ICES Journal of Marine Science, 2025, Vol. 82, Issue 9, fsaf168
https://doi.org/10.1093/icesjms/fsaf168
Received: 19 February 2025; **revised:** 11 July 2025; **accepted:** 31 August 2025
Original Article

Automated fish detection in videos to support commercial fishing sustainability and innovation in the Alaska walleye pollock (*Gadus chalcogrammus*) trawl fishery

Katherine C. Wilson [1,*], **Moses Lurbur** [2,3], **Noëlle Yochum** [1,4]

[1]Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, WA 98133, USA
[2]Pacific States Marine Fisheries Commission, Under contract to Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, WA 98133, USA
[3]Present address: University of Washington, Paul G. Allen School of Computer Science & Engineering , Seattle, WA 98195, USA
[4]Present address: Currently at: Trident Seafoods, Fishing Innovation and Sustainability , Seattle, WA 98107, USA

*Corresponding author. Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, WA 98133, USA.
E-mail: Katherine.wilson@noaa.gov

Abstract

Bycatch reduction devices (BRDs) are used in the Alaska walleye pollock (*Gadus chalcogrammus*) fishery to reduce Pacific salmon (*Oncorhynchus* spp.) bycatch. Evaluation of BRD effectiveness often requires people to process collected or live-feed video. Deep learning can be used to detect and classify fish in video to support BRD and other fisheries bycatch work. We fine-tuned and evaluated the detection model EfficientDet and YOLO11 to find salmon and pollock in videos collected inside a trawl using 11 572 salmon and 73 394 pollock annotations from 16 989 video frames. We evaluated model performance across all data and during high abundances of krill, varying fish density, camera occlusions, low lighting, and combinations of these using five-fold cross validation. The best performing model was further evaluated by applying it to videos from three fishing tows not used for model training, using it in a salmon presence algorithm that was developed to assess whether an efficient semi-automated video review process was feasible, and comparing it with the performance of a salmon-only detection model. We found that the YOLO models performed better than EfficientDet and on average detected 90% of salmon and pollock with 72% accuracy using a 50% detection overlap threshold. The YOLO models performed comparably to annotators for fish detection: the detection performance was higher for pollock and lower for salmon than the variability measured between annotators. Model performance across trawl and video conditions was more variable for salmon and generally lowest during high fish densities. The YOLO salmon and pollock model performed better than the salmon-only model when using an optimal confidence score threshold. When applied to full fishing tows, the YOLO salmon and pollock model incorrectly detected Pacific herring (*Clupea pallasi*) as salmon, and correctly predicted 99.3% of salmon presences while reducing the number of video frames needing to be reviewed by 85%. Overall, the models detected salmon and pollock well inside a pollock trawl, but camera placement, lighting, and occlusions presented challenges. We provide our annotated dataset, salmon presence algorithm, and recommendations for optimising video quality in trawls.

Keywords: deep learning; computer vision; bycatch reduction; fisheries sustainability; trawl gear

Introduction

In commercial fisheries, video collection has increased over recent decades with the goal of improving fishing operational awareness, meeting regulatory requirements, and enabling research to improve fishing efficiency and sustainability. This increase in video collection has primarily been driven by advancements in camera technologies that have produced a large selection of relatively low-cost, high quality, and readily available camera systems that can be deployed in many different environments. The rise of video collection has increased the need for resources to monitor and review footage, which can be a time-consuming, tedious, and expensive task for people. These challenges can lead to delays in results and innovation. Fortunately, over the last few decades, there have also been major advancements in computer vision and deep learning tools to automate video analysis and reduce the time and cost of generating large video datasets (Zhang et al. 2021).

Deep learning is a subset of machine learning inspired by how the human brain works that uses multiple layers of simpler representations to learn and form hierarchical representations of complex data from experience, which then allows computers to recognize complicated patterns and make predictions (Goodfellow et al. 2016). Deep learning models can be used to automate the detection, classification, and tracking of subjects in images and videos. Convolutional neural networks (CNNs), a class of deep learning models, are widely used for image processing and have been applied in various industries to semi-automate or fully automate video and imagery analysis tasks. They have been used to detect product defects in manufacturing (Wang et al. 2019) and diseases in medical imaging (Liu et al. 2019), to allow vehicles, equipment, and robots to navigate autonomously (Grigorescu et al. 2020), and to reduce the time for reviewing videos for surveillance (Sreenu & Durai 2019) and commercial fishing electronic monitoring applications (Tseng and Kuo 2020). Some notable examples of detection CNNs include EfficientDet (Tan et al. 2020), You Only Look Once (YOLO; Redmond et al.), and Faster region-based CNN (R-CNN; Ren et al. 2017)

For fisheries research and other marine applications, there have been many studies that successfully used CNNs to classify or detect marine fish species to assist with imagery analysis in unconfined open-water conditions (Salman et al. 2016, Qin et al. 2016, Ditria et al. 2020, Ovchinnikova et al. 2021, Alaba et al. 2022). However, the background conditions and number of fish present in the imagery for many of these studies are less challenging than what can occur in high-volume commercial trawl fisheries. Deep learning has also been successfully used at aquaculture farms, which are confined areas with high densities of fish, to monitor fish for disease, feeding behaviour, and more (Zhao et al. 2021). Zhang et al. (2020) used CNNs to detect fish in images from an Atlantic salmon (*Salmon salar*) aquaculture farm in China with a 95% detection accuracy. The Tidal project, originally by X, also used deep learning to track Atlantic salmon for Norwegian aquaculture farms and have had enough success to garner support for commercializing their aquaculture platform (https://x.company/projects/tidal/). Despite having high densities of fish, the object detection task for aquaculture farms do not have to distinguish between different species of fish which is necessary for bycatch detection in trawls.

Applications of deep-learning for trawl surveys have more similarities to the conditions observed in commercial trawl fisheries and have shown promise for automating imagery analysis. Garcia et al. (2020) and Allken et al. (2021) evaluated deep-learning models respectively for fish segmentation and detection in imagery from Deep Vision (Scantrol, Norway; Rosen et al. 2013, Underwood et al. 2014), a trawl camera system developed to identify and measure fish during pelagic trawl surveys. Garcia et al. (2020) used a Mask R-CNN to segment fish and found that the model could detect 96% of non-overlapping fish and 79% of the overlapping or occluded fish, but they did not evaluate species identification. Allken et al. (2021) trained RetinaNet to identify blue whiting (*Micromesistius poutassou*), Atlantic herring (*Clupea harengus*), Atlantic mackerel (*Scomber scombrus*), and mesopelagic fishes in images with no krill present and achieved approximately 85% prediction accuracy when requiring a 50% or greater overlap of predictions and annotations. However, the amount of blue whiting, and Atlantic herring and mackerel caught during the survey tows used in this study were low (947 or less fish) compared to high-volume commercial trawl fisheries like the pollock fishery whose tows usually contain one to two orders of magnitude more pollock (i.e. hundreds of metric tonnes of pollock).

One driver for video collection and the use of deep learning in commercial fisheries is to develop and evaluate the efficacy of new, innovative technologies and methods to improve fishing performance or bycatch reduction. As a step towards developing real-time catch information for the demersal *Nephrops* (Norway lobsters; *Nephrops norvegicus*) fishery, a Mask R-CNN model was used to detect and count Norway lobsters and three other categories of fish with 75% prediction accuracy and 84% detection rate (Sokolova et al. 2021). To support the development of commercial trawl systems capable of in-situ release of bycatch (i.e. active selection), Yi et al. (2024) developed a Coordinate-Aware Mask R-CNN (CAM-RCNN) to increase performance generalisation for different sets of imagery (e.g. imagery from different vessels). The CAM-RCNN had 31% and 57% prediction accuracy respectively for imagery from a new source and imagery from the source used for training. An Institut Français de Recherche pour l'Exploitation de la Mer project (Euronews 2022) and a Heriot-Watt University and Fisheries Innovation and Sustainability partnership project (Hollely 2023) are also working to integrate deep learning into trawl systems to enable automatic detection, sorting, and retention or release of fish.

Many trawl fisheries could benefit from real-time catch information or active selection gear, including the commercial walleye pollock (*Gadus chalcogrammus;* hereafter referred to as "pollock") fishery in Alaska where Pacific salmon (*Oncorhynchus* spp.; hereafter referred to as "salmon") bycatch is a concern. The pollock fishery is the second largest fishery in the world, with 3.4 million tonnes captured in 2022 (FAO 2024). The Alaskan fishery accounts for approximately 40% of these landings (Fissel et al. 2016). If the prohibited species catch limits for Chinook salmon (*O. tshawytscha*) are reached, the fishery is closed by sector and season (Ianelli et al. 2019). BRDs known as "salmon excluders" were developed to reduce salmon bycatch in this fishery. The efficacy of salmon excluders has often been evaluated by reviewing video collected to monitor these devices during fishing (Gauvin and Paine 2004, Gauvin et al. 2011, Gauvin et al. 2013, Gauvin et al. 2015, Gauvin 2016, Lomeli and Wakefield 2019, Yochum et al. 2021). There are also research efforts to develop active selection BRDs that use live-stream video and remotely operated devices within the trawl for this fishery (Rose and Barbee 2022). Developing CNNs to automate video processing to detect fish for these applications as well as other tools to support bycatch mitigation could provide great benefits for these research efforts and the pollock fishing industry.

Despite on-going efforts in other fisheries to develop deep-learning models for bycatch applications, no similar work has been done in the Alaska pollock fishery and there are currently no deep-learning models trained to recognise pollock and salmon. Additionally, as far as we know, none of the previous studies have evaluated the performance of fish detection in a high-volume, pelagic commercial trawl fishery. Therefore, we selected two open-source object detection models with high benchmark dataset performance, EfficientDet and YOLO11, to fine-tune and evaluate for salmon and pollock detection in the pollock fishery in Alaska. We also developed a detection-based salmon presence prediction algorithm to evaluate the feasibility of semi-automated video review for bycatch reduction efforts. We used videos collected during experimental trials of a salmon excluder (Yochum et al. 2021) to train and evaluate the detection models to identify salmon and pollock in the trawl for 10 different trawl and video conditions: krill presence, varying fish density, camera occlusions, low lighting, and combinations of these. A single class salmon-only model was also trained and evaluated in the same manner as one of the multi-class models to assess whether a single-class model could perform better at the task of salmon detection than a multi-class model. The detections from the multi-class and single-class models and their respective salmon presence predictions were also compared for three fishing tows with known salmon occurrences.

Materials and methods

Video collection and review

We used videos that were collected in the pelagic trawl of the F/V *Pacific Explorer* in 2019 and 2020 in Alaska's Eastern Bering Sea during the pollock commercial fishing season (Fig. 1A) to evaluate the performance of a salmon excluder (Yochum et al. 2021). For both years, video was collected dur-

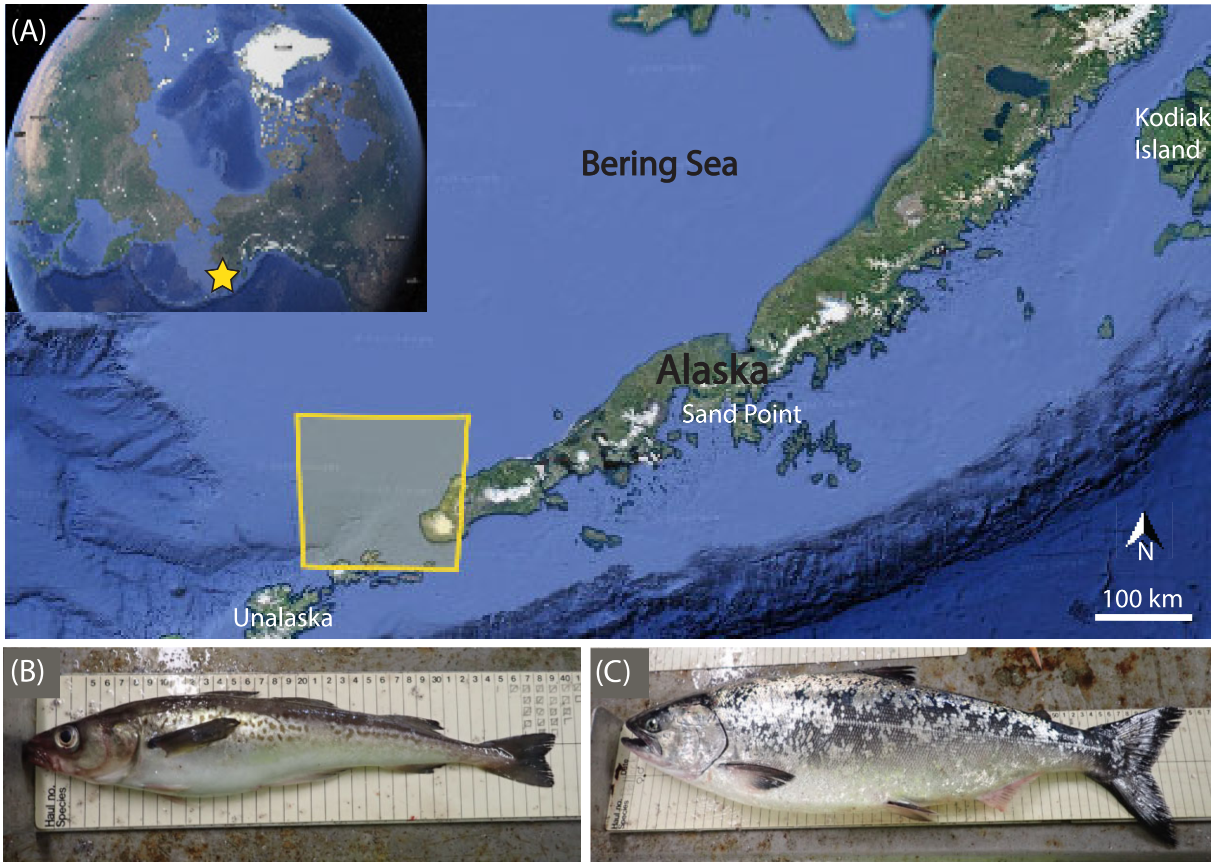

Figure 1: Map of Alaska showing (A) the eastern Bering Sea area of data collection in the yellow box with an inset showing the location in the northern hemisphere marked by a star. Below the map are examples of (B) walleye pollock (*Gadus chalcogrammus*) and (C) Pacific salmon (*Oncorhynchus* spp.) from the catch.

ing multiple research tows in the summer season. The Sexton Corporation's (Salem, OR) trawl camera systems, equipped with a Mobius (Rotterdam, The Netherlands) high-definition action camera and white LED lighting that illuminated from both sides of the camera, were placed at the entrance to the excluder on the inside of the top netting panel facing aft to view the fish at the entrance of the excluder (Fig. 2). The LED lighting system was capable of producing a minimum of 2028 lumens when operated at full brightness. Video was collected at 30 frames per second (fps) at 720p in 2019 and 1080p in 2020. Additionally, for all tows, the vessel's captain estimated the total pollock catch for the tow. The catch estimate and information about when the net was placed in the water, fishing began and ended, and the net was hauled back were recorded.

Yochum et al. (2021) quantified salmon escapement rates and the number of salmon that entered the excluder for the 2019 tows by having trained personnel review the videos and record when salmon were first seen. The 2020 tows were also reviewed by trained personnel for the same purpose, but the time that salmon were last seen in the video was also recorded to support additional analysis.

Annotation

Four people annotated a random selection of the video footage that had previously been reviewed and had records of when salmon were detected to train and evaluate models. The salmon records from four fishing tows from 2019 and three tows from 2020 were used to randomly select 168 clips of footage with salmon present and extract frames that corresponded with these times to create videos for annotation. We used footage from seven tows across both years to capture variability in fishing conditions. For 2019 data, which only had records of when salmon were first seen, two second clips were used due to the absence of information about when salmon were last seen in the videos. A smaller selection of 16 clips without salmon or any fish were also randomly selected and included to ensure the dataset was representative of the most possible scenarios during fishing operations for model training and evaluation. In total, our dataset consisted of a 184 video clips.

All annotators were trained to use the Computer Vision Annotation Tool (CVAT; https://www.cvat.ai/) to annotate the tracks of two classes of objects, salmon and pollock, using rectangular bounding boxes with fish attributes and tag frames for different trawl and video conditions. For tracks, CVAT assigns a unique identifier to each tracked object, in this case every salmon and pollock in the video, and the identifier is associated with all the respective bounding boxes. CVAT can linearly interpolate bounding boxes for tracks based on provided bounding boxes. Annotators were trained to check and, if needed, correct the predicted boxes for each frame if this feature was used.

Subsets of the same data were annotated by the four annotators to assess accuracy and consistency as part of the training process. The lead annotator reviewed these annotations, determined when each annotator was ready to annotate other

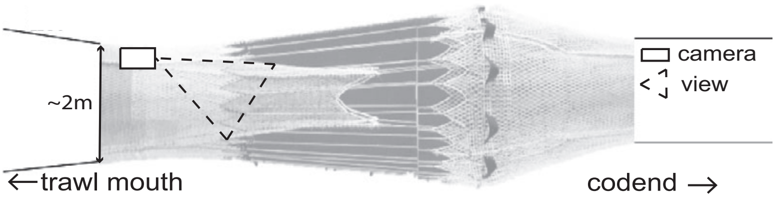

Figure 2: Example image of the salmon excluder (Yochum et al. 2021) shown in the last taper section of the net with the camera placement (white box) and the approximate field of view (dashed triangle) shown. The diameter of the net is approximately 2 m at the beginning of the excluder.

data used in this study, selected a single set of training annotations to include in the final annotation dataset, and reviewed and modified these and all further annotations to ensure consistency and accuracy.

We used several video tags and annotation attributes to differentiate among common trawl and video conditions. These tags and attributes enabled more granular model training and evaluation, and they were all clearly defined to provide consistency among annotations. Video tags were used to indicate six background conditions: (1) the presence of krill (*Euphausiacea*); (2) camera occlusions; (3) low lighting or poor visibility; and (4–6) low, medium, and high fish density (predominantly pollock; Fig. 3). Krill tags were used when the presence of krill impacted an annotator's ability to distinguish the edges of fish. Camera occlusions were defined as having 50% or more of the camera view blocked by an object and tagged accordingly. Low lighting tags were used if 50% or more of the netting normally visible in the frame was too dark to see the individual meshes. The fish density was defined by the proportion of the netting that was visible over a single video annotation clip: 0% to 25% of the netting visible was high fish density, 26% to 50% was medium density, and 51% or more of the net visible was low density. Fish annotations were marked as occluded when 50% or more of the view of the fish was blocked or with the poor visibility attribute if a fish's eyes or fins were not visible. Additionally, annotation attributes were used to mark occluded or poorly visible fish annotations.

All salmon and pollock were annotated from the time they could first be identified as a fish entering the video scene (i.e. the camera's field of view) until they were no longer distinguishable as fish as they either left the scene or began to vanish from view as they moved out of the camera's detection range. Only a small portion of the fish needed to be visible to allow people to identify them. Fish tracks that began to vanish out of the camera's detection range were ended when the fish first became indistinguishable from the background, and the fish was no longer annotated if it became visible in the detection range again unless it had clearly made a permanent change in swimming direction and was moving forward in the net. If fish momentarily became fully occluded by other objects or left the scene, the original fish track was continued if annotators were confident it was the same fish. If it was ambiguous for any reason, a new fish track was started.

For each tow, we calculated (1) the total number of frames; (2) frames without salmon and pollock; and (3) salmon and pollock tracks and bounding boxes included in the final annotation dataset used for model training and evaluation. We also calculated the total number of frames tagged as the three different levels of fish densities (low, medium, or high) with no additional tags, and tagged as these fish densities with any of the two other camera and background condition tags (krill, occluded or low light) or the combination (krill and occluded or low light) for a total of 12 possible conditions.

Annotator performance baseline

The annotator training dataset and Microsoft Common Object in Context (COCO) metrics, a standard metric for measuring object detection performance, were used to measure variability among annotators and provide a baseline from which to compare the performance of people and automated detectors. An unknown level of variability was expected due to subjective choices about bounding box size and classification, and the start and end of track annotations. The dataset had 587 to 902 frames of overlap between only four pairs of annotators due to one annotator not having overlapping training data with two of the other annotators.

We calculated the COCO metrics Average Precision (AP), mean Average Precision (mAP), Average Recall (AR) and mean Average Recall (mAR) (COCO Consortium 2015) between annotator-pairs using the COCO application programming interface for Python (PyPI 2018). AP and AR are for single classes, and the mAP and mAR are respectively the mean AP and AR for all classes for multi-class models.

The AP is a measure of precision (Eq. 1) that conveys the accuracy of detections. It is a measurement of the number of correct (true positives; *TP*) detections relative to the total correct and incorrect detections (false positives; *FP*). When applied to annotators, it is a measure of how many annotations are the same compared to the total number of annotations of the annotator being evaluated. COCO's AP and mAP calculate the average precision across 101 recall values that range from zero to one and 10 intersection over union (IoU) thresholds that range from 0.5 to 0.95. The IoU threshold is the percentage of overlap required for a detection or, in this case the annotation, to be treated as a match with an annotated ground-truth object.

Recall (Eq. 2) measures the ratio of annotated objects detected by a model compared to the total number of annotations, which includes missed detections (i.e. false negatives; *FN*). When applied to annotators, it is a measure of how many annotations are the same relative to the total number of ground-truth annotations. AR (Hosang et al. 2015) and mAR is the recall value averaged over the same IoU thresholds used for AP and mAP.

$$precision = \frac{TP}{TP + FP} \tag{1}$$

$$recall = \frac{TP}{TP + FN} \tag{2}$$

The COCO metrics use the confidence scores assigned by detection models and range from 0 to 1 (i.e. 0% to 100% confidence) to rank all detections and calculate the precision

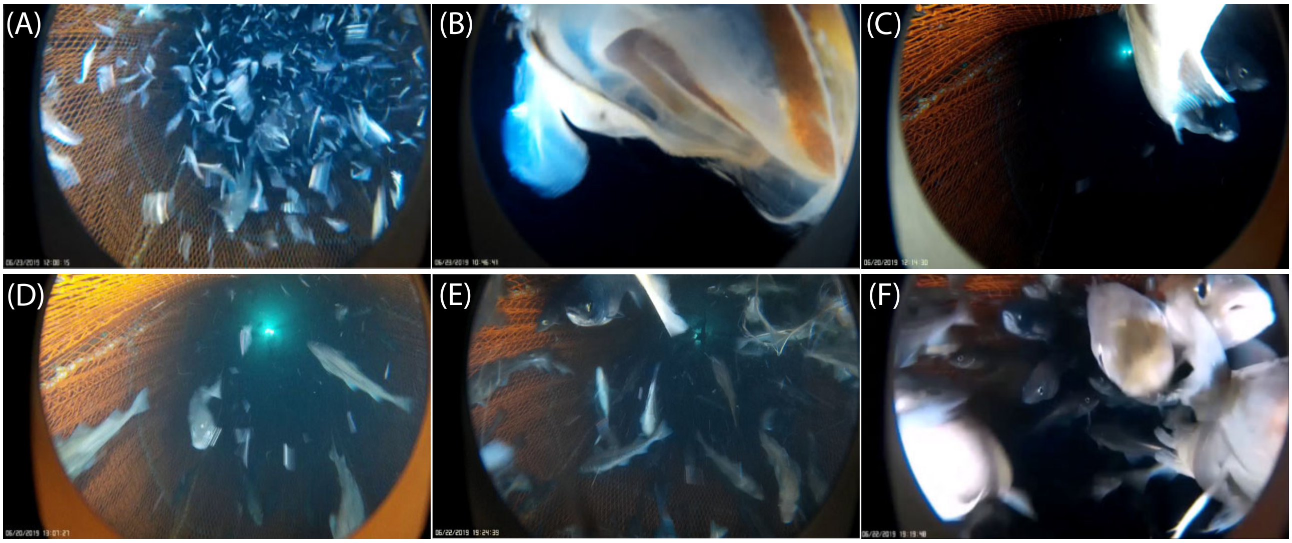

Figure 3: Example video frames of when tags would be used to mark (A) the presence of krill, (B) occlusions, (C) low lighting, and to define fish density as (D) low, (E) medium, or (F) high. A jellyfish is occluding the camera in (B). A single Pacific salmon (*Oncorhynchus* spp.) is present in (C) and (F), and two are present in (E). The remaining fish in the images are walleye pollock (*Gadus chalcogrammus*).

and recall for each detection using the pre-defined IoU thresholds. The AP, mAP, AR, and mAR for all detections is then calculated using these values. These metrics range from zero (i.e. all detections were incorrect or all annotated objects missed) to one (i.e. all detections correct or all annotated objects were detected).

To use the COCO metrics to compare annotators, we calculated the mean and standard deviation of AP, mAP, AR, and mAR, and the AP, mAP, AR, and mAR for bounding boxes with 50% or greater overlap (respectively $AP_{0.5}$, $mAP_{0.5}$, $AR_{0.5}$, $mAR_{0.5}$) between pairs of annotators. These metrics were calculated twice for each pair of annotators: first with one set of annotations acting as the ground-truth and again with the other set of annotations acting as the ground-truth. The mean and standard deviation of the two values were reported as the metric for each pair. The annotations that were treated as predicted detections were all assigned a confidence score of one.

Model training and evaluation

Model selection

We evaluated an EfficientDet (Tan et al. 2020) and YOLO11 (Khanam and Hussain 2024) model for salmon and pollock detection. These two models were selected because at the time of this study they performed better on the COCO 2017 dataset than other open-source and widely used models. COCO 2017 is an object detection benchmarking dataset that allows people to compare the performance of different models by providing a consistent training and evaluation dataset. The COCO 2017 dataset consists of more than 200 000 labelled images with 1.5 million objects that represent 80 object classes (Lin et al. 2014, Suppl. Figure 1). For this study, we selected pre-trained EfficientDet D2 and YOLO11n models that were trained with and optimized for the COCO 2017 dataset. Optimisation involves tuning configurable hyperparameters (model parameters that are set before training) to achieve higher performance. Pre-trained models were chosen instead of using untrained models to reduce the time and computational resources required for training and possibly produce a model with greater prediction generalisation. The EfficientDet model by the Tensor Flow Model Garden (Yu et al. 2020) and the YOLO11 model by Ultralytics (Glenn et al. 2023) were used.

Out of eight available EfficientDet models designed and pre-trained for different image sizes, EfficientDet D2 (hereafter referred to as EfficientDet) for 768 × 768 pixel images was chosen because its image size was the closest match to our lower resolution videos (720p, 1280 × 720 pixels). We chose the smallest and most computationally efficient of five available YOLO11 models, YOLO11n (hereafter referred to as YOLO11), since all were pre-trained for 640×640 pixel images.

Cross-validation dataset

We used five-fold cross-validation to train and evaluate the object detection models as multi-class models for salmon and pollock classes (hereafter referred to as the multi-class model). For this cross-validation, our 184 annotated video clips were randomly assigned to one of five data subsets and, when necessary, clips were re-assigned to balance the number of clips of the different trawl and video conditions across subsets. Five different models were trained and evaluated with all of these data subsets. For example, one model was trained with subsets 1, 2, 3, and 4 and evaluated on subset 5 while another model was trained with subsets 2, 3, 4, and 5 and evaluated on subset 1.

Model fine-tuning

We fine-tuned the pre-trained EfficientDet and YOLO11 models with our annotated dataset to detect salmon and pollock respectively using TensorFlow 2 and the Ultralytics package and most of the default hyperparameters that were provided for these models. We used default hyperparameters to keep the fine-tuning process simple and evaluate the performance achievable with hyperparameters optimized for a large and diverse image dataset. The batch size was reduced to four to accommodate the GPU and memory limitations of our computing resources, and each model's default optimizer's

Table 1. EfficientDet D2 and YOLO11n hyperparameters for training, evaluation, and prediction.

Type	Hyperparameter	EfficientDet D2	YOLO11n
Training	Optimizer	momentum	auto (SGD + Adam)
	Initial cosine learning rate	0.001*	0.001*
	Final cosine learning rate	NA	1E-05
	Momentum	0.9	0.937
	Weight decay	NA	5E-4
	Warm-up epochs	0.046*	0.05*
	Warm-up momentum	NA	0.9*
	Warm-up learning rate	5E-4*	5E-4*
	Box loss	weighted smooth L1	complete IoU
	Classification loss	weighted sigmoid focal (γ=1.5, α=0.5)	binary cross-entropy
	Box loss gain	1	7.5
	Classification loss gain	1	0.5
	Distribution focal loss gain	NA	1.5
	Image size	768 × 768	768 × 768*
	Batch size	4*	4*
	Data Augmentation	None	None
Evaluation and prediction	NMS IoU	0.5	0.5*
	NMS score threshold	1E-8	1E-8*
	Max detections per image	100	100
	Image size	768 × 768	768 × 768*

The first column indicates whether the hyperparameter is for training or evaluation and prediction. NMS refers to non-maximum suppression, SGD is stochastic gradient descent, and NA is used for model hyperparameters that were not used or not user-defined. Settings changed from the pre-trained model's default values are marked by a "∗.".
Insert Source Here

learning rates were adjusted to account for the lower batch size and to use similar settings (Table 1). Optimizers are used to minimize the loss functions during training. EfficientDet uses a momentum optimizer (Polyak 1964) and YOLO11 uses Ultralytic's auto optimizer (combination of AdamW (Loshchilov and Hunter 2017) and Stochastic Gradient Descent (Amari 1993) optimizers). YOLO's post processing non-maximum suppression (NMS) settings, which helps to eliminate duplicate detections, and image size were set to defaults used for EfficientDet. For all other EfficientDet and YOLO hyperparameters, the default values of the pre-trained models were used, including the maximum detections default of a 100 detections per image, which is higher than the number of objects typically present in our video frames.

Using a small subset of annotated data, we explored data augmentation methods with EfficientDet, and determined they did not provide performance gains. Therefore, data augmentation was not used for fine-tuning EfficientDet or YOLO11. The data augmentation methods tested included random image colour distortion, horizontal flips, scaling and cropping, and brightness and contrast adjustments.

The EfficientDet models were trained for 15.4 epochs and the YOLO11 models were trained between 186 and 314 epochs. The number of epochs for the EfficientDet models was determined by monitoring model loss values while training and finding the approximate location where the loss had decreased to its lowest value prior to steadily increasing. The training epoch at this loss minimum should optimize performance while limiting overfitting of the data. Training for the YOLO11 models was automatically stopped after 100 epochs if the model validation metrics did not improve. We used the best saved models for evaluation. All training and training evaluation were performed using a Google Cloud virtual instance with four NVIDIA T4 graphical processing units.

Model performance

After training was complete, AP, mAP, AR, and mAR, and AP$_{0.5}$, mAP$_{0.5}$, AR$_{0.5}$, and mAR$_{0.5}$ for salmon and pollock were calculated for each of the multi-class model's respective evaluation datasets to evaluate overall performance. The general toolbox for identifying object detection errors (TIDE; Boyla et al. 2020) was used to categorize and estimate the percentage of errors that were false positives or negatives and which were caused by classification or localization errors and background or missed detections. False positive detections were further evaluated by applying the models to the video frames where no fish were present using five confidence score thresholds (0.5 to 0.9).

The AP$_{0.5}$ and AR$_{0.5}$ for each class of the best performing multi-class model was calculated across all confidence score thresholds to determine the threshold that produced equal AP$_{0.5}$ and AR$_{0.5}$ (hereafter referred to as the optimal confidence score threshold) prioritizing these two metrics equally, and compared with the performance variability of annotators. The model with the best trade-off between lower false positive detections and higher AP$_{0.5}$ and AR$_{0.5}$ was selected as the best performing multi-class detection model.

Performance variability was evaluated for all multi-class models using the optimal confidence thresholds for AP$_{0.5}$ and AR$_{0.5}$ for the best performing model. The AP$_{0.5}$ and AR$_{0.5}$ for salmon and pollock were calculated for each of the trawl and camera conditions present in each model's respective evaluation datasets.

Salmon-only model

A single-class salmon-only model (hereafter referred to as the single-class model) was also fine-tuned and the model's overall performance, optimal confidence threshold, and performance variability was evaluated in the same manner as the multi-class models. The single-class model was trained and evaluated with only salmon annotations and all pollock annotations omitted using the data subset from the five-fold cross validation that produced the best overall model performance. The model was trained for 250 epochs with the same hyperparameters as the selected best performing multi-class model.

Model validation on fishing tows

Fishing tows

Video from three tows from those collected in 2019 (tows 16, 18, and 20) were selected to further evaluate the detection performance for salmon. The three tows had varying catch rates for pollock and salmon and were not used for model training or evaluation. The same NMS hyperparameters used to evaluate the models were used to run the models for predictions on full length fishing tows (Table 1).

The salmon previously indicated in the Yochum et al. (2021) analysis were reviewed for these three tows and the start and end frames of each individual salmon presence was recorded to create a dataset that included frame information for all salmon. Using an open-source video review software, DJV, a reviewer skipped to the video times where salmon were indicated and recorded the first and last frame that each salmon was identified in the video. Salmon were considered present when the reviewer could distinguish the fish as salmon, including when fish were partially occluded. Salmon were considered not present when they could no longer be recognised because they had become too small or had faded into the background of the video. To ensure the accuracy of individual salmon presence frame data, another person reviewed and recorded the start and end frames of all previously indicated salmon for a random 10% of the videos from all three tows as well as all salmon indicated in the original review sheet that our initial reviewer could not find. Any discrepancies beyond minor differences in start and end frames were discussed and resolved.

Fish detections

The best performing multi-class model and the single-class model were applied to the videos from the three full length fishing tows. All detections with a confidence score equal to or greater than the optimal confidence score threshold for each class were retained to evaluate detection performance for pollock and salmon. The average number of salmon and pollock detections per second was calculated by applying a 30-frame moving average to the total detections per frame of each. The total number of salmon present for each frame of the tow videos was determined from the individual salmon presence frame data. These totals were plotted with the average detections per second for salmon and the total salmon detections per frame to understand the performance of the models and any possible issues with false positive and negative detections. The detections for a small subset of frames from these tows were also plotted to visualize and understand both types of detection errors.

Salmon presence

We established a detection-based salmon presence algorithm to predict general salmon presence and to assess the feasibility of using detections to indicate general salmon presence in support of semi-automated video review. Multiple variables were included to allow us to require consecutive frames of salmon detections or multiple salmon detections per frame to possibly improve salmon presence prediction when salmon false positives or missed detections occurred. The code and models used for salmon presence prediction are publicity available on our Github repository.

The salmon presence algorithm, S, was calculated using three thresholds: the minimum confidence of salmon detections C, the minimum number of salmon detections per frame M, and the minimum number of consecutive frames containing salmon detections N. Salmon presence, S, was predicted as true for a given set of consecutive frames of length N if for every frame there are at least M salmon detections with a confidence score greater than or equal to C.

Salmon presence, S, for a given set of consecutive frames F_k, where k is the set or prediction occurrence number, is determined by the following equations:

$$\left\{d_{ij} \in D_i | d_{ij} \geq C\right\} \tag{3}$$

$$\left\{ f_{i,i+1,i+2,\ldots,i+n} \in F_k || D_i| \geq M, \forall f_i \right\} \tag{4}$$

$$S\left(F_k\right) = \begin{cases} 1, & |F_k| \geq N \\ 0, & |F_k| < N \end{cases} \tag{5}$$

Where D_i is the set of all salmon detections d_{ij} with a confidence score greater than or equal to C for frame f_i where i and j are respectively the frame number and detection number (Eq. 3). A consecutive set of frames was included in the set F if all frames have salmon detections greater than or equal to M (Eq. 4). For each set in F, S is true if the size of the set is greater than or equal to N (Eq. 5).

To determine the best range of values for thresholds C, M, and N, we performed parameter optimisation using the LIPO global optimization algorithm (Malherbe and Vayatis 2017). A random selection of 20% of the videos from each tow were used to optimise parameters. Two metrics were used to choose the preferred values for these parameters: proportion of included frames and presence recall.

The proportion of included frames, I, was defined as the fraction of total frames where salmon presence was predicted (Eq. 6) and is indicative of the proportion of video that a person would need to review. This metric gives an approximation of the potential time savings from a semi-automated review process.

$$I = \frac{\sum_{S(F_k)=1}^{k} |F_k|}{total\ frames} \tag{6}$$

Presence recall, P, was defined as the proportion of general salmon presence occurrences that had at least one salmon presence prediction. We chose this definition of successful presence detection because the prediction algorithm would correctly indicate the video location of salmon for a person conducting a semi-automated review process. If T is a set containing the general ground-truths of salmon occurrences (a consecutive set of frames where salmon is present) where m indicates the set or occurrence, then we calculated presence recall by counting the number of sets in T that intersect with any set in F (i.e. they share at least one frame) and the intersecting F set has $S(F)$ equal to one. We then divided the total correctly predicted salmon presence occurrences by the total true number of occurrences (Eq. 7).

$$P = \frac{\sum_{T_m \cap\ F\ \wedge\ S(F)=1}^{m} 1}{|T|} \tag{7}$$

Two other metrics, frame recall and precision, were used with *I* and *P* to evaluate the optimal salmon presence algorithm. Frame recall was calculated as the number of frames where salmon presence was correctly predicted (true positives), divided by the total number of frames where salmon presence was true (Eq. 2). Frame precision was calculated as the number of frames where salmon presence was correctly predicted, divided by the total number of frames where salmon presence was predicted (Eq. 1).

The optimal salmon presence algorithm determined by the parameter optimisation with a confidence threshold that was selected empirically to maximize presence recall while still eliminating a high proportion of frames for review was run on all three full-length tow videos and evaluated using these four metrics: I, P, frame recall, and frame precision. General salmon presence, rather than individual salmon presence, was determined from the individual salmon presence frame data and used to calculate these metrics.

Results

Annotation dataset

The annotated dataset included 16 989 frames with 219 salmon and 3091 pollock tracks that consisted of 11 575 salmon and 73 394 pollock bounding box annotations, respectively (Table 2, Fig. 4). There were 1059 frames where salmon and pollock were not present and the remaining 15 930 frames included salmon and pollock annotations for nine of 12 background conditions. Approximately 85.1%, 7.9%, and 0.7% of the dataset was respectively classified as low, medium, and high fish densities. High densities of krill were present in 14.2% of the frames, and 5.8% had the camera occluded or had low illumination. Three other possible background conditions were not present in the randomly selected data: the presence of high densities of krill during high fish density and the presence of both high densities of krill and camera occlusions during medium and high fish densities.

For the seven tows that annotation video clips were selected from, there were a total of 690 chum salmon (*O. keta*) and 4 Chinook salmon (*O. tshawytscha*) caught and tow speeds ranged from 1.6 to 2.2 m/s. The annotated data included 16% of the salmon present in the four tows used from 2019 and 44% of the salmon present in the three tows from 2020. A higher percentage of salmon in the 2020 tows were used for annotation because there were fewer salmon present in these tows.

Annotator performance baseline

The maximum variability between salmon and pollock annotations completed by different annotators was approximately 30% and pollock accounted for most of this. The average $mAP_{0.5}$ was 0.7 ± 0.08, indicating that between annotators an average of 70% of the annotated salmon and pollock matched with 50% or greater overlap. The average $mAP_{0.5}$ ranged from 0.57 to 0.79, while the average $mAR_{0.5}$ was 0.81 ± 0.07 and ranged from 0.74 to 0.87. Higher AP and AR showed there was greater consistency among salmon annotations (approximately 85–90% agreement) compared to the pollock annotations (approximately 60%–77% agreement) (Table 3). Annotator variability was greatest for pollock on the bottom

Table 2. Summary of the annotated dataset by year and tow from which the videos were collected, including the total number of frames and the number of bounding boxes and tracks for salmon and pollock included from each tow. This is followed by the number of frames by fish density (No Fish, and Fish low, med, and high) with background conditions of no krill (Clear), high density of krill present (Krill), camera occlusions or low light (Occluded or low light), and high density of krill and camera occlusions or low light (Krill & Occluded).

			Salmon		Pollock			Fish low				Fish med			Fish high	
Year	Tow #	Frames	Tracks	Boxes	Tracks	Boxes	No Fish	Clear	Krill	Occluded or low light	Krill & Occluded	Clear	Krill	Occluded or low light	Clear	Occluded or low light
2019	15	2460	39	1532	116	4075	307	2124	0	29	0	0	0	0	0	0
	17	4080	42	1802	371	10 154	420	1294	2048	268	50	0	0	0	0	0
	21	1050	22	686	751	12 989	17	474	0	49	0	374	0	16	34	86
	22	5664	92	4846	1670	39 911	0	4114	268	378	0	754	56	94	0	0
2020	1	728	8	503	113	2875	43	624	0	0	0	61	0	0	0	0
	5	393	3	329	13	784	21	361	0	11	0	0	0	0	0	0
	6	2623	13	1874	57	2606	251	2369	0	3	0	0	0	0	0	0
Total		16 998	219	11 572	3091	73 394	1059	11 360	2316	738	50	1189	56	110	3 34	86

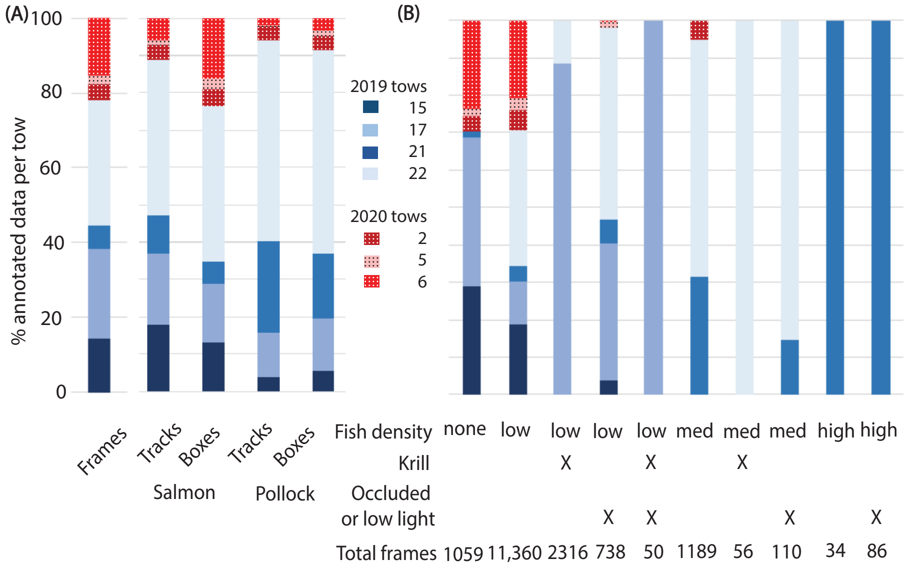

Figure 4: The percentage of annotations from each tow. (A) The percentage of total frames, bounding boxes and tracks of salmon and pollock. (B) The percentage of frames for four levels of fish density (none, low, med, high) and different background conditions present in the annotations are shown. The table below the bar plot shows the fish density in the top row and the total number of frames for each condition or combination of conditions in the bottom row. If krill (Krill) or either camera occlusions or low light (Occluded or low light) were present, it is indicated by an "X" respectively in the second and third rows of the table. 2019 tows are shown in shades of blue and 2020 tows in shades of red with white or black dots.

and sides of the net and entering and leaving the camera scene.

Model training and evaluation

Cross-fold validation dataset

Separating annotations by video clips led to variability in the number of frames, annotations, and tracks included in each of the five data subsets used for cross-fold model evaluation and training. Each data subset included a range of 16% to 24% of the annotated frames (2670 to 4133 frames), 17% to 26% of the salmon annotations (1998 to 3001 frames), 19% to 21% of the salmon tracks (42 to 45 tracks), 17% to 23% of the pollock annotations (12 471 to 16 524 frames), and 16% to 28% of the pollock tracks (497 to 871 tracks) (Table 4).

Model performance

YOLO11 outperformed EfficientDet for all metrics we evaluated (Table 5). YOLO11 achieved a $\mathrm{mAP}_{0.5}$ of 0.72 ± 0.083 and $\mathrm{mAR}_{0.5}$ of 0.90 ± 0.014 for salmon and pollock detection; whereas, EfficientDet achieved only a $\mathrm{mAP}_{0.5}$ of 0.54 ± 0.068 and $\mathrm{mAR}_{0.5}$ of 0.83 ± 0.059. The mean $\mathrm{AP}_{0.5}$ and $\mathrm{AR}_{0.5}$ for models was higher for pollock for both YOLO11 and EfficientDet than for salmon. The multi-class model that was trained and evaluated using data subset 5 had much lower salmon $\mathrm{AP}_{0.5}$ than the other four models for both YOLO11 (approximately 30% lower) and EfficientDet (approximately 11% lower). For YOLO11, the $\mathrm{AP}_{0.5}$ of the other four multi-class models and the single-class model were similar and ranged from 0.7 to 0.8. The single-class model's salmon $\mathrm{AR}_{0.5}$ was higher (0.983) than all the YOLO11 multi-class model's $\mathrm{AR}_{0.5}$ (max 0.88).

For the multi-class models', false positives were associated with an average of 18% of the YOLO11 and 29% of the EfficientDet D2 performance error, while respectively approximately 11% and 12% of these were due to other objects in the scene or the background being detected as either salmon or pollock. False negatives contributed to an average of 7.5% of the model error for the YOLO11 multi-class models compared to 10% for the EfficientDet models. These false negatives accounted for an average of 0.8% of annotations missed by YOLO11 and 5% missed by EfficientDet. For the single-class model, false positives accounted for a similar percentage of error as the multi-class model's averages (approximately 19% total and 11% background). However, the single-class model had a lower percentage of false negatives (1.3%) and missed detections (0.3%).

Low confidence score thresholds led to low numbers of false positive salmon and pollock detections for the multi-class models and no false positive salmon detections for the single-class model when neither salmon or pollock were present. The number of false positive fish detections for the 1059 frames with no fish was variable for the five multi-class models, ranging from 2 to 66 detections (approximately 0.1–6.2% of frames) when a confidence score threshold of 0.5 was used to 0 detections for a threshold of 0.9 (Figure 5A). The 66 false positive fish detections for the 0.5 confidence threshold was an anomaly produced by high pollock false positives for one of the trained models. The model with the second highest false positive detections for the 0.5 confidence threshold had only 18 false detections (approximately 1.7% of frames).

When using the optimal confidence score thresholds, the best performing multi-class model precision and recall for pollock detection was higher than the performance variability measured for our annotators. However, the detection performance for salmon precision and recall for the multi-class and single-class models was lower than our annotator variability (Fig. 5B). The salmon precision for the multi-class and single-class models were similar across confidence score thresholds, but for confidence scores greater than 5e-3 the multi-class model had higher salmon recall than the single-class model. The confidence score that optimized $\mathrm{AP}_{0.5}$ and $\mathrm{AR}_{0.5}$ for pol-

Table 3. COCO metrics for annotator analysis, including mean Average Precision (mAP$_{0.5:0.95}$) for both classes (salmon and pollock) and Average Precision (AP$_{0.5:0.95}$) for each class for ten intersection of union thresholds greater than 0.5 up to 0.95. COCO's mAP$_{0.5}$, mean Average Recall (mAR$_{0.5}$), Average Precision (AP$_{0.5}$), and Average Recall (AR$_{0.5}$) for each class for annotator annotations with an intersection over union of 0.5 or greater follow. The comparison of annotator 2 with annotators 3 and 4 was not possible because there were no overlapping annotations between these annotators, indicated by "NA" (not applicable).

Annotator pair		Total frames	mAP$_{0.5}$	mAP$_{0.5:0.95}$	mAR$_{0.5}$	AP$_{0.5}$		AP$_{0.5:0.95}$		AR$_{0.5}$	
						Salmon	Pollock	Salmon	Pollock	Salmon	Pollock
1	2	587	0.57 ± 0.02	0.22 ± 0.01	0.74 ± 0.07	na	0.57 ± 0.02	na	0.22 ± 0.01	na	0.73 ± 0.06
1	3	742	0.79 ± 0.04	0.38 ± 0.04	0.87 ± 0.04	0.86 ± 0.04	0.75 ± 0.04	0.47 ± 0.06	0.32 ± 0.04	0.91 ± 0.04	0.85 ± 0.05
1	4	902	0.69 ± 0.04	0.34 ± 0.01	0.81 ± 0.08	0.83 ± 0.02	0.58 ± 0.12	0.46 ± 0.02	0.26 ± 0.07	0.89 ± 0.03	0.74 ± 0.14
3	4	742	0.69 ± 0.07	0.34 ± 0.04	0.82 ± 0.07	0.87 ± 0.02	0.55 ± 0.17	0.45 ± 0.02	0.25 ± 0.11	0.91 ± 0.01	0.73 ± 0.16

Table 4. The number of frames, annotations, and clips in the five data subsets used for five-fold cross validation model training and evaluation.

Data subset:	1	2	3	4	5
Frames	2994	3547	2670	4133	3654
Annotations	16 015	18 628	16 221	14 594	19 525
Salmon annotations	2019	2434	1998	2123	3001
Pollock annotations	13 996	16 194	14 223	12 471	16 524
Salmon tracks	42	44	43	45	45
Pollock tracks	510	606	607	497	871
No fish clips	6	5	6	6	6
Fish low clips	32	32	33	32	32
Fish med clips	4	4	5	4	4
Fish high clips	0	0	1	0	1
Occlusion or low light clips	7	6	6	6	6
High density krill clips	4	5	5	4	4

lock and salmon for the multi-class model was approximately 0.1 and 0.23 respectively, and it was 0.06 for the single-class salmon-only model (Fig. 5B).

The evaluation of detection performance across the nine different trawl and video conditions showed there was greater variability among model performance for salmon, that detection performance was generally lower during high fish density, and that salmon detection was better than pollock detection when high densities of fish or krill were present (Fig. 6). When using the optimal confidence score thresholds, the median AP$_{0.5}$ ranged from 0.29 to 0.87 and the median AR$_{0.5}$ ranged from 0.4 to 0.88 for salmon and pollock across conditions. The median AP$_{0.5}$ and AR$_{0.5}$ were highest during medium fish density with high abundnace of krill present and were lowest during high fish density conditions and when a high abundance of krill and camera occlusions or low light were present at the same time. Salmon median AP$_{0.5}$ and AR$_{0.5}$ were lower than pollock excpet when high densities of fish or krill were present.

Model validation on fishing tows

Fishing tows

The tows selected for model validation (2019 tows 16, 18, and 20) included 52 videos and had a range of salmon occurrences and estimated pollock catch. The duration of video collected for each tow was 2.5, 2.1, and 2.1 hours, the salmon occurrence rates were 63.2, 51.4, and 160.5 individuals per hour, and the estimated pollock catch rates were 6, 1.4, and 27.8 metric tonnes per hour respectively for tows 16, 18, and 20 (Yochum et al. 2021). The total number of salmon presence occurrences for all tows was 664 (tow 16 = 168, tow 18 = 134, and tow 20 = 362) and the number of recorded general salmon occurrences was 655 (tow 16 = 165, tow 18 = 131, and tow 20 = 359) due to instances when the same salmon came in and out of frame several times.

The rate of agreement between the two people who determined the individual salmon presence was 84.5% for the random 10% sample of tow videos. During the review of all salmon that had been previously recorded for these tows, we identified instances where previously recorded salmon could not be identified in videos (n = 6), instances where salmon were present in the videos but not indicated in the review sheet (n = 7), as well as two salmon that were missed by the first reviewer and caught during the second review. These discrepancies were corrected for in our individual salmon presence

Table 5. Model performance metrics for the evaluation datasets for all multi-class model runs of the five-fold cross validation and the YOLO11 single-class salmon-only model.

		Both			Pollock			Salmon			TIDE			
Model	Data subset	$mAP_{0.5:0.95}$	$mAP_{0.5}$	$mAR_{0.5}$	$AP_{0.5:0.95}$	$AP_{0.5}$	$AR_{0.5}$	$AP_{0.5:0.95}$	$AP_{0.5}$	$AR_{0.5}$	FP	FN	Background	Missed
EfficientDet D2	1	0.259	0.497	0.775	0.295	0.595	0.844	0.222	0.399	0.705	27.54	13.67	10.65	6.44
	2	0.341	0.645	0.917	0.368	0.711	0.943	0.313	0.579	0.891	27.08	5.64	8.08	2.27
	3	0.296	0.567	0.833	0.301	0.592	0.852	0.29	0.541	0.815	26.48	11.24	11.84	6.67
	4	0.246	0.486	0.777	0.266	0.532	0.783	0.227	0.44	0.77	29.12	13.79	11.49	8.43
	5	0.26	0.496	0.856	0.362	0.704	0.924	0.159	0.288	0.789	35.56	6.72	18.57	3.18
YOLO11	**1**	**0.51**	**0.801**	**0.922**	**0.493**	**0.819**	**0.967**	**0.528**	**0.783**	**0.877**	**11.48**	**6.84**	**4.56**	**0.55**
	2	0.481	0.748	0.904	0.472	0.695	0.97	0.49	0.801	0.837	14.79	8.00	7.91	0.73
	3	0.464	0.726	0.891	0.45	0.748	0.968	0.478	0.703	0.815	16.06	9.19	9.73	0.65
	4	0.478	0.745	0.909	0.453	0.762	0.966	0.503	0.727	0.853	16.13	7.57	9.11	0.88
	5	0.355	0.579	0.888	0.453	0.76	0.959	0.257	0.398	0.817	30.25	6.02	23.63	1.11
Salmon only	NA	NA	NA	NA	NA	NA	NA	0.512	0.786	0.983	19.39	1.31	10.78	0.3

For each model, the evaluation data subset (Data subset) and the applicable COCO metrics are provided: mean Average Precision ($mAP_{0.5:0.95}$) for both classes (pollock and salmon) and Average Precision ($AP_{0.5:0.95}$) for each class for ten intersection of union thresholds greater than or equal to 0.5 up to 0.95 and mean Average Precision ($mAP_{0.5}$) and mean Average Recall ($mAR_{0.5}$) for both classes and the Average Precision ($AP_{0.5}$) and Average Recall ($AR_{0.5}$) for each class for an intersection over union threshold of 0.5 or greater. Metrics from the general toolbox for identifying object detection error (TIDE; Boyla et al. 2020), including an estimate of the error percentage due to false positive (FP), false negative (FN), background, and missed detections (Missed) are also given. The results of the five-fold cross validation model with the overall best performance are bold. "NA" shows the metrics that are not applicable for the single-class model.

frame data and accounted for less than 3% of all presence data used during our analysis.

Fish detections

The multi-class model detected more of the known salmon than the single-class model with the optimal confidence score threshold. However, salmon false positives were higher for the multi-class model than the single-class model (Fig. 7). Salmon false positives occurred more often when Pacific herring (*Clupea pallasi*, hereafter referred to as "herring") and krill were present, both of which could be seen in detections for tow 20 (Fig. 7, 8). Detection performance for tows 16 and 18 was similar to tow 20, but since no krill or herring were present in these tows there were less salmon false positives evident. Overall, the detection results were the best for tow 18, which had the lowest pollock catch rates leading to lower fish densities throughout the tow.

Examples of detections from each model suggested that fish detection performed well with these videos, but error occurred due to fish missed in the background and sides and bottom of the net (Fig. 8D); fish in the far background being detected (Fig. 8A); herring detected as salmon and pollock (Fig. 8B, C, and F); and jellyfish detected as pollock. Salmon in the back of the net or on the sides and bottom of the net were occasionally not detected by both models.

Salmon presence

On average our optimized salmon presence algorithm, using detections from the multi-class model, included 15% of all frames and had a mean presence recall of 99.3%, frame precision of 22.7%, and frame recall of 79.3% (Table 6). For the 15% proportion of included frames, 22% contained salmon. Of all frames where salmon were present, 77% were predicted correctly by the algorithm. Of the 664 true salmon presence occurrences across all tows, only five were missed by the algorithm (99.3% presence recall). Three missed presences were in tow 20 and two were in tow 18. The performance of the salmon presence algorithm varied across tows based on the conditions present (Fig. 9).

When the single-class model was used, fewer frames were included on average and the frame precision was higher than when the multi-class model was used, but the mean presence recall and frame recall were lower (Table 6). The lower recall was due to the algorithm missing 25 salmon presences when the single-class model was used.

The optimal values chosen for the parameters of the salmon presence algorithm were: $C = 0.2$, $M = 1$, and $N = 2$. The LIPO optimisation results indicated that requiring detections across consecutive frames ($N > 1$, Eq. 5) helped reduce presence false positives, but multiple salmon detections per frame ($M > 1$, Eq. 4) did not improve the performance of presence prediction. Increasing the minimum confidence threshold of detections reduced presence recall and the proportion of included frames (i.e. increased the proportion of eliminated frames) (Fig. 10).

Discussion

In this study, we evaluated salmon and pollock detection accuracy in trawl videos for two object detection models: EfficientDet D2 and YOLO11n. Using a dataset of almost 17 000 frames with 11 572 salmon and 73 394 pollock annotations and five-fold cross validation the models were trained

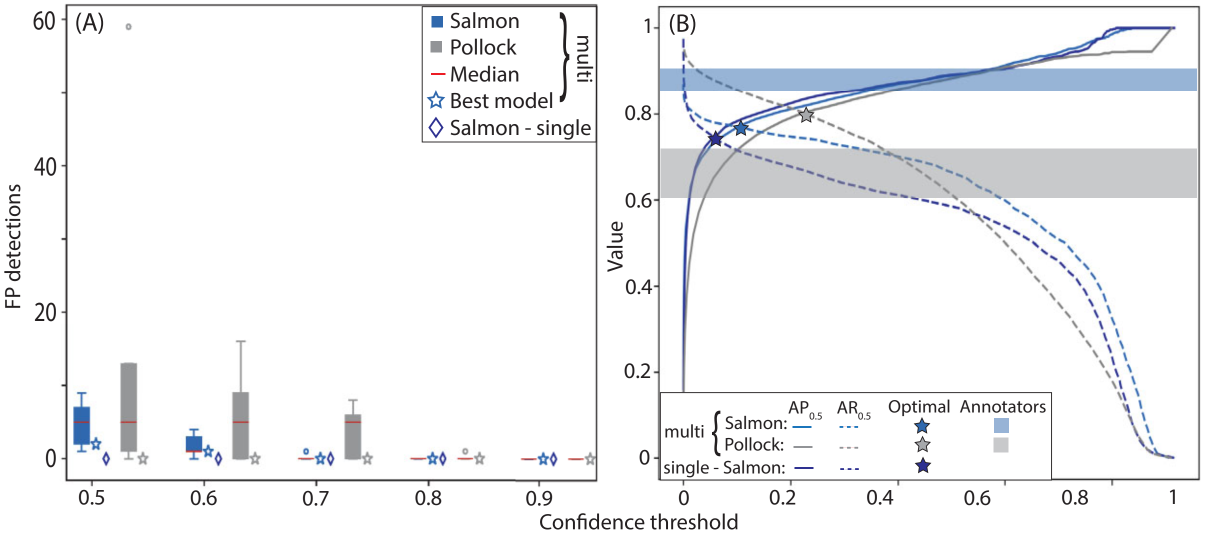

Figure 5: YOLO11n model performance for different confidence score thresholds when no fish are present and fish are present. (A) Boxplots of the multi-class models' number of false positive (FP) salmon (light blue) and pollock (grey) detections for 1 059 frames with no fish present for five confidence thresholds ranging from 0.5 to 0.9. The red line in the middle of each box, the bottom and top of the boxes, the whiskers, and the open circles show the values of the median, the first and third quartiles, the minimum and maximum, and outliers, respectively. The star and diamond (dark blue) to the right of each box respectively show the values of the best multi-class model and the single-class salmon-only model. (B) COCO's Average Precision (AP$_{0.5}$, solid line) and Average Recall (AR$_{0.5}$, dashed line) for an intersection over union threshold of 0.5 for the best multi-class model's salmon (light blue) and pollock (grey) detections and the single-class model's salmon (dark blue) detections are shown for all confidence score thresholds. The shaded blue and grey areas show our annotators' performance range for salmon and pollock respectively. The lower bound of the shaded areas is the mean AP$_{0.5}$, and the upper bound is the mean AR$_{0.5}$ achieved between annotators. The stars show the confidence score that optimises the models AP$_{0.5}$ and AR$_{0.5}$.

as multi-class models to detect salmon and pollock. Each model's overall performance was evaluated with COCO and TIDE metrics, and the performance variability across ten different trawl and video conditions was examined. A single-class salmon-only model was also trained and evaluated with the same dataset used for the best performing multi-class model. The best performing multi-class and the single-class models were compared with the detection variability we measured for our annotators and further tested on full fishing tows to assess performance on a greater amount of data and to evaluate the feasibility of streamlining video review using a detection-based salmon presence prediction algorithm that we developed.

Our analysis showed that YOLO11n performed better than EfficientDet D2 at detecting salmon and pollock in trawl videos and within the range of the variability we measured between annotators when considering the performance of both classes together. At optimal confidence score thresholds, the multi-class performed better at salmon detection than the single-class model. The detection performance across models was more variable for salmon than pollock across different trawl and video conditions and it was generally lowest during high fish densities. For full fishing tows, the multi-class model detected more salmon and salmon presences than the single-class model. The YOLO11n multi-class detection model's ability to predict salmon presence would support a semi-automated video review process that would be more efficient than a fully manual review.

Annotation dataset

We used multiple tows from two different years to create a diverse annotation dataset, but it had weaknesses such as few examples of Chinook salmon, young and smaller salmon, high fish density conditions, and no examples of herring or different trawls. The video used to create the annotation dataset was collected from a single vessel during the summer pollock fishing season when the salmon bycatch is predominately chum salmon (*O. keta*) as opposed to the winter season when Chinook salmon (*O. tshawytscha*) are more prevalent including younger, smaller adults (Witherell et al. 2002, Tucker et al. 2011, Stram and Ianelli 2015). More examples of Chinook salmon are needed to know how well the models can detect this species. The full fishing tow detection results showed that herring were incorrectly detected as salmon and this may have been reduced if video clips with herring had been included in the annotation dataset. A lack of data from multiple vessels could also cause the developed models to not generalise well to video collected from other vessels, and we were unable to evaluate this with our dataset.

The annotation dataset was also imbalanced and included fewer salmon annotations than pollock annotations which likely contributed to the lower detection performance for salmon than pollock. The dataset also had fewer frames with medium and high fish density compared with low fish density and low numbers of frames with medium or high fish densities with camera occlusions, low light, or high krill abundance present. This led to the exclusion of different conditiond in some of the cross-validation data subsets. For example, our dataset only included two video clips with high fish density and, therefore, only two of the five cross-validation subsets had examples of high fish density. This also led to some of the different trawl and video conditions having low frame sample sizes, models with evaluation datasets that did not include any examples of some conditions, and high model performance variability across these conditions. Furthermore, the data subset that was used for training and evaluating the best perform-

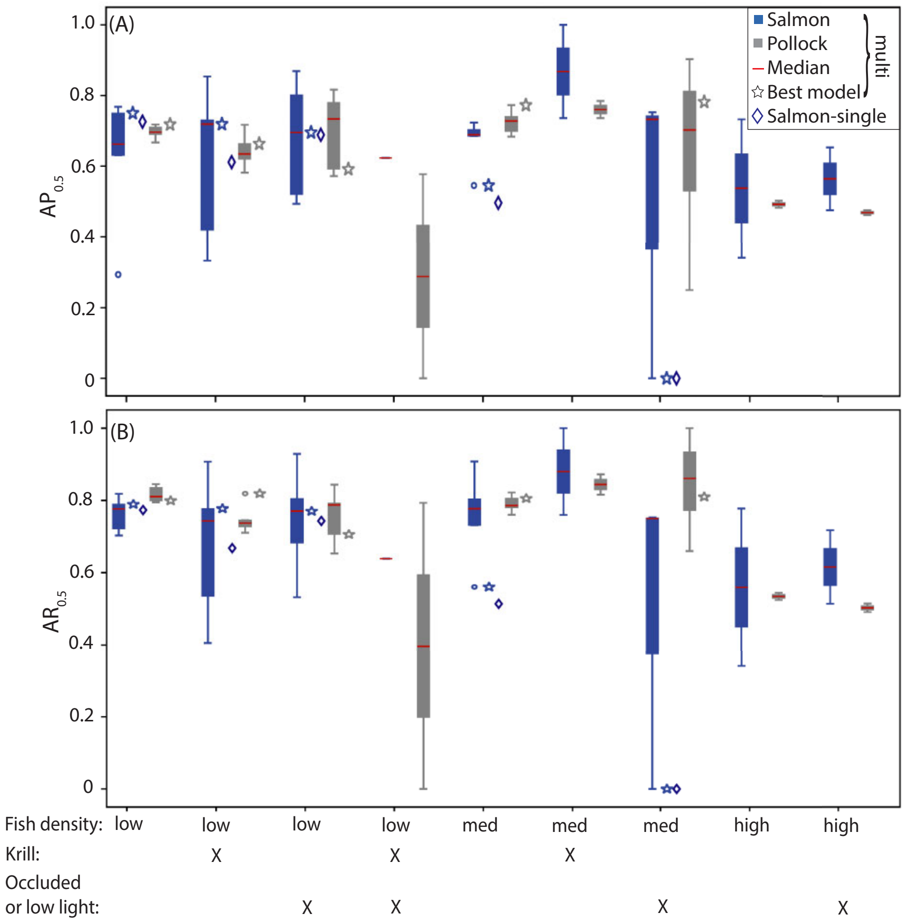

Figure 6: Boxplots of COCO's (A) average precision ($AP_{0.5}$) and (B) average recall ($AR_{0.5}$) for intersection over union of 0.5 or greater for the five YOLO11n multi-class models. The $AP_{0.5}$ and $AR_{0.5}$ values for salmon (blue) and pollock (grey) are shown for nine different background conditions indicated in the table along the x-axis. Fish density was either low, medium (med), or high, and is shown along the top row of the table. A "X" is used to mark if other conditions were present in these data in the lower rows. See Figure 5 for boxplot and marker descriptions.

ing multi-class model was missing examples of low fish density with high abundance of krill and camera occlusions or low light present, medium fish density with high abundance of krill present, and high fish density. This may have led to the model's higher performance metrics, and it is possible that one of the other models may have greater generalisation and perform better on the full fishing tows.

The addition of data that contains herring and other conditions that had low representation in our dataset (e.g. presence of krill or high fish densities) would likely increase model performance since more data often leads to increased model performance (Goodfellow et al. 2016). Balancing the number of annotated video clips across cross-validation data subset would improve the cross-validation evaluation and is achievable. However, a fully balanced dataset for these videos is likely not achievable due to higher occurrences of low fish density compared to medium and high fish density and pollock compared to salmon.

Annotator performance baseline

The majority of annotator variability was due to differences among pollock annotations. Given the high number of pollock present in these videos this was not surprising. The camera setup and illumination used to collect the videos produced a mostly aft-facing, tunnel view that led to fish fading into the background and provided limited visibility of the bottom of the net. Both of these situations created challenges for and discrepancies among annotation. Fish, especially smaller fish like some of the pollock, were hard to distinguish on the bottom and deciding when to end a fish track as it faded into the background was more subjective. When high densities of krill were present it was also hard to distinguish fish and lead to discrepancies in the annotations. Often the salmon present in our videos were larger than the pollock and, therefore, were easier to distinguish on the bottom and when high densities of krill were present.

Model training and evaluation

YOLO11 performed better on this dataset than EfficientDet, and both models did better at predicting pollock than salmon. Furthermore, YOLO11's fish detection for these trawl videos was comparable to the variability measured for people's ability to detect fish. However, the performance for pollock detection was higher and the performance for salmon detection was lower than the variability measured for our annotators. The largest contributor to the false positive error for YOLO11 was the detection of other objects as salmon and pollock (i.e. background detections), and fewer errors were due to salmon being classified as pollock or vice versa. The opposite was true for EfficientDet. The lower performance for salmon was likely

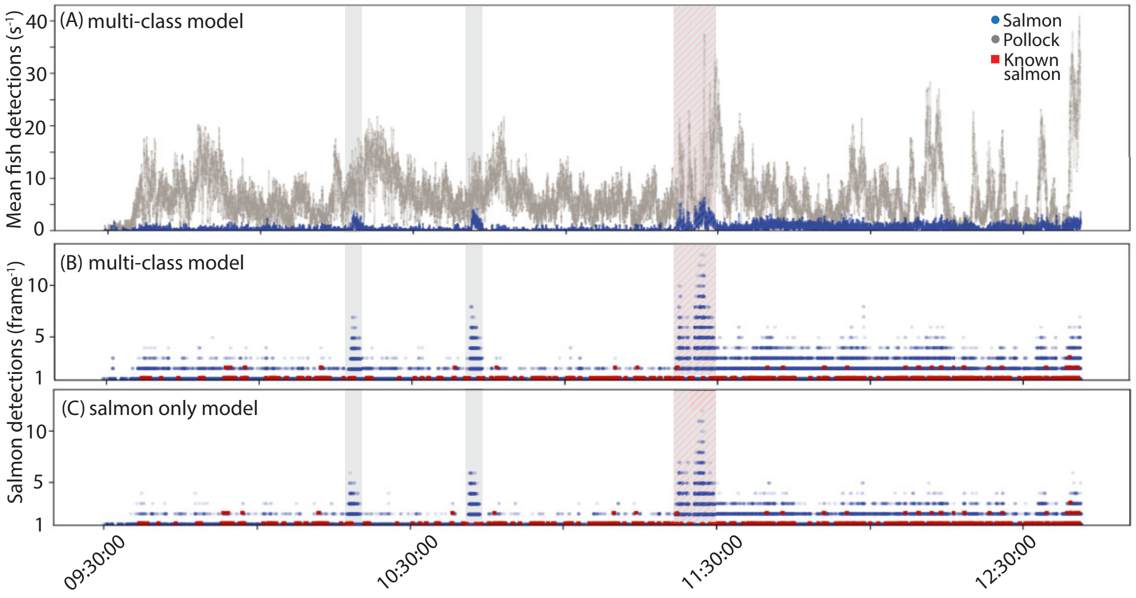

Figure 7: Fish detections from the multi-class and single-class YOLO11n models for 2019 tow 20 compared with the salmon detected by people for these tows. (A) The multi-class model's mean number of salmon (blue) and pollock (grey) detections per second. The times that salmon were detected by people (red) are overlaid and shown for the multi-class model's and the single-class model's salmon detections per frame in (B) and (C), respectively. The light grey background in the plots shows times that herring were common and the light red background shows when high abundances of krill and herring were present. In A-C partially transparent markers are used to plot the mean fish detections per second and salmon detections per frame so they appear darker when more frames in a time period have detections and lighter when there are fewer detections.

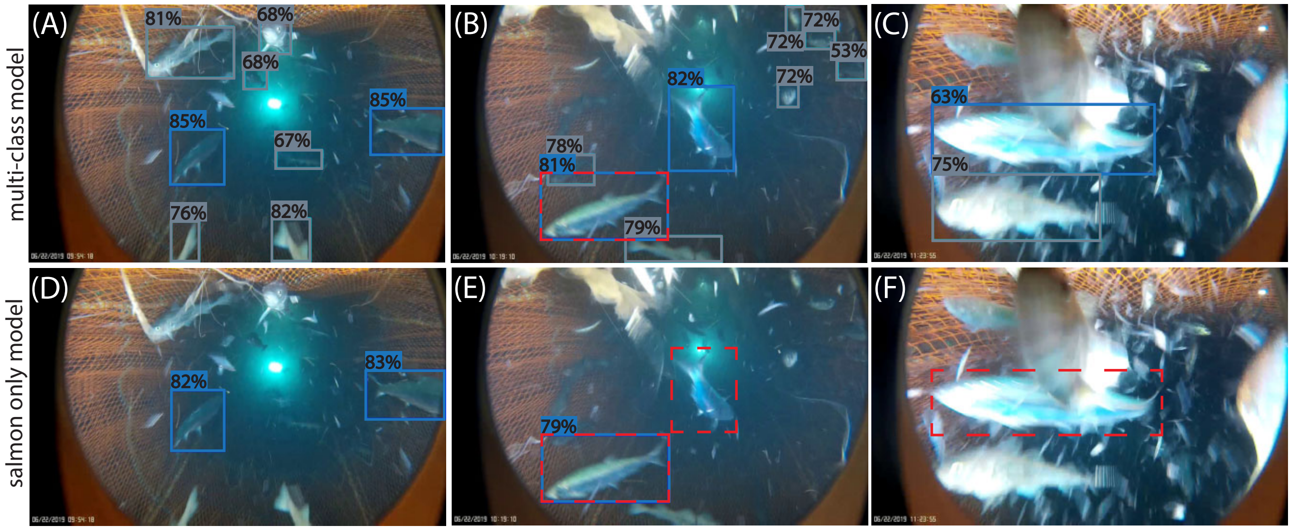

Figure 8: Examples of fish detections with confidence scores greater or equal to 0.5 from three frames for the multi-class (A-C, top images) and single-class model (D-F, bottom images). Salmon (blue boxes) and pollock (grey boxes) detections are shown with the confidence score as a percentage at the top of the boxes. Red dashed boxes show incorrect or missed detections of salmon.

Table 6. The multi-class (salmon and pollock) and single-class (salmon only) model salmon presence algorithm results. The 2019 fishing tow (Tow), presence recall (P), frame precision and recall, proportion of included frames (I), percentage of frames with salmon present from reviewer records (% of frames presence known), the number of salmon occurrences that were correctly predicted (Correct predicted occurrences), and the number of salmon occurrences that were missed (Missed occurrences) are shown.

Model	Tow	*P*	Frame precision	Frame recall	*I*	% of frames presence known	Correct predicted occurrences	Missed occurrences
Salmon & pollock	16	100%	17%	87%	15%	3%	168	0
	18	99%	27%	78%	8%	3%	132	2
	20	99%	24%	73%	22%	7%	359	3
Salmon only	16	98%	25%	77%	9%	3%	165	3
	18	97%	54%	66%	3%	3%	129	5
	20	93%	40%	57%	10%	7%	345	17

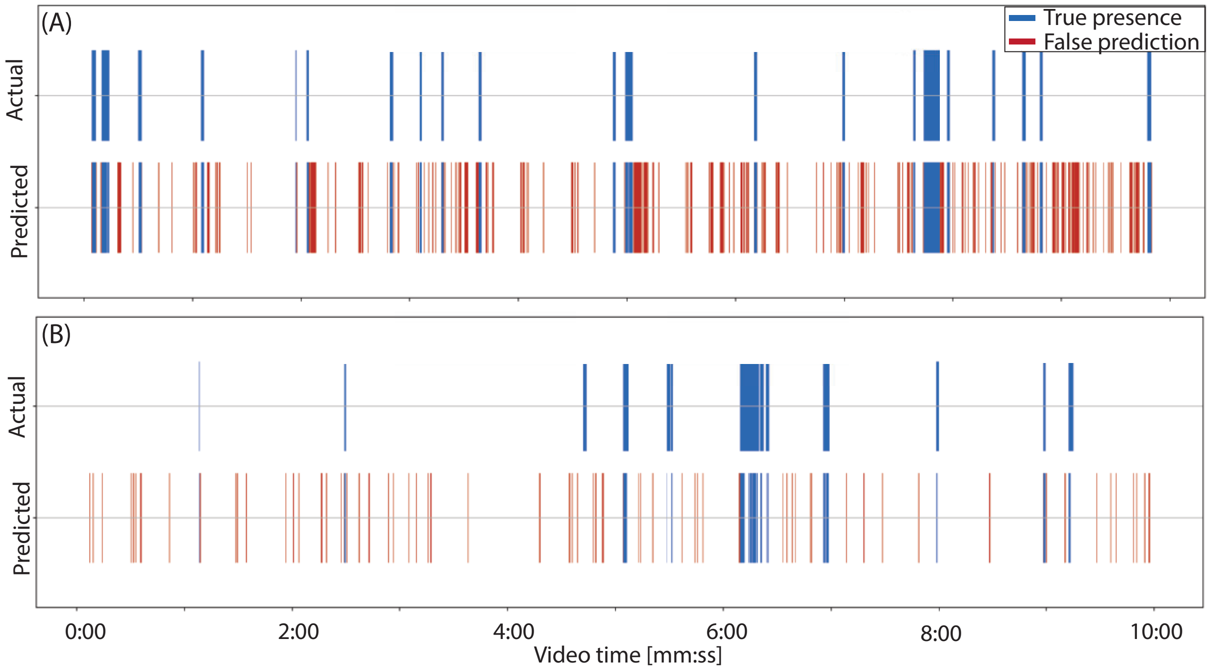

Figure 9: Predicted and actual salmon presence for a single video from (A) tow 16 and (B) tow 20 using the multi-class pollock and salmon detection model. False positive predictions are shown in red and the actual presence and correct predictions are in blue. All occurrences of salmon presence were correctly predicted in A. The first occurrence of salmon presence in B was missed and was the only missed occurrence in the plot.

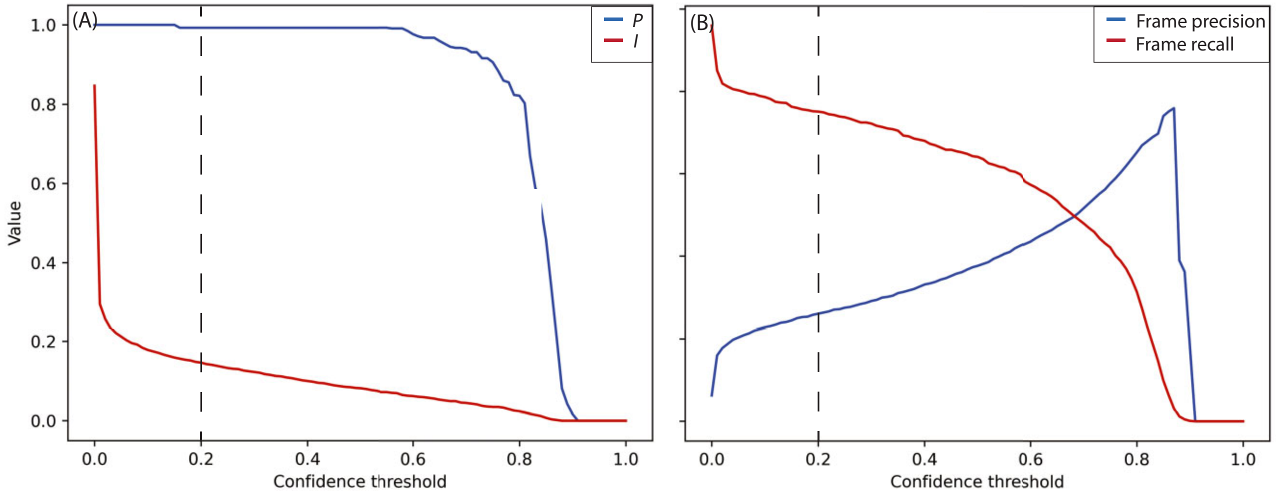

Figure 10: Metrics for evaluating the salmon presence algorithm when using the multi-class pollock and salmon detection model plotted across the range of object detection confidence thresholds. In (A) presence recall (*P*, red, Eq. 7) and proportion of included frames (*I*, blue, Eq. 6) and in (B) frame recall (red, Eq. 2) and precision (blue, Eq. 1) for salmon presence prediction are shown. The black dashed line shows the confidence threshold used for presence prediction.

due to fewer annotated salmon available for training compared to pollock.

One of the five multi-class models for both YOLO11 and EfficientDet had relatively low average precision compared to the other models for each detector, which was due to the dataset used for evaluating these models. The evaluation data subset used had one relatively long clip (1171 frames, 33% of the data subset) of a single salmon near the end of a tow and another short clip with 10 salmon during high fish density. Errors in salmon detection for the high number of salmon annotations in these clips likely led to the lower performance metrics for salmon for these models.

The evaluation of model performance across different confidence score thresholds highlighted important aspects of detection models: they can be adjusted to prioritize either fewer missed or fewer incorrect detections. In this study, we used the confidence threshold where precision and recall are equal to prioritize both equally. With this confidence score threshold, the best performing YOLO11 multi-class model performed better than the single-class model. However, the single-class YOLO11 model achieved higher overall $AR_{0.5}$ than the multi-class models when no confidence score threshold was used due to higher recall values for confidence scores below 5e-3.

The models' detection performance across the nine different trawl and video conditions revealed expected and unexpected findings. Lower performance during high fish density was expected due to this condition causing partially occluded fish that can make detection and classification more challenging. The greater variability of model performance for salmon was also expected due to fewer salmon annotations included in our dataset. The higher detection performance for medium fish density with high abundance of krill and the higher median $AP_{0.5}$ and $AR_{0.5}$ for salmon compared to pollock during high fish and krill density were unexpected. We believe this may be due to fewer fish being detectable and annotated by people during these conditions. During clear video conditions, fish barely visible in our videos were annotated on the sides, bottom, and farther back in the net and often missed by the model. When high densities of krill were present in the videos, people did not detect fish in these locations often leading to fewer annotations for the models to possibly miss.

Our YOLO11 model detected pollock and salmon in these videos from an Alaska commercial pollock trawl, a high-volume pelagic trawl fishery, in a comparable fashion to what was observed for object detection performance in the demersal *Nephrops* fishery (Sokolova et al. 2021). The detection prediction accuracy was approximately 10% lower than what Allken et al. (2021) achieved for fish detection for Deep Vision imagery collected in a trawl survey where the abundance of fish was much lower, images with krill present were excluded, and the camera capture environment was more controlled.

Overall, the precision achieved by YOLO11 was relatively high despite the challenges present in these videos, the differences between our imagery and the COCO dataset used to pre-train the models (Suppl. Figure 1), and our lack of hyperparameter optimization. The high densities and overlap of fish in some of our video and the non-canonical camera views presented additional object detection challenges that may not be present in much of the COCO dataset. We chose not to optimise hyperparameters to balance our objectives to establish a baseline understanding of how well the available pre-trained models performed on our dataset. A better way to compare the performance between these models is to conduct hyperparameter optimization for each.

Model validation on fishing tows

The multi-class YOLO11 model did better than the single-class model at detecting the salmon present in the full fish tows, but salmon false positives were common, especially in the multi-class model, due to misclassification of pollock and herring. The increase in salmon detection for the multi-class model compared to the single-class model was slightly evident in the detections across full tows and more evident when the detections were used to infer general salmon presence. The salmon false positives due to misclassification of pollock and herring highlight the challenges of classifying fish species with morphological similarities when the imagery is not ideal. Including video clips with herring in the annotated dataset and collecting better video imagery would help these models differentiate between these fishes. Furthermore, since the multi-class model differentiated between salmon and pollock well, training a model to detect all three could be promising and could benefit the pollock fishery since herring also has catch limits in the Alaska pollock fishery (NPFMC 2024). However, increasing the number of classes can decrease other aspects of detection performance (Dean et al. 2013), and these trade-offs should be evaluated.

Despite object detection error, predicting general salmon presence using our YOLO11 model and our presence prediction algorithm showed the feasibility for a semi-automated review to reduce the time for analysis for videos collected in high-volume trawl fisheries. By performing detection filtering using thresholds for the detection confidence score and consecutive frames with detections, which were both shown to be important, we were able to exclude over 80% of video frames from potential review while missing fewer than 1% of salmon. Furthermore, our code and models processed tows in just hours, with an approximate 90 fps processing speed, compared to an experienced video reviewer needing a few days to weeks to review a tow. The actual time savings from using our model and salmon presence prediction algorithm cannot be estimated because our metrics are imperfect indicators of the semi-automated review process. The proportion of included frames being 15% does not necessarily mean that video review would be 85% faster, but we believe it will be less time than reviewing 85% more frames. It may be possible to improve the presence prediction algorithm by considering other features for filtering, such as bounding box size, and implementing the algorithm with better performing detectors.

In addition to hyperparameter optimization and expanding our annotated dataset, other options that may improve detection model performance include increasing the annotated data using data augmentation and evaluating different methods for species classification or new object detection models. In this study, only a few data augmentation techniques were evaluated with a small data subset for EfficientDet and no performance increases were achieved. Further evaluation of data augmentation options with a larger dataset could lead to improved performance or better performance when applying these models to new video. For example, using an automated approach to find optimal data augmentation options has increased object detection performance across different datasets, dataset sizes, backbone architectures and detection algorithms (Zoph et al. 2020). Another option to possibly increase detection performance is to use more information from tracking or multiple detectors. If reliable object tracking was implemented, weighing the classifications of all detections in a track could result in better fish classification (Dawkins et al., 2024). Also, weighing the predictions of multiple or different types of classifiers may improve fish classification accuracy (Xie et al. 2019). Lastly, new object detection models are released often and models that achieve higher performance on benchmark datasets may be able to achieve higher performance on our dataset.

The best option for increasing object detection performance is improving the quality of video collected in the trawl. The camera setup and capture environment used to collect these videos were the most significant factors that limited the detection performance. The illumination and placement of the camera made it challenging to see and distinguish identifying features of salmon or pollock on the bottom and sides of the net or in the background. It was challenging for people familiar with identifying salmon to review these videos and recognise every salmon. We found some discrepancies in the salmon records when the salmon presence frame data was created, and Yochum et al. (2021) acknowledged that the limited visibility and the presence of high densities of pollock or krill likely led to salmon being missed and underestimated by people. Furthermore, correctly classifying some fish as salmon or pollock was difficult and took input from multiple people to arrive at a consensus. Some modifications to the camera or net setup that might help with these challenges include the use of additional lighting, which may have impacts on fish behaviour; use of additional cameras to view fish at the bottom of the net better and provide more opportunity to see all fish when density is high; positioning cameras to have a perpendicular view of fish to allow morphological differences to be more apparent and reduce issues with background fish; adding solid, high-contrast material around the net or a full compartment (e.g. DeepVision; https://www.deepvision.no/) to enhance illumination and object contrast; or modifying the net to control the flow and location of fish past the camera.

Application in fishery

Our object detection model and salmon presence algorithm could be used to reduce the time to generate data from video footage collected in the Alaska commercial pollock fishery to evaluate salmon BRDs and accelerate the pace of salmon bycatch reduction research for this fishery. The salmon presence prediction results showed the potential for a semi-automated video review. However, to gain the greatest reduction in video review time, more work is needed to build user-friendly tools that allow for models like ours to be used effectively by video reviewers.

In addition to semi-automating and increasing the efficiency of video review to support bycatch reduction research, object detection could be used in the Alaska pollock fishery to inform vessels of salmon bycatch (e.g. an alarm) and target catch in real-time or to support active bycatch reduction devices that are currently being researched. However, for the fishery to ever adopt these tools, further steps are needed to address issues with our annotated dataset, conduct further model evaluation, and develop a model with better salmon detection performance.

In the Alaska pollock fishery, relatively low numbers of Chinook salmon bycatch can shut down the fishery, so the detection of all Chinook salmon is important. False positive salmon detections would be preferable over missed salmon, but obtaining a relatively low number of false positives would be ideal. A different camera implementation will likely be needed to achieve this high level of performance, but simply using our model with a higher confidence score threshold would reduce the salmon false positives while increasing the likelihood of missing salmon. A model that meets or exceeds the performance of people could be acceptable to use in the fishery and further model development may make this possible.

When using object detection models on fishing vessels, model performance should be evaluated initially and then periodically as a quality control measure due to differences between vessel cameras, nets, and tows, which cause natural real-world variations that impact performance (Hendrycks et al. 2021). Model performance may vary for different vessels, trawls, catch compositions, and tow speeds. A single detection model might generalise well for different vessels if the model was trained with imagery captured from many vessels and fishing environments, which may be achievable over time with data sharing and continual model training. However, it may be necessary to train custom models for each vessel using imagery captured from only that vessel to achieve the highest performance. If that is the case, each vessel's data could be used to fine-tune and optimise performance of a single model like we did in this study.

For most commercial fishery research and applications, the total number of fish or biomass of a species is more useful than relative information about fish detections. To achieve this, these detections could be used to track and count individuals, or possibly converted to a biomass. For accurate biomass estimates, a measure of model performance, fish length frequency distribution, and possibly the rate that fish flow through the net would need to be accurately estimated. The camera angle used in this study and the high densities of fish that can be present would make tracking and counting challenging, but it may be possible with suitable approaches. For biomass estimates, fish length frequency distributions can be estimated from the catch (Gulland and Rosenberg 1992), or stereo-cameras could be used to measure this directly from the videos (Williams et al. 2010, Rosen et al. 2013). If fish flow is needed, the detections or tracks of fish could provide estimates of this, or water flow measurements could suffice for passive swimming fish.

Deep learning is a powerful tool for automating imagery analysis and our results suggest that this technology has tremendous potential, but should be used thoughtfully and with a thorough understanding of its strengths, limitations, and the scope of data used for training and evaluating. Understanding the limitations of detection models is critical for operational use for bycatch reduction in commercial fisheries. Also, it is essential to consider all aspects of the task to be automated before deciding to use deep learning methods as the solution. Developing reliable automated methods can require great initial effort in data collection and curation, annotation, and model development (Goodfellow et al. 2016) and should be weighed against potential time savings compared to traditional methods. As shown here and in other studies, deep learning methods may ultimately still require people in the processing or analysis loop to verify detections, tracks, or counts of objects (Wilchek et al. 2023). However, we have shown that YOLO11n has the potential to reduce the time needed to review videos collected in the trawls of the Alaska pollock fishery and provide salmon awareness for salmon bycatch reduction methods in this fishery. Given the performance we observed for the pollock fishery, we believe other high-volume, pelagic trawl fisheries could use deep-learning object detection methods to assist with video review or monitoring to support bycatch reduction efforts.

Conclusions

The open-source pre-trained object detection models EfficientDet D2 and YOLO11n were trained and evaluated using a subset of annotated video collected in the trawl of an Alaska pollock fishing vessel to determine if deep-learning could be used to support salmon bycatch reduction in this fishery and possibly other high-volume, pelagic trawl fisheries. Using an annotated dataset of almost 17 000 frames with 11 572 salmon and 73 394 pollock annotations to train and evaluate models, we found that YOLO11n performed better than EfficientDet D2 at detection of salmon and pollock in trawl videos and within the range of the variability we measured between annotators. For the evaluation dataset, the YOLO11n multi-class models detected on average 90% of the annotated salmon and pollock, and 72% of the models' predictions were correct when using an IoU threshold of 0.5. The YOLO11n multi-class model for salmon and pollock also performed better than a single-class salmon model.

When using the YOLO11n multi-class detection model with a salmon prediction algorithm we developed to analyse full fishing tows, salmon presence was predicted for 15% of the total video frames and only 5 of 664 salmon presences were missed. This level of performance would support a semi-automated video review process that would be more efficient than a fully manual review and assist in expediting video analysis to evaluate salmon bycatch reduction devices in the Alaska pollock fishery. The YOLO11n model also shows promise for being able to predict salmon for bycatch reduction methods that require real-time monitoring, but the models will need further development and evaluation to achieve the performance level required for these applications. Improvements to the trawl camera setup such as a high contrast background, a perpendicular flow view, and additional lighting could increase object detection performance most by providing higher quality of imagery data.

This work showed that deep learning object detection methods are accessible and robust. With just a few changes to model hyperparameters and the use of a relatively small annotation training dataset compared to the COCO dataset used for pre-training the models, we were able to detect fish at comparable rates to the variability measured between people conducting this task. The methods that we used to evaluate our model performance provided valuable insight into performance objectives, trade-offs, and improvements needed to use these models to support bycatch reduction efforts in high-volume trawl fisheries. We believe this information and our annotations, video imagery, and models that we provided can support the continued development of automated video and image processing methods for bycatch mitigation innovation in the Alaska pollock fishery.

Acknowledgements

This work was supported by efforts from the 2019 and 2020 F/V *Pacific Explorer* captain and crew, David Bryan, Matthew Callahan, Connor Fox, Katherine Hellen-Schneider, and Deborah Sharpe, and we are grateful for their help. We also thank Dr Shari Matzner (Pacific Northwest National Laboratory), Dr Kresimir Williams (NOAA Alaska Fisheries Science Center), and many anonymous reviewers for providing feedback on this manuscript.

The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the authors and do not necessarily reflect those of NOAA or the US Department of Commerce. Reference to trade names does not imply endorsement by the National Marine Fisheries Service, NOAA.

Author contributions

Katherine Wilson (Conceptualization [equal], Data curation [supporting], Formal Analysis [lead], Investigation [lead], Methodology [lead], Project administration [lead], Resources [lead], Software [lead], Supervision [lead], Validation [lead], Visualization [lead], Writing – original draft [lead]), Moses Lurbur (Formal Analysis [supporting], Methodology [supporting], Software [supporting], Writing – review & editing [equal]), Noëlle Yochum (Conceptualization [equal], Data curation [lead], Funding acquisition [lead], Project administration [supporting], Writing – review & editing [equal]).

Supplementary data

Supplementary data is available at *ICES Journal of Marine Science* online.

Conflict of interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This work was supported by the National Cooperative Research Program (2019, 2020, 2021), which provided funding for data collection and ML's work on the project. NOAA's Alaska Fisheries Science Center supported KW's and NY's work on the project.

Data availability statement

Links to the imagery and annotation datasets generated and analysed for this study, the final weights of our pollock and salmon EfficientDet D2 and YOLO11n models, and the code for salmon presence prediction are all available at NOAA Fisheries InPort AFSC/RACE/MACE: 2019-2020 Salmon and pollock object detection.

References

Alaba SY, Nabi MM, Shah C *et al.* Class-aware fish species recognition using deep learning for an imbalanced dataset. *Sensors* 2022;**22** 8268. https://doi.org/10.3390/s22218268 (26 January 2022, date last accessed).

Allken V, Rosen S, Handegard NO *et al.* A deep learning-based method to identify and count pelagic and mesopelagic fishes from trawl camera images. *ICES J Mar Sci* 2021;78:3780–92. https://doi.org/10.1093/icesjms/fsab227

Amari SI. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 1993;5:185–96. https://doi.org/10.1016/0925-2312(93)90006-O

Bolya D, Foley S, Hays J *et al.* Tide: a general toolbox for identifying object detection errors. In: *European Conference on Computer Vision, Part III 16*. p. 558–73. Glasgow, UK, 2020. https://doi.org/10.1007/978-3-030-58580-8_33

COCO Consortium. 2015 detection-eval. https://cocodataset.org/#detection-eval (12 December 2023, date last accessed).

Dawkins M, Prior J, Lewis B *et al.* FishTrack23: an Ensemble Underwater Dataset for Multi-Object Tracking. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. p. 7167–76. Waikoloa, Hawaii, 2024. https://doi.org/10.1109/WACV57701.2024.00701

Dean T, Ruzon MA, Segal M *et al.* Fast, accurate detection of 100,000 object classes on a single machine. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition* . p. 1814–21. Portland, OR, USA, 2013. https://doi.org/10.1109/CVPR.2013.237

Ditria EM, Lopez-Marcano S, Sievers M *et al.* Automating the analysis of fish abundance using object detection: optimizing animal ecology with deep learning. *Frontiers in Marine Science* 2020;7,429https://doi.org/10.3389/fmars.2020.00429

Euronews 2022 The 'Game of Trawls': smart fishing nets could save millions of sea creatures. https://www.euronews.com/green/2021/12/05/the-game-of-trawls-smart-fishing-nets-could-save-millions-of-sea-creatures (05 December 2021, date last accessed).

FAO 2024 The State of World Fisheries and Aquaculture 2024: blue Transformation in action. https://doi.org/10.4060/cd0683en

Fissel BE, Dalton M, Felthoven RG *et al.* 2016 Stock Assessment and Fishery Evaluation Report for the Groundfishes Fisheries of the Gulf of Alaska and Bering Sea/Aleutian Island Area: economic Status of the Groundfish Fisheries off Alaska, 2015. https://repository.library.noaa.gov/view/noaa/18817 (07 February 2024, date last accessed).

Garcia R, Prados R, Quintana J *et al.* Automatic segmentation of fish using deep learning with application to fish size measurement. *ICES J Mar Sci* 2020;**77**:1354–66. https://doi.org/10.1093/icesjms/fsz186

Gauvin J, Gruver J, McGauley K *et al.* 2013 Salmon Excluder EFP 11-01. https://media.fisheries.noaa.gov/dam-migration/efp-salmonby-final-report-0111.pdf (26 January 2022, date last accessed).

Gauvin J, Gruver J, McGauley K. 2015 Central Gulf of Alaska Salmon Excluder EFP 13-01. https://media.fisheries.noaa.gov/dam-migration/efp-salmonby-final-report-0113.pdf (26 January 2022, date last accessed).

Gauvin J, Gruver J, Rose C. 2011 Final report for EFP 08-02 to explore the potential for flapper-style Salmon excluders for the Bering Sea Pollock Fishery. https://media.fisheries.noaa.gov/dam-migration/efp-salmonby-final-report-0208.pdf (26 January 2022, date last accessed).

Gauvin J. Bering Sea Salmon Excluder EFP 15-01 Final Report. 2016http://meetings.npfmc.org/CommentReview/DownloadFile?p=a94e693a-f95d-4e32-9c42-2dc2cb63efab.pdf&fileName=D3%20Salmon%20Excluder%20EFP.pdf (26 January 2022, date last accessed).

Gauvin JR, Paine B. 2004 EFP 03-01: test of a Salmon excluder device for the pollock trawl fishery January 2003 through March 2004. https://media.fisheries.noaa.gov/dam-migration/efp-salmonby-final-report-0103.pdf (26 January 2022, date last accessed).

Glenn J, Ayush C, Jing Q. 2023 Ultralytics YOLO. https://github.com/ultralytics/ultralytics (09 January 2025, date last accessed).

Goodfellow I, Bengio Y, Courville A. *Deep learning*. Cambridge, MA, USA: MIT press, 2016.

Grigorescu S, Trasnea B, Cocias T *et al.* A survey of deep learning techniques for autonomous driving. *J Field Rob* 2020;**37**:362–86. https://doi.org/10.1002/rob.21918

Gulland JA, Rosenberg AA. A review of length-based approaches to assessing fish stocks. *FAO Fisheries Technical Paper. No. 323*. Rome:FAO, 1992, 100. https://www.fao.org/4/t0535e/t0535e00.htm (03 June 2025, date last accessed).

Hendrycks D, Basart S, Mu N *et al.* 2021 The many faces of robustness: a critical analysis of out-of-distribution generalization. In: *Proceedings of the IEEE/CVF international conference on computer vision*. p. 8340–9. https://doi.org/10.1109/ICCV48922.2021.00823

Hollely B. 2023 Smartrawl: aI-empowered fishing net to help prevent marine bycatch. https://fisorg.uk/smartrawl-ai-empowered-fishing-net-to-help-prevent-marine-bycatch/ (15 May 2024, date last accessed).

Hosang J, Benenson R, Dollár P *et al.* What makes for effective detection proposals?. *IEEE Trans Pattern Anal Mach Intell* 2015;**38**:814–30. https://doi.org/10.1109/TPAMI.2015.2465908

Ianelli J, Fissel B, Holsman K *et al.* 2019 *Assessment of the walleye pollock stock in the Eastern Bering Sea: 2019 Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Bering Sea and Aleutian Islands Regions*, North Pacific Fishery Management Council, Anchorage, AK, USA.https://apps-afsc.fisheries.noaa.gov/refm/docs/2019/EBSPollock.pdf (07 February 2024, date last accessed).

Khanam R, Hussain M. Yolov11: an overview of the key architectural enhancements. *arXiv preprint*. 2024https://doi.org/10.48550/arXiv.2410.17725

Lin TY, Maire M, