Detecting Underwater Discrete Scatterers in Echograms with Deep Learning-Based Semantic Segmentation

Rhythm Vohra*, Femina Senjaliya*, Melissa Cote*, Amanda Dash*†, Alexandra Branzan Albu*,
Julek Chawarski†, Steve Pearce†, Kaan Ersahin†
* Electrical and Computer Engineering, University of Victoria, Victoria, Canada
† ASL Environmental Sciences, Victoria, Canada

Email: {rhythmvohra, feminasenjaliya, mcote, aalbu}@uvic.ca, {adash, jchawarski, spearce, kersahin}@aslenv.com

Abstract

This paper reports on an exploratory study of the automatic detection of discrete scatterers in the water column from underwater acoustic data with deep learning (DL) networks. Underwater acoustic surveys using moored singlebeam multi-frequency echosounders make environmental monitoring tasks possible in a non-invasive manner. Discrete scatterers, i.e., individual marine organisms, are particularly challenging to detect automatically due to their small size, sometimes overlapping tracks, and similarity with various types of noise. As our interest lies in identifying the presence and general location of discrete scatterers, we propose the use of a semantic segmentation paradigm over object detection or instance segmentation, and compare several state-of-the-art DL networks. We also study the effects of early and late fusion strategies to aggregate information contained in the multi-frequency data. Experiments on the Okisollo Channel Underwater Discrete Scatterers dataset, which also include schools of herring and juvenile salmon, air bubbles from wave and fish school activity, and significant noise bands, show that late fusion yields higher metrics, with DeepLabV3+ outperforming other networks in terms of precision and intersection over union (IoU) and Attention U-Net offering higher recall. The detection of discrete scatterers is a good example of a problem for which exact annotations cannot be reached due to various reasons; in several cases, network outputs seem visually more adequate than the annotations (which contain inherent noise). This opens up the way for utilizing actual detection results to improve the annotations iteratively.

1. Introduction

This paper deals with the automatic detection of underwater discrete scatterers, i.e., single marine organisms, from multi-frequency echograms, through an exploratory study of the deep learning (DL)-based semantic segmentation paradigm.

1.1. Context

Underwater acoustic surveys allow for the collection of high spatio-temporal resolution data that enable marine biologists and oceanographers to perform a variety of non-invasive tasks crucial for environmental monitoring. Echosounders measure acoustic backscatter from the water column. Multi-frequency echosounders ping the water column by emitting series of acoustic pulses at different frequencies, listening for the echoes from potential targets between the pulses. The basic idea is that aquatic organisms with an acoustic impedance different from that of the surrounding body of water, when subjected to a pressure wave, scatter the wave in a characteristic way [16] according to the pulse frequency. Several factors influence backscatter characteristics, including the acoustic instrument, environment, and target size, shape, and material properties [16].

Data from moored multi-frequency singlebeam echosounders are typically visualized as sets of single-frequency 2D images called echograms, in which the x axis represents different pings over time (temporal unit) for one given frequency and the y axis represents the depth or range from the instrument in the water column (distance unit). The pixel intensity is color-coded to represent the amplitude of reflected echoes at that frequency at a given time in a given (small) volume (although residual echoes from other frequencies may also be present, appearing as bands of noise [33]), generally computed as the volume backscattering strength (called S_v). Echograms are, to this day, mostly interpreted with manual or semi-automatic methods based on statistical characteristics of aggregations of organisms [36] using commercial software (e.g., Echoview [9]), a time-consuming and error-prone process. There is a critical need for automatic processing and analysis methods of echograms with respect to efforts in species abundance

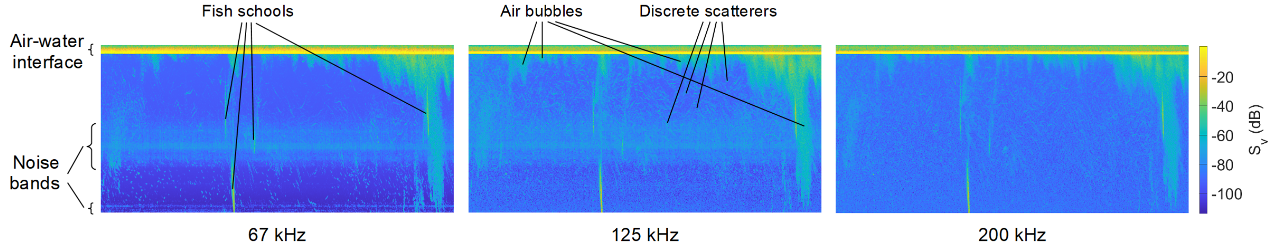

Figure 1. Sample one-hour multi-frequency echogram illustrating some of the challenges in detecting discrete scatterers. The yellow band represents the air-water interface, fish schools appear as yellowish vertical aggregations, air bubbles as greenish/cyan aggregations near the air-water interface or fish schools, and numerous discrete scatterers, more intense here at 125 and 200 kHz, as cyan small objects. Noise bands, visible mostly at 67 and 125 kHz, also appear in cyan. Volume backscattering strength (S_v) is displayed in the "parula" colormap.

tracking and environmental monitoring.

In this paper, we are interested in identifying discrete scatterers in echograms using DL, in particular semantic segmentation networks, which assign a class label to each image pixel. "Discrete scatterers" here refers to individual organisms, as opposed to aggregations (be they fish schools or cloud-like aggregations of zooplankton). They differ on several aspects in terms of their (small) size, morphology, density, and distribution, and thus present specific challenges. In particular, they tend to yield similar backscattering strength compared to some types of noise, such as residual echoes from other frequencies or sidelobe artifacts (see [21] for information on sidelobes), school contours partially enclosed by the echosounder beam, as well as air bubbles from wave and fish school activity. Fig. 1 illustrates this challenge with an example of a noisy multi-frequency echogram containing discrete scatterers (small cyan objects) as well as fish schools and air bubbles.

1.2. Usecase: Jellyfish

This study focuses on jellyfish as a usecase for the detection of discrete scatterers. Insights into jellyfish distribution is of economic and ecological significance, as they are important consumers in pelagic food webs and in some cases, pose risks to finfish aquaculture. Measuring them acoustically is appealing as there are no standardized methods for sampling them, in part due to the difficulty of capturing them with nets. Jellyfish are considered as "weak scatterers" [7] due to their gelatinous bodies similar to the surrounding sea-water in density/sound speed. Their backscattering strength is generally weaker than that of swim-bladdered fish due to their high water content [26]. Their individual tracks are typically recognizable as long, thin, and faint, since jellyfish move slowly and remain in the echosounder beam for a long time. Jellyfish are also typically entrained in currents, so they follow similar motions across depths. The opening and closing of their bell during swimming create a sizable change in their backscattering strength [7, 26]. We hypothesize that (some) semantic segmentation networks can cope with the variability and small size of jellyfish tracks, given the proper training data.

1.3. Contributions

Our contributions are two-fold: 1) We propose the use of a semantic segmentation paradigm for the pixel-level detection of discrete scatterers in noisy echograms and present an experimental design that compares state-of-the-art DL architectures, outlining their strengths and weaknesses for this challenging application. 2) We generate fused inputs from multi-frequency volume backscattering strength data and study the effects of early vs. late fusion on the pixel-level detection of discrete scatterers in echograms.

To the best of the authors' knowledge, this is the first work reporting on discrete scatterer detection in echograms utilizing DL methods. This problem constitutes a good example for a class of problems in which exact annotations cannot be reached for various reasons, such as the discrete scatterers' characteristics, significant noise in the data, and the absence of actual ground truth. Our annotations are of high enough quality for DL training purposes; however, during the experimental evaluation, we need to carefully compare the outputs with their corresponding annotations, as some divergence may be motivated by errors/noise in the annotation process. Ongoing work is looking at how to take advantage of both human expertise and DL to improve upon the annotation and validation processes.

The remainder of this paper is as follows: Sec. 2 reviews related works on the detection of marine species from acoustic data, Sec. 3 presents our proposed methodology including details on the dataset, Sec. 4 discusses experimental results, and Sec. 5 provides concluding remarks.

2. Related works

Traditionally, acoustic classification of (mostly aggregations of) pelagic species from echograms rely on characteristics that can be morphometric (i.e., related to the geometry of the aggregations), bathymetric (i.e., related to the position in the water column), and/or energetic (i.e., related to the signal properties) [15, 30, 36]. Hand-crafted features derived from a combination of those characteristics are then fed to machine learning classifiers (e.g., [25, 29, 33]). Other conventional multi-frequency approaches focus solely on energetic characteristics and rely on the relative or differential/combined frequency response of different species (e.g., [8, 34]). Specific to jellyfish, conventional methods include those of Mutlu [26], whose experiments were designed to estimate the target strength of the common jellyfish in the Black Sea with a dual-frequency echosounder, of Brierley et al. [3], who worked on estimating the target strength of two species of tethered and free-swimming jellyfish in Namibian waters with single- and multi-frequency echosounders, respectively, and of Colombo et al. [7], who focused on establishing a link between four jellyfish species and sound-scattering layers in echograms.

DL-based semantic segmentation has a proven track record in various applications in the visible spectrum and beyond, notably in natural and medical image segmentation [1]. Acoustic classification from echograms using DL-based semantic segmentation has only recently been gaining some traction. Brautaset et al. [2] proposed a semantic segmentation network based on the architecture of the popular U-Net model [31] for the detection of schools of sandeel. They trained their model with crops from multi-frequency echograms to classify each pixel as sandeel, background, or other, improving upon the traditional school classification algorithm of Korneliussen et al. [17]. Ordoñez et al. [28], re-using Brautaset et al.'s model, examined preparation strategies of echosounder data for DL, and found that providing the network with auxiliary information related to the range improved classification performance. Recently, Choi et al. [6] also focused on the identification of schools of sandeel with U-Net, however within a semi-supervised framework. Leveraging a small amount of annotated data via supervised DL and a large amount of readily available unannotated data via unsupervised DL, they proposed two objective functions – an unsupervised clustering objective and a supervised segmentation objective – to alternately optimize the network at the end of the decoder part, achieving a performance comparable to that of the fully supervised method with 40% of the annotated data. Slonimer et al. [35] utilized data from the same acoustic surveys available from Fisheries and Oceans Canada as our dataset (see Sec. 3.1) to detect schools of herring and juvenile salmon. Their end goal was not semantic segmentation of the echograms, but rather the detection of school instances; as such, they proposed a two-stage approach in which the first stage makes use of U-Net-like networks to classify pixels. In addition to four frequency channels, they also input two simulated channels (water depth and solar elevation angle) to encode spatial and temporal information, which improve the performance. Marques et al. [23] also tackled the detection of schools of herring and of juvenile salmon from echograms (single frequency channel) with a framework based on the pixel-level instance segmentation Mask R-CNN [13] network. They argued that pixel-level detection, compared to object detection, opens up possibilities for automatic biological analyses due to the finer delimitation of schools.

Another relevant work is that of French et al. [11], who proposed JellyMonitor, a system for detecting jellyfish from multibeam sonar imagery using an older convolutional neural network (CNN) architecture that classifies image patches selected from a blob extraction and tracking mechanism based on Gaussian and Kalman filters. One important difference with respect to their data is that their multibeam sonar yields 2D images with a significantly better resolution, operating at 3 MHz, but a significantly shorter max range, compared to our singlebeam echosounder, which operates at frequencies ranging from 67 to 455 kHz and yields 1D images (that we concatenate over time to generate 2D images). For DL-based acoustic identification of marine species from multibeam sonar imagery, we refer the interested reader to the survey paper by Wei et al. [40].

Detecting pixels of interest in echograms using DL-based semantic segmentation has the advantage, over traditional methods, of automatically determining the best features from the data, alleviating the need for carefully handcrafted features. Our main goal is to identify the presence and general location of discrete scatterers and, due to the small, varying and sometimes overlapping nature of their tracks, in particular those of jellyfish, we favor this paradigm over object detection or instance segmentation. Existing works on semantic segmentation of echograms focus on aggregations of marine organisms; by focusing on discrete scatterers, we explore here a new and relevant application of semantic segmentation networks.

3. Methodology

Fig. 2 shows the flowchart of the proposed approach for discrete scatterer detection. Echograms are fed as input to a DL semantic segmentation network (three different networks are included in the experiments) to produce pixel-level detections. Two fusion strategies for dealing with the multi-frequency nature of the data are explored: 1) early fusion (fusion of the inputs, Fig. 2(a)), in which the contents of all frequency channels are fused into single input images and the network outputs fused results seamlessly; 2) late fusion (fusion of the outputs, Fig. 2(b)), in which single-frequency images (at all frequencies) are directly input to the network and single-frequency outputs are then fused to yield the final results. We also test using no fusion at all (i.e., providing results for each individual frequency). The remainder of this section details the dataset used in the experiments, the early and late fusion approaches, as well as the compared semantic segmentation networks.

3.1. OCUDS Dataset

The Okisollo Channel Underwater Discrete Scatterers (OCUDS) dataset is composed of 125 one-hour multi-

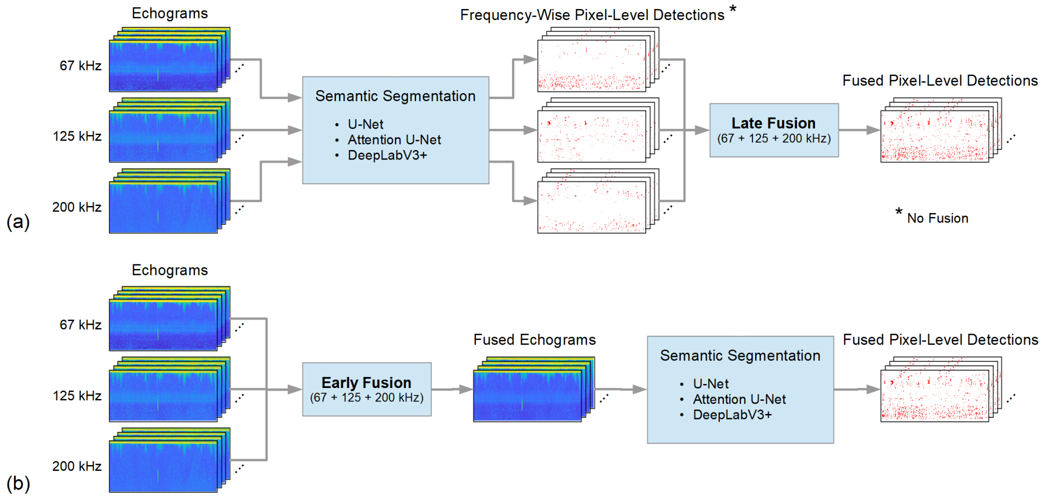

Figure 2. Flowchart of the proposed method. In (a), single-frequency images are fed to a DL semantic segmentation model, yielding frequency-specific results that are aggregated in a late fusion mode to obtain the final detections. In (b), in an early fusion mode, fused inputs combining the contents of all frequencies are fed to a DL semantic segmentation model, directly yielding the final detections. We also test the case of no fusion, with frequency-wise detections marked by an asterisk (*) in the upper branch used as final results.

frequency echograms generated from a moored upward-looking autonomous echosounder at the Venture point in the Okisollo Channel, a sheltered body of water separating the islands of Sonora and Quadra in the Discovery Passage of Vancouver Island, BC, Canada in 2015. The mooring embodied a bottom-mounted Acoustic Zooplankton Fish Profiler (AZFP) [18] echosounder that measured data at four frequencies (67, 125, 200, and 455 kHz) from the bottom depth of 55 m. Prior to deployment, the echosounder was calibrated by the manufacturer. Primarily deployed to study the migration timing and dynamics of juvenile salmon in the area [32], the AZFP collected data for their migration period, which typically extends from early May to July. The collected data are visualized as 571×1200-pixel echograms, where each pixel represents approximately 10 cm depth resolution (height-wise) through 3 s of time (width-wise).

