Instance Segmentation-based Identification of Pelagic Species in Acoustic Backscatter Data

Tunai Porto Marques*, Melissa Cote*, Alireza Rezvanifar*, Alexandra Branzan Albu*,
Kaan Ersahin†, Todd Mudge† and Stéphane Gauthier§
* Electrical and Computer Engineering, University of Victoria, Canada
† ASL Environmental Sciences, Victoria, Canada
§ Fisheries and Oceans Canada, Sidney, Canada

Email: {tunaip, mcote, arezvani, aalbu}@uvic.ca, {kersahin, tmudge}@aslenv.com, stephane.gauthier@dfo-mpo.gc.ca

Abstract

This paper addresses the automatic identification of pelagic species in acoustic backscatter data. Large quantities of data acquired during underwater acoustic surveys for environmental monitoring and resources management, visualized as echograms, are typically analyzed manually or semi-automatically by marine biologists, which is time-consuming and prone to errors and inter-expert disagreements. In this paper, we propose to detect pelagic species (schools of herring and of juvenile salmon) from echograms with a deep learning (DL) framework based on instance segmentation, allowing us to carefully study the acoustic properties of the targets and to address specific challenges such as close proximity between schools and varying size. Experimental results demonstrate our system's ability to correctly detect pelagic species from echograms and to outperform an existing object detection framework designed for schools of herring in terms of detection performance and computational resources utilization. Our pixel-level detection method has the advantage of generating a precise identification of the pixel groups forming each detection, opening up many possibilities for automatic biological analyses.

1. Introduction

This paper focuses on the detection of pelagic species in acoustic backscatter data. Multi-frequency echosounders enable the acquisition of time series of acoustic backscatters during underwater acoustic surveys. Echosounder data can be visualized in 2D images known as echograms, where the vertical axis shows the depth or range in the water column, and the horizontal axis represents the time. The intensity of each pixel corresponds to the reflected echo amplitude or intensity, generally computed as the volumetric backscatter strength. Different species, geological formations, and various phenomena (e.g., supsended sediments) produce different echoes. Thus, an echogram will display a variety of structures and patterns that may be indicative of the seabed, the water-air interface, and the presence of one or many biological objects in the water column, usually schools of fish and/or planktonic organisms [27].

Underwater acoustic surveys allow marine biologists to gather large quantities of data that enable them to perform a variety of tasks crucial for environmental monitoring, such as species identification, biodiversity mapping, and animal behaviour studies, in a non-invasive manner. Detecting pelagic species, including schools of herring and juvenile salmon, over large periods of time constitutes an important part of fisheries and ocean resources management as well as valuable information towards a better understanding of the effects of climate change on the oceans. Echograms are commonly interpreted with manual or semi-automatic methods, using commercial software like Echoview[1]. Given the shear size of the data to analyze, this is a time-consuming process, prone to errors and inter-expert disagreements. Indeed, echogram analysis can be challenging due to many factors, including the varying size and acoustic properties of the targets, significant inter-class similarities, and the specific context of the data acquisition. For instance, a marine biologist will determine a type of fish based on location, time, behaviour, acoustic backscatter strength, differences in acoustic backscatter strength from different frequencies, and additional data from net tows and/or underwater cameras [23].

In this paper, we propose to detect pelagic species from echograms using a deep learning (DL) framework based on *instance segmentation*. More specifically, we aim to detect schools of herring and of juvenile salmon, which can be found concurrently in the same geographic locations (i.e., within the same echograms), but typically at differ-

[1] https://www.echoview.com/

ent depths. In addition to the more general challenges of echogram analysis mentioned above, from a computer vision viewpoint, challenges related to the identification of pelagic species include the potential close proximity of different schools, making it harder to detect and distinguish between them; the possible close proximity of schools of juvenile salmon to the surface, which may then overlap with the turbulence of the water-air interface; the potential presence of bubbles around a school, altering the apparent morphology of the school; the possible small size of schools of juvenile salmon compared to the size of the echograms, which may make the feature extraction process for identification purposes less reliable. There are several available image analysis paradigms for the identification of pelagic species from echograms: image classification, semantic segmentation, object detection, and instance segmentation. We elect to use an instance segmentation paradigm here, which assigns an instance label locally to each pixel in the echogram, tackling precisely (at the pixel level) the "where" and "what" and distinguishing between different instances of the same class, in our case different schools of the same species. This allow us to address many of the aforementioned challenges related to overlaps and morphology, in addition to allowing for more precise biological analyses.

Our contributions are two-fold. 1) From a methodological viewpoint, we provide a comprehensive experimental design considering diverse feature extraction backbones within an instance segmentation framework adapted for echograms, taking advantage of the most powerful state-of-the-art DL architectures. 2) From a practical viewpoint, we show that instance segmentation networks are more suitable and accurate for the detection of pelagic species than object detection networks; the proposed instance segmentation framework offers a unique opportunity for automatic biological analyses based on a precise identification, at the pixel level, of schools of herring and of juvenile salmon. To the best of our knowledge, the proposed framework is the first of its kind in fisheries and acoustics (see Sec. 2.2). It is capable of detecting pelagic schools with an accuracy that closely matches that of human operators (see Sec. 4.4). Differently from what an expert can do manually, it also specifies the pixels associated with each detection allowing for a precise estimation of the number of specimens per school (see Sec. 4.6). Other advantages over object detection networks include the ability to better distinguish between schools in close proximity and the additional information on the often complex morphology of each output at the pixel level, instead of simple bounding boxes. Our method can be easily reproduced in other layouts/datasets and is also inherently scalable due to the use of a DL-based instance segmentation network: it can identify new species as long as training pixel-level annotations are provided.

The remainder of the paper is divided as follows. Sec. 2 reviews related works on marine species detection in echograms. Sec. 3 details the PLHS (Pixel-Level Herring and Salmon) dataset and our proposed instance segmentation framework. Sec. 4 discusses experimental results, including a comparison with the object detection framework of [23]. Sec. 5 presents concluding remarks.

2. Related works

The detection of pelagic species, and more generally of marine species from echograms can be categorized into classical machine learning (ML)- and DL-based approaches. Both categories are reviewed next.

2.1. Classical machine learning-based approaches

Classical approaches to echogram analysis make use of hand-crafted features focused on statistical characteristics of organism aggregations. We find three different groups of characteristics in the literature [32, 12, 26]: 1) positional/bathymetric characteristics, which relate to the position in the water column; 2) morphometric characteristics, linked to the school height, width, and perimeter; and 3) energetic characteristics, pertaining to the backscattered signal properties. Hand-crafted features are typically extracted using commercial software tools (e.g., Echoview), based on the above domain-dependent taxonomy [23].

Hand-crafted feature-based methods for marine species detection in echograms generally rely on classical ML methods for feature classification. To detect various fish schools (Bonaerensis anchovy, Patagonian anchovy, rough scad, sprat, longtail hoki, and blue whiting), Cabreira *et al.* [2] compared three types of artificial neural network (ANN) architectures and found that for asymmetrical numbers of input data per species, the best ANN differed from one species to another, while for symmetrical data, self-organizing maps (SOMs) yielded the best performance. This work led to ECOPAMPA [33], a recent tool for automatic fish schools detection and assessment from echo data based on the same ANN architectures. Also comparing different types of ANNs and support vector machines (SVMs), Robotham *et al.* [29] classified schools of anchovy, common sardine, and jack mackerel. They found that multilayer perceptrons (MLPs) and SVMs performed better for multi-class classifications. Working with high-resolution echograms, LeFeuvre *et al.* [15] detected Atlantic cod and capelin using a Mahalanobis distance classifier. Also using Mahalanobis distance information, Charef *et al.* [3] identified three broad fish groups using a discriminant function analysis. Focusing on the classification of six mesopelagic fish groups, Gauthier *et al.* [7] proposed a decision model based on an objective classification decision tree. Fallon *et al.* [5] favored random forests (RFs) to classify Southern Ocean krill and icefish echoes. More recently, Proud *et al.* [24] also proposed to use RFs to detect schools of silver

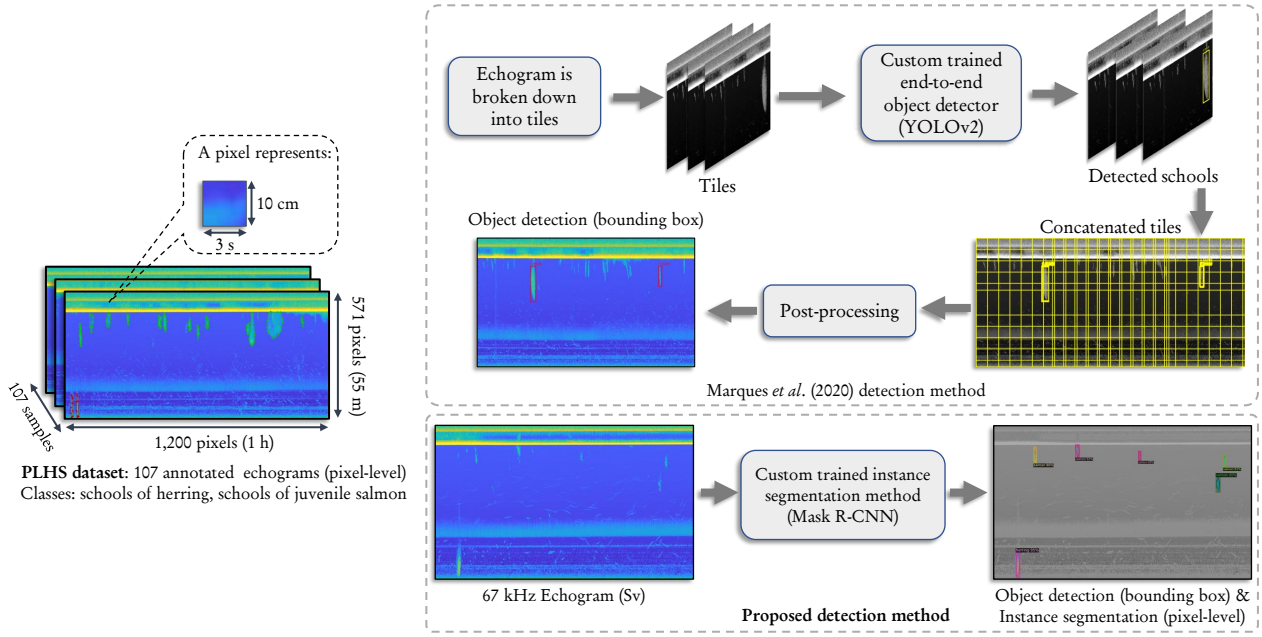


Figure 1: Proposed method and comparison baseline. PLHS (left) is a 107-sample dataset where schools of herring and of juvenile salmon are annotated on a pixel level. Using PLHS, we train the proposed, fully-custom instance segmentation-based detector (bottom right) as well as the method proposed by Marques *et al.* [23] (top right).

cyprinid from echograms for generating consistent biomass time series. RFs have also been used by Mannocci *et al.* [21] in the context of tropical tuna purse seine fisheries. The authors trained RFs to differentiate between high and low bycatch occurrence in data collected by echosounder buoys attached to drifting fish aggregating devices.

