 4.3) is then discussed.

4.1 Two-Step Classification and Segmentation Process

Figure 4.1 shows the workflow of our proposed method.

The process begins with the introduction of raw echogram images into the Deep Learning Image Classification Model which is discussed in Section 4.2. The advantage of this model lies in its capacity to interpret and categorize the complex acoustic signatures in the echograms. This model, trained on a diverse dataset, recognizes different surface environmental conditions, each represented by a specific class. These classes include open water, windy open water, ice with keels, ice without keels, slushy conditions, and mixed conditions as described in Section 3.2.1. This classification is a critical preliminary step, as it segregates the echograms into homogenous groups that exhibit similar features and acoustic behaviors. Such grouping allows for a more targeted and effective subsequent analysis, setting a clear path for the segmentation phase.

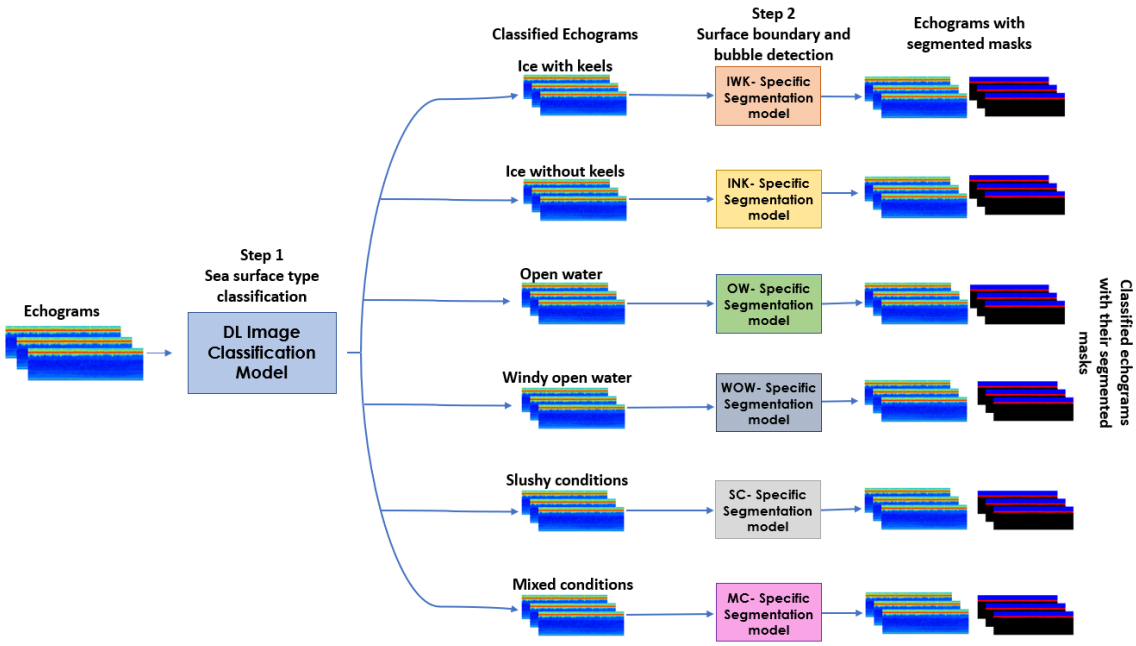

Figure 4.1: *Diagram of the two-step classification and segmentation process for analyzing ocean boundary phenomena in echograms. This flowchart illustrates the initial classification of echograms into six sea surface conditions using a DL Image Classification Model. Following the classification, dedicated DL Image Segmentation Models for each condition produce segmented masks, with a particular focus on enhancing bubble detection in classes where they are prevalent, thereby addressing class imbalance and improving segmentation quality.*

Upon the completion of classification, the echograms are channeled into the second phase of the methodology—segmentation. At this stage, each class of echogram encounters a DL Image Segmentation Model specifically trained and refined for that particular surface condition. The details of these segmentation models are discussed in Section 4.3.

This two-step process results in a collection of classified echograms, each matched with a segmented mask that accurately reflects the specific environmental conditions observed. By sequencing through this process, the methodology not only enhances the precision of the segmentation but also ensures that the responses are highly efficient, and capable of processing vast quantities of data without compromising on the quality of the analysis because of dedicated class-specific segmentation models.

4.2 Sea-Surface Boundary Classification Approach

Echograms are converted into visual representations, or images, from acoustic reflections, or S_v (volume backscattering strength) values (Chapter 3). These echogram images capture unique patterns and traits intrinsic to different maritime settings. Just like how image classification models use visual features to recognize and sort objects in pictures [74], these models can do the same within an echogram by identifying and categorizing them based on their acoustic characteristics.

For the classification task within our end-to-end pipeline, we test a number of advanced deep-learning frameworks, known for their effectiveness in a variety of applications: DenseNet-201 [4], ResNet-101 [2], DarkNet 53 [5], and Inception-v3 [61] on our CBASSA Dataset for Sea-Surface Type Classification (see Section 3.2). These models perform better when transfer learning is applied. Transfer Learning makes use of a network pre-trained on a large dataset like ImageNet [75] (image classification of 1000 classes of natural image objects) to leverage features that have already been learned from a large and diverse dataset. This is very useful for addressing the class-imbalance problem in our small dataset, which is characterized by a large discrepancy in the class representation (e.g., 211 instances of MC against 882 instances of WOW). To further address this, we use class imbalance management approaches.

4.2.1 Architectures Selected for Classification

This section gives more details of the architectures that we selected to work with: DenseNet-201 [4], ResNet-101 [2], DarkNet 53 [5], and Inception-v3 [61].

ResNet-101

ResNet-101 [2] uses residual learning to speed up the training of complex networks. It resolves the vanishing gradient issue by employing skip connections, or shortcut connections. In this network, gradients flow directly through these connections, skipping one or more layers by performing identity mapping. This architecture greatly simplifies the backpropagation of the gradients to the earlier layers; as a result, even the first layers in this network can learn effectively. The main innovation is its residual blocks, where the output from one layer is summed to the output of the following layer, formulated as $x_l = F(x_{l-1}) + x_{l-1}$. This helps the network in the learning process as it learns the residual functions with reference to the input and not the unreferenced function which enables the training of depths of the

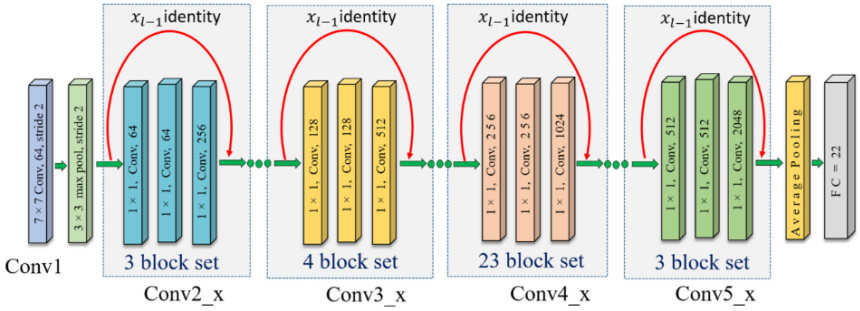

Figure 4.2: *ResNet-101 [2] Architecture (figure reproduced from [3]).*

networks that were not thought to be practical before [3].

The architectural design of ResNet-101 [2], as shown in Figure 4.2, contains 5 convolution stages that are followed by a fully-connected layer and an average pooling layer. A softmax classifier is present in the end to process the extracted features for the classification task. ResNet-101 [2] has 100 convolutional layers and therefore has a deeper representation of features, supported by skip connections which help prevent the decline of model performance due to enhanced depth.

DenseNet-201

The DenseNet-201 [4] structure, as discussed by Huang et al., is a deep convolutional network in which every layer is directly connected to every other layer in a feed-forward fashion. The '201' in DenseNet-201 refers to the total number of layers within the network, indicating its depth and complexity. The word "dense" indicates the dense connection pattern, which guarantees a smooth feature flow throughout the network. This architecture is comprised of the blocks that are densely connected forming feature reusing across the network and the transition blocks which help to control the feature-map size and to reduce computational power. DenseNet-201 [4] architecture is shown in Figure 4.3. The mathematical formulation representing the input to the l^{th} layer, $x_l = H_l([x_0, x_1, x_2, ..., x_{l-1}])$, illustrates the concatenation of feature maps produced by all preceding layers, where H_l is a composite function of operations including batch normalization, ReLU activation, and convolution. All the feature maps acquired by preceding layers were analyzed individually at every layer and given as input after being concatenated into a single tensor. A global average pooling is connected to the last dense block together with a softmax classifier at the end. The trained labels are then classified [3]. When working with very big datasets or when computational resources are restricted, the dense connection might result in greater

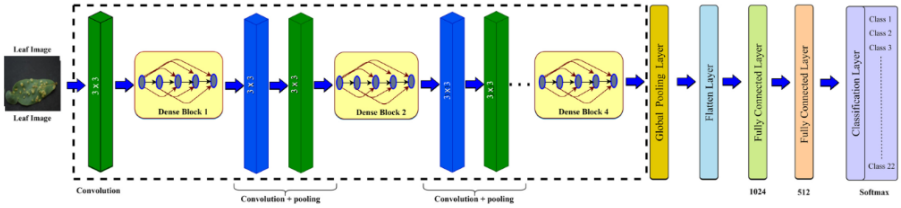

Figure 4.3: *DenseNet-201 [4] Architecture (figure reproduced from [3]).*

memory usage during training but this is not the case with our data.

DarkNet-53

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

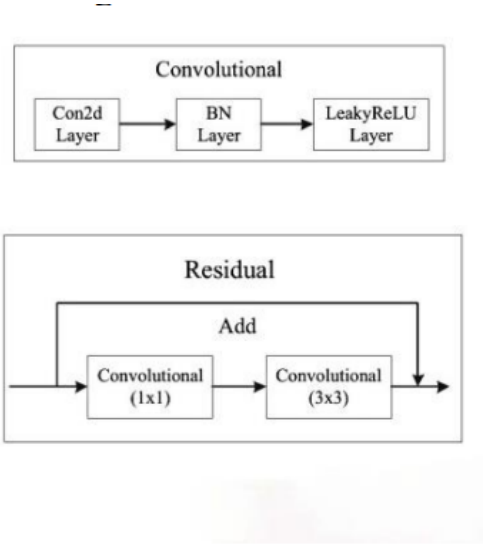

Figure 4.4: *Darknet-53 [5] Structure (figure reproduced from [6]).*

DarkNet-53 [5] tries to find the balance of architectural depth and computational efficiency. The architecture is named ”DarkNet-53” due to its utilization of 53 convolutional layers, which form the backbone of this model. The structure is shown in Figure 4.4. The network features several convolution layer sizes, but uses mostly 1x1 convolutions for the channel reduction while 3x3 convolutions are employed for spatial features extraction and shortcut connections similar to the ones in ResNet. The shortcut connections address the gradient vanishing problem, allowing efficient training of deep networks because gradient flow is observed across the layers. DarkNet-53’s [5] architecture is characterized by successive convolutional blocks that have increasing channel depth and by periodically reducing spatial dimensions through strided convolutions, so the feature maps are concentrated.

The key advantage of this network architecture is that as spatial information becomes increasingly dense, there are no significant compromises in the network's representational capacity. Moreover, DarkNet-53 [5] does not make use of pooling layers, but only uses convolutions with strides to reduce the dimension of the feature map, which is believed to preserve more information. The network model is optimized for performance, with the main focus on increasing throughput and minimizing latency, which is the most important factor in real-time object detection tasks.

