ICES Journal of Marine Science

ICES Journal of Marine Science (2021), https://doi.org/10.1093/icesjms/fsab227

Original Article

A deep learning-based method to identify and count pelagic and mesopelagic fishes from trawl camera images

Vaneeda Allken [1,*], Shale Rosen [1], Nils Olav Handegard [1], and Ketil Malde [1,2]

[1] *Institute of Marine Research, Bergen, Norway*
[2] *Department of Informatics, University of Bergen, Bergen, Norway*

* *Corresponding author: Tel: +4748293794; e-mail: vaneeda@hi.no*

Allken, V., Rosen, S., Handegard, N. O., and Malde, K. A deep learning-based method to identify and count pelagic and mesopelagic fishes from trawl camera images. – ICES Journal of Marine Science, 00: 1–13.

Received 6 May 2021; revised 1 October 2021; accepted 20 October 2021.

Fish counts and species information can be obtained from images taken within trawls, which enables trawl surveys to operate without extracting fish from their habitat, yields distribution data at fine scale for better interpretation of acoustic results, and can detect fish that are not retained in the catch due to mesh selection. To automate the process of image-based fish detection and identification, we trained a deep learning algorithm (RetinaNet) on images collected from the trawl-mounted Deep Vision camera system. In this study, we focused on the detection of blue whiting, Atlantic herring, Atlantic mackerel, and mesopelagic fishes from images collected in the Norwegian sea. To address the need for large amounts of annotated data to train these models, we used a combination of real and synthetic images, and obtained a mean average precision of 0.845 on a test set of 918 images. Regression models were used to compare predicted fish counts, which were derived from RetinaNet classification of fish in the individual image frames, with catch data collected at 20 trawl stations. We have automatically detected and counted fish from individual images, related these counts to the trawl catches, and discussed how to use this in regular trawl surveys.

Keywords: acoustic-trawl survey, deep learning, deep vision, fish abundance estimation, fish classification, fish detection, image analysis, object detection, RetinaNet.

Introduction

Trawl sampling is an established method for obtaining biological data on marine ecosystems, and can reveal important information on species composition, population parameters such as age, size, and maturity, as well as predator–prey interactions. In particular, trawl surveys are used extensively by fisheries biologists across a wide range of ecosystems, and constitute a key data source for managing human impact on the marine environment and ecosystems (Evans and Grainger, 2002; Johnsen *et al.*, 2019). Trawl sampling is also an important part of acoustic-trawl surveys, where the species composition of the trawl catches provide information for assigning the acoustic backscatter to taxa, the length distribution of individual fish for conversion of acoustic energy into fish abundance or biomass, and information about age composition (Simmonds and MacLennan, 2005). Recent work has shown that a trawl equipped with underwater cameras can provide much of the same information without the need to capture fish (Williams *et al.*, 2010; Rosen and Holst, 2013). While species and sizes can be resolved from images, some information, like determining age and diet, still depends on physical sampling. In contrast, time- and depth-referenced camera images provide the fine-scale distribution within the volume trawled, whereas the trawl catch data represents only the aggregate distribution. This increase in spatial resolution is particularly useful when combined with acoustics, since it allows species to be allocated with increased precision to specific regions of the water column. The acoustic data is typically presented as an echogram, where the reflected echo energy is presented by depth and time (or distance sailed), which aligns well with the time- and depth-referenced images.

Trawl surveys are typically set up to catch a narrow range of target species, and codend mesh sizes are optimized for that purpose.

Small non-target fishes (juveniles or small species) pass through the meshes and are absent from or under sampled in the catch. In-trawl camera systems can identify and quantify juvenile fish as small as young of the year individuals (Underwood *et al.*, 2014). This is relevant for both registering the youngest year classes of commercially important species and for assessing non-commercial species or fish stocks where fisheries are being developed. Mesopelagic fish is an example of an emerging fishery where data are scarce and highly uncertain (Irigoien *et al.*, 2014; St. John *et al.*, 2016; Proud *et al.*, 2019), in part because the mesh size of the net typically used in trawl survey retains a biased sample of these species due to their size range (~2–20 cm). Here, optical systems can provide much needed data on species abundance and distribution.

A serious obstacle in using camera systems effectively is the amount of time required to manually review the images (Underwood *et al.*, 2014). This "analysis bottleneck" is a major challenge for deployment of high-volume sensors in general, and affects a range of marine science applications (Malde *et al.*, 2019). Automated image analysis is needed to make effective use of large-scale data, and algorithms have been developed that automate the process of identifying and measuring fish from conveyor belts images (White *et al.*, 2006), within trawls (Williams *et al.*, 2016), in aquaculture (Zion *et al.*, 2007), and in open marine environments (Shafait *et al.*, 2016).

In the past decade, the field of computer vision has progressed rapidly, driven by the introduction of deep learning (LeCun *et al.*, 2015) algorithms. Of particular interest here is the class of object detection algorithms, systems that identify and locate individual objects in images. Two-stage object detectors work by firstly identifying a number of candidate object locations (regions of interest) in an image, and then processing those regions to identify and classify actual objects (Girshick, 2015). In contrast, one-stage object detectors (Liu *et al.*, 2016; Redmon *et al.*, 2016) locate and classify objects in a single operation, and are often more suitable for real time analysis.

Object detection algorithms based on deep learning are increasingly being used for underwater imagery analysis, including fish detection and classification (Moniruzzaman *et al.*, 2017). Jalal *et al.* (2020) used a YOLO (You Only Look Once) deep neural network with temporal information acquired *via* Gaussian mixture models and optical flow to detect and classify fish in unconstrained underwater videos, and Ditria *et al.* (2020) showed that a deep learning model trained to estimate fish abundance can outperform humans by up to 13.4% on single image datasets.

These methods typically require large annotated datasets to train the algorithms. Such training sets are challenging to build as they require huge volumes of data to be scrutinized by experts, which is typically a severe drain on resources and is often prone to human-error. An alternative and more efficient approach is to use synthetic data to train a convolutional neural network (CNN) and test on real data. This requires far less manually annotated training data and has successfully been used to separate images containing pelagic fish species in a previous study (Allken *et al.*, 2018) showing that training on a combination of real and synthetic images generated using only 70 fish cutouts per fish species resulted in up to 94% accuracy in image classification.

The objective of this study is to train a deep neural network with a small annotated dataset augmented using synthetic images to automatically identify and count individual fish from trawl camera images, and to explore whether automated analysis of images can in part or fully replace physical sampling from the trawl catch in scientific surveys. We use data from surveys targeting important pelagic species in the Norwegian Sea, specifically blue whiting (*Micromesistius poutassou*), Atlantic herring (*Clupea harengus*), Atlantic mackerel (*Scomber scombrus*) and mesopelagic fishes (Mueller's pearlside, *Maurolicus muelleri* and glacier lanternfish, *Benthosema glaciale*). We used a manually annotated set of trawl camera images to estimate the accuracy of our object detection methods, and develop a statistical model for comparing the aggregate predictions to actual catch data registered in the surveys.

Material and methods

Image data

We use a publicly available data set of in-trawl images collected using the Deep Vision trawl camera system (Scantrol Deep Vision A/S, Bergen, Norway) during two cruises in May 2017 and 2018 in the Northeast Atlantic Ocean (Allken *et al.*, 2021). The in-trawl camera system was placed between the extension and the codend of a Multipelt 832 pelagic sampling trawl (Egersund Trål AS, Egersund, Norway) used for surveying small pelagic species in the North-East Atlantic. The trawl has an opening 25 m high x 60 m wide with mesh size grading from 16 m in the wings and forward section to 22 mm in the codend. The path of the trawl typically sampled multiple depth layers during a haul, and images were taken at an interval of 200 ms (100 ms during a hardware test at one station). In total, we collected 1266397 stereo image pairs from the 2017 cruise and 782618 images from the 2018 cruise. Cameras and lighting were upgraded between the 2017 and 2018 cruises, resulting in sharper but slightly lower final resolution images for the 2018 data (1228 x 1027 pixels downsampled from 2456 x 2054 pixels native camera resolution) as compared with the 2017 data (1392 x 1040 pixels, native camera resolution). More technical details on the Deep Vision camera system are reported in Allken *et al.* (2021).