An important drawback of hand-crafted feature-based methods is that new sets of features need to be engineered for each species, making the methods not easily and readily scalable to various and diverse marine species. This drawback is mitigated by DL-based approaches.

2.2. Deep learning-based approaches

DL-based approaches have been shown to achieve excellent results for a variety of computer vision-related applications in the visible spectrum, including for object detection. To this date, there is still only a handful of works utilizing DL beyond the visible spectrum for the detection of marine species from acoustic backscatter data.

In a hybrid study involving both classical ML and DL methods, Shang and Li [31] studied echo features and classification methods of fish using simulated data. They experimented with three different features (all based on backscatters, i.e., echo waveforms, echo spectrograms, echo spectra) and four different classification methods (decision tree, adaboost, ANNs, and convolutional neural network (CNN)), and found that the best combination was CNN with echo spectrograms. Hirama *et al.* [11] detected five fish species (yellowtail, salmon, squid, sardine, and juvenile tuna) from echosounder data in a set-net using a CNN. With this image classification-based approach, the echograms have to be divided in a rough set of anisotropic non-overlapping tiles classified individually, assuming only one class of fish per tile. In a slightly different direction more in line with natural language processing, Måløy [20] focused on the spatiotemporal properties contained in echograms. The author proposed a transformer-based approach that interprets the spatiotemporal dynamics of echograms through attention mechanisms to classify fish behavior and detect the onset of pancreas disease in farmed Atlantic salmon. Closer to our work, Brautaset *et al.* [1] focused on acoustic classification in multifrequency echosounder data for sandeel detection. They proposed a semantic segmentation CNN based on the U-Net [30] architecture. Their method yielded a substantially higher performance compared to that of Korneliussen *et al.* [14], who used a "traditional automated processing pipeline" to detect sandeel. Rezvanifar *et al.* [28] proposed a hybrid approach to detect schools of herring from echograms, in which regions of interest (ROIs) are first extracted based on schools' intensity and morphology and then classified via a DL classifier. The authors compared three popular CNN architectures for the feature extraction and classification task and found that DenseNet [13] achieved the best overall performance. One drawback is that the ROI extraction is species-specific and cannot be straightforwardly extended to other species. In a follow-up paper by Marques *et al.* [23], the authors provided a comparative study covering the entire spectrum of learning

approaches, from traditional and hybrid methods to complete end-to-end DL object detection networks. Focusing on schools of herring, they concluded that the latter are preferable to other learning approaches, providing comparable or better results than traditional methods even with limited training data. A limited number of papers focus on DL methods for the detection of marine species from sonar data, which generally have a higher resolution compared to echograms. They tackle the detection of jellyfish [6] and fish count/concentration [19, 8]. Neupane and Seok [22] provide a review of DL-based approaches for the more general topic of automatic target recognition from sonar data.

DL approaches are becoming more popular in fisheries acoustics, but instance segmentation is notably absent: recent works [1, 23] focused on semantic segmentation and object detection, respectively.

3. Methodology

Our proposed pelagic schools detector is composed of a custom, DL-based instance segmentation framework that was trained using PLHS, a novel proprietary dataset of schools of herring and juvenile salmon (see Fig. 1 and Sec. 3.1). The recent method proposed by Marques *et al.* [23] closely relates to the task our system attempts to perform, thus it is used as the main baseline of comparison (see Sec. 4). Fig. 1 gives an overview of our framewok and compares it to the framework proposed by Marques *et al.* [23]. As seen on the bottom right of Fig. 1, our proposed method is able to: 1) identify schools of juvenile salmon and of herring; 2) specify the pixels composing each instance; 3) provide bounding boxes around each detection.

3.1. PLHS dataset

The Pixel-Level Herring and Salmon (PLHS) dataset consists of 107 echograms with pixel-level annotations indicating the presence of schools of herring and of juvenile salmon. It contains 153 instances of schools of herring and 252 instances of schools of juvenile salmon, covering schools with different morphologies, positions in the water column, and biological densities. The annotations follow the MS-COCO format [18]. Particularities of this dataset include the fact that it can be used to train models aiming to perform multi-class object detection, but also semantic or instance segmentation tasks, due to the granularity of its annotations. It also identifies two distinct pelagic species that are often difficult to differentiate even by specialists, resulting in an equally challenging automatic detection task. Fig. 2 shows sample annotated echograms in PLHS.

The acoustic data used to create the echograms in PLHS were obtained from Canada's Department of Fisheries and Oceans (DFO) and acquired using AZFP echosounder instruments [16]. These AZFPs were deployed by DFO in fixed positions close to the sea bottom looking upward under water columns of approximately 55 m. The instruments were located at the Okisollo channel, off the coast of British Columbia, Canada, between the months of May and October of 2015 and 2016. Each AZFP measurement is done at four frequencies: 67, 125, 200, and 455 kHz. The measurements at each frequency are visualized as a 571×1200-pixel echogram (water column depth x time) that represents one hour. Thus each pixel of a PLHS sample represents roughly 3 s throughout a depth resolution of 10 cm.

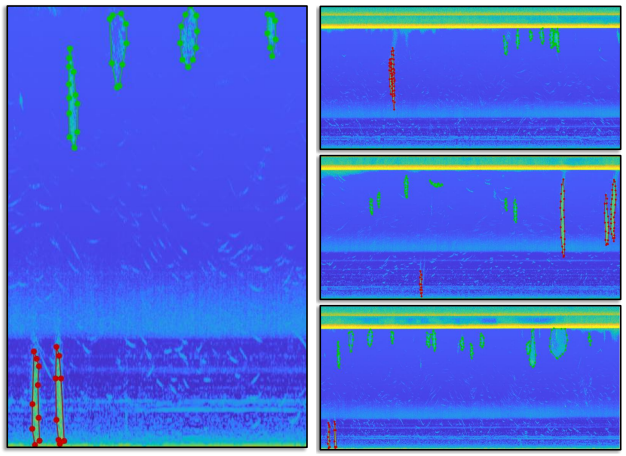


Figure 2: The Pixel-Level Herring and Salmon (PLHS) dataset. Left: a zoomed-in region showing annotations of schools of herring (red) and juvenile salmon (green). Right: three samples illustrating its diverse nature: most samples include multiple instances of schools of herring and salmon.

Since the acoustic response of schools of herring and of juvenile salmon is expected to be more pronounced at lower frequencies, we only consider the 67 kHz channel from the multifrequency data that AZFPs capture. We create standard echograms that display volume backscattering strength (S_v). The S_v representation of acoustic data, often used to detect the presence of marine species, reflects the sum of all the acoustic response within a volume scaled to 1m^3. Given raw acoustic data from AZFPs, the S_v representation can be calculated as follows [16]:

$$S_v = EL_{max} - \frac{2.5}{a} + \frac{N}{26216a} - SL + 20\log R + 2\alpha R - 10\log(\frac{c\tau\Psi}{2}) \quad (1)$$

where EL_{max} represents the acoustic input (in dB re 1μPa) that the transducer has to receive to produce a full-scale output on the 16-bit A/D converter, a the slope of the detector response in units of volts/dB, N the number of "counts" (raw value) obtained from the instrument, SL the source level (dB re 1μPa at 1m), R the range of the instrument (m), α the absorption coefficient (dB/m), c the speed of sound (m/s), τ the transmit pulse length (s), and Ψ the equivalent

solid angle that the transducer beam creates. The specific values of these parameters are available as metadata associated with each AZFP deployment. The AZFP instruments are calibrated by the manufacturer before each deployment.

Before carrying out the manual annotation process, we consulted with specialists from DFO, who provided important biological cues, such as: 1) schools of herring typically appear as elongated shapes in the vertical axis with a strong acoustic echo in the center of the school; 2) there are particular periods of the year (August-September) when the frequency with which schools of juvenile salmon are detected is expected to be reduced significantly; 3) schools of juvenile salmon often appear as smaller morphological structures than those representing schools of herring in echograms; 4) schools of herring are not typically travelling in close proximity to those of juvenile salmon; 5) schools of salmon usually travel closer to the surface. It is important to note that these biological cues might not necessarily be valid in other geographical regions. Fig. 2 (left) illustrates a scenario where these cues were paramount to the annotation process: since the two schools closer to the sea bottom are easily identified as schools of herring, the smaller, sparser schools located at the top of the image are likely from juvenile salmon (as reflected by the annotations). Fig. 2 (right) shows three samples from the PLHS dataset.

3.2. Instance segmentation framework

Our detection system is based on Mask-RCNN, the state-of-the-art instance segmentation method of He *et al.* [9]. The official pre-trained implementation of Mask R-CNN [2] is trained on a dataset of natural images (COCO [18]) that structurally differ from our visual targets. Therefore, we initially re-trained all parameters from the Mask-RCNN architecture using the PLHS dataset to assess its ability to identify pelagic species in echograms. The performance observed in these initial experiments was rather low, likely due to the small size of the PLHS dataset and the complexity of the Mask R-CNN architecture.

CNNs are able to automatically extract meaningful visual features from images of diverse natures. The fully connected networks (FCN) of CNNs combine these features into "templates" that are representative of the different classes from a given dataset. The feature extraction and template creating capabilities obtained with the pre-trained version of Mask R-CNN using COCO proved to be extremely useful to our application. We use transfer learning on the official implementation of Mask R-CNN to take advantage of these capabilities and fine-tune the framework to fit the two classes of the PLHS dataset. In particular, we freeze the updating of parameters of the first block of Mask R-CNN ("stem") as well as its first residual block [10].

[2] https://github.com/facebookresearch/detectron2

Our proposed system was trained using a number of *backbone* models. Each model employs a different strategy for the extraction of visual features and requires an exclusive training process. We experiment with nine different combinations of backbones: **1)** ResNet-101 [10] with a learning schedule (LS) of $3x$; **2)** ResNet-101 with Feature Pyramid Networks (FPN) [17] and LS = $3x$; **3)** ResNet-50 [10] with LS = $1x$; **4)** ResNet-50 with LS = $3x$; **5)** ResNet-50 with deformable convolutions (DC) [4] and LS = $1x$; **6)** ResNet-50 with DC and LS = $3x$; **7)** ResNet-50 with FPN and LS = $1x$; **8)** ResNet-50 with FPN and LS = $3x$; **9)** ResNeXt-101 [34] with FPN and LS = $3x$. The "learning schedule" refers to the number of times that the original dataset (COCO [18]) was visited during pre-training (epochs): LS of $1x$ equates to approximately 12 COCO epochs, and $3x$ to approximately 37 COCO epochs. Feature pyramid networks are a mechanism proposed by Lin *et al.* [17] to represent feature maps at different scales, ultimately allowing for the identification of targets with significantly distinct dimensions. Deformable convolutions were introduced by Dai *et al.* [4] to help CNNs better adapt to possible geometric transformations of the visual targets.

4. Experimental results and discussion

4.1. Comparison baseline

We compare the performance of the proposed system on the PLHS dataset with that of the object detection framework by Marques *et al.* [23]. The method in [23] works with data obtained with an AZFP (similarly to our method), considers schools of herring as visual targets, and outputs detections as bounding boxes. We retrained the model of [23] with PLHS to include the "school of juvenile salmon" class and allow for a direct comparison with our work.

4.2. Training considerations

In the training routine of both our method and the comparison baseline, we used the PLHS dataset with a division of 73% for training/validation and 27% for testing. All models (see Sec. 3.2) are trained using a single NVIDIA™ GeForce GTX 1660 Ti GPU. We used the same set of hyper-parameters for the training of the proposed method in all configurations: 300 iterations, 2 images per batch, base learning rate of 0.02 (this learning rate drops linearly during training), 256 ROIs per image, and Stochastic Gradient Descent (SGD) with 0.9 of momentum as an optimizer. This particular set of hyper-parameters does not necessarily yield to an optimal performance for all backbones; some larger models could likely benefit from longer training and from considering additional images per batch in a more robust hardware setting.

4.3. Quantitative evaluation

Table 1 presents the performance of the proposed method along with that of the comparison baseline [23] for the various configurations/models, in terms of mean average precision (mAP). Both methods are evaluated exclusively on the test set. As the comparison baseline does not provide pixel-level detection, we also report the performance of our method for bounding boxes ("Object Detection" column) for a direct comparison. Bold font indicates the best results for each metric. For instance segmentation and an Intersection-over-Union (IoU) threshold of 0.5, the best backbone configuration of our method is #4, which includes Mask R-CNN with a ResNet-50 backbone, no FPN and LS = $3x$. Its performance is closely followed by that of configuration #2, which differs in terms of backbone model (ResNet-101) and in the usage of FPN. When looking at mAP for IoU thresholds $\in [0.5 : 0.05 : 0.95]$, the situation is reversed, with configuration #2 yielding the best performance followed by #4. The worst performances are linked with the use of deformable convolutions (i.e., #5 and #6), which would require longer training routines. A similar performance is observed when considering the mAP for object detection (i.e., using bounding boxes). Our method outperforms the comparison baseline significantly for object detection, by approximately 35 points (configuration #2 and IoU=0.5) and 34 points (configuration #4 with IoU $\in [0.5 : 0.05 : 0.95]$); and by approximately 15 points and 9 points for the worst-performing configurations (i.e., #6).

Aside from its superior detection performance and more granular output, our method also executes about 10x faster than that of Marques *et al.* [23]. Our method processes each echogram in ~ 0.4 s (with small variations for each configuration) versus ~ 4 s per sample for [23]. This difference is mainly due to the overlapping tiling strategy of [23], which requires a full YOLOv2 inference for each tile.

4.4. Qualitative evaluation

Fig. 3 shows representative detection results of the best-performing configurations of both the proposed method (i.e., configuration #4) and that of the comparison baseline [23] (i.e., configuration #11). Although the baseline's results are qualitatively excellent on a first analysis, upon a closer inspection we identified two reasons explaining its significantly lower performance metrics (see Table 1). First, the baseline often brakes a valid school down into two or more detections, creating false positives (see Fig. 3a). Second, the baseline produces bounding boxes with a considerably worst fit with the ground truth than those generated by the proposed method (see Fig. 3b). Despite these performance-lowering characteristics, we consider that the baseline method generated correct detections that carry significant scientific value in most of the test samples.

Despite our system's high performance (see Table 1), we consider that the metrics are still under-representing its actual capabilities. When qualitatively analyzing the predictions of our system, we notice a number of scenarios where detections of schools of salmon were triggered in regions not annotated as such in the dataset, but that closely resembled valid schools. Some of these "gray area" scenarios (as discussed in Sec. 3.1) could reasonably be considered as true positives, and would likely lead to different annotations if interpreted by different specialists. Fig. 4a illustrates this phenomenon: in this particular region, only three schools of juvenile salmon were annotated. The proposed method correctly identified these schools, but also indicated the presence of a fourth one (yellow arrow), which could have been considered as valid in the ground truth, based in part on the subjective analysis of the scientist annotating the dataset. Regardless, this "incorrect" detection hinders the performance of our system as reported in Table 1. A similar scenario is depicted in Fig. 4b, where the two leftmost schools of herring are identified as four instances by the system. While this result could be interpreted as valid and is extremely useful for the timely analysis of echograms by scientists, these two extra detections are classified as false positives for performance evaluation purposes.

4.5. Class-specific performance analysis

We observed that schools of juvenile salmon are particularly challenging to annotate because their morphology and acoustic echo vary significantly across echograms. Conversely, schools of herring typically present easy-to-identify characteristics (i.e., vertically-elongated shapes with strong intensities), leading to an overall easier annotation process (see Fig. 2 left). This phenomenon is echoed by the class-specific performances of our method. While Table 1 presents aggregate results that consider all classes of the dataset, we also compute the class-specific detection performance. Table 2 highlights the fact that the instance segmentation of schools of juvenile salmon is particularly difficult to perform, given that their morphology changes abruptly across samples. It also shows that the choice of configuration plays an important role on the system's capabilities. For instance, configuration #2 is preferable if the identification of schools of salmon is the focus of a study, while configuration #4 yields the best herring-specific performance.

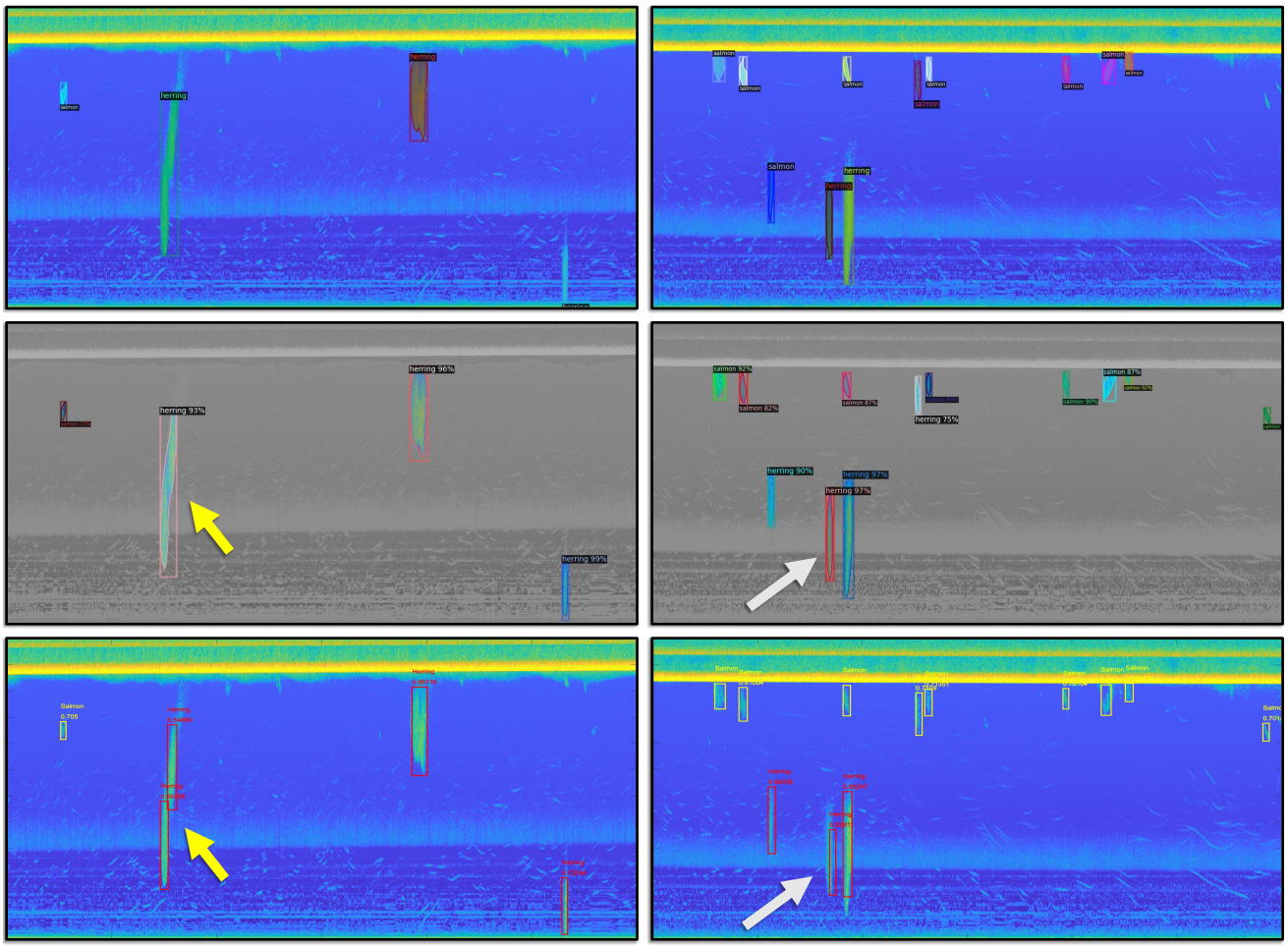
4.6. Instance segmentation vs. object detection

Instance segmentation methods are able to provide detailed information about their detection output; not only a list of pixels composing each detection is generated, but also a distinction between intra-class instances (e.g., school of salmon "A", school of salmon "B"). This ability allows for a precise estimation of populations associated with each detection, as illustrated in Fig. 5, which is not possible via object detection. Consider, for example, that each pixel in a

#	Configuration	Backbone	FPN[5]	LS[6]	Notes	mAP (Instance Segmentation) IoU=0.5	mAP (Instance Segmentation) IoU=0.5:0.05:0.95	mAP (Object Detection) IoU=0.5	mAP (Object Detection) IoU=0.5:0.05:0.95
1	Mask R-CNN [1]	ResNet-101[2]	N	3x		87.72	44.00	86.85	45.05
2	Mask R-CNN [1]	ResNet-101[2]	Y	3x		90.35	**52.79**	**90.15**	48.01
3	Mask R-CNN [1]	ResNet-50[2]	N	1x		89.63	46.08	89.63	47.76
4	Mask R-CNN [1]	ResNet-50[2]	N	3x		**92.12**	50.19	89.12	**50.48**
5	Mask R-CNN [1]	ResNet-50[2]	N	1x	Deformable convolutions [4]	73.8	29.61	70.79	26.99
6	Mask R-CNN [1]	ResNet-50[2]	N	3x	Deformable convolutions [4]	73.95	25.93	70.63	24.74
7	Mask R-CNN [1]	ResNet-50[2]	Y	1x		90.05	49.56	87.11	49.69
8	Mask R-CNN [1]	ResNet-50[2]	Y	3x		89.69	45.92	88.20	43.90
9	Mask R-CNN [1]	ResNeXt-50[4]	Y	3x	Aggregated residual transforms [34]	87.49	43.49	87.49	38.96
10	YOLOv2[3,7]	Darknet-53[3]	N	N/A	Tiling strategy [23]	N/A	N/A	41.69	11.63
11	YOLOv2[3,7]	ResNet-50[2]	N	N/A	Tiling strategy [23]	N/A	N/A	55.67	16.04

[1,2,3,4,5,6]: Mask R-CNN [9], ResNet-50 [10], YOLOv2 [25] and ResNeXt [34], Feature Pyramid Networks [17] and "learning schedule", respectively.
[7]: A custom-trained version of the method proposed by Marques *et al.* [23].

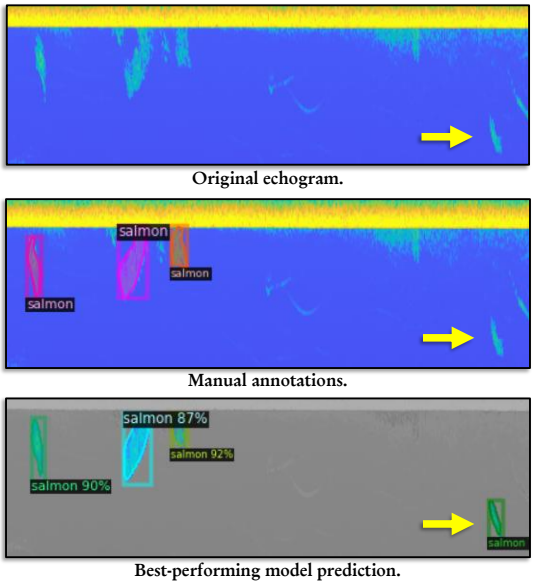
Table 1: Mean average precision (mAP) comparison for the detection results on the test set of PLHS. Configurations 1-9 represent the instance segmentation-based method proposed, while the remaining layouts use the comparison baseline [23]. Best results are highlighted in bold.



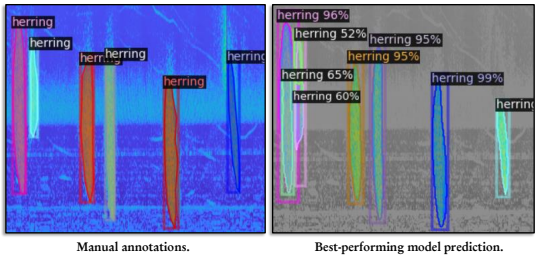
(a) A single school of herring is divided into two detections by the comparison baseline (yellow arrow).

(b) The bounding boxes created by our best-performing model (configuration #4 in Table 1) better fits the ground truth (gray arrow).

Figure 3: Qualitative comparison between the ground truth annotations (first row), best-performing configuration of the proposed method (second row) and best-performing version of the comparison baseline [23] (third row).



(a) The proposed system identifies possibly valid schools of salmon (yellow arrows) that are not annotated as such.



(b) An instance where the proposed system broke two correct detections down into four distinct objects (leftmost schools).

Figure 4: Two scenarios where the incorrect detections of the proposed system could be argued as valid.

		AP (Inst. Segm.)		AP (Object Det.)	
#	Configuration (see Table 1)	Herring	Salmon	Herring	Salmon
4	ResNet-50 $3x$	60.18	40.19	59.05	41.92
2	ResNet-101 FPN $3x$	58.29	47.3	47.61	48.42
7	ResNet-50 FPN $1x$	55.77	43.34	50.20	49.18

Table 2: Class-specific Average Precision (AP) for instance segmentation and object detection considering IoU $\in [0.5 : 0.05 : 0.95]$. Only the results for the three best-performing configurations detailed in Table 1 are presented.

PLHS sample representing a school of herring contains approximately α specimens. The bounding box produced as the output of the object detection method in Fig. 5 contains approximately $7,000$ pixels, while the manual annotation and instance segmentation outputs depicted in this same Fig. have roughly $3,500$ and $4,200$ pixels, respectively. In this illustrative example, the instance segmentation output would result in a significantly better estimation of herring population (an error of 700α fish), while the bounding box-based object detection would have an estimation error of $3,500\alpha$ herring. The precise morphology of a detection, as offered by the proposed method, might carry vital information about schools of fish such as grouping and movement patterns, predation-related movements, environmental and anthropogenic stress, among others, which is not available via the bounding boxes of object detection methods.

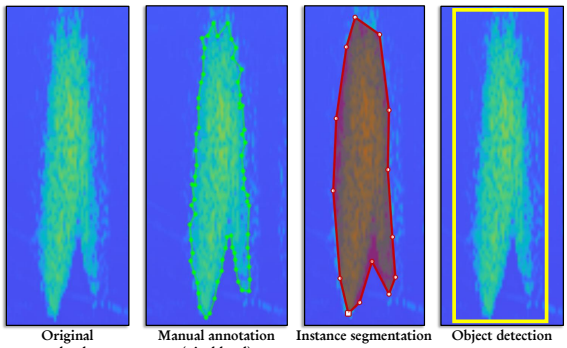


Figure 5: Illustration of different outputs and their influence on biological analyses. The precise morphology obtained with the output of instance segmentation methods allows for a better estimation of specimens count.

5. Conclusion

We propose a system that allows for a timely and precise identification of pelagic species (schools of herring and of juvenile salmon) from acoustic backscatter images (echograms). The proposed system uses a deep learning-based instance segmentation framework, the first of its kind in fisheries and acoustics, to generate not only bounding boxes around objects, but also the identify the groups of pixels that form each detection. This opens up many possibilities in terms of automatic biological analyses from underwater acoustic survey data. Our method comfortably outperforms the object detection framework proposed by Marques *et al.* [23] while providing more information (i.e., pixel-level data) as output in shorter processing times. The training and evaluation is done using PLHS, a novel dataset of pixel-level annotations of schools of herring and of juvenile salmon in echograms. Future work will involve a standardization module that allows for echograms coming from multiple instruments and deployment layouts to be used as input, as well as a semantic segmentation-based module dedicated to the identification of krill and hake.

Acknowledgments

This work was supported by NSERC Canada and ASL Environmental Sciences through the Alliance Grants program. The authors would like to thank Steve Pearce at ASL for his help with Sec. 3.1.

References

[1] Olav Brautaset, Anders Ueland Waldeland, Espen Johnsen, Ketil Malde, Line Eikvil, Arnt-Børre Salberg, et al. Acoustic classification in multifrequency echosounder data using deep convolutional neural networks. *ICES Journal of Marine Science*, 2020.

[2] Ariel G Cabreira, Martín Tripode, and Adrián Madirolas. Artificial neural networks for fish-species identification. *ICES Journal of Marine Science*, 66(6):1119–29, 2009.

[3] 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):1–11, 2010.

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 764–773, 2017.

[5] 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.

[6] Geoff French, Michal Mackiewicz, Mark Fisher, Mike Challiss, Peter Knight, Brian Robinson, et al. JellyMonitor: automated detection of jellyfish in sonar images using neural networks. In *IEEE International Conference on Signal Processing (ICSP)*, pages 406–12. IEEE, 2018.

[7] 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.

[8] Dmitry Glukhov, Rykhard Bohush, Juho Mäkiö, and Tatjana Hlukhava. A joint application of fuzzy logic approximation and a deep learning neural network to build fish concentration maps based on sonar data. In *International Workshop on Computer Modeling and Intelligent Systems (CMIS)*, pages 133–42, 2019.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2961–2969, 2017.

[10] 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.

[11] 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 *Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES)*, pages 112–5. IEEE, 2017.

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

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[14] 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.

[15] 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):137–47, 2000.

[16] 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 *2012 Oceans*, pages 1–6. IEEE, 2012.

[17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV*, pages 740–755. Springer, 2014.

[19] Liang Liu, Hao Lu, Zhiguo Cao, and Yang Xiao. Counting fish in sonar images. In *IEEE International Conference on Image Processing (ICIP)*, pages 3189–93. IEEE, 2018.

[20] Håkon Måløy. Echobert: A transformer-based approach for behavior detection in echograms. *IEEE Access*, 8:218372–218385, 2020.

[21] Laura Mannocci, Yannick Baidai, Fabien Forget, Mariana Travassos Tolotti, Laurent Dagorn, and Manuela Capello. Machine learning to detect bycatch risk: Novel application to echosounder buoys data in tuna purse seine fisheries. *Biological Conservation*, 255:109004, 2021.

[22] Dhiraj Neupane and Jongwon Seok. A review on deep learning-based approaches for automatic sonar target recognition. *Electronics*, 9(11), 2020.

[23] Tunai Porto Marques, Alireza Rezvanifar, Melissa Cote, Alexandra Branzan Albu, Kaan Ersahin, Todd Mudge, and Stephane Gauthier. Detecting marine species in echograms via traditional, hybrid, and deep learning frameworks. In *International Conference on Pattern Recognition (ICPR)*, 2020.

[24] 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.

[25] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7263–7271, 2017.

[26] 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.

[27] David G Reid. Report on echo trace classification. *ICES Cooperative Research Report*, (238), 2000.

[28] 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. In *NeurIPS Workshop Tackling Climate Change with Machine Learning*, 2019.

[29] 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):115–22, 2010.

[30] 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.

[31] Yue Shang and Jianlong Li. Study on echo features and classification methods of fish species. In *International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 1–6. IEEE, 2018.

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

[33] 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):e05906, 2021.

[34] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1492–1500, 2017.