Based on the YOLOv3 Object Detection system, this model is specially designed for efficiency, enabling the fast processing of images without compromising precision or depth. On the other hand, this focus for speed may decrease the ability to carry out detailed feature extraction, which might cause limitations in effectiveness in instances where the details are necessary for precise classification and segmentation.

Inception-v3

Adaptive inception modules that capture information at many scales within a single layer are the feature that sets Inception-v3 [61] apart from other approaches. These modules capture information at various scales at the same time within a single layer, which is a key feature that inception-v3 is distinguished from traditional CNNs. The architecture is shown in Figure 4.5. Each inception module consists of parallel paths that use convolutions of different sizes (1x1, 3x3, and 5x5) and a 3x3 max pooling, all operating on the same input level. This type of structure makes the network scalable so that it can perform well for tasks that involve features of different sizes. For instance, marine acoustic imaging where the subjects' sizes are highly variable. The 1x1 convolutions also act as bottleneck layers, reducing dimensionality and thus computational cost before processing by larger convolutions. Inception-v3 [61] also involves factorized convolutions which split the n x n convolution network into smaller asymmetric operations (such as a 3x3 convolution can be split into two pieces: 3x1 and 1x3), saving much space without sacrificing the feature of modeling complex patterns. The addition of batch normalization in both auxiliary classifiers as well to Inception modules themselves assists in better training convergence through the normalization of inputs layer by layer. It also uses label smoothing that helps to improve model generalization and decrease model overfit on training data.

The capacity to adapt to different sizes and scales present in marine acoustic imaging makes Inception-v3 [61] a flexible model. For some applications, particularly those involving highly specialized or unusual data features, Inception-v3's [61] modular design may

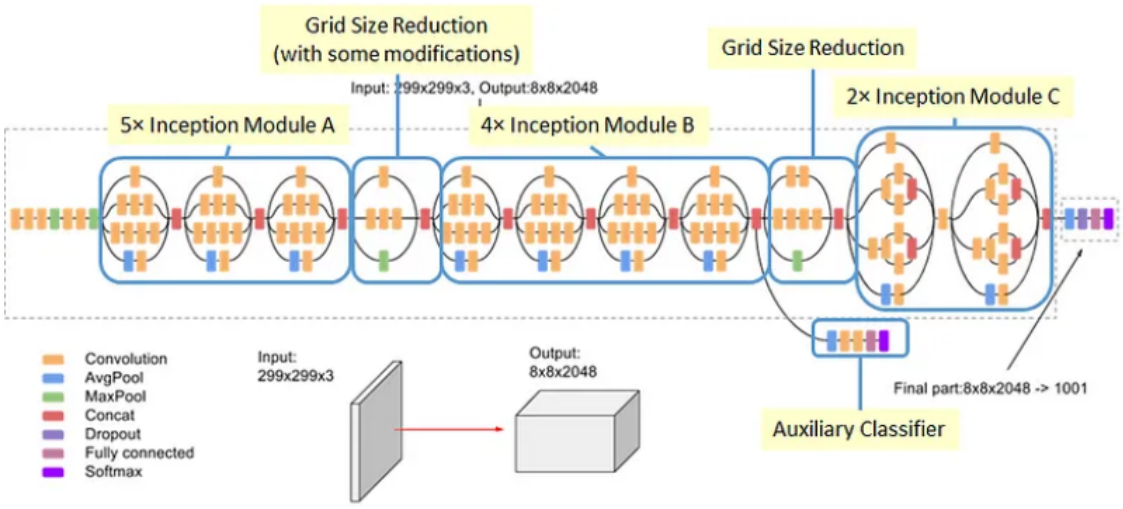

Figure 4.5: *Inception-v3 Architecture, Batch Norm and ReLU are used after Conv (figure reproduced from [7]).*

need more fine-tuning even if it offers flexibility and computational efficiency.

We ended up using the Inception-v3 [61] model as the architecture for the classification task after rigorous testing with all the above-mentioned architectures. This choice is backed up by the quantitative results discussed in Section 5.1.1.

4.2.2 Transfer Learning

The process of pre-training a model which is general to a vast dataset, like ImageNet, and transferring it to a new but related job is called transfer learning (TL). By implementing this approach, one achieves two objectives. Firstly, TL allows us to cut down training time and computer resources. Secondly, it uses the features that the model has learned through previous training on large datasets [76]. TL therefore helps to achieve excellent feature representation for image classification tasks.

Because of pre-training, the model has already acquired a feature hierarchy that is useful for echogram classification. Features that allow the model to identify, for example, textures, edges, and forms in common items are also relevant when identifying patterns and boundaries in echogram images. We aim to adjust these models to the unique characteristics of our acoustic data while preserving their high degree of accuracy and efficiency by fine-tuning them on our echogram dataset [77].

Recent work has improved model adaptability and efficacy to data from different dis-

tributions in the context of transfer learning [78]. For instance-based transfer, algorithms such as TrAdaBoost [79] use training data to identify the most pertinent examples from the source that will help with the target job and lower the error rate. Furthermore, to address domain adaptation, Deep Adaptation Networks [80] and Selective Adversarial Networks [81] have been proposed. These networks allow for the learning of domain-invariant features and partial transfer when only subsets of data are relevant to the source and target domains. While accepting these developments we apply the standard approach to transfer learning with the aid of robust feature extraction features of pre-trained models like Inception-v3. This method, which adds the adaptability of transfer learning with the modification of the final layers of the models, is a basic technique in machine learning. It utilizes the general applicability of features discovered from large datasets and, thus, it is among the essential strategies for achieving high performance with minimal data in new domains.

4.2.3 Class-Weighting

In addressing the challenge of class imbalance within our training data, we adopt a cost-sensitive class weighting approach. This strategy mitigates the disproportionate influence of the majority class [82].

Our approach diverges from traditional methods that primarily focus on modifying the training set. As highlighted by Provost [83], modifying the distribution of the data with no changes in the model to counteract the influence of the imbalance on the classification results risks presenting misleading information. The latest works underpin the strength of weighting or applying thresholds to the continuous classifier output, whose outputs are support function or class probability estimates, as opposed to data resampling methods. Such a methodology could increase the accuracy of a number of conventional types of classifiers without the need of changes to the data distribution of training data [84].

To increase the importance of the positive (minority) class, recalibration of the learning or decision-making process is performed. This recalibration often entails raising the decision threshold to offset the bias towards the negative (majority) class or updating algorithms to include class penalties or weights [82].

Penalties are allocated to each class using a cost matrix in our cost-sensitive learning paradigm. Reduced probability of learners incorrectly categorizing examples from the underrepresented group is the goal of this modification [84]. The minority class's lower representation is counterbalanced by imposing a higher cost on misclassifications of it, which encourages a more equitable learning process.

A major challenge in the process of cost-sensitive learning is establishing a good cost matrix. The cost matrix is based on empirical data, which shows historical data and outcomes, or defined by experts in the field who have deep understanding of the problem space [82]. For our deep learning model, we've taken an empirical approach to design the cost-sensitive class weighting, analyzing our dataset to determine the frequency of each class and adjusting the weights accordingly. This strategy reflects the inherent value and rarity of each class within the dataset, which subsequently allows our model to effectively address the class imbalance problem without skewing the underlying data distribution directly.

4.2.4 Implementation Details

Echograms are categorized into six predetermined categories using the Inception-v3 architecture as part of a transfer learning technique. To ensure a thorough framework for evaluation, the CBASSA-classification dataset (see Section 3.2) is first divided into training (70%), validation (10%), and testing sets (20%). In order to accommodate the input requirements of the Inception-v3 model, images are resized to a total of 299×299 pixels in the three different color channels or RGB.

To enhance the training dataset, data augmentation techniques are utilized, which include random reflections along the x-axis and translations within a [-30, 30] pixel range for both the x and y axes. The goal of this is to increase the model's capacity for generalization. In order to modify the model, two new layers must be integrated: one for total connectedness and the other for classification. In order to meet the particular needs of our classification task, this procedure entails replacing the preexisting "predictions" layer with the newly constructed fully connected layer and replacing the original classification layer with a customized one that takes class weights into account. To improve model accuracy across different classes, the classification layer uses computed class weights to counteract the dataset's class imbalance.

A few variables need to be defined. N represents the total number of images in the dataset; n_i represents the number of images in class i, where i varies between 1 and C; C represents the total number of classes. After adding up all of the images in each class, we get the total number of images N:

$$N = \sum_{i=1}^{C} n_i \tag{4.1}$$

Every class's frequency f_i is obtained by dividing the total number of images by the

number of images in class i.

$$f_i = \frac{n_i}{N} \tag{4.2}$$

Using this frequency as a base, the inverse of the frequency is used to get the class weight w_i for each class i.

$$w_i = \frac{1}{f_i} = \frac{N}{n_i} \tag{4.3}$$

In addition to the architecture discussed above, the other networks discussed in 4.2.1 were also evaluated using transfer learning and class weighting approaches to discriminate echograms into their classes. The comparative analysis of these networks and the rationale behind selecting Inception-v3 as the optimal choice for our application will be discussed in the chapter 5.

4.3 Echogram Segmentation for Near-Surface Phenomena

For the purpose of training, we are employing the CBASSA Dataset for Sea-Surface Boundary and Near-Surface Bubbles Detection, which is covered in Section 3.3. Our specific aim is to identify bubbles and outline the sea surface boundary inside the echograms. We investigate the efficacy of several cutting-edge deep-learning architectures: U-Net [51], Attention U-Net [64], DeepLab-v3 [11], and U-Net++ [65].

Given their demonstrated efficacy in several studies [19, 14, 23, 17], we particularly concentrate on U-Net-like architectures. Our echogram analysis is a good fit for these architectures because they have shown remarkable performance in a variety of segmentation tasks. In addition to U-Net-like architectures, DeepLab-v3—which employs a distinct technique for semantic segmentation—was another architecture we investigated.

As shown in Section 3.3, boundary areas are seen in all echograms, but bubbles are limited to only two of the six classes. In section 3.3.3, we also discuss bubbles, which exhibit a quasi-random occurrence that is difficult to follow because of their fluid nature and form diversity. Out of all these models, we select the one that uses the transfer learning covered in 4.2.2 to achieve better performance.

In order to tackle these issues and guarantee uniform performance in every category, we employ transfer learning and class weighting techniques to correct the distribution of classes. We also use a zoomed-in tiling strategy and custom loss function to accurately capture the bubble shapes given the quasi-randomness of bubbles.

We use the same transfer learning strategy as for classification (see Section 4.2.2) for these segmentation networks, allowing them to gain from the abundance of information stored in the pre-trained models.

4.3.1 Semantic Segmentation Network Architectures

U-Net

In order to produce accurate segmentations with less training photos, U-Net [51] was initially proposed for medical image segmentation. The architecture is encoder-decoder and it is based on the idea of Fully Convolutional Networks (FCNs) [85]. The architecture, in Figure 4.6, showcases a symmetric encoder-decoder structure designed for precision. The shape of the network resembles the letter 'U', therefore, it is called U-Net.