Annotated datasets (D1, D2, and D3)

From these data, a data set was constructed by selecting and annotating images from 20 trawl stations from each year (Allken *et al.*, 2021). Individual-labelled images were organized into the following categories: (i) blue whiting, (ii) Atlantic herring, (iii) Atlantic mackerel, (iv) mesopelagic fishes (Mueller's pearlside and glacier lanternfish), or (v) mixed, if more than one of the four above species / groups was represented in the image. The two mesopelagic fishes, Mueller's pearlside and glacier lanternfish, are of similar small size (< 10 cm), shape and colouration and could not be reliably distinguished in the Deep Vision images and were therefore grouped. The larger size of blue whiting, herring and mackerel (20–35 cm) makes it much easier to detect characteristic features such as the number and placement of dorsal and ventral fins and body patterning to reliably differentiate them from one another. Each image was annotated with the species/group name and the image coordinates of the bounding boxes that encapsulated each fish (see Allken *et al.*, 2021, section 2.3). Images were grouped into three sets. One set with fully annotated images in addition to fish crops of individual fish from the first four categories a-d) was used to augment the training data (see e.g. Allken *et al.*, 2018), and two separate sets (*D*2 and *D*3, 1536 images in total) of fully annotated images without crops, including mixed images, used for training/validation and testing, respectively. The dataset *D*1 refers to the source-train2017-annotations.csv and source-train2018-annotations.csv in Allken *et al.* (2021), and *D*2

and *D*3 refers to the val_annotations.csv and test_annotations.csv data sets, respectively.

Synthetic datasets ($D1_m$ and $D1_{ms}$)

A previous study (Allken *et al.*, 2018) showed that including synthetic data in the training set increases classification accuracy when training data is limited. We adopt a similar approach here and reserve most (82%) of the real annotated images for validation and testing, using the same augmentation techniques to extract as much information as possible from a relatively small number (343 out of 1879) of training images.

Synthetic images containing between one and ten randomly selected fish crops, originating from *D*1 (353 individual fish crops extracted from 343 images) were generated for each year. After a number of transformations including resizing, flipping or rotation, the fish crops were pasted on background images (20 per year) at random positions selected so that at least one-third of each fish was visible from the edges. The Python script and procedure used to generate synthetic images is described in further details in Allken *et al.* (2018, 2021). We found in prior experiments that training CNNs on synthetic images created using backgrounds and crops from different years resulted in a poorer classification performance than when each synthetic image was formed using crops and backgrounds from the same year. The explanation is likely to be related to upgraded lighting and cameras between the 2 years (2017 and 2018), which resulted in sharper images and brighter, more even lighting with a more neutral colour in the 2018 dataset (Allken *et al.*, 2021, section 2.1.1). Some fish crops also contain a part of the original background that would contrast sharply with a background from a different year.

Fish tend to aggregate by species, and individual fish observed in real images are more likely to belong to the same species. To simulate this, we generated two synthetic datasets, $D1_m$ and $D1_{ms}$, each composed of 20000 synthetic images (10000 per year) with corresponding annotations. In $D1_m$, all of the 20000 synthetic images are composed of crops randomly selected from all four species, whereas in $D1_{ms}$, only $\frac{1}{5}$ of the dataset is composed of mixed species images while the remaining $\frac{4}{5}$ of the dataset is composed of single species images (4000 mixed images and 16000 single species images of blue whiting, herring, mackerel and mesopelagic fishes, respectively). This allows us to evaluate whether training on a dataset that reflects this distribution influences the performance of the network.

Datasets used for training, validation, and testing

Only real (non-synthetic) images were used for validation (*D*2) and testing (*D*3). Different combinations of real (*D*1 and *D*2) and/or synthetic ($D1_m$ and $D1_{ms}$) images were used for training. By varying the number of real images between 0 and 652 ($D1 + 0.5 \times D2$) and the number of synthetic images between 0 and 20000, 41 different training datasets were created. Models were validated after each epoch on the remaining images from *D*2 and for each dataset, the best trained model was evaluated on *D*3.

Network architecture, training procedure, and performance evaluation

One common approach to object detection uses a separate region proposal process to identify putative objects and a convolutional network to evaluate each proposed region. Since the initial R-CNN (Girshick *et al.*, 2014), new and improved region proposal methods have been developed, resulting in Fast R-CNN (Girshick, 2015) and Faster R-CNN (Ren *et al.*, 2015). As an alternative approach, single-shot object detectors incorporate region proposals and classification in a single process. This often leads to fast processing times Redmon *et al.* (2016). Similar to region proposal methods, there has been a series of improvements raising speed and accuracy from the early models like SSD (Liu *et al.*, 2016) and YOLO (Redmon *et al.*, 2016) to recent models like YOLO v3 (Redmon and Farhadi, 2018) and RetinaNet (Lin *et al.*, 2017b). One challenge with single-shot detectors is that for object detection tasks, the number of negatives (i.e. locations with no object) outnumbers positive cases by a large margin, leading to a severe class imbalance. In addition, many datasets are dominated by examples (positive and negative) that are easy to classify. RetinaNet is shown to outperform Faster R-CNN and YOLO v3 (Redmon and Farhadi, 2018), and it uses a technique called focal loss that reduces the loss for well-classified samples. This emphasizes information from samples that are difficult to classify, accelerating learning for difficult cases. RetinaNet also incorporates multi-resolution classification using a feature pyramid network (Lin *et al.*, 2017a). The Feature Pyramid Network detects objects at different scales by constructing a set of multiscale feature maps from each input image, and using nine translation-invariant anchors at each position of the feature map. This gives the network more flexibility when object sizes vary. For our data, fish vary in size from the small mesopelagic (< 10 cm in length) to the larger pelagic species (20–30 cm in length), and in apparent size due to variable lens proximity. Different species can often be difficult to distinguish, especially when viewed partially or from the ventral side. We therefore believe that RetinaNet is an appropriate architecture for this task.

We use the Keras implementation of RetinaNet, initialized with weights pre-trained on ImageNet, and using default hyperparameters (Adam optimizer, learning rate $= 10^{-5}$, iteration steps per epoch $= 10000$). We trained the network over 50 epochs using different combinations of real and synthetic images and then validated the network using real images from D2. All data used for training were subject to standard augmentation methods as implemented by RetinaNet.

Mean average precision

The metric typically used to evaluate the performance of an object detection model is the mean average precision (mAP). For each image, the model generates a set of predictions, each of which is associated with a prediction score that indicates the confidence in the prediction. The intersection between the bounding boxes of each detection and a ground-truth annotation is calculated. If the intersection over union (IOU) is over the threshold (set here at 0.5), the prediction is considered a true positive. The precision is the number of true positives divided by the number of detections and the recall is the number of true positives divided by the number of annotations. By selecting a confidence threshold, we can trade off precision for recall. The average precision (AP) is calculated by varying this threshold, and measuring the precision for a number of different recall values for each class, then computing the area under the precision-recall curve. The mAP is the mean of the APs over all classes.

For each of the training data sets, the weights that gave the highest mAP values (on the validation dataset) were used to evaluate the test set *D*3.

Using model for predictions

In order to use the model for predictions on unannotated data, the model was converted to an inference model. In the default configuration of RetinaNet for inference models, each class (of fish) is processed separately, which can result in multiple independent predictions for the same fish. To avoid this, we use the –no_class_specific_filter option when converting the model.

```
python keras_retinanet/bin/convert_model.py
–no_class_specific_filter path/to/snapshot/ path/to/inference/model
```

The confidence score threshold used for predictions was derived empirically, using the F1 score, which is the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} = \frac{2TP}{2TP + FP + FN}, \quad (1)$$

where TP, FP, and FN are the number of true positives, false positives, and false negatives, respectively. The denominator is equal to the sum of the number of predictions (TP+FP) and the number of annotations (TP+FN). The number of true positives and the number of predictions are a function of the score threshold. The score threshold corresponding to the maximum F1 score, was used as optimal score threshold and used for model predictions.

Estimating fish abundance and distribution across entire trawls

In contrast to the 1879 images used for training/testing, the images from the trawl hauls were not annotated individually, but the catch (number of individuals per species) was recorded for each trawl haul in 2018. The sum of predicted fish counts by species was compared with the catch data for each trawl haul. Trawl sampling in 2017 was carried out using a codend with an open seam to limit catch sizes and was therefore not quantitatively assessed.

A large proportion of the 2049016 images collected by the Deep Vision camera were empty, especially at the beginning and end of the trawl haul where the trawl is at the surface and the trawl opening is collapsed. We used the Deep Vision software (Deep Vision Analysis version 3.3, Scantrol Deep Vision, Bergen, Norway) to remove images containing air bubbles or netting from the beginning and end of the trawl haul and to identify images containing fish. But an examination of a set of images identified as active by the software revealed that many of those images were still either empty or only contained krill. Running the object detection algorithm on those images resulted in of false positives as the model is only trained to recognize the four species described previously and may identify any artefact in an ”empty” image or krill as one of those. We, therefore, trained a simple classification neural network to identify images that were empty or contained only krill (see Supplementary materials). We assembled a random selection of images from the 2017 and 2018 trawl surveys and sorted them into ”empty”, ”fish” and ”krill” images. An empty image consisted of any image not containing fish or krill. Images containing fish were labelled as ”fish” images even if they also contained krill, whereas ”krill” images only contained krill. Out of the 5589 images thus sorted (3378 empty, 1292 fish, and 919 krill images), 30% were held out for validation/testing and the rest was used for training. We ran our best model (classification accuracy of 90% on the test set) on all the trawl images and only 16.4 % of the images from the 2018 survey were predicted to contain fish. We subsequently ran our best object detection algorithm on the ”fish” images for predictions. While the classification model may misclassify a percentage of fish images as ”empty” or ”krill”, we estimated that the number of false positives (empty images classified by RetinaNet as one of the four species) outnumbered the fish images missed by the classification network.

Comparing model predictions with catch data

The fish distribution for the 2018 cruise data was estimated by sampling the catch at each trawl station using established protocols (Mjanger *et al.*, 2017). The weight of the entire catch was measured, then a randomized sub-sample was taken and total weight and length distribution for each species was measured in the subsample and scaled up to the entire catch. We can, thus, compare catch data estimates for the larger species (blue whiting, herring, and mackerel) with our predictions. For the mesopelagic fishes (Mueller’s pearlsides and glacier lanternfish) we have no reliable catch data. These are small enough to escape through the 22 mm (diamond) meshes in the trawl’s codend, and were not registered in the catch at any of the stations. The manual review of the images determined mesopelagic fishes were present at all but four of the 20 stations.

Fish often take longer than the 200 ms interval between frames to pass through the field of view. Thus, the predicted count is inflated by multiple images taken of the same fish. To compare our results to the catch estimate, three linear multiple regression models were fitted to the estimated catches (one model for each species). We expect the Deep Vision counts to be linearly related to the catch, and that the slope will be a measure of how many times a fish is imaged by the system.

We hypothesize that the intercept is zero, i.e. there are no density dependent effects so the number of fish has no effect on the accuracy of the counts. We further assume that a high catch of the other species will cause the Deep Vision counts to increase due to misclassifications. We also include an interaction term since we also expect that when the species in focus increases, there will be misclassifications that work the other way, i.e. the target species will be misclassified as the other species, and at some point this will counter the linear term of the other species.

Taking herring as an example, we fit a model where the Deep Vision counts per trawl station for herring was predicted using the herring catch estimates and the sum of mackerel and blue whiting catches as predictors.

Results

Performance based on training data composition

We trained RetinaNet on 41 separate datasets composed of different combinations of real ($D1$ and $D2$) and/or synthetic images ($D1_m$ and $D1_{ms}$; Table 1) for 50 epochs and for each combination, saving the model producing the best validation set mAP. Each trained model was then evaluated on the same test set, $D3$ of 918 images (Figure 1).

On average, training a model for 50 epochs took 24 h and the evaluation of $D3$ took 97 s (i.e. an average of 0.106 s per image) on an NVIDIA GeForce RTX 2080 Ti graphics card. A model trained using only 343 real images from $D1$ resulted in a test mAP of 0.717. Training RetinaNet exclusively on synthetic images based on the

Table 1. Composition of real and synthetic images. All images in set *D*1, *D*2, and *D*3 are manually labelled with bounding boxes surrounding each individual fish. The *D*1 set is also used as the source for individual fish crops for generating synthetic images for training. *D*2 are real images split between training and validation, and *D*3 are solely used for testing.

Datasets	Real images					Synthetic images					Total
	BW	H	Ma	Me	Mix	BW	H	Ma	Me	Mix	
*D*1 (tr)	87	97	77	82	–	–	–	–	–	–	343
$D1_m$ (tr)	–	–	–	–	–	669	675	678	62	17 316	20 000
$D1_{ms}$ (tr)	–	–	–	–	–	4 000	4 000	4 000	4 000	4 000	20 000
*D*2 (tr+val)	158	145	107	102	106	–	–	–	–	–	618
*D*3 (test)	237	214	158	151	158	–	–	–	–	–	918

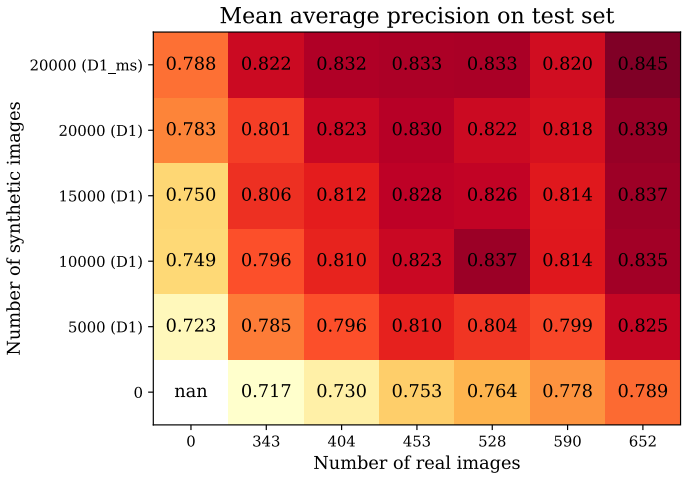

Figure 1. Mean average precision (mAP) on test set *D*3 when RetinaNet was trained on datasets with different configurations of real (*x*-axis) and synthetic images (*y*-axis). The top row $D1_{ms}$ illustrates the effect of training on a synthetic dataset with a balanced distribution of single (4000 per species) and mixed species (4000) images, see Section 2.1.2.

$D1_m$ dataset, improved the mAP by up to 9.2% (for 20000 synthetic images). Using a dataset composed of real and synthetic images improved performance further (see Figure 1, columns 2–7). With images from D1 and 20000 images generated using crops from D1, we obtained a mAP of 0.801 (a 11.7% increase in performance when compared with training only with real images), showing that a relatively small number of annotations (652 images) is sufficient to train an object-detection network adequately when using the right augmentation techniques. In contrast, nearly doubling the number of real images in the training set from 343 to 652 (by adding up to 50% of *D*2) improved the mAP by 10.1% (mAP = 0.789). A model trained on 652 real images and 20000 synthetic images resulted in a mAP of 0.839.

Prior knowledge about image composition improves performance

Training the model on the synthetic dataset, $D1_{ms}$ where the fish composition was more realistic, i.e. 80% of the images contained fish of the same species, resulted in a further increase in performance (Figure 1, upper row), showing that the model benefited from prior knowledge related to the likely distribution of species composition within an image. Substituting $D1_m$ by $D1_{ms}$ in a dataset composed of 652 real images and 20000 synthetic images, resulted in an increase in mAP score from 0.839 to 0.845.

Optimal score threshold

The score threshold determines the precision and the recall. There is a trade-off effect between those two performance metrics as increasing the score threshold increases precision Figure 2a) but reduces the recall (Figure 2b). We use the F1 score, which is a performance metric that incorporates precision and recall in a single measure to determine the optimal score threshold. We derived the confidence threshold for our predictions empirically using the best inference model (mAP = 0.843) and the corresponding validation data. In this case, the best model was trained on 20000 synthetic and 652 real images (train+50% val), we used the remaining 309 images from the validation dataset to evaluate precision, recall and F1 score for score thresholds ranging from 0.05 to 1. The maximum F1 score was obtained at 0.48 for blue whiting, 0.47 for herring, 0.53 for mackerel, and 0.43 for mesopelagic fish (Figure 2c). The maximum F1 score for all species together was obtained at a score threshold of 0.47 (Figure 2d).

Evaluating the predictions

The breakdown of the predictions on the test set *D*3 (when the score threshold was set to 0.47) for the best model compared to the ground-truth is shown in Table 2. Mesopelagic fish were not misidentified as other fish species but were missed in 11% of cases. Herring, which is the species most highly represented (39%) in the

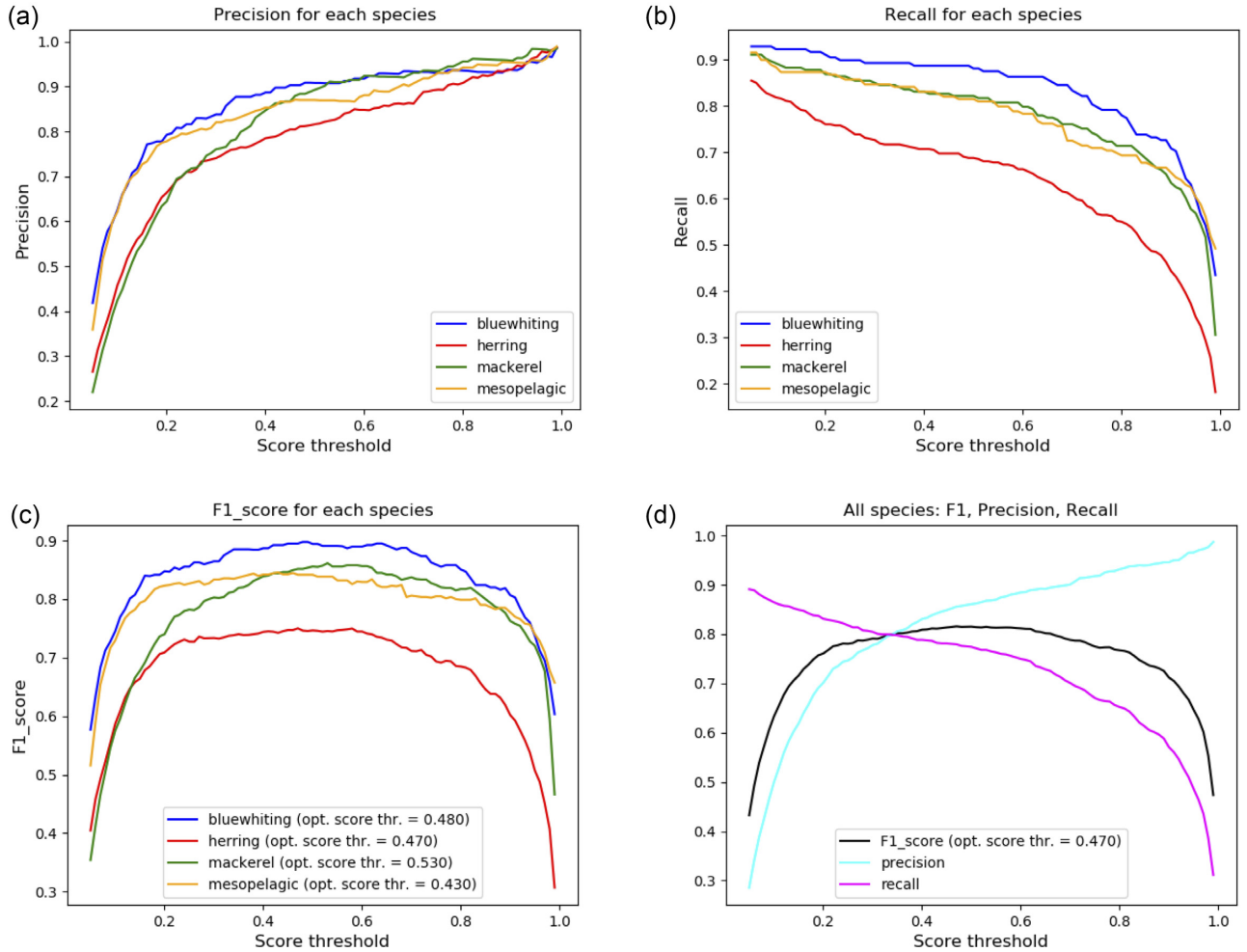

Figure 2. Precision (top left), recall (top right), and F1 score (bottom left) for each species and for all species together (bottom right) on a dataset of 309 images (50% of dataset *D2*) for score thresholds between 0 and 1. The maximum F1 score for all species was obtained at a score threshold of 0.47.

Table 2. Confusion matrix (score threshold = 0.47) where the first four rows show the total number of fish annotated for that species and the distribution of the predictions, including missed detections. For example, out of 470 instances of blue whiting (BW), 385 were correctly identified, while 36 were misidentified as herring (H), two as mackerel (Ma), three as mesopelagic fishes (Me), and 44 were missed. Numbers in bold indicate correct identifications. The fifth row shows the false positives (excluding fish misidentified as one of the other three species) for each species.

Species	BW	H	Ma	Me	Missed	Total
BW	**385**(82%)	36 (8%)	2 (<1%)	3(<1%)	44 (9%)	470
H	24 (3%)	**673**(72%)	28 (3%)	0	208 (22%)	933
Ma	2 (<1%)	33 (6%)	**455**(81%)	0	68 (12%)	559
Me	0	0	0	**360**(89%)	44 (11%)	404
False pos.	58	153	79	50		
Total (pred.)	469	895	582	413		

test data, is also most frequently missed (22% of all herrings) by the model. In this dataset, the total number of false positives makes up for the false negatives, so that there is little difference between the predicted and real counts. For example, 8% of the blue whiting and 6% of the mackerel are misidentified as herring. This, along with other instances of false positives, reduces the difference between the total predicted herring count (895) and the real count (933).

Predicting fish distribution and abundance across trawl stations

The trained model was used to predict the number of fish (by species/group) in all images (including un-annotated images) from all trawl stations. The model outputs were combined with the position where the images were taken and acoustic data to reconstruct the distribution of the different species in relation to the measured acoustic backscatter. For example, at station 364 (Figure 3),

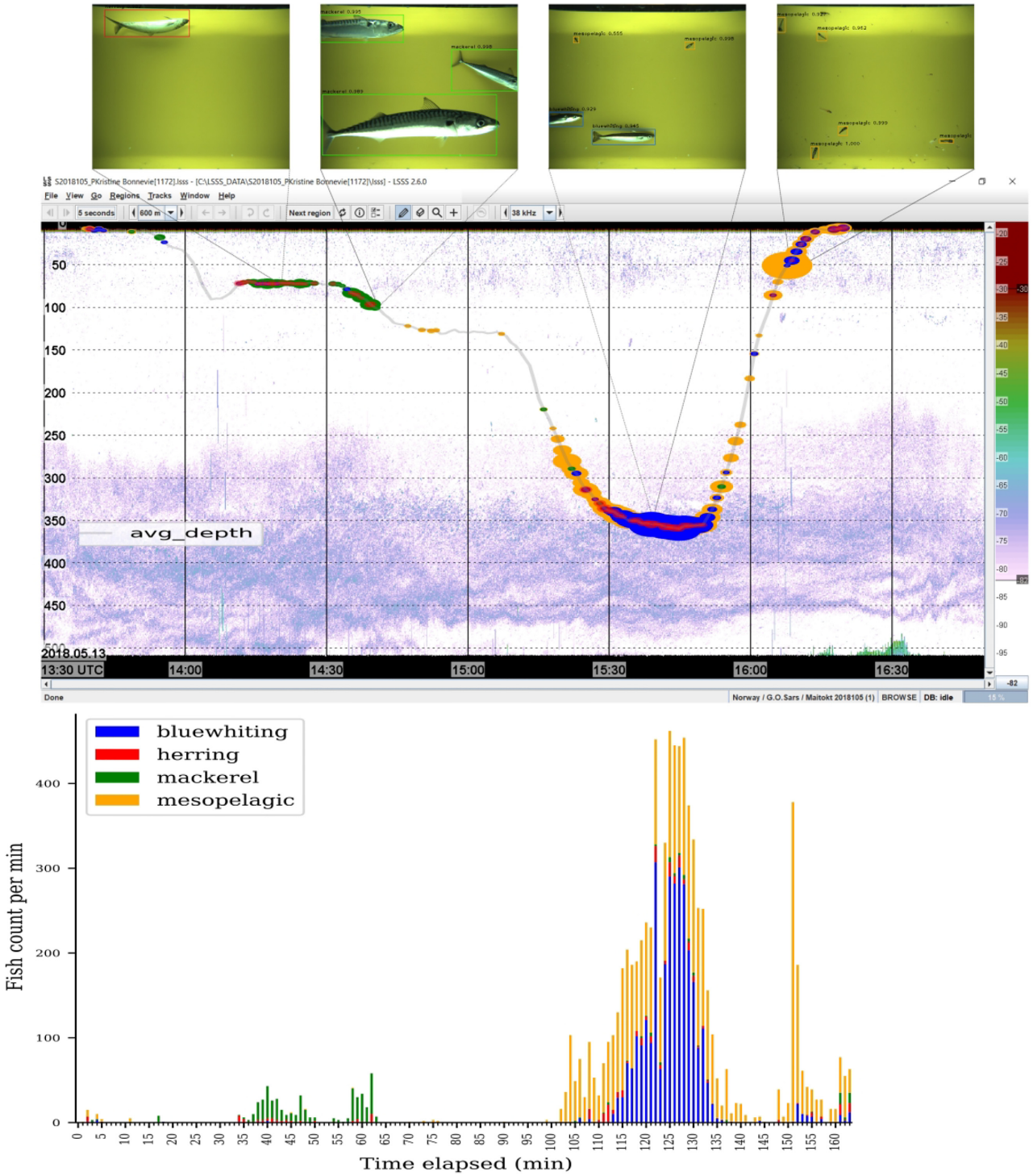

Figure 3. Echogram of station 364 (38 kHz) with depth profile and predictions of blue whiting (blue), herring (red), mackerel (green), and mesopelagic fishes (orange). Images at top are from the positions indicated along the trawl's path and show the bounding boxes calculated by the object detection model. The size of the bubbles in the centre panel are proportional to the number of fish predicted per minute. The stacked histogram at bottom similarly shows the number of each fish species per minute but avoids the problem of the blue whiting symbol obscuring mesopelagic fishes.

Table 3. Catch and Deep Vision (DV) predictions for left and right images per trawl station. No mesopelagic fishes were registered in the catch at any of the 20 stations.

Sta	Blue whiting			Herring			Mackerel			Mesopelagic	
	Catch	DV count		Catch	DV count		Catch	DV count		DV count	
		Left	Right		Left	Right		Left	Right	Left	Right
343	887	9 519	9 811	40	1 421	1 371	0	242	250	1 396	1 465
344	28	392	432	19	1 124	1 119	288	9 872	10 234	190	237
346	0	4	2	1	4	4	1	4	5	41	46
347	0	980	941	0	140	144	127	1 018	1 058	84 626	87 493
348	0	19	12	2	35	27	3	16	15	566	623
349	0	1 902	2 261	1 398	18 262	18 002	947	40 615	40 736	797	1 132
350	0	104	123	110	1 168	1 141	42	1521	1 546	34	38
351	598	3 296	3 288	2	343	295	0	83	76	64	218
352	8	45	41	46	278	273	10	86	90	1 360	1 632
353	2	362	417	416	7 761	8 028	0	460	539	255	399
354	1	69	74	63	707	733	0	71	87	324	383
355	0	397	464	564	12 253	12 627	0	661	702	481	1 771
356	15	399	466	794	7 919	8 211	0	374	396	1 397	1 599
357	0	456	555	818	9 728	9 616	0	523	565	789	983
358	0	861	1 056	4 527	77 320	77 904	0	2 370	2 690	511	700
359	0	123	154	1 054	4 370	4 486	0	217	202	54	80
360	0	284	332	473	11 927	12 251	0	512	557	352	537
361	183	1 581	1 754	422	9 955	10 560	0	514	568	547	568
362	105	796	827	4	208	181	0	157	138	7 027	7 041
364	614	3 077	3 265	9	288	269	90	493	495	4 189	4 296

the mackerel were encountered at relatively shallow depth (70–100 m) while herring were encountered both in the shallow layer mixed with the mackerel and deeper than 250 m in association with mesopelagic fishes and blue whiting. These distributions match backscatter layers visible in the echogram.

Comparing image prediction counts and trawl catches

The prediction counts were compared with the trawl catches. The Deep Vision system has a stereo camera, and the images comes in pairs. While only left images were used in the annotated training, validation and testing datasets, we ran the model on both left and right images when predicting on unannotated images from the trawl stations. The total counts from all non-empty left and right images from all the trawl stations (Table 3) were compared to check for prediction consistency. There was very little difference between counts by species from the left and right images: average CV between right and left images for each species (weighted by count at each station) was 4.8% for blue whiting, 1.3% for herring, 1.4% for mackerel, and 4.1% for mesopelagic fishes.

The number of times an individual fish is captured in consecutive images may be species dependent. To check this, we first summed the Deep Vision counts (average of count in left and right images) across all stations for a given species (Table 3) and divided by the summed catch. The ratios were 10.4, 15.4, and 40.0 for blue whiting, herring, and mackerel, respectively. This is consistent with subsequent analyses of the factors contributing to the number of duplicate times a fish was imaged. Those results showed that species and orientation were significant while size, time, and the number of other fishes present were uncorrelated (Westergerling, 2021).

We further used regression models to predict the DV counts based on the catches. Typically, counts are modelled using a poisson, quasi poisson, or negative binomial model. The initial data exploration showed that the data were overdispersed preventing us from using the poisson distribution. Using a model based on the quasi poisson model resulted in highly heteroscedastic residuals. The negative binomial model did not converge (using R). Valid models were constructed using a log–log model adding one to the catches and DV counts to avoid zero inflation.

Using these transformations, a simple linear regression was performed to predict the summed Deep Vision counts for each station. The total number of pelagic fish caught across all species was used as a predictor variable. When no fish were caught, we would expect the model to predict zero counts, but the intercept was found to be 1.85 (Figure 4a) and significantly different from zero (t-test, $p < 0.001$, see Supplementary material for details).

To evaluate the correlation between species catch and Deep Vision counts, multiple linear regression models were built. For each species, two model types were considered: (i) reduced models predicting each species' Deep Vision count based on the species' catch alone and (ii) full models predicting each species' Deep Vision count using species' catch and the sum of catches from the other two species as covariates.

For all species, the full model (including an interaction term) was found to have a significantly better fit than the reduced model (F-tests, all tests $p < 0.05$, see Supplemantary material). When including the other species as covariates, none of the models had an intercept different from zero (t-tests, all $p < 0.05$, see Appendix). The models show that there is an increase in Deep Vision counts when the catches of the other species are high (Figure 4). The model coefficients are 0.74, 0.62, and 0.84 for the log-transformed sum of the other species for blue whiting, herring, and mackerel, respectively. When the catch of another species is high, a small labelling error of the predominant species will lead to an inflation in the counts for

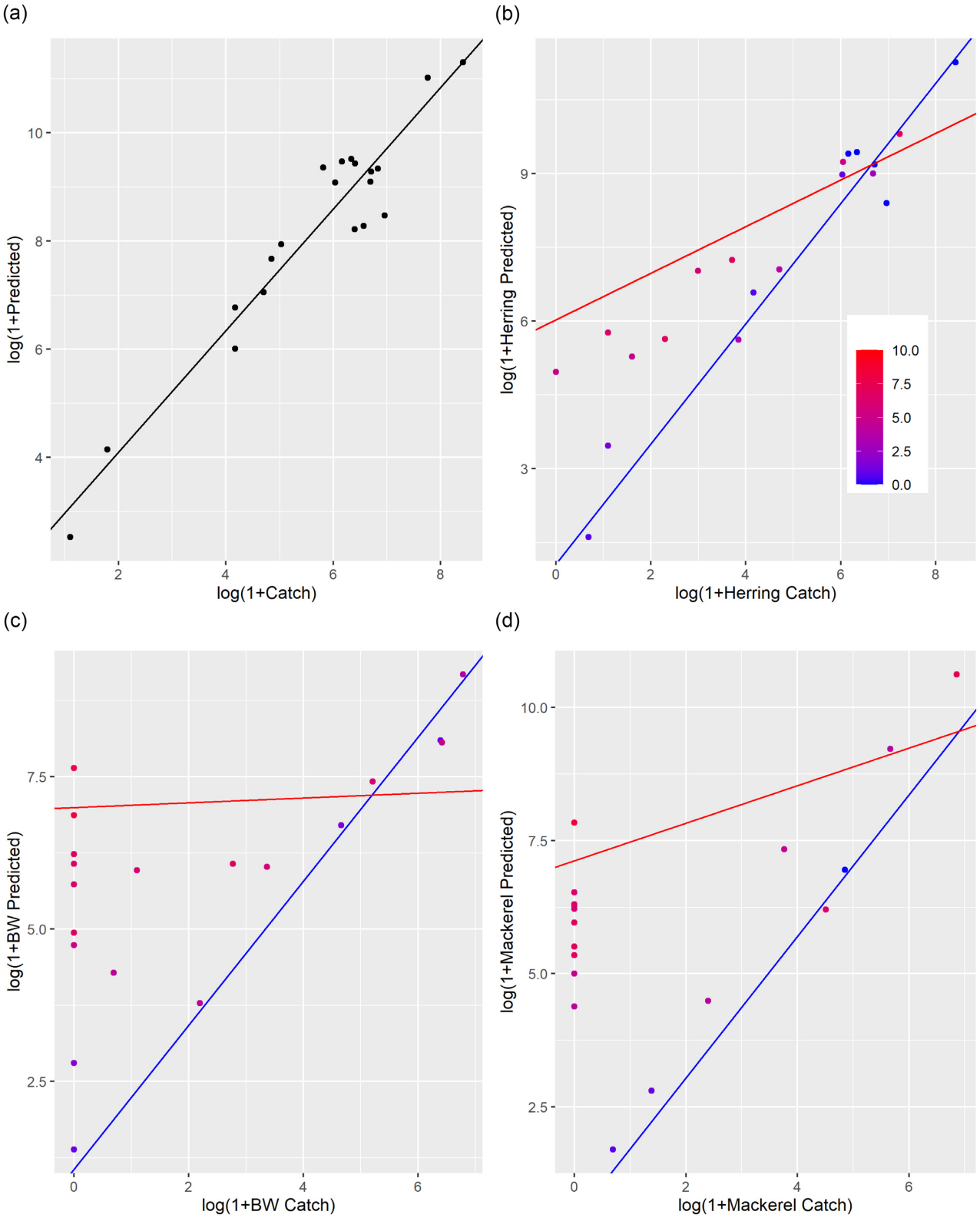

Figure 4. Trawl catches vs. Deep Vision predictions. (a) The log-transformed sum of the catches vs. predictions for herring, blue whiting, and mackerel combined. The dots are the individual data points per station, the black curve is prediction from the linear regression models (B, C, and D). The log-transformed predictions as a function of catches for herring, blue whiting (BW), and mackerel, respectively. The colour of the dots denote the log-transformed sum of the other species, e.g. for the herring plot (B) this sum is the log-transformed sum of the blue whiting and mackerel catch from that station. The blue and red lines are the regression lines from the multiple regression model when the log-transformed sum of the other species are zero and eight, respectively.

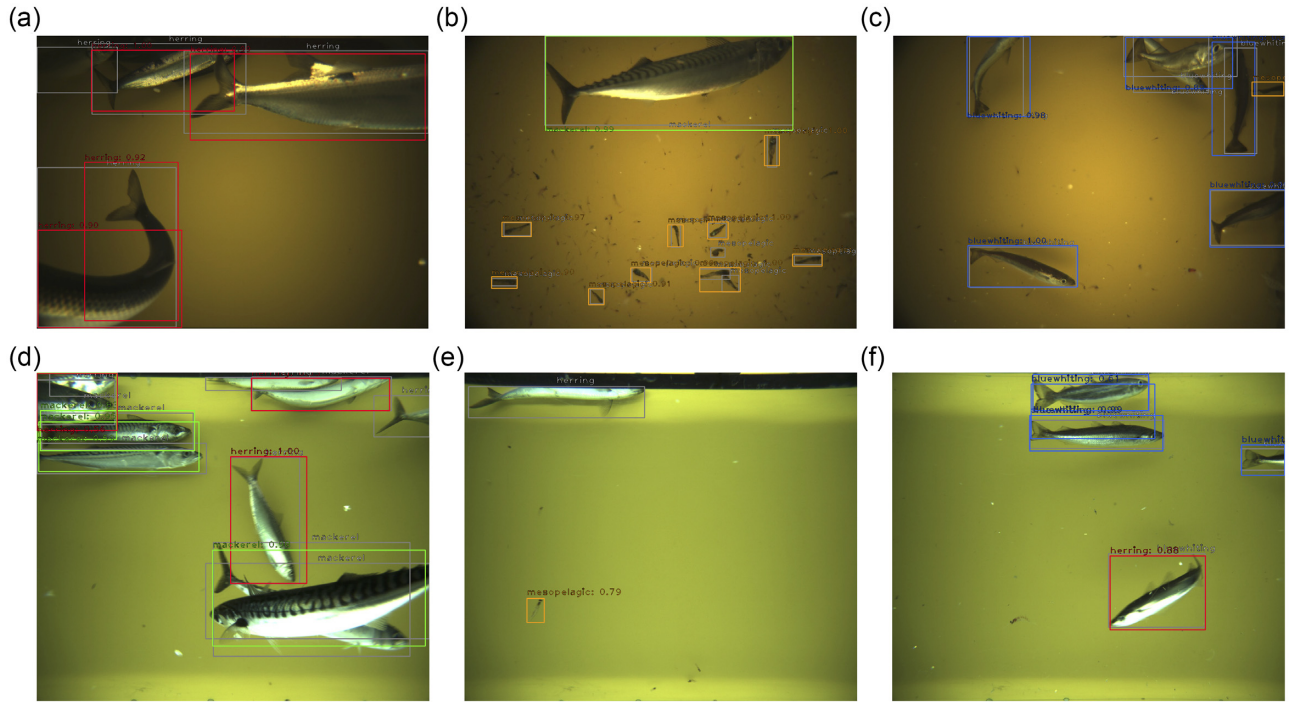

Figure 5. Wrong/missed predictions. All annotation boxes are drawn in grey while the predictions boxes are red for herring, green for mackerel, blue for blue whiting, and orange for mesopelagic fishes. Examples of false positives in (a) where a herring is counted more than once; (c) and (e) where the visible part of a fish and a krill respectively are misclassified as mesopelagic fish; and (f) where a blue whiting is wrongly classified as herring. Examples of false negatives in (a) where one partial herring is missed; (b) where one instance of a mesopelagic fish is missed; (c) where two blue overlapping blue whiting are counted as one; (d) where a partial herring is missed and two overlapping mackerels are counted as one; and (e) where a herring in the top of the image is missed. Top row images are from 2017 and bottom row from 2018 showing the effect of upgraded cameras and lights.

the minor species. This leads to overcounts in cases where the catch of the lesser species is low or even zero when the other species has a high catch. This effect is seen for all three species (Figure 4b–d).

There is also an interaction effect between the catch of a species and the sum of the other species. The (negative) model coefficients for the interaction terms are −0.14, −0.094, and −0.13 for the blue whiting, herring, and mackerel, respectively (see Supplemantary material). This indicates that when the catch of the target species is large, the misclassification works in the other direction, i.e. fish are erroneously labelled as the other species. The interaction effect counters the effect of the other species for high catches only (Figure 4b–d).

Discussion

Using an implementation of RetinaNet, we have demonstrated how an object detector can be used to process trawl camera images, automatically locating fish individuals and identifying their species. This allows us to observe and quantify fish before they are captured, which has important applications in fisheries oceanography, fisheries management, and for developing selective fishing methods.

Model considerations

Training data for deep learning

A major challenge with deep learning is the need for large amounts of labelled data for the training step, and the annotation process which, in the case of object detection tasks, involves manually drawing boxes around objects is particularly time-consuming. Using synthetic data provides us with a quick and easy way to generate large amounts of clean, automatically annotated images. Whilst doubling the number of real images used for training improved the performance by 10% (from 0.717 to 0.789), this requires an investment of effort and expense, a greater improvement in model performance (14.6%) was achieved through the use of synthetic data (from 0.717 to 0.822, which is an automated and easy task. This suggests that better exploitation of existing data should be considered before the acquisition of new labelled data. To further improve performance, the distribution of the simulated data should follow the distribution of real data. In particular, pelagic fish often school, leading to observations that are dominated by a single species. Simulating data by drawing from a uniform distribution of species was found to be inferior to using a distribution that favoured dominant species observations.

Model limitations

A visual inspection of missed or wrong predictions in the test set images shows that there are several causes for mis-classifications. Reasons for false positives (predictions > counts) include misidentification due to poor lighting or partial fish (Figure 5c), fish being misidentified (Figure 5f), or small objects like krill being wrongly identified as mesopelagic fishes (Figure 5e) or the same fish being counted twice (Figure 5a). False negatives are frequently observed in cases where there are several fish overlapping and not all are counted (Figure 5b–d). We also have false negatives in cases

where only part of a fish is visible in an image (Figure 5a and d) or when the lighting is suboptimal (Figure 5a–c).

The average score threshold used (0.47) is lower than the optimal score threshold for mackerel (0.53) and higher than that of mesopelagic fishes (0.43), which may cause us to overpredict mackerel and underpredict mesopelagic fish. However, an analysis of the test results (see Table 2) shows that the false positives tend to compensate for the false negatives, so that the predicted count differs at most 4% from the true count. Overall, the sum of total fish predictions (2359) is just 0.3% less than the total number of fish in the images (2366) as overestimation for one species compensates for underestimation for other species.

When using the model for predictions on real unannotated images, it is likely that mesopelagic fish are over-counted. While our approach includes a pre-processing step where images predicted to contain krill are excluded from the images used for predictions, a proportion of the remaining images may still contain objects that the model was not trained to recognize. Small non-fish objects, such as bubbles, when the trawl is near the surface and scales, krill, and fragments of larger fishes when at the fishing depth tend to be misidentified as mesopelagic fishes. The model may, therefore, not be adequately equipped to distinguish small objects from mesopelagic fish.

Application in fisheries oceanography surveys

Trawl sampling is used extensively in fisheries oceanography and fisheries management, and long time series exist for a wide range of regions. Trawl catches can be used directly in swept-area or swept-volume surveys to calculate abundance or in combination with acoustics in order to verify the species and sizes present for apportioning acoustic backscatter.

Increased spatial resolution from trawl data

The most immediate application of data from an in-trawl camera system is likely improved interpretation of acoustic data. The improved spatial resolution from the trawl can also support studies of species overlap and trophic interactions.

Presently, assigning acoustic backscatter to species and sizes is accomplished by trained experts examining reflected energy over multiple frequencies (De Robertis *et al.*, 2010; Korneliussen and Ona, 2003) and assigning the presence and relative abundance of species aided with data from the trawl catch. However, there is a mismatch in spatial resolution between acoustic data, which are recorded at scales of < 10 m along the vessel's path and catch in the trawl, which is integrated over the kilometer-scale path trawled. Time- and depth-referenced images provide a fine-scale three-dimensional record of where each object was encountered. While a traditional catch sample yields a mixed catch integrated over the entire volume sampled, depth- and time-referenced images can be linked directly to the observed layers.

While it is possible to examine images manually, using our method for identifying fish automatically, we can easily display the observed fish over the echogram (Figure 4). The Large Scale Survey System (LSSS) software (Korneliussen *et al.*, 2016) is commonly used to analyse echosounder data, and currently supports manual inspection of trawl camera images, which are placed in the corresponding location on the echogram based upon depth and time stamp indicated in the Deep Vision metadatafile. Our goal is to integrate the classification system presented here to automate the process of species identification and abundance estimation. Techniques to extract length measurements automatically from the images, such as those presented by Garcia *et al.* (2019), could be similarly integrated to provide information for target strength conversions in order to estimate biomass from the acoustic data.

Quantitative estimates of fish abundances

Swept-area or swept-volume surveys use trawl catches directly as a relative measure of abundance. This is most common for surveys of demersal fishes, but is also used for some pelagic surveys such as the mackerel swept surface survey in the Norwegian sea (Nøttestad *et al.*, 2015).

We believe the species and size composition measured from images taken inside the trawl may provide a more complete record of the range of species and sizes present than the physical catch retained in the cod end. Trawls are generally size-selective (Wileman *et al.*, 1996), with most selection occurring in the codend where fish accumulate. Individuals smaller than the codend meshes may pass through and be either underrepresented or completely missing from the catch. This dataset provides a clear example of this challenge: the mesopelagic fishes, glacier lanternfish, and Mueller's pearlsides were verified in the images at 16 of the 20 trawl stations from the 2018 survey but were never recorded in the catches. Mesopelagic fish is an emerging fishery, and abundance estimates are uncertain (Irigoien *et al.*, 2014; Proud *et al.*, 2019). Using the Deep Vision system on conventional fisheries oceanography surveys can provide much needed data on the species abundance and distribution.

A frame rate of 200 ms causes individual fish to be imaged several times as they pass the camera system. For this reason we cannot use the Deep Vision counts directly as a measure of abundance (c.f. Table 3). The number of duplicate images per individual appears to be species dependent. The sum over all stations of Deep Vision counts divided by the catches are 10.4, 15.4, and 40.0 for blue whiting, herring, and mackerel, respectively. These trends are reflected in the different swimming capacities of these species (Videler and Wardle, 1991; Sambilay, 1990), with speed and endurance of blue whiting $<$ herring $<$ mackerel. Thus, we believe the species dependent effect is most likely due to different residence times inside the field of view of the camera. However, even if the mackerel is a fast swimmer, 40 frames for each individual seems high. Another explanation is that the counts are affected by misinterpretations from the other species, see discussion of the model below. For our method to be applied to trawl surveys and for maintaining established time series, we need to correct for these differences.

One way to mitigate multiple observations is to track each individual fish between subsequent frames. The advantage of this approach is that it can be generally applied, but requires a robust tracking algorithm. As fish densities are sometimes high, this either requires a robust method to identify and separate different fish individuals (rather than just species), or a high enough frame rate that fish positions are close enough between frames to resolve ambiguities. It is also likely that the tracking would be biased for high counts, since tracking in high densities increase the probability of track association errors (Blackman and Popoli, 1999).

Instead of tracking we have modelled the relationship between the Deep Vision counts and the catch data using a simple linear regression model. This requires concurrent catch data from calibration and does not directly generalize across species and surveys. Our model does not currently include other factors that can affect

the swimming behaviour and the corresponding residence time, such as fish size, orientation, water temperature, and fish activity (feeding, spawning, migrating, and so on), as these datasets are still under development. Given enough data, some of these could be added as covariates to the model.

As shown in Figure 5, our model worked well in our test case. Overall catches and counts from images were correlated along a 1:1 linear regression in a log–log plot. For the cases where we use catches of the other species as covariates, no intercepts were significantly different from zero (t-test, $p > 0.05$, Supplementary material). There is, however, a positive trend in the intercepts suggesting that there could be a small density dependence, which potentially caused the RetinaNet object detector to miss fish due to occlusion in crowded images.

The sum of catch from the other species is a larger and significant (t-test, $p < 0.01$ for all cases) effect compared to the density dependent effect, especially for blue whiting and mackerel (Figure 5b-d). As an example, the non-significant (t-test, $p > 0.05$) intercept for blue whiting is 1.06, which corresponds to $\exp(1.06) + 1 \sim 4$ fish, and the effect of the sum of the other species is 1.18, which will affect the prediction a lot more than the intercept when the log of the sum of the other catches are above 1. This effect is pronounced even with a low false positive rate when the abundance of one species is much higher than the target species.

Finally, we note that the performance of the trawl may change when using an open cod-end due to changes in hydrodynamics. If the method is to be used on an existing survey with an open cod-end, the trawl performance needs to be monitored. One way to address this is to conduct every other haul with a closed cod end and check if there are any biases in the Deep Vision counts between a closed and an open cod-end.

Concluding remarks

We have developed a method that can reliably count and identify individual fish from in-trawl images. The challenge is more related to the data than to the model architecture, and it is important to ensure that the data is appropriately collected and labelled.

To implement the method on a fisheries oceanography survey, we recommend the following steps: First, a set of training data needs to be established to train the deep learning network. This includes empty background images, crop outs of species of interest as well as fully annotated data set split into training, validation, and testing. Second, for using the method on swept area/swept volume surveys, trawl hauls with catches and in trawl images need to be recorded to establish the correspondence between counts from images and trawl catches. Third, if the objective is to use the system with an open cod end, it is recommended to initially do every other haul with a closed cod end to detect potential effects of opening the cod end. And finally, it is recommended to keep collecting at least some closed cod end trawl hauls to monitor the performance of the system and to ensure that there is no drift in the performance. Continued sampling of physical catch is also likely necessary in order to collect life-history parameters such as diet, age, maturity, and condition which cannot be assessed from images.

Supplementary data

Supplementary material is available at the *ICESJMS* online version of the manuscript.

Data availability

The data used to train and test the models in the article and the code used to generate synthetic data is described in more detail at https://doi.org/10.1002/gdj3.114. The data are available in the Norwegian Marine Data Centre at https://doi.org/10.21335/NMDC-551736490. More information about the survey data can be found in the Supplementary material.

Acknowledgements

We acknowledge partial support from the COGMAR, CRISP, and CRIMAC projects funded by the Research Council of Norway (grants 270966O70, 203477, and 309512) and the Machine learning project and the REDUS projects funded by the Norwegian Ministry of Trade, Industry and Fisheries. We thank Sindre Vatnehol for helping with the data collection during the 2018 survey.

References

Allken, V., Handegard, N. O., Rosen, S., Schreyeck, T., Mahiout, T., and Malde, K. 2018. Fish species identification using a convolutional neural network trained on synthetic data. ICES Journal of Marine Science, 76: 342–349.

Allken, V., Rosen, S., Handegard, N. O., and Malde, K. 2021. A real-world dataset and data simulation algorithm for automated fish species identification. Geoscience Data Journal, 00: 1–11. https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/gdj3.114 (last accessed 10 November 2021).

Blackman, S. S. and ,Popoli, R. 1999. Design and Analysis of Modern Tracking Systems. Radar Library.

De Robertis, A., McKelvey, D. R., and Ressler, P. H. 2010. Development and application of an empirical multifrequency method for backscatter classification. Canadian Journal of Fisheries and Aquatic Sciences, 67: 1459–1474.

Ditria, E. M., Lopez-Marcano, S., Sievers, M., Jinks, E. L., Brown, C. J., and Connolly, R. M. 2020. Automating the analysis of fish abundance using object detection: optimizing animal ecology with deep learning. Frontiers in Marine Science, 7: 429.

Evans, D., and Grainger, R. 2002. Gathering Data for Resource Monitoring and Fisheries Management, chapter 5, pp. 84–102. John Wiley & Sons, Ltd.https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470693919.ch5 (last accessed 10 November 2021).

Garcia, R., Prados, R., Quintana, J., Tempelaar, A., Gracias, N., Rosen, S., Vågstøl, H. *et al.* 2019. Automatic segmentation of fish using deep learning with application to fish size measurement. ICES Journal of Marine Science, 77: 1354–1366.

Girshick, R. 2015. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. *In* Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 580–587.

Irigoien, X., Klevjer, T. A., Røstad, A., Martinez, U., Boyra, G., Acuña, J. L., Bode, A. *et al.* 2014. Large mesopelagic fishes biomass and trophic efficiency in the open ocean. Nature communications, 5: 1–10.

Jalal, A., Salman, A., Mian, A., Shortis, M., and Shafait, F. 2020. Fish detection and species classification in underwater environments using deep learning with temporal information. Ecological Informatics, 57: 101088.

Johnsen, E., Totland, A., Holmin, A. J., and Skålevik, A. 2019. Stox: an open source software for marine survey analyses. Methods in Ecology and Evolution, 10: 1523–1528.

Korneliussen, R. J., Heggelund, Y., Macaulay, G. J., Patel, D., Johnsen, E., and Eliassen, I. K. 2016. Acoustic identification of marine species using a feature library. Methods in Oceanography, 17: 187–205. Special

section on Novel instrumentation in Oceanography: a dedication to Rob Pinkel.

Korneliussen, R. J. and ,Ona, E. 2003. Synthetic echograms generated from the relative frequency response. ICES Journal of Marine Science, 60: 636–640.

LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep learning. Nature, 521: 436–444.

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. 2017a. Feature pyramid networks for object detection. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollar, P. 2017b. Focal loss for dense object detection. *In* Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. and Berg, A. C. 2016. Ssd: Single shot multibox detector. *In* Lecture Notes in Computer Science, Springer. 21–37.

Malde, K., Handegard, N., Eikvil, L., and Salberg, A. B. 2019. Machine intelligence and the data-driven future of marine science. ICES Journal of Marine Science, 04: 1274–1285.

Mjanger, H., Svendsen, B., Fotland, Å., Mehl, S., and Salthaug, A. 2017. Håndbok for prøvetaking av fisk og krepsdyr (in norwegian). Technical Report Versjon 4.0, Institute of Marine Research, Bergen, Norway.

Moniruzzaman, M., Islam, S. M. S., Bennamoun, M., and Lavery, P. 2017. Deep learning on underwater marine object detection: A survey. *In* Advanced Concepts for Intelligent Vision Systems, Ed by Blanc-Talon, J., Penne, R., Philips, W., Popescu, D., and Scheunders, P., pp. 150–160. Springer International Publishing. Cham.

Nøttestad, L., Utne, K. R., Óskarsson, G. J., Jónsson, S. Þ., Jacobsen, J. A., Tangen, Ø., Anthonypillai, V. *et al.* 2015. Quantifying changes in abundance, biomass, and spatial distribution of northeast atlantic mackerel (*Scomber scombrus*) in the nordic seas from 2007 to 2014. ICES Journal of Marine Science, 73: 359–373.

Proud, R., Handegard, N. O., Kloser, R. J., Cox, M. J., and Brierley, A. S. 2019. From siphonophores to deep scattering layers: uncertainty ranges for the estimation of global mesopelagic fish biomass. ICES Journal of Marine Science, 76: 718–733.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. 2016. You only look once: Unified, real-time object detection. *In* Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788.

Redmon, J. and ,Farhadi, A. 2018. Yolov3: An incremental improvement. Cite arxiv:1804.02767Comment: Tech Report. http://arxiv.org/abs/1804.02767 (last accessed 10 November 2021).

Ren, S., He, K., Girshick, R., and Sun, J. 2015. Faster r-cnn: towards real-time object detection with region proposal networks. Cite arxiv:1506.01497Comment: Extended tech report. http://arxiv.org/abs/1506.01497 (last accessed 10 November 2021).

Rosen, S., and Holst, J. C. 2013. DeepVision in-trawl imaging: Sampling the water column in four dimensions. Fisheries Research, 148: 64–73.

Sambilay, J. V.C. 1990. Interrelationships between swimming speed, caudal fin aspect ratio and body length of fishes. Fishbyte, 8: 16–20.

Shafait, F., Mian, A., Shortis, M., Ghanem, B., Culverhouse, P. F., Edgington, D., Cline, D. *et al.* 2016. Fish identification from videos captured in uncontrolled underwater environments. ICES Journal of Marine Science, 73: 2737–2746.

Simmonds, J. and MacLennan, D. 2005. Fisheries Acoustics. Theory and Practice. Blackwell Science, Oxford.

John, St. M. A., Borja, A., Chust, G., Heath, M., Grigorov, I., Mariani, P., Martin, A. P. *et al.* 2016. A dark hole in our understanding of marine ecosystems and their services: Perspectives from the mesopelagic community. Frontiers in Marine Science, 3: 31.

Underwood, M. J., Rosen, S., Engås, A. and Eriksen, E. 2014. Deep vision: an in-trawl stereo camera makes a step forward in monitoring the pelagic community. PLOS ONE, 9: 1–8.

Videler, J., and Wardle, C. 1991. Fish swimming stride by stride: speed limits and endurance. Reviews in Fish Biology and Fisheries, 1: 23–40.

Westergerling, E. 2021. A Comparison of an In-Trawl Camera System to Acoustic and Catch Results for Small Pelagic and Mesopelagic Fish. Masters thesis, Ghent University, Ghent, Belgium.

White, D. J., Svellingen, C., and Strachan, N. J. C. 2006. Automated measurement of species and length of fish by computer vision. Fisheries Research, 80: 203–210.

Wileman, D. A., Ferro, R. S. T. Fonteyne, R. and Millar, R. B. 1996. Manual of methods of measuring the selectivity of towed fishing gears. Technical Report 215, ICES. https://www.ices.dk/sites/pub/Publication%20Reports/Cooperative%20Research%20Report%20(CRR)/CRR%20215.pdf (last accessed 10 November 2021).

Williams, K., Lauffenburger, N., Chuang, M.-C., Hwang, J.-N. and Towler, R. 2016. Automated measurements of fish within a trawl using stereo images from a camera-trawl device (camtrawl). Methods in Oceanography, 17: 138–152. Special section on Novel instrumentation in Oceanography: a dedication to Rob Pinkel.

Williams, K., Rooper, C., and Towler, R. 2010. Use of stereo camera systems for assessment of rockfish abundance in untrawlable areas and for recording pollock behavior during midwater trawls. Fishery Bulletin, 108: 352–362.

Zion, B., Alchanatis, V., Ostrovsky, V., Barki, A., and Karplus, I. 2007. Real-time underwater sorting of edible fish species. Computers and Electronics in Agriculture, 56: 34–45.

Handling Editor: David Demer