The echograms display the standard volume backscattering strength (S_v), calculated from the raw acoustic data. S_v, reflecting the sum of all the acoustic response within a volume scaled to 1 m^3, can be calculated from the deployment metadata and is given by the following [18]:

$$S_v = EL_{max} - \frac{2.5}{a} + \frac{N}{26214a} - SL + 20\log R + 2\alpha R - 10\log(\frac{c\tau\Psi}{2}). \quad (1)$$

Here, EL_{max} expresses the echo level (in dB re 1μPa) necessary to saturate the 16-bit A/D converter; N is the instrument-provided "counts" from the raw data which is linearly related to the logarithm of the received voltage that is amplified, band-pass filtered and "detector" passed; a is the gradient of the detector response (volts/dB); α is the seawater absorption coefficient (dB/m); R represents the range from the instrument (m); SL is the source level of the instrument (dB re 1μPa at 1m); c is the sound speed (m/s); τ gives the length of the transmit pulse (s); and Ψ is the two-way solid angle of the beam.

In order to visualize the echograms, S_v values, typically ranging from around -125 to 0 dB in this case, are converted to red-green-blue (RGB) integers using a proper color map. The popular "jet" color map is appealing for its full visual spectrum showing large changes in chroma and luminance [37]. However, jet highlights even the smallest existing image features with high contrast and may accentuate unwanted information (e.g., noise); thus, we instead use "parula", a perceptually uniform designed multi-hue color map [37] (see Fig. 1 for examples). Since parula has less color contrast, it helps focus on salient information. It is also perceptually free from ambiguity, color-blind friendly, and overall offers a better understanding of the echograms.

The echograms in OCUDS contain many schools of herring and salmon; moreover, they strongly indicate the presence of a large number of discrete scatterers consistent with individual tracks of marine organisms such as jellyfish. They also contain significant noise in the form of horizontal bands (residual echoes from other frequencies or sidelobe) and sometimes vertical bands (most likely electric noise, see Fig. 3). We discard the 455 kHz data to focus on the lower frequency channels (67, 125, and 200 kHz) as the 455 kHz channel does not provide reliable measurements past 30-40 m, and the acoustic response of the discrete scatterers is inconspicuous at that frequency. The following biological cues, native to our data, were considered in the annotation process: 1) discrete scatterers are small objects with a relatively weak backscattering strength; 2) they can appear at any depth in the water column; 3) their backscattering strength can be similar to that of air bubbles, which can be differentiated as typically forming a cloud-like structure of even smaller objects near the surface or near schools; 4) jellyfish have a specific behavior described in Sec. 1.2, i.e., they may appear elongated (but not purely vertically) as due to their slow movement, they tend to stay a long time in the echosounder beam; 5) jellyfish tend to be entrained by internal waves, therefore several individuals in close proximity may appear to follow similar trajectories; 6) schools of herring and salmon (not discrete scatterers) present vertical elongated shapes and have a strong backscattering strength.

Our semi-automatic annotation process, based on standard image processing techniques and the above cues, is as follows. 1) RGB echograms are converted to grayscale images and thresholded with a manually selected threshold to segment potential marine organisms (which will include the boundaries of aggregations, discrete scatterers, and sometimes air bubbles) from the background. 2) The area of each segmented object is computed, and all segments smaller than a manually selected value are kept as annotation candidates. 3) To deal with noise, some of which is typically included in the annotation candidates, a post-processing step is applied in a process similar to that of the noise removal in [10]. For each particularly noisy region in a given echogram, the S_v values are summed across all frequencies and thresholded to obtain a noise mask. This process allows us to roughly separate returned echoes that appear in all frequencies from those typically appearing in only one or two, such as horizontal noise bands. 4) The final annotations are obtained by removing any likely noise pixels from the annotation candidates using the noise mask. Due to the discrete scatterers' characteristics and the noise in the echograms, in addition to some uncertainty in the absence of actual ground truth, these annotations are imperfect, yet still useful for training and for comparing the performance of semantic segmentation networks. The top rows of the echograms are excluded to remove the water-air interface (yellow band) and anything above it.

OCUDS follows a standard 80%/20% partitioning for training/testing, which corresponds to 100 and 25 multi-frequency echograms for training and testing, respectively.

3.2. Early fusion vs. late fusion

Multi-frequency echosounders capture complementary information coming from each frequency, as marine organisms and physical phenomena respond differently to each one according to their acoustic properties and size. Each frequency-specific image represents the same set of scatterers (in range and time), but looks different from images at other frequencies. There are two main explicit strategies for dealing with multi-frequency data: early data fusion, i.e., merging the input data as a pre-processing step prior to any analysis, or late data fusion, i.e., as a post-processing step to merge single-frequency results. We experiment with both strategies, along with no fusion at all (i.e., providing results for each individual frequency), to determine which one yields the best performance in our application. An implicit strategy would involve letting the network deal with all frequency channels at once; this has been done for instance in [35], which mapped the S_v values from four frequencies to four grayscale images to form a 4-channel input.

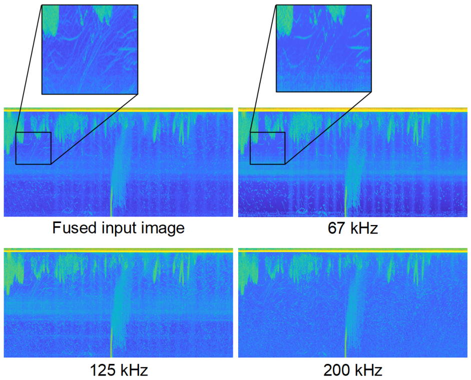

Figure 3. Sample fused input image generated from three single-frequency images.

Early fusion (Fig. 2(a)) is akin here to pixel-level image fusion, which combines multiple input images into a fused image. A fused image is expected to be more informative for human or machine perception in comparison to any of the single images [19]. Considering how single-frequency RGB images are created from S_v values using the parula color map (see Sec. 3.1) and that we are interested in returned echoes at all frequencies, we use a simple fusion method that sums the S_v values pixel-wise and then converts the summed S_v values to the parula color map. The conversion process is thus carried out on a larger S_v value range of about -375 to 0 dB (compared to single-frequency images), which effectively averages the responses from all frequencies. Fig. 3 shows an example of a fused echogram obtained from the 67, 125, and 200 kHz images, in which we can see the returned echoes from all frequencies (mostly visible in the enlarged region). One downside of the fused input images is that since a larger range of S_v values is mapped to the same RGB range, the loss of precision in the conversion process is greater than for single-frequency images. One advantage is that the fusion tends to lessen the noise (horizontal and vertical bands in Fig. 3), which is often amplified at narrow frequency ranges. Annotations are obtained following a similar process: single-frequency annotations are fused using a bit-wise OR operation.

Late fusion (Fig. 2(b)), which happens at the results stage, has the advantage of providing more training data to the networks, as each single-frequency image is a distinct input. However, each input image is less informative. A fused output is generated from the single-frequency results with a bit-wise OR operation, similarly to how annotations are created in the early fusion strategy, as both annotations and outputs are binary (discrete scatterers vs. background).

3.3. Semantic segmentation networks

The goal of semantic segmentation is to partition an image into mutually exclusive subsets representing a meaningful region of the original image [12]. Semantic segmentation networks thus perform a dense prediction, i.e., a pixel-level classification, yielding an output segmentation map of the same dimensions as the input image. In this study, we experiment with several popular DL-based semantic segmentation networks with a proven track record in various applications: U-Net [31], Attention U-Net [27], and DeepLabV3+ [5]. We also explored a very different network, YOLOv7 [39] (the latest in the popular object detection series which includes provisions for pixel-level classification), but segmentation architectures that use an initial bounding-box detection phase like YOLO have difficulty with small objects (the bounding box regression will experience gradient explosions and the loss going to infinity). Given the size of the discrete scatterers, we encountered this issue and removed YOLOv7 from the final experiments.

U-Net [31] was first proposed for medical image segmentation. Building upon Fully Convolutional Networks (FCNs) [22], it aims to require fewer training images and yield more precise segmentations. Following an encoder-decoder architecture, it has three main parts: 1) the encoding/contracting path, a typical CNN with repeated applications of convolutions followed by a rectified linear unit (ReLU), batch normalization and max pooling, which progressively reduces the spatial information while increasing the feature information; 2) the bottleneck, used for dimension reduction; 3) the decoding/expanding path, complementary to the contracting path, which combines spatial and feature information through a series of upsampling transposed convolutions, concatenated with the corresponding feature maps from the contracting path via skip connections. This allows the network to make local predictions that respect the global structure of the image content.

Attention U-Net [27] adds an attention gate mechanism to U-Net to focus on specific elements of interest. Skip connections in U-Net may propagate redundant low-level information due to a poor feature representation in the initial layers. Attention gates filter features propagated through the skip connections to actively suppress activations in irrelevant regions, thus reducing redundant features and highlighting useful salient features. Attention U-Net utilizes additive soft attention and applies weights to different regions of the image; which regions get larger weights (and thus more attention) is learned as the model is trained.

DeepLabV3+ [5] is a semantic segmentation encoder-decoder architecture using the Atrous Spatial Pyramid Pooling (ASPP) module that incorporates cascading atrous convolutions to model multi-scale context after the downsampling backbone (i.e., ResNet-50 [14]). It improves over DeepLabV3 [4] by adding a decoder module after the ASPP. To reduce the model's size, bilinear interpolation is used to adjust the output size instead of transposed convolutions as in U-Net. This architecture allows for accurate segmentations with a smaller number of parameters.

4. Experimental results

This section discusses experimental results for the detection of discrete scatterers from multi-frequency echograms for all three compared semantic segmentation networks, which were implemented in Python using the PyTorch [38] framework and trained and tested on the OCUDS dataset.

For each model, we train from scratch (without the use of any pre-trained weights) for 500 epochs with an Adam optimizer, a learning rate of $1e^{-5}$ and no weight decay. Random horizontal flipping and ImageNet normalization augmentations are used, after which the images are resized to 400×800 pixels. Batch sizes of 2 and 4 are used for U-Net/Attention U-Net and for DeepLabV3+, respectively. Given the minute nature of the discrete scatterers and unbalanced class distribution, we use Dice Loss [24] (with Binary Cross Entropy) and Focal Loss [20] functions for training U-Net/Attention U-Net and DeepLabV3+, respectively, which we empirically found to perform best.

4.1. Quantitative evaluation

Table 1 compares the quantitative performance of the three semantic segmentation networks evaluated on the test set of OCUDS, for the early, late, and no fusion (with single-frequency results) strategies. As this is a pixel-level binary classification problem, we report the following standard evaluation metrics: precision, recall, and intersection over union (IoU). Precision showcases the proportion of detections that are actual discrete scatterers, while recall emphasizes the proportion of actual discrete scatterers that are detected. Precision and recall are often a trade-off between the two, and which one is favored depends on the context. For instance, precision can be favored when the extraction of definite discrete scatterers needs to be accurate or noise needs to be eliminated, even at the expense of the exclusion of low-confidence scatterers. In cases where noise in the results (false positives) does not have a significant impact but detecting each and every discrete scatterer is of importance, recall is the most relevant metric. IoU complements precision and recall by measuring the similarity between the set of detected pixels and the set of annotated pixels, comparing the size of the intersection with that of the union of the two sets. Here, these metrics are computed echogram-wise then averaged over the entire test set.

From Table 1, across all strategies, DeepLabV3+ outperforms both U-Net and Attention U-Net in terms of precision and IoU, and Attention U-Net yields the best recall. The best precision (0.685 ± 0.070 for DeepLabV3+) and recall (0.381 ± 0.063 for Attention U-Net) are obtained for

			Network	
Stragegy	Metric	U-Net	Attention U-Net	DeepLabV3+
Early Fusion	Precision	0.553 ± 0.059	0.449 ± 0.077	0.651 ± 0.085
	Recall	0.276 ± 0.039	0.380 ± 0.042	0.315 ± 0.085
	IoU	0.223 ± 0.059	0.256 ± 0.029	0.265 ± 0.060
Late Fusion	Precision	0.649 ± 0.061	0.529 ± 0.072	**0.685 ± 0.070**
	Recall	0.253 ± 0.072	**0.381 ± 0.063**	0.334 ± 0.083
	IoU	0.219 ± 0.053	0.281 ± 0.037	**0.285 ± 0.058**
No Fusion 67 kHz	Precision	0.548 ± 0.100	0.456 ± 0.108	0.603 ± 0.135
	Recall	0.216 ± 0.081	0.322 ± 0.084	0.237 ± 0.085
	IoU	0.180 ± 0.058	0.227 ± 0.053	0.205 ± 0.074
No Fusion 125 kHz	Precision	0.299 ± 0.100	0.457 ± 0.100	0.605 ± 0.115
	Recall	0.219 ± 0.082	0.323 ± 0.084	0.310 ± 0.134
	IoU	0.179 ± 0.059	0.228 ± 0.049	0.243 ± 0.082
No Fusion 200 kHz	Precision	0.546 ± 0.100	0.449 ± 0.097	0.585 ± 0.057
	Recall	0.216 ± 0.082	0.322 ± 0.085	0.310 ± 0.084
	IoU	0.179 ± 0.059	0.225 ± 0.049	0.250 ± 0.055

Table 1. Performance (mean ± standard deviation) of compared semantic segmentation networks, computed echogram-wise, on the OCUDS test set for early, late, and no fusion (best results in bold).

the late fusion strategy. This seems to indicate that having additional training images (one for each frequency), each displaying less information, is beneficial in this case. For the no fusion strategy, there is no clear tendency in terms of one frequency outperforming others; compared to early and late fusion, no fusion yields lower metric values. Looking at the standard deviation values, there is again no clear tendency in terms of which network is more consistent from one echogram to the next. All networks yield higher precision and lower recall, being more selective in predicting pixels that are discrete scatterers. This is appealing for our goal of focusing on the presence and general location of the discrete scatterers: the networks do not detect all of them, but what they detect is in accordance with the annotations. One caveat is that due to the annotations not being 100% accurate, these metrics alone cannot completely capture the quality of the results, which we evaluate visually next.

4.2. Qualitative evaluation

Fig. 4 shows representative results from all compared semantic segmentation networks, taken at different times in the day (10 am and 8 pm) on different days, for the two best fusion strategies, i.e., early and late. Additional visual results can be found in the supplementary material. From Fig. 4(a) and (b), it is clear that Attention U-Net tends to yield a larger number of detected pixels (in black) with many smaller detections. This observation is supported by the higher recall of Attention U-Net (see Table 1). The difference between U-Net and DeepLabV3+ is not as clear, as both tend to yield less detections. There is a tendency across all networks to yield additional and smaller detections in the late fusion case. Generally, all three networks generate less detections compared to the annotations, except in the case of late fusion Attention U-Net, which detects additional discrete scatterers in the noisy middle region. *From a visual inspection, we can see that on several occasions, the networks outperform the annotations.* On one hand, annotated pixels that likely correspond to noise (e.g., in Fig. 4(b), the annotations include parts of the vertical noise band near the left border) or to a small fish school (e.g., in Fig. 4(a) the 67 kHz annotations include a vertical aggregation near the top about 1/5 from the left side, which is most likely an aggregation of juvenile salmon) are not detected as such by the networks (e.g., by DeepLabV3+ in these two cases). A possible explanation is that semantic segmentation networks tend to have trouble with very small objects; the deeper into the network, the larger the receptive field and thus the less influence a single pixel has on later filters. Also, networks learn the probability of certain local patterns; as such, noise present in the annotations is less likely to be modelled as it has no known distribution/pattern. Another consideration is that outliers (like the juvenile salmon aggregation case) are insufficient in numbers for the networks to model their distribution properly, and are thus not learned. On the other hand, some unannotated pixels in the middle region that are obscured by noise (and are thus too aggressively removed in steps 3 and 4 of the annotation process, see Sec. 3.1), are detected as discrete scatterers, especially in the late fusion strategy; such detections make sense from a biological viewpoint. This can be explained in part by the fact that in the late fusion case, some discrete scatterers may be more visible at one frequency compared to their fainted version present in the fused input echograms, and are thus detected more easily in the single-frequency inputs. These results, while useful as is for biologists to speed up their analysis tasks, open up the possibility of improving the annotations in an iterative manner using actual detection results.

5. Conclusion

This study explores the use of semantic segmentation networks for the pixel-level detection of discrete scatterers in noisy multi-frequency echograms. Several state-of-the-art DL architectures are compared on the challenging OCUDS dataset for the jellyfish usecase, which contain echograms that are particularly difficult to correctly annotate on a pixel level. Experiments with early and late fusion strategies, along with no fusion at all, reveal that, quantitatively, late fusion is preferable in terms of precision, recall, and IoU. They also show that DeepLabV3+ tends to have more precise detections, while Attention U-Net tends to miss less detections. Interestingly, the networks appear to qualitatively outperform the annotations on several occasions, being able to detect some discrete scatterers in noisy region while at the same time not detecting annotated scatterers that are more in line with noise or small aggregations of fish. The problem that we are trying to solve is a good example of a class of problems where perfect ground truth (expressed via annotations) cannot be reached due to mul-

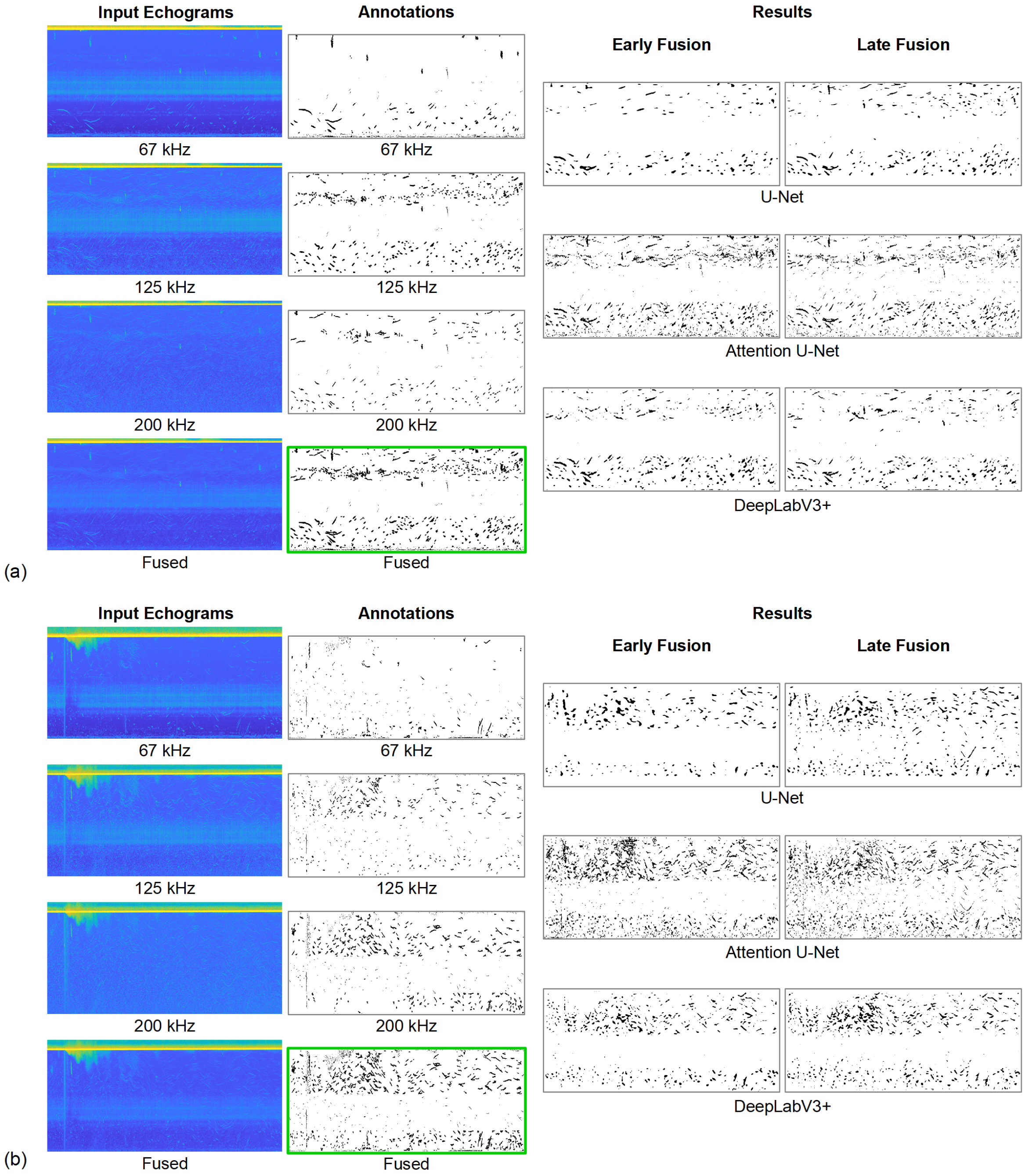

Figure 4. Sample results for two echograms from the OCUDS test set, for the best performing strategies (early and late fusion): 26-JUL-2015 20:00-21:00 (a), and 31-JUL-2015 10:00-11:00 (b). All results should be compared to the fused annotations (with green bounding box). In the annotations and results, black pixels represent discrete scatterers.

tiple reasons. Ongoing work focuses on how to integrate annotation, testing, and validation into an iterative, convergent loop which takes advantage of both human expertise and DL. Future work will look into conducting extensive surveys among marine biologists to assess their preferred outputs in terms of usefulness for environmental monitoring tasks, as well as exploring the use of soft labels to account for the uncertainty in the annotations.

Acknowledgments

This work was enabled by NSERC Canada and ASL Environmental Sciences through the Alliance-Mitacs Grants program and in part by support provided by the Digital Research Alliance of Canada.

References

[1] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, 54:137–178, 2021. 3

[2] 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*, 2020. 3

[3] Andrew S Brierley, Bjørn Eric Axelsen, David C Boyer, Christopher P Lynam, Carol A Didcock, Helen J Boyer, Conrad AJ Sparks, Jennifer E Purcell, and Mark J Gibbons. Single-target echo detections of jellyfish. *ICES Journal of Marine Science*, 61(3):383–393, 2004. 3

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 6

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, pages 801–818, 2018. 6

[6] 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*, 2023. 3

[7] Gustavo Alvarez Colombo, Hermes Mianzan, and Adrian Madirolas. Acoustic characterization of gelatinous plankton aggregations: four case studies from the argentine continental shelf. *ICES Journal of Marine Science*, 60(3):650–657, 2003. 2, 3

[8] 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. 3

[9] Echoview Software Pty Ltd. Hydroacoustic Data Processing - Echoview — Echoview. https://echoview.com/. Accessed: 2023-02-20. 1

[10] Paul G Fernandes. Classification trees for species identification of fish-school echotraces. *ICES Journal of Marine Science*, 66(6):1073–1080, 2009. 5

[11] Geoff French, Michal Mackiewicz, Mark Fisher, Mike Challiss, Peter Knight, Brian Robinson, and Angus Bloomfield. Jellymonitor: Automated detection of jellyfish in sonar images using neural networks. In *14th IEEE International Conference on Signal Processing (ICSP)*, pages 406–412. IEEE, 2018. 3

[12] Shijie Hao, Yuan Zhou, and Yanrong Guo. A brief survey on semantic segmentation with deep learning. *Neurocomputing*, 406:302–321, 2020. 6

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[15] John K Horne. Acoustic approaches to remote species identification: A review. *Fisheries Oceanography*, 9(4):356–71, 2000. 2

[16] Rolf J Korneliussen. Acoustic target classification. Technical Report No. 344, International Council for the Exploration of the Sea (ICES), 2018. 1

[17] Rolf J Korneliussen, Yngve Heggelund, Gavin J Macaulay, Daniel Patel, Espen Johnsen, and Inge K Eliassen. Acoustic identification of marine species using a feature library. *Methods in Oceanography*, 17:187–205, 2016. 3

[18] 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. 4

[19] 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. 5

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 6

[21] Hongxia Liu, Fanlin Yang, Shuangqiang Zheng, Qianqian Li, Donghui Li, and Hongchun Zhu. A method of sidelobe effect suppression for multibeam water column images based on an adaptive soft threshold. *Applied Acoustics*, 148:467–475, 2019. 2

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 6

[23] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4378–4387, 2021. 3

[24] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016. 6

[25] 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. 2

[26] Erhan Mutlu. Target strength of the common jellyfish (Aurelia aurita): a preliminary experimental study with a dual-beam acoustic system. *ICES Journal of Marine Science*, 53(2):309–311, 1996. 2, 3

[27] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Atten-

tion U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 6

[28] Alba Ordonez, Ingrid Utseth, Olav Brautaset, Rolf Korneliussen, and Nils Olav Handegard. Evaluation of echosounder data preparation strategies for modern machine learning models. *Fisheries Research*, 254:106411, 2022. 3

[29] 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. 2

[30] D Reid, C Scalabrin, P Petitgas, J Masse, R Aukland, P Carrera, et al. Standard protocols for the analysis of school based data from echo sounder surveys. *Fisheries Research*, 47(2-3):125–36, 2000. 2

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015. 3, 6

[32] 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. 4

[33] S Rousseau, S Gauthier, C Neville, S C Johnson, and M Trudel. Acoustic classification of juvenile Pacific salmon (Oncorhynchus spp) and Pacific herring (Clupea pallasii) schools using random forests. *Frontiers in Marine Science*, 9:857645, 2022. 1, 2

[34] Mei Sato, John K Horne, Sandra L Parker-Stetter, and Julie E Keister. Acoustic classification of coexisting taxa in a coastal ecosystem. *Fisheries Research*, 172:130–136, 2015. 3

[35] 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 *19th Conference on Robots and Vision (CRV)*, pages 8–15. IEEE, 2022. 3, 5

[36] Timothy K Stanton. 30 years of advances in active bioacoustics: a personal perspective. *Methods in Oceanography*, 1:49–77, 2012. 1, 2

[37] Michael Stoelzle and Lina Stein. Rainbow color map distorts and misleads research in hydrology–guidance for better visualizations and science communication. *Hydrology and Earth System Sciences*, 25(8):4549–4565, 2021. 4

[38] The Linux Foundation. PyTorch. https://pytorch.org/. Accessed: 2023-02-27. 6

[39] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022. 6

[40] Yaoguang Wei, Yunhong Duan, and Dong An. Monitoring fish using imaging sonar: Capacity, challenges and future perspective. *Fish and Fisheries*, 23(6):1347–1370, 2022. 3