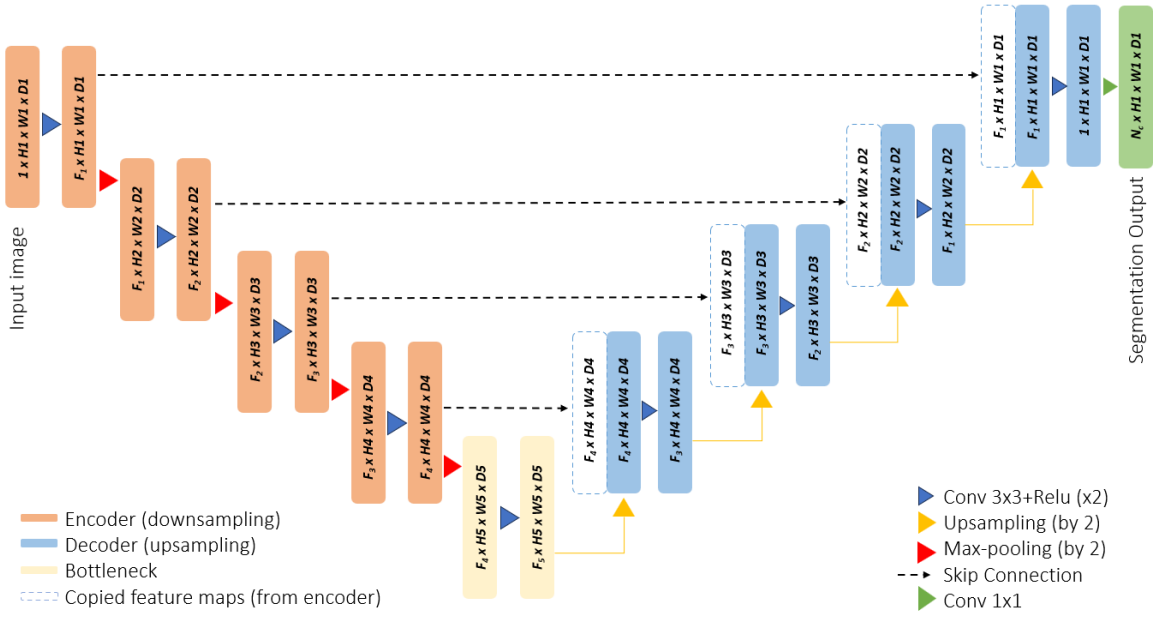

Figure 4.6: *The U-Net architecture, modified from [8, 9], showing the encoder-decoder structure. Throughout the architecture, the notation FxHxWxD describes the dimensions of multi-channel feature maps: "F" stands for the number of feature map channels, while "HxWxD" denotes the spatial dimensions—height, width, and depth—of the image. "Nc" in the final layer indicates the total number of target classes.*

The encoder, or contracting path, consist of two 3x3 convolutions (each followed by ReLU and batch normalization), combined with 2x2 max pooling operations for down-sampling. Each down-sampling step doubles the number of feature channels and reduces

spatial dimensions, encoding the input image into feature representations at multiple levels of abstraction. The bottleneck, an important part of the architecture, compressing and subsequently expanding the feature representations after the encoder has condensed the spatial dimensions. This component acts as a harmonizer of spatial and feature information and encapsulates both the localized and overarching contextual details of the input image. The decoder, or expansive path, uses transposed convolutions to upsample the feature maps, restoring the spatial dimensions. As shown in Figure 4.6, critical to U-Net's [51] design are the skip connections that bridge the encoder and decoder. These connections concatenate the upsampled feature maps with the correspondingly sized feature maps from the contracting path, reintroducing localized spatial features lost during down-sampling. In the end, a 1x1 convolution is utilised to convert the feature vectors from decoder layers into a probability map with N dimensions (N stands for the number of classes to be segmented into). U-Net [51] performs very well even when there is scarcity of data because of it's skip connections at every layer and symmetric structure which is able to preserve high resolution information that might typically be lost through the downsampling process and finer details that could have been lost during the upsampling process.

Attention U-Net

Attention U-Net [64] expands on U-Net by adding attention gate mechanisms to concentrate on particular components of interest. In U-Net, skip connections have the potential to spread unnecessary low-level information, but attention gates actively block activations in unimportant regions, eliminating duplicate features and emphasizing important ones.

Attention mechanisms are present right before the concatenation of the information from encoder layer to decoder layer at each level. These gates are responsible for controlling the flow of information by learning to adapt to important and relevant spatial regions. As illustrated in Figure 4.7, at the decoder, a gating signal, which encompasses context-aware information at a broader scale, is merged with feature maps from the encoder. This combined data then proceeds through an attention gate where it calculates attention scores, α , ranging between 0 and 1. These scores highlight important areas within the feature maps. The feature maps that carry higher-level spatial details are pixel-wise multiplied by these attention scores, thereby sharpening the focus on significant spatial regions of the image.

During training, Attention U-Net [64] learns to give more weight (and therefore more attention) to relevant regions by applying additive soft attention and weighting distinct

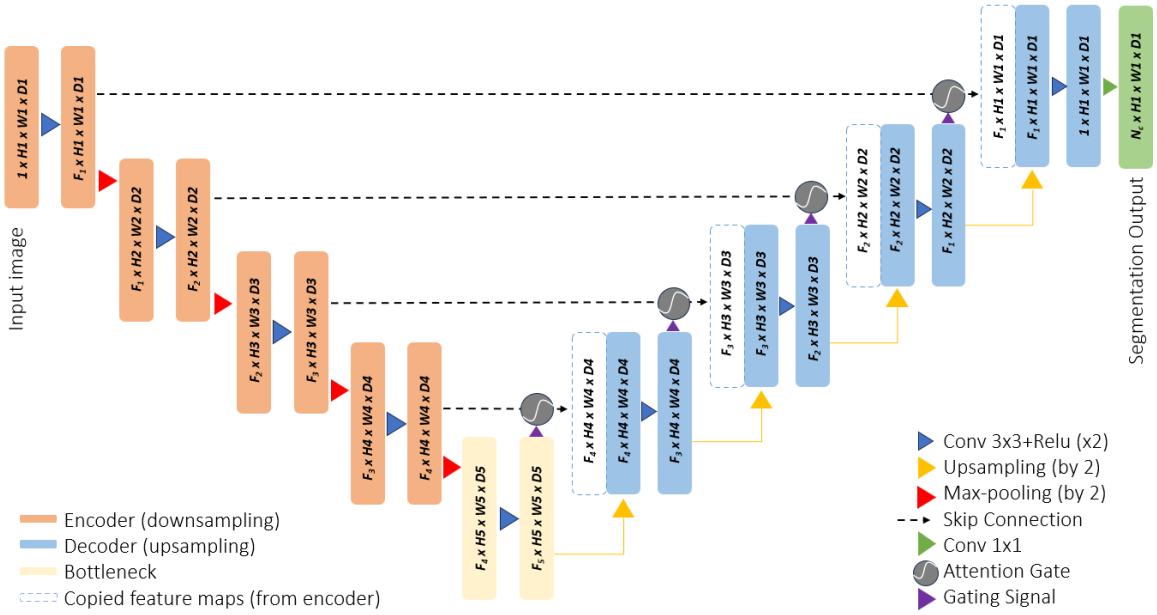

Figure 4.7: *The diagram of Attention U-Net architecture, adapted from [9]. The architecture is similar to the U-Net architecture with the addition of attention mechanism right before concatenation of information from encoder in the skip connections at each level*

sections of the image.

U-Net++

U-Net++ [65] goes a step further than U-Net by including both nested and dense skip connections through its architecture. These particular items allow the model to have to extract multi-scale features. The purpose of the embedded skip connections is the merging of features of different levels of the network hierarchy which makes closer attention to details and overall context possible for the better understanding of the image. Unlike the traditional skip connections in the original U-Net architecture, U-Net++ consists of series of interconnected layers, or convolutional bridges, that give it the nested structure as shown in Figure 4.8. This is primarily to bridge the semantic gap between the encoder and decoder feature maps prior to the fusion.

Moreover, U-Net++ takes advantage of the concept of dense connectivity, which is used by DenseNet, in which each layer receives information from all previous layers and sends its own feature maps to all next layers. This results in a high-dimensional feature space, which combines both high resolution details and semantic contexts. In U-Net++, every skip pathway contributes to the final segmentation map, providing multi-scale information that

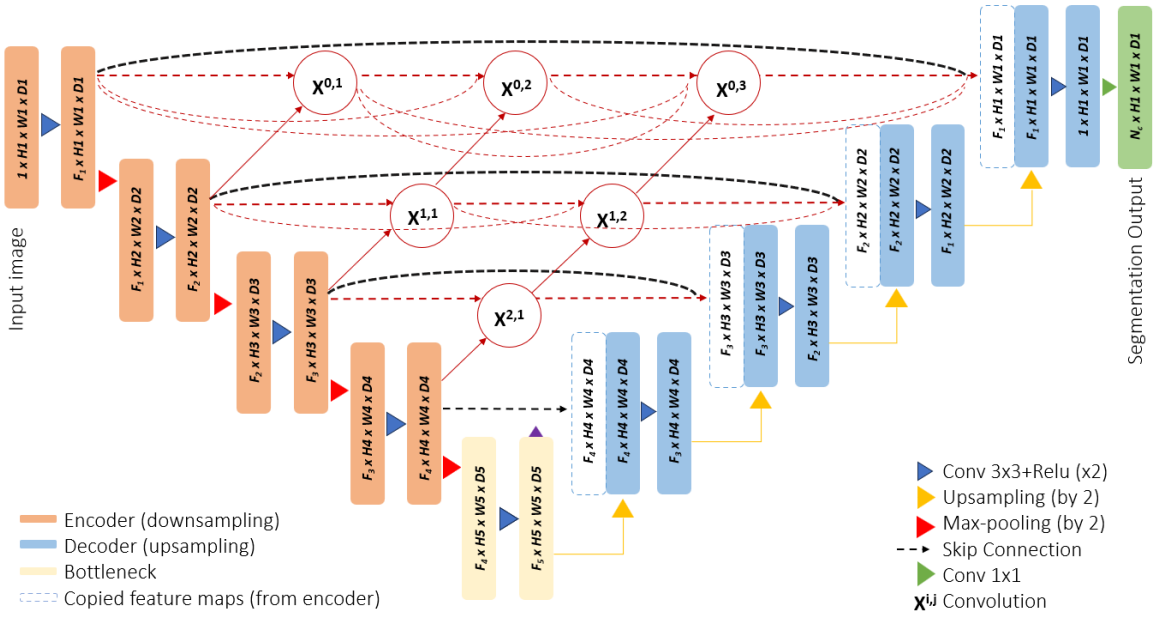

Figure 4.8: *The architecture of UNet++ demonstrating the nested skip connections contributing the information from all the encoder layers to the decoder layers (modified from [10]). The black dotted skip connection are the original skip connection present in U-Net architecture while the red dotted skip connection indicated the newly introduced nested skip connections.*

enhances the network's ability to distinguish between objects of varying sizes. Moreover, the dense skip connections enhances feature reuse and the attainment of feature propagation throughout the network. Such improvements lead to better segmentation, especially in scenarios where the image content is complex or the object size variation is considerable.

DeepLabV3

DeepLabV3 [11] is designed specifically for semantic image segmentation tasks. It is an advanced architecture within the DeepLab series. It performs best when it makes use of atrous convolutions, which improve context capture at various scales without appreciably raising computing complexity. Atrous convolutions, also known as dilated convolutions, expand the receptive field of filters by introducing 'holes' in between filter values. Traditional convolution operations perform a direct element-wise multiplication of the filter over the input feature map; atrous convolutions, however, introduce gaps in the input feature map, covered by the filter to incorporate a wider context without increasing the number of parameters. This technique enables the network to incorporate context from a larger area of

the image, enhancing its ability to understand the relationship between objects at different scales without the computational cost typically associated with enlarging the filter size.

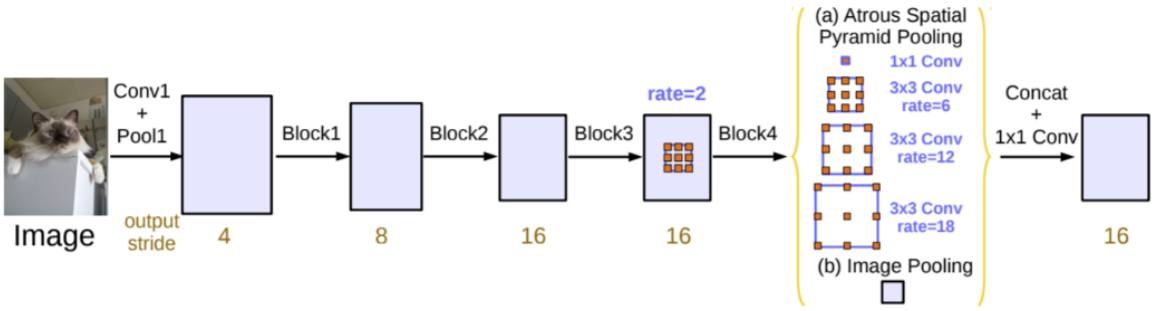

Figure 4.9: *DeepLab-v3 model with ASPP module (sourced from [11]).*

The Atrous Spatial Pyramid Pooling (ASPP) module is also incorporated by DeepLabV3 [11] to enhance its contextual knowledge even more. It systematically applies a set of filters, or kernels, to the input feature map at different scales or rates. , capturing information at multiple scales and then aggregating the resulting feature maps to form a richer semantic feature representation.

The architecture, as seen in Figure 4.9 is basically structured around a backbone network such as VGG [55], DenseNet [4], or ResNet [2], which extracts initial features. Atrous convolutions are then used in the terminal blocks to efficiently gather contextual details without making the model more complex. The Atrous Spatial Pyramid Pooling (ASPP) network is further characterized by accurate pixel level classification across various scales and therefore solving object detection of objects of different sizes. The following 1x1 convolution reshapes the output by matching the initial image size, resulting in a fine output mask. All things considered, DeepLabV3 [11] is well-liked for its outstanding performance in a range of applications needing accurate object identification and visual context understanding.

4.3.2 Zoomed-in Tiling Strategy

Accurate boundary detection is crucial for the segmentation of near-surface events. The model guarantees accurate segmentation of each region by precisely learning the boundaries of the bubble and sea surface. This allows for precise and detailed interpretation of the visual data.

When resizing images from their original 712x201 pixel size to the 768x224 pixel size,

we use a particular tiling technique to improve the accuracy during image segmentation. Each image is divided into six non-overlapping, side-by-side tiles, each measuring 128x128 pixels.

Through focusing on the specific zoomed-in tiles the model is equipped to better learn the nuances of edges and boundaries of the sea surface and near-surface bubbles, both of which are essential aspects of the segmentation process.

To further refine the model's learning efficiency, regions of the image that do not contribute to the analysis of near-surface phenomena are excluded from the training process. Specifically, a strip of the image measuring 768x96 pixels, situated at the bottom, is treated as non-essential and is disregarded when the entire image is presented to the model during the training process. This exclusion not only prevents the model from expending computational resources on irrelevant data but also sharpens its capacity to learn from the most informative parts of the image.

The advantages of tiling are many. First of all, high-resolution images can be handled by the model, thereby resolving the problem of inadequate processing owing to hardware capabilities. The model is exposed to the details in the finer scale which improves its precision. However, this approach can complicate tile management and stitching, possibly affecting context continuity and feature alignment. It is essential to complement the tiling strategy with full images afterward for contextual understanding and to learn from global features that span across the entire image. This subsequent learning also helps overcome fragmentation issues where important features or objects are split between tiles.

4.3.3 Class Weighting Strategy for Bubble and Sea Surface Boundary Classes

We propose a customized class weighting strategy to address the imbalances between the occurrence of bubble and boundary classes. Our strategy combines presence-based and pixel-based techniques.

Pixel Counting for Each Class

Firstly, the pixel count for Class i (c_i) represents the total number of pixels in the image that are part of the class i, where i should be an integer value 1,2,3, respectively, for blue (surface), red (bubble), and black (background). Afterwards, if the image is in RGB format, it can be represented as follows:

$$N = \sum c_i \tag{4.4}$$

The pixel count for class i divided by the overall pixel count (N) yields the class frequency (f_i), which is the frequency of class i in the image. It is represented as follows:

$$f_i = \frac{c_i}{N} \tag{4.5}$$

The class weight (w_i) is then calculated as the inverse of the class i frequency. The computation involves dividing the total number of pixels (N) by the number of pixels for class i, as shown below:

$$w_i = \frac{N}{c_i} \tag{4.6}$$

Lastly, we normalize the weight of class i, and the Normalized Class Weight (w_i^{norm}) is calculated by dividing it by the total of all class weights. The normalized class weight is shown as follows:

$$w_i^{\text{norm}} = \frac{w_i}{\sum w_i} \tag{4.7}$$

Because the weights are scaled proportionately and add up to one when they are normalized, comparing different classes is made simpler and the weights' relative relevance is more easily understood.

Image Counting with Presence of Each Class

The quantity of images in which class i is present is known as the presence count for class i (p_i). The total number of images in the dataset is M We calculate the class presence frequency (f_{p_i}), which is an indicator of how frequently class i appears in the photos.

$$f_{p_i} = \frac{p_i}{M} \tag{4.8}$$

Next, we compute the class weight (w_{p_i}), which is the inverse of the class i presence frequency. The total number of images (M) divided by the presence count for class i is used to calculate it if the presence count for that class is greater than zero; if not, it is set to zero. The following equation represents this:

$$w_{p_i} = \begin{cases} \frac{M}{p_i} & \text{if } p_i > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{4.9}$$

The weight of class i normalized by the sum of all class weights is computed as the normalized class weight (w_{p_i}norm).

$$w_{p_i}^{\text{norm}} = \frac{w_{p_i}}{\sum w_{p_i}} \tag{4.10}$$

Combining Both Sets of Weights

For class i, the final weight (w_i^{final}) is determined by taking the maximum of the normalized presence weight and the normalized pixel count weight.

$$w_i^{\text{final}} = \max(w_{p_i}^{\text{norm}}, w_i^{\text{norm}}) \tag{4.11}$$

The max function involved in calculating the final weight of the class focuses on the disparate nature of class imbalance by taking into account the pixel scarcity as well as the rarity of occurrence in different classes. Thus, it ensures that the training process focuses on the most identifiable form of imbalance for every class and in turn, directly influences what the model learns. By dynamically adjusting to the form of imbalance that poses the greater challenge to model accuracy, the max function serves as a critical mechanism for enhancing detection and classification of sparse features within high-dimensional data spaces, crucial for tasks requiring fine-grained discrimination among classes. This approach in conjunction with a weighted loss function such as cross-entropy will optimize the gradient update process, which shifts towards minimizing error among the under-represented classes directly impacting their specificity and sensitivity.

These final weights are then used by deep learning models in a weighted loss function, like cross-entropy loss.

4.3.4 Custom Boundary Loss

Our goal is to improve the segmentation accuracy by enhancing the model's capacity to collect fine-grained features at class boundaries with the introduction of a custom boundary loss. This custom loss function allows for quantifying the alignment between the true and anticipated border locations.

We use the Sobel operator, one-hot encoding, and edge map calculations. The Sobel operator used for edge detection consists of two 3×3 kernels that are convolved with the image. For a given image A, the horizontal (G_x) and vertical (G_y) derivatives are calculated as follows:

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * A, \quad G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} * A \tag{4.12}$$

where $*$ denotes the convolution operation.

One-Hot Encoding: The predicted classes and the ground truth mask labels are one-hot encoded. For a predicted class c in pixel (i, j), the one-hot encoding vector is e_c where $e_{c_k} = 1$ if $k = c$, otherwise $e_{c_k} = 0$.

One-hot encoding converts categorical class data into a binary format that can be utilized for edge detection using the Sobel operator.

Edge Map Calculation: Edge maps are calculated by convolving the one-hot encoded predicted classes and ground truth mask labels with the Sobel kernels:

$$\begin{aligned} \text{edge\_pred\_x} &= G_x * \text{predicted\_classes\_one\_hot}, \\ \text{edge\_pred\_y} &= G_y * \text{predicted\_classes\_one\_hot}, \\ \text{edge\_mask\_x} &= G_x * \text{mask\_labels\_one\_hot}, \\ \text{edge\_mask\_y} &= G_y * \text{mask\_labels\_one\_hot}. \end{aligned} \tag{4.13}$$

Magnitude of Gradients: The magnitude of the gradient for each pixel is computed, which corresponds to the strength of the edges at that pixel:

$$\begin{aligned} \text{edge\_pred} &= \sqrt{(\text{edge\_pred\_x})^2 + (\text{edge\_pred\_y})^2}, \\ \text{edge\_mask} &= \sqrt{(\text{edge\_mask\_x})^2 + (\text{edge\_mask\_y})^2}. \end{aligned} \tag{4.14}$$

Boundary Loss Calculation: Finally, the boundary loss is computed as the mean absolute error (L1 loss) between the edge magnitudes of the predicted and the ground truth mask:

$$\text{Boundary\_loss} = \frac{1}{N} \sum_{i=1}^{N} \left| \text{edge\_pred}_i - \text{edge\_mask}_i \right| \tag{4.15}$$

where N is the total number of pixels in the image.

Having a custom boundary loss enables a more refined optimization process, improving the segmentation model's capacity to precisely identify class boundaries.

Accordingly, the total loss is the sum of the cross-entropy loss and the custom boundary loss that will contribute to the model optimization with an improved pixel-wise classification and boundary delineation effect and more detailed segmentation outputs.

To represent the combined loss function in mathematical format, we can denote the total loss as $\mathcal{L}_{\text{total}}$. This total loss comprises two components: the cross-entropy loss ($\mathcal{L}_{\text{CE}}$) and the custom boundary loss ($\mathcal{L}_{\text{CB}}$). The cross-entropy loss is weighted by the class weights (w^{final}).

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CB}} + \mathcal{L}_{\text{CE}} \tag{4.16}$$

where

$$\mathcal{L}_{\text{CB}} = \text{CustomBoundaryLoss} = \frac{1}{N}\sum_{i=1}^{N}\left|\text{edge\_pred}_i - \text{edge\_mask}_i\right| \tag{4.17}$$

$$\mathcal{L}_{\text{CE}} = \text{CrossEntropyLoss}(\text{weight} = w^{\text{final}}) \tag{4.18}$$

Further expanding, the cross-entropy loss $\mathcal{L}_{\text{CE}}$ with weights can be expressed as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{M}\sum_{i=1}^{M} w_{y_i} \cdot \log\left(\frac{\exp(x_{i,y_i})}{\sum_{c=1}^{C}\exp(x_{i,c})}\right) \tag{4.19}$$

The weighted cross-entropy loss L_{CE} is computed for an entire batch of image data, where M represents the total number of pixels within that batch. Each pixel's contribution to the loss is adjusted by a weight w_{y_i}, which correlates to the class label y_i of the i-th pixel. The model outputs logits $x_{i,c}$ for each class c at each pixel i, which are the raw predictions prior to applying the softmax function. The actual class label for each pixel is denoted by y_i, and C signifies the total number of classes. This formulation ensures that the loss calculation directly reflects the accuracy of class predictions across all pixels, weighted appropriately to enhance model performance on imbalanced data.

The weights w_{y_i} are computed based on the class weights w_j^{final} as described in the equation:

$$w_{y_i} = w_{y_j}^{\text{final}} \tag{4.20}$$

where,

$$w_{y_j}^{\text{final}} = \left[\max(w_{p_j}^{\text{norm}}, w_j^{\text{norm}})\right] \tag{4.21}$$

(from equation 4.11)

Here, $w_{y_j}^{\text{final}}$ represents the final class weight associated with the ground truth class label y_i where i represents i-th pixel and j represents index of class. w_{y_i} the weight for the i-th pixel's class label is derived directly from a precomputed array of final class weights $w_{y_j}^{\text{final}}$, which is obtained by selecting the maximum value between the normalized presence weight $w_{p_j}^{\text{norm}}$ and the normalized pixel count weight w_j^{norm} (see section 4.3.3). The maximum function is used to ensure that the class weight chosen for each class is the highest between the pixel weight and the occurrence weight.

4.3.5 Implementation

We experimented and compared all the models described in section 4.3.1. However, following preliminary testing, U-Net++ showed the most encouraging outcomes. As a result, we moved forward with U-Net++ as our base architecture and applied novel strategies to increase confidence in the results obained from our two-step pipeline. We provide a more thorough justification of our choice of U-Net++ in chapter 5. We used an 80%–20% split between training and testing data while processing the CBASSA Dataset for Sea-Surface Boundary and Near-Surface Bubbles Detection. We also applied random horizontal flipping as a data augmentation technique. We start with a pre-trained U-Net++ model (pretrained on the ImageNet dataset [75]) and then apply a two-step transfer learning (TL) procedure to improve its performance. First, as described in subsection 4.3.2, we improve the performance of the model by employing the zoomed tiling learning technique for images and masks from the CBASSA Dataset for Sea-Surface Boundary and Near-Surface Bubbles Detection. This enables the model to recognize fine details within smaller image tiles. Increased sensitivity to the fine-grained features in the data is made possible by this targeted training. The model is then exposed to full-sized images and masks as we move on to the second stage of transfer learning. By doing this, we increase the model's comprehension to include the images' larger spatial context. By utilizing the knowledge acquired during the tiling learning stage, we adjust the model to comprehend and evaluate the entire image. This step-by-step training technique, which starts with a pretrained U-Net++ architecture and advances through tiling learning to full-size image learning, makes it easier to create a segmentation model that is both extremely flexible and efficient. The model

demonstrates a feature that enables both high-level contextual awareness and low-level detail recognition, leading to remarkable abilities in segmenting images of intricate structures and elements.

Furthermore, we used the class weighting strategy and our custom boundary loss to optimize the model's performance. This way, multiple strategies get incorporated into the model to maximize its strength and reliability. As we discussed, the class weighting approach (CW) entails assigning different weights to various classes depending on their frequency or significance in the data set. These weights are fed into the cross-entropy loss function which makes the loss function concentrate on the classes that are less represented and, at the same time, improve the overall segmentation accuracy.

Chapter 5

Results and Discussion

This section reports the experimental findings related to identifying sea surface boundaries and detecting near-surface bubbles within underwater Arctic echograms. It encompasses an evaluation of the sea surface type classification phase, comparing different architectural models, and similarly evaluates the echogram segmentation phase across various models. It also presents the overall results of end-to-end pipeline and includes findings from an ablation study to assess the impact of individual components. For the classification of sea surface types, the implementation was carried out in MATLAB, utilizing the Stochastic Gradient Descent with Momentum (SGDM) as the optimization technique. The training process was performed with a mini-batch size of 10, being limited to a maximum of 20 epochs, and the initial learning rate was set to 1×10^{-4}. The segmentation was implemented in Python with the PyTorch framework by applying the Adam optimizer algorithm and an initial learning rate set on 1×10^{-3}. This task was taken up considering a batch size of 2, for 200 epochs. When in use, the Zoomed-in tiling strategy covered the first 50 epochs. The restriction of the zoomed-in tiling strategy to the first 50 epochs optimizes the model's learning phase, capturing the important details, and avoiding the overfitting risk observed on 100 epoch tiling. This timeframe balances feature absorption details and general understanding of image context, ensuring that the model doesn't get overly focused on non-generalizable details.

5.1 Quantitative evaluation

The classification task is assessed using the standard accuracy, recall, and precision metrics. The evaluation of the segmentation task involves using both standard metrics (intersection

over union (IoU), recall, precision) to compare the predicted segmentation masks against the ground truth masks, and specific metrics that give deeper insights into how accurately the model traces the sea surface boundary and tracks bubble distribution over time. These include the overall mean vertical distance (OMVD), which measures the average vertical gap in pixels between the predicted and actual boundaries (with each pixel equal to 25.3 cm), and the relative error (RE), which normalizes this gap by the predicted mask's thickness. Additionally, false positives time-wise (FP-T) and false negatives time-wise (FN-T) measure the model's accuracy over time by counting the time instances (or pings, corresponding to the number of columns in the echogram) where the model either incorrectly identifies a segment or misses one, respectively, highlighting the model's ability to maintain consistency across temporal sequences. For example, an FP-T score of 13 for bubbles means the model wrongly detected the presence of bubbles in 13 out of 712 pings (with 712 pings representing an hour's worth of data or the echogram's width). An FN-T score of 13 for bubbles means the model failed to detect the presence of bubbles in 13 out of 712 pings, even though there was at least one pixel indicating bubbles in each of those pings. Intersection over Union (IoU) provides a ratio of correctly segmented areas, essential for validating the model's efficacy in distinguishing between bubbles, surfaces, and background. Lastly, precision ensures the relevance of detected segments, and recall confirms the model's ability to capture all pertinent features, crucial for comprehensive detection of bubbles and boundaries within the echogram data. Here, these metrics are computed echogram-wise and then averaged over the entire test set for each task.

5.1.1 Sea surface type classification

Table 5.1 shows the performance of various image classification architectures (ResNet-101 [2], Darknet-53 [5], DenseNet-201 [4], and Inception-v3 [61]) for the sea surface type classification problem on the test set, for each of the six classes and overall. ResNet-101 excels in the OW class with the highest precision and accuracy, and it leads in accuracy and recall for SC. Darknet-53, while competitive, does not secure the top position in any case. DenseNet-201 achieves the highest recall for OW, and it stands out with the highest precision for WOW and SC. Inception-v3 outperforms others in IK, INK, and MC across all three metrics, indicating its robustness in complex classifications. Additionally, Inception-v3 achieves the overall highest scores in recall, precision, and accuracy, with a notable second-highest performance in accuracy and recall for OW and accuracy and recall for WOW, making it a clear choice for step 1 of our end-to-end pipeline.

	Metric	OW	WOW	IK	INK	SC	MC	Overall
RN101	Acc ↑	**0.943**	**0.966**	0.942	0.955	**0.946**	0.965	0.858
	Rec ↑	0.791	**0.943**	0.870	0.832	**0.854**	0.738	0.838
	Prec ↑	**0.887**	0.922	0.865	0.848	0.793	0.689	0.834
DN53	Acc ↑	0.916	0.950	0.952	0.963	0.930	0.963	0.835
	Rec ↑	0.814	0.875	0.929	0.832	0.378	0.643	0.805
	Prec ↑	0.750	0.922	0.861	0.848	0.776	0.692	0.814
DN201	Acc ↑	0.917	0.955	0.952	0.966	0.944	0.970	0.852
	Rec ↑	**0.899**	0.875	0.955	0.812	0.698	0.714	0.826
	Prec ↑	0.716	**0.939**	0.845	0.943	**0.902**	0.769	0.852
IncV3	Acc ↑	0.930	0.956	**0.967**	**0.977**	0.943	**0.976**	**0.877**
	Rec ↑	0.837	0.926	**0.974**	**0.842**	0.777	**0.837**	**0.853**
	Prec ↑	0.794	0.911	**0.888**	**1.000**	0.825	**0.821**	**0.873**

Notes: OW: open water, WOW: windy open water, IK: ice with keels, INK: ice without keels, SC: slushy conditions, MC: mixed conditions, Acc: accuracy, Rec: recall, Prec: precision, RN101: ResNet-101 [2], DN53: Darknet-53 [5], DN201: DenseNet-201 [4], IncV3: Inception-v3 [**?**].

Table 5.1: Performance evaluation of various image classification architectures for the sea surface type classification problem on the test set. Best results in bold font, selected architecture underlined.

5.1.2 Echogram segmentation

Table 5.2 shows the per-class performance of various semantic segmentation architectures (U-Net [51], Attention U-Net [64], DeepLabV3 [11], and UNet++ [65]) for the echogram segmentation problem on the WOW test set. In order to select the best architecture for the proposed full end-to-end pipeline, the experiments targeted WOW conditions as these include all pixel classes. UNet++ distinguishes itself, especially in segmenting surfaces, with the lowest OMVD and RE, indicating precise boundary capture and shape consistency. It also excels with the highest IoU and precision for bubbles and highest IoU and recall for surfaces. Interestingly, U-Net performs very differently for surfaces and bubbles, with the highest precision and lowest recall for surfaces, and the highest recall and lowest precision for bubbles. Attention U-Net does not yield any of the best metrics, but generally improves upon U-Net's performance for surfaces. DeepLabV3 performs best only in terms of FP-T for bubbles. The bubble class appears harder to segment compared to the surface class with overall lower metrics values; this is also illustrated by the fact that all architectures were able to prevent false detection time wise (FP-T and FN-T) for surfaces. Overall, UNet++'s comprehensive performance across key metrics positions it as the optimal model for step 2 or our end-to-end pipeline.

Class	Metric	UNet [51]	A-UNet [64]	DLV3 [11]	UNet++ [65]
Surf	OMVD ↓	10.603	1.304	0.920	**0.571**
	RE ↓	0.382	0.032	0.022	**0.014**
	FP-T ↓	**0.000**	**0.000**	**0.000**	**0.000**
	FN-T ↓	**0.000**	**0.000**	**0.000**	**0.000**
	IoU ↑	0.751	0.969	0.978	**0.986**
	Recall ↑	0.751	0.980	0.986	**0.995**
	Prec ↑	**1.000**	0.989	0.992	0.991
Bub	OMVD ↓	6.712	4.432	2.039	**1.112**
	RE ↓	0.300	0.416	0.363	**0.294**
	FP-T ↓	**0.366**	8.666	132.030	25.710
	FN-T ↓	210.150	177.430	**0.560**	25.300
	IoU ↑	0.243	0.417	0.424	**0.605**
	Recall ↑	**0.960**	0.817	0.470	0.663
	Prec ↑	0.243	0.445	0.576	**0.773**

Notes: Surf: sea surface, Bub: bubble, OMVD: overall mean vertical distance, RE: relative error, FP-T: false positives time-wise, FN-T: false negatives time-wise, IoU: intersection over union, Prec: precision, A-UNet: Attention U-Net, DLV3: DeepLabV3.

Table 5.2: Performance evaluation of various architectures for the echogram segmentation problem on the WOW test set. Best results in bold font, selected architecture underlined.

5.1.3 Single vs. multiple segmentation models

Table 5.3 compares the performance of the six sea surface type-specific segmentation models (per model, overall utilizing the sea surface boundary classification ground truth (called "Step 1 GT"), and overall utilizing the full pipeline, (called "End-to-end, proposed")), to that of a single global model trained on all data (all sea surface types at once), for the echogram segmentation problem on the test set. All models are built upon the Inception-v3 and UNet++ frameworks and incorporate the "Zoom" (zoomed-in tiling), "BL" (custom boundary loss), and "CW" (class-weighting) strategies. The "Step 1 GT" scenario eliminates the potential for errors carried over from the initial sea surface type classification, while the 'End-to-end, proposed' scenario demonstrates the final output. The table indicates that the models perform exceptionally well for classes not involving bubbles, achieving high scores in recall, precision, and IoU. For bubble-associated classes (WOW and MC), the task is more complex, as bubbles are harder to segment (detailed in Sec. 5.1.2). The models excel in the WOW category, showcasing low RE and satisfactory OMVD, coupled with strong recall—our primary metric to ensure no bubble is missed—and supported by reasonable IoU and precision. Segmentation results by the MC model are less remark-

Model	Pixel Class	OMVD ↓ (pixels)	RE ↓	FN-T ↓ (columns)	FP-T ↓ (columns)	IoU ↑	Recall ↑	Precision ↑
OW-specific	Surface	0.423	0.010	0.000	0.000	0.990	0.996	0.995
WOW-specific	Surface	0.592	0.014	0.000	0.000	0.986	0.990	0.996
	Bubble	1.093	0.197	7.030	54.080	0.663	0.908	0.700
IK-specific	Surface	0.945	0.021	0.000	0.000	0.979	0.992	0.987
INK-specific	Surface	0.423	0.010	0.000	0.000	0.990	0.997	0.993
SC-specific	Surface	0.615	0.015	0.000	0.000	0.985	0.997	0.988
MC-specific	Surface	0.596	0.014	0.000	0.000	0.985	0.991	0.994
	Bubble	1.329	0.306	24.930	41.760	0.533	0.777	0.650
Overall	Surface	0.413	0.010	0.000	0.000	0.990	0.995	0.995
(Step 1 GT)	Bubble	1.170	0.232	13.000	49.970	0.620	0.864	0.683
Overall	Surface	**0.650**	**0.015**	**0.000**	**0.000**	**0.985**	**0.993**	**0.992**
(End-to-end, proposed)	Bubble	**1.231**	**0.223**	**29.730**	41.680	**0.598**	**0.813**	**0.663**
Single	Surface	2.380	0.054	**0.000**	0.336	0.972	0.983	0.989
	Bubble	6.319	1.946	222.358	**8.896**	0.243	0.270	0.610

Notes: OW: open water, WOW: windy open water, IK: ice with keels, INK: ice without keels, SC: slushy conditions, MC: mixed conditions, GT: ground truth annotations, OMVD: overall mean vertical distance, RE: relative error, FN-T: false negatives time-wise, FP-T: false positives time-wise, IoU: intersection over union.

Table 5.3: Performance evaluation of the six sea surface type-specific models per model, overall (utilizing sea surface boundary classification ground truth) and overall (full end-to-end pipeline, proposed) vs. a single global model for the echogram segmentation problem on the test set. Best results for each pixel class, between overall (proposed) and single, shown in bold font.

able, attributable to the complexity of learning a class combining multiple conditions.

The Step 1 GT approach demonstrates impressive results, significantly outperforming the single-model approach, particularly in the recall. This indicates the strength of type-specific models in handling segmentation when the correct classification is assured. The metrics for the proposed end-to-end pipeline, especially for bubbles, exhibit minor variations (2-3 points) in IoU and precision from the "Step 1 GT" approach, with a noticeable adjustment in recall due to the misclassification within the test set, underscoring the challenge of accurately classifying various sea conditions. However, for surfaces, the metrics remain consistent between "Step 1 GT" and the end-to-end pipeline, with a negligible difference (0.05 points). We conclude that there is an error propagation effect, with errors in sea surface type classification from step 1 affecting the end results of the proposed pipeline, but of limited scope. Utilizing a single model applied globally without the classification step shows a significant drop in performance, with OMVD increasing fivefold, decreases in recall and IoU (significant for bubbles), and a marked increase in FN-T for bubbles, compared to the proposed end-to-end approach. This illustrates the advantages of our end-to-end approach's specialized models over a generalized single model, ensuring more a accurate echogram segmentation.

5.1.4 Comparative Performance Analysis

Exp	Zoom	BL	CW	OMVD ↓ (pixels)	RE ↓	FN-T ↓ (columns)	FP-T ↓ (columns)	IoU ↑	Recall ↑	Precision ↑
1				1.112	0.294	25.710	**25.300**	0.605	0.663	0.773
2	✓			1.076	0.257	21.230	25.880	0.659	0.765	**0.804**
3		✓		**1.036**	0.245	23.860	28.150	0.628	0.716	0.797
4			✓	4.102	0.541	2.166	169.58	0.561	0.856	0.591
5	✓	✓		1.102	0.243	23.310	29.980	0.647	0.737	0.765
6	✓		✓	1.060	0.250	28.35	28.65	0.636	0.731	0.771
7		✓	✓	1.149	0.198	15.066	47.280	0.622	0.809	0.676
8 (proposed)	✓	✓	✓	1.093	**0.197**	**7.030**	54.08	**0.663**	**0.908**	0.700

Notes: Exp: experiment, Zoom: learning features on zoomed-in echogram tiles first, BL: custom boundary loss, CW: class weighting, OMVD: overall mean vertical distance, RE: relative error, FN-T: false negatives time-wise, FP-T: false positives time-wise, IoU: intersection over union.

Table 5.4: Comparative Performance Analysis of the proposed segmentation model on the WOW test set for the bubble pixel class. Best results in bold font. Experiment 1: baseline UNet++, experiment 8: proposed approach.

Table 5.4 details an exhaustive Comparative Performance Analysis on the proposed echogram segmentation model, focusing on the challenging task of identifying bubbles in the WOW test dataset. The study spans eight experiments, each testing different combinations of three key strategies in the enhanced UNet++ framework: pre-training on zoomed-

Exp	Zoom	BL	CW	OMVD ↓ (pixels)	RE ↓	FN-T ↓ (columns)	FP-T ↓ (columns)	IoU ↑	Recall ↑	Precision ↑
1				0.571	0.014	**0.000**	**0.000**	0.986	**0.995**	0.991
2	✓			**0.465**	**0.011**	**0.000**	**0.000**	**0.989**	**0.995**	0.994
3		✓		0.569	0.014	**0.000**	**0.000**	0.986	0.993	0.993
4			✓	2.528	0.040	**0.000**	**0.000**	0.957	0.990	0.966
5	✓	✓		0.485	0.012	**0.000**	**0.000**	0.988	**0.995**	0.994
6	✓		✓	0.521	0.013	**0.000**	**0.000**	0.988	**0.995**	0.993
7		✓	✓	0.686	0.017	**0.000**	**0.000**	0.984	0.987	**0.996**
8 (proposed)	✓	✓	✓	0.592	0.014	**0.000**	**0.000**	0.986	0.990	**0.996**

Notes: Exp: experiment, Zoom: learning features on zoomed-in echogram tiles first, BL: custom boundary loss, CW: class weighting, OMVD: overall mean vertical distance, RE: relative error, FN-T: false negatives time-wise, FP-T: false positive time-wise, IoU: intersection over union.

Table 5.5: Comparative Performance Analysis for windy open water conditions, *sea surface pixel* class, on the test sets. Best results shown in bold font. Experiment 1: original UNet++, experiment 8: proposed approach.

in echogram sections (“Zoom”), integrating a custom boundary loss into the loss function (“BL”), and implementing a class weighting system (“CW”). Experiment #1 serves as the baseline with the standard UNet++ setup, and experiment #8 incorporates all strategies, embodying the full proposed method. The baseline model (#1) is effective in reducing false positives over time (“FP-T”), “Zoom” (#2) leads to higher precision, “BL” (#3) achieves the best boundary delineation (“OMVD”), and “CW” (#4) initially decreases performance. Nonetheless, combining “CW” with “Zoom” and “BL” significantly improves results (notable when compared to #5, which excludes ’CW’), positioning our full method (#8) as the most effective, with the highest scores in relative error (RE), false negatives over time (FN-T), intersection over union (IoU), and recall. The table confirms the superiority of our comprehensive method over the basic UNet++ model for segmenting complex features in echograms, especially for detecting bubbles—a critical aspect of analyzing Arctic underwater imagery in detail. The surface class results in table 5.5, stable across experiments, highlight Experiment #8’s superior precision. The metrics FP-T and FN-T would always be zero since they indicate that the surface was consistently present, and none of the pings showed an absence of surface in either the ground truth or the predictions. Therefore, these metrics do not contribute to understanding the model’s performance for this class. Although slightly trailing in recall and IoU, these figures support the model’s robustness, with bubble segmentation metrics proving pivotal.

5.2 Qualitative Evaluation

Figure 5.1 displays outcomes from a sea surface type classification task within Arctic underwater echograms, processed through the Inception-v3 model. The echograms demonstrate the model's high accuracy, with the predicted sea surface conditions aligning with the ground truth for the majority of the images. However, there is an exception where the model incorrectly identifies 'Slushy Conditions' as 'Windy open water,' likely due to the echogram's surface appearing fuzzy and irregular, which is characteristic of both conditions and may confuse the model.

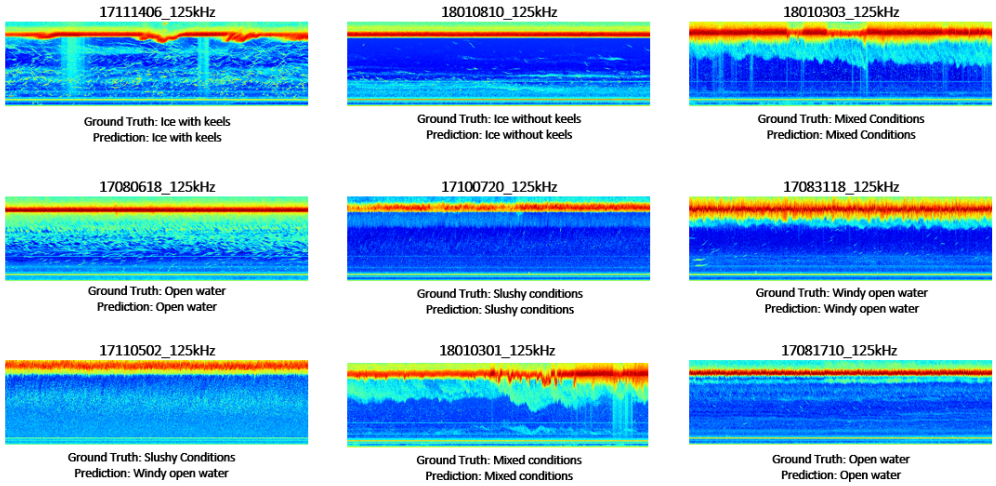

Figure 5.1: *Classification of arctic sea surface conditions using Inception-v3 model on underwater echograms. The images depict accurate model predictions for various conditions.*

Figure 5.2 shows typical segmentation results of fused multi-frequency echograms for all compared architectures and for the proposed approach. Compared to other models, ours exhibits the closest alignment with the ground truth, particularly in the accurate delineation of boundaries, essential for the purpose of finding the sea surface line and bubble extent. U-Net tends to underdetect surface pixels and overdetect bubbles. In contrast, the Attention-UNet shows a more nuanced segmentation performance, but with less precision in the boundary. This indicates a potential compromise between detecting a higher number of pixels and maintaining segmentation integrity. Compared to the UNet and Attention-UNet, the UNet++ model achieves a significantly higher IoU and precision for bubbles, indicating a marked improvement in segmentation visually. However, the proposed UNet++ model edges out slightly in terms of finer boundaries, as reflected in the IoU, RE, and OMVD metrics for bubbles in Table 5.2. DLV3, when compared to the proposed UNet++,

visually demonstrates improved performance over the A-UNet but does not quite match the proposed model's performance. The qualitative analysis positions UNet++ as the leading model, with its segmentation fidelity and quantitative results confirming its effectiveness.

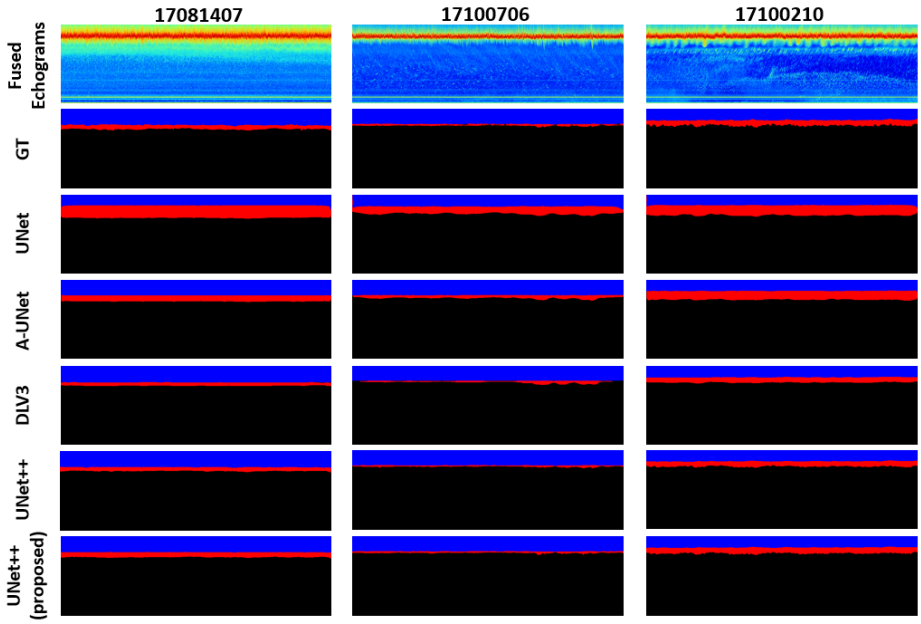

Figure 5.2: Sample segmentation results (rows 3 to 7) for fused echograms (row 1) across models, with ground truth (GT) masks in row 2. Color code of pixel masks: background (black), blue (surface), red (bubbles). Timestamp: YYMMDDHH.

Our end-to-end pipeline, as depicted in the figure 5.3, demonstrates exceptional proficiency in segmenting and classifying various sea ice conditions as seen in 5.3(a), 5.3(c), 5.3(e), and 5.3(d) as well as complex scenarios like "Windy open water" and "Mixed conditions" with accuracy, as seen in 5.3(b), 5.3(f), and 5.3(g) supported by the ablation studies in tables 5.4 and 5.5. Classification results are also precise, matching ground truth for challenging labels such as "Mixed Conditions" and "Slushy Conditions" in 5.3(e) and 5.3(f), confirmed by Table 5.1. However, Figure 5.3(h) presents a rare but noteworthy exception, where a misclassification occurs, leading to an inaccurate segmentation output. This instance highlights a limitation of our pipeline when encountering dissimilar classes. When a misclassification happens, especially between categories with significant differences such as those with bubbles (windy open water and mixed conditions) and those without (ice with keels, ice without keels, open water, slushy conditions), the segmentation model is deprived of the opportunity to correctly identify and segment bubbles. This is because the model tailored for non-bubble classes lacks the necessary training to recognize and segment bubble features, leading to a complete omission of these critical elements in the output. Similarly,

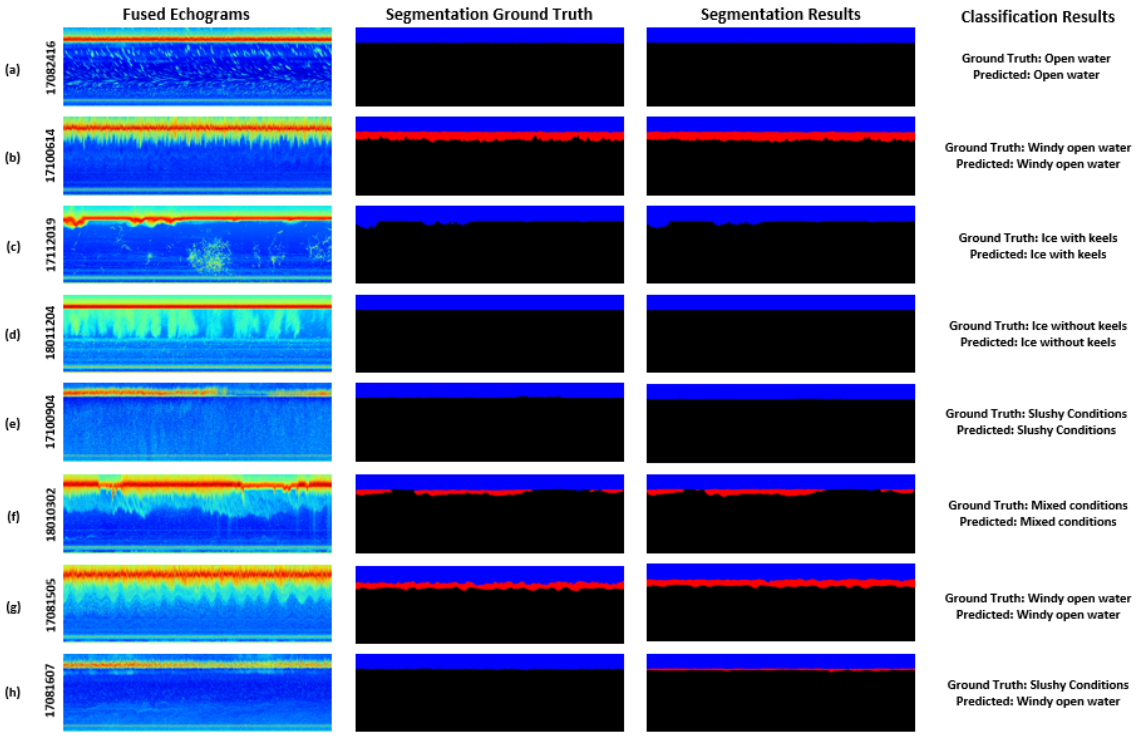

Figure 5.3: Additional sample results of the proposed method for both classification (GT: ground truth, pred: predicted) and segmentation, for various sea surface types. Color code of pixel masks: background (black), blue (surface), red (bubbles). Timestamps (YYMMDDHH): (a) 17082416, (b) 17100614, (c) 17112019, (d) 18011204, (e) 17100904, (f) 18010302, (g) 17081505, (h) 17081607. OW: open water, WOW: windy open water, IK: ice with keels, INK: ice without keels, SC: slushy conditions, MC: mixed conditions.

if a non-bubble class is incorrectly identified as one containing bubbles, the segmentation model may erroneously attempt to segment bubbles where none exist. Despite this, the infrequency of such errors underscores the reliability and overall accuracy of the pipeline in the challenging task of learning sea boundary phenomena.

This combination of the Inception-v3 model's feature extraction capabilities and the proposed UNet++ model's segmentation accuracy results in a robust pipeline adept at both segmenting and classifying echograms from diverse marine environments.

Chapter 6

Conclusions and Future Work

Researching biological and physical phenomena in oceans relies heavily on echosounders. These instruments are essential to oceanographic research and fish stock management, and they are used by both commercial fisheries and scientific ocean observatories. The significant amount of manual labor required to choose and analyze the data is what, significantly slows down the process.

Using two specialized models—Inception-v3 for identifying different types of sea surfaces and the proposed UNET++ for accurate segmentation of sea surfaces and bubbles in echograms—this research provided an end-to-end DL pipeline that addresses the challenge as a two-step procedure. Using Inception-v3, distinct conditions such as open water, windy open water, ice with keels, ice without keels, mixed condition, and slushy condition were recognized in the echograms. Simultaneously, the segmentation of echograms into categories such as bubbles, sea surface, and background was achieved using multi-frequency fused data through the application of the proposed UNet++ model. This model incorporates three distinct learning strategies. Firstly, a class weighting strategy was employed to address the issue of class imbalance ensuring that minority classes receive appropriate attention during the learning process. Secondly, a zoomed-in tiling strategy is utilized, focusing on the fine details of segment boundaries to improve the model's accuracy in delineating these areas. Lastly, a custom loss function is designed to minimize the discrepancies between the predicted boundaries and the actual ground truth, thereby refining the model's segmentation capabilities. The CBASSA dataset is used to compare several state of the art deep learning models, such as Inception-v3 [61] and UNet++ [65].

The two-step integrated approach is designed such that the initial step lays the foundational groundwork, setting the stage for a more detailed and focused investigation in the subsequent phase. This method addresses the challenges encountered in the analysis

of echograms, particularly where differentiating between visually similar sea states proves difficult, and where distinguishing bubbles from marine life near the surface is challenging due to their similar signal features and structures. It also solves the time-consuming issue of manual annotations and analysis by biologists which involve the elimination of personal subjectivity in the annotation of these echograms.

Another potential future research matter can be addressed through investigating mixed sea surface conditions which are observed in the form of the blending of various states and often remain underrepresented in data sets. Collecting more data in these co-occurring scenarios and applying ground truth refinement, utilizing aspects of model outputs for annotation, could substantially improve the efficiency of the pipeline in general.

Often, when bubbles and biological signals are detected together, the latter can wrongly be excluded from biomass calculations, leading to less reliable biological assessments. One of the possible future targets for exploration could be correctly identifying the near-surface life forms and bubbles. By analyzing complex patterns and textures unique to biological entities intertwined with bubbles. This approach would enable the calculation of more reliable biomass estimates and support biological studies.

Bibliography

[1] S Rousseau, S Gauthier, S Johnson, C Neville, and M Trudel. Juvenile salmon acoustic monitoring in the Discovery Islands, British Columbia. Technical Report No. 3277, Fisheries and Oceans Canada - Pêches et Océans Canada, 2018.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] Vaibhav Tiwari, Rakesh Chandra Joshi, and Malay Kishore Dutta. Deep neural network for multi-class classification of medicinal plant leaves. *Expert Systems*, 39(8):e13041, 2022.

[4] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[5] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[6] Zhao Wei, Yan Xu Zhu, Qi Xuan Li, and Cai Wang. Improved smoking target detection algorithm based on yolov3. In *Journal of Physics: Conference Series*, volume 1883, page 012052. IOP Publishing, 2021.

[7] S. H. Tsang. Review: Inception-v3 — 1st runner up of image classification in ilsvrc 2015. https://sh-tsang.medium.com/review-inception-v3-1st-runner-up-image-classification-in-ilsvrc-2015-17915421f77c, 2015. Accessed: 2023-03-17.

[8] 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.

[9] Yanbo Gao, Jiping Guan, Fuhan Zhang, Xiaodong Wang, and Zhiyong Long. Attention-unet-based near-real-time precipitation estimation from fengyun-4a satellite imageries. *Remote Sensing*, 14(12):2925, 2022.

[10] Al-Akhir Nayan, Boonserm Kijsirikul, and Yuji Iwahori. Mediastinal lymph node detection and segmentation using deep learning. *IEEE Access*, 10:89289–89307, 08 2022.

[11] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[12] David Lemon, Paul Johnston, Jan Buermans, Eduardo Loos, Gary Borstad, and Leslie Brown. Multiple-frequency moored sonar for continuous observations of zooplankton and fish. In *IEEE Oceans*, pages 1–6. IEEE, 2012.

[13] Xavier Lurton. *An Introduction to Underwater Acoustics: Principles and Applications*. Springer Science Business Media, 2002.

[14] R. Vohra, F. Senjaliya, M. Cote, A. Dash, A. B. Albu, J. Chawarski, and K. Ersahin. Detecting underwater discrete scatterers in echograms with deep learning-based semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 375–384, 2023.

[15] Hydroacoustic Data Processing - Echoview. https://echoview.com/, 2023. Accessed: 2023-02-20.

[16] Timothy K Stanton. 30 years of advances in active bioacoustics: a personal perspective. *Methods in Oceanography*, 1:49–77, 2012.

[17] Tunai Porto Marques, Melissa Cote, Alireza Rezvanifar, Alex Slonimer, Alexandra Branzan Albu, Kaan Ersahin, and Stéphane Gauthier. U-msaa-net: A multiscale additive attention-based network for pixel-level identification of finfish and krill in echograms. *IEEE Journal of Oceanic Engineering*, 2023.

[18] Alireza Rezvanifar, Tunai Porto Marques, Melissa Cote, Alexandra Branzan Albu, Alex Slonimer, Thomas Tolhurst, Kaan Ersahin, Todd Mudge, and Stéphane Gauthier. A deep learning-based framework for the detection of schools of herring in echograms. *arXiv preprint arXiv:1910.08215*, 2019.

[19] Scott C Lowe, Louise P McGarry, Jessica Douglas, Jason Newport, Sageev Oore, Christopher Whidden, and Daniel J Hasselman. Echofilter: A deep learning segmention model improves the automation, standardization, and timeliness for post-processing echosounder data in tidal energy streams. *Frontiers in Marine Science*, 9:867857, 2022.

[20] Jean-Michel A Sarr, Timothee Brochier, Patrice Brehmer, Yannick Perrot, Alassane Bah, A Sarré, MA Jeyid, M Sidibeh, and S El Ayoubi. Complex data labeling with deep learning methods: Lessons from fisheries acoustics. *ISA transactions*, 109:113–125, 2021.

[21] 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.

[22] Changkyu Choi, Michael Kampffmeyer, Nils Olav Handegard, Arnt-Børre Salberg, and Robert Jenssen. Deep semisupervised semantic segmentation in multifrequency echosounder data. *IEEE Journal of Oceanic Engineering*, 48(2):384–400, 2023.

[23] Alex L Slonimer, Stan E Dosso, Alexandra Branzan Albu, Melissa Cote, Tunai Porto Marques, Alireza Rezvanifar, Kaan Ersahin, Todd Mudge, and Stéphane Gauthier. Classification of herring, salmon, and bubbles in multifrequency echograms using u-net neural networks. *IEEE Journal of Oceanic Engineering*, 2023.

[24] Mark V Trevorrow, Svein Vagle, and David M Farmer. Acoustical measurements of microbubbles within ship wakes. *The Journal of the Acoustical Society of America*, 95(4):1922–1930, 1994.

[25] Joseph D Warren, Timothy K Stanton, Peter H Wiebe, and Harvey E Seim. Inference of biological and physical parameters in an internal wave using multiple-frequency, acoustic-scattering data. *ICES Journal of Marine Science*, 60(5):1033–1046, 2003.

[26] Janet Coetzee. Use of a shoal analysis and patch estimation system (shapes) to characterise sardine schools. *Aquatic Living Resources*, 13(1):1–10, 2000.

[27] Ketil Malde, Nils Olav Handegard, Line Eikvil, and Arnt-Børre Salberg. Machine intelligence and the data-driven future of marine science. *ICES Journal of Marine Science*, 77(4):1274–1285, 2020.

[28] Wu-Jung Lee and Valentina Staneva. Compact representation of temporal processes in echosounder time series via matrix decomposition. *The Journal of the Acoustical Society of America*, 148(6):3429–3442, 2020.

[29] Zhilun Zhang, Yining Yu, Mohammed Shokr, Xinqing Li, Yufang Ye, Xiao Cheng, Zhuoqi Chen, and Fengming Hui. Intercomparison of arctic sea ice backscatter and ice type classification using ku-band and c-band scatterometers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2021.

[30] Timothy K Stanton. 30 years of advances in active bioacoustics: A personal perspective. *Methods in Oceanography*, 1:49–77, 2012.

[31] D Reid, Carla Scalabrin, Pierre Petitgas, Jacques Massé, Richard Aukland, Pablo Carrera, and S Georgakarakos. Standard protocols for the analysis of school based data from echo sounder surveys. *Fisheries Research*, 47(2-3):125–136, 2000.

[32] John K Horne. Acoustic approaches to remote species identification: a review. *Fisheries oceanography*, 9(4):356–371, 2000.

[33] Rolf J Korneliussen and Egil Ona. An operational system for processing and visualizing multi-frequency acoustic data. *ICES Journal of Marine Science*, 59(2):293–313, 2002.

[34] Alex De Robertis, Denise R McKelvey, and Patrick H Ressler. Development and application of an empirical multifrequency method for backscatter classification. *Canadian Journal of Fisheries and Aquatic Sciences*, 67(9):1459–1474, 2010.

[35] RJ Korneliussen, E Ona, I Eliassen, Y Heggelund, R Patel, OR Godø, C Giertsen, D Patel, E Nornes, T Bekkvik, et al. The large scale survey system-lsss. In *Proceedings of the 29th Scandinavian Symposium on Physical Acoustics, Ustaoset*, volume 29, 2006.

[36] J Michael Jech and William L Michaels. A multifrequency method to classify and evaluate fisheries acoustics data. *Canadian Journal of Fisheries and Aquatic Sciences*, 63(10):2225–2235, 2006.

[37] IR Higginbottom, TJ Pauly, and DC Heatley. Virtual echograms for visualization and post-processing of multiple-frequency echosounder data. In *Proceedings of the Fifth European Conference on Underwater Acoustics, ECUA*, pages 1497–1502, 2000.

[38] Stéphane Gauthier, Johannes Oeffner, and Richard L O'Driscoll. Species composition and acoustic signatures of mesopelagic organisms in a subtropical convergence zone, the new zealand chatham rise. *Marine Ecology Progress Series*, 503:23–40, 2014.

[39] Roland Proud, Richard Mangeni-Sande, Robert J Kayanda, Martin J Cox, Chrisphine Nyamweya, Collins Ongore, Vianny Natugonza, Inigo Everson, Mboni Elison, Laura Hobbs, et al. Automated classification of schools of the silver cyprinid rastrineobola argentea in lake victoria acoustic survey data using random forests. *ICES Journal of Marine Science*, 77(4):1379–1390, 2020.

[40] 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.

[41] P LeFeuvre, GA Rose, R Gosine, R Hale, W Pearson, and R Khan. Acoustic species identification in the northwest atlantic using digital image processing. *Fisheries Research*, 47(2-3):137–147, 2000.

[42] Aymen Charef, Seiji Ohshimo, Ichiro Aoki, and Natheer Al Absi. Classification of fish schools based on evaluation of acoustic descriptor characteristics. *Fisheries Science*, 76:1–11, 2010.

[43] Annalisa Minelli, Anna Nora Tassetti, Briony Hutton, Gerardo N Pezzuti Cozzolino, Toby Jarvis, and Gianna Fabi. Semi-automated data processing and semi-supervised machine learning for the detection and classification of water-column fish schools and gas seeps with a multibeam echosounder. *Sensors*, 21(9):2999, 2021.

[44] M Woillez, PH Ressler, CD Wilson, and JK Horne. Multifrequency species classification of acoustic-trawl survey data using semi-supervised learning with class discovery. *The Journal of the Acoustical Society of America*, 131(2):EL184–EL190, 2012.

[45] Sebastián A Villar, Adrián Madirolas, Ariel G Cabreira, Alejandro Rozenfeld, and Gerardo G Acosta. Ecopampa: A new tool for automatic fish schools detection and assessment from echo data. *Heliyon*, 7(1), 2021.

[46] Jie Yang, Jeffery A Nystuen, Stephen C Riser, and Eric I Thorsos. Open ocean ambient noise data in the frequency band of 100 hz–50 khz from the pacific ocean. *JASA Express Letters*, 3(3), 2023.

[47] 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.

[48] Anas Yassir, Said Jai Andaloussi, Ouail Ouchetto, Kamal Mamza, and Mansour Serghini. Acoustic fish species identification using deep learning and machine learning algorithms: A systematic review. *Fisheries Research*, 266:106790, 2023.

[49] Yudai Hirama, Soichiro Yokoyama, Tomohisa Yamashita, Hidenori Kawamura, Keiji Suzuki, and Masaaki Wada. Discriminating fish species by an echo sounder in a set-net using a cnn. In *2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES)*, pages 112–115. IEEE, 2017.

[50] Changkyu Choi, Michael Kampffmeyer, Nils Olav Handegard, Arnt-Børre Salberg, Olav Brautaset, Line Eikvil, and Robert Jenssen. Semi-supervised target classification in multi-frequency echosounder data. *ICES Journal of Marine Science*, 78(7):2615–2627, 2021.

[51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[52] Alex L Slonimer, Melissa Cote, Tunai Porto Marques, Alireza Rezvanifar, Stan E Dosso, Alexandra Branzan Albu, Kaan Ersahin, Todd Mudge, and Stéphane Gauthier. Instance segmentation of herring and salmon schools in acoustic echograms using a hybrid u-net. In *2022 19th Conference on Robots and Vision (CRV)*, pages 8–15. IEEE, 2022.

[53] Savannah J Sandy, Seth L Danielson, and Andrew R Mahoney. Automating the acoustic detection and characterization of sea ice and surface waves. *Journal of Marine Science and Engineering*, 10(11):1577, 2022.

[54] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[55] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[56] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[57] Junyi Chai, Hao Zeng, Anming Li, and Eric WT Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134, 2021.

[58] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, et al. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

[59] Jeff Heaton. Ian goodfellow, yoshua bengio, and aaron courville: Deep learning: The mit press, 2016, 800 pp, isbn: 0262035618. *Genetic programming and evolvable machines*, 19(1-2):305–307, 2018.

[60] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[61] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[62] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022.

[63] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[64] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018.

[65] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.

[66] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. arxiv 2018. *arXiv preprint arXiv:1804.03999*, 1804.

[67] Nasa worldview. https://worldview.earthdata.nasa.gov/, 2024. Accessed: [Insert date here].

[68] Shutao Li, Xudong Kang, Leyuan Fang, Jianwen Hu, and Haitao Yin. Pixel-level image fusion: A survey of the state of the art. *information Fusion*, 33:100–112, 2017.

[69] Rolf Adams and Leanne Bischof. Seeded region growing. *IEEE Transactions on pattern analysis and machine intelligence*, 16(6):641–647, 1994.

[70] Nameirakpam Dhanachandra, Khumanthem Manglem, and Yambem Jina Chanu. Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54:764–771, 2015.

[71] Lifeng He, Xiwei Ren, Qihang Gao, Xiao Zhao, Bin Yao, and Yuyan Chao. The connected-component labeling problem: A review of state-of-the-art algorithms. *Pattern Recognition*, 70:25–43, 2017.

[72] Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, et al. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2023.

[73] Igor Ryazanov, Amanda T Nylund, Debabrota Basu, Ida-Maja Hassellöv, and Alexander Schliep. Deep learning for deep waters: an expert-in-the-loop machine learning framework for marine sciences. *Journal of Marine Science and Engineering*, 9(2):169, 2021.

[74] Òscar Lorente, Ian Riera, and Aditya Rana. Image classification with classic and deep learning techniques, 2021.

[75] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[76] Abeer Saber, Mohamed Sakr, Osama M Abo-Seida, Arabi Keshk, and Huiling Chen. A novel deep-learning model for automatic detection and classification of breast cancer using the transfer-learning technique. *IEEE Access*, 9:71194–71209, 2021.

[77] Mohammadreza Iman, Hamid Reza Arabnia, and Khaled Rasheed. A review of deep transfer learning and recent advancements. *Technologies*, 11(2):40, 2023.

[78] Cheng Wang, Delei Chen, Lin Hao, Xuebo Liu, Yu Zeng, Jianwei Chen, and Guokai Zhang. Pulmonary image classification based on inception-v3 transfer learning model. *IEEE Access*, 7:146533–146541, 2019.

[79] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, pages 193–200, 2007.

[80] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[81] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Partial transfer learning with selective adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2724–2732, 2018.

[82] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

[83] Foster Provost. Machine learning from imbalanced data sets 101. 2008.

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

[85] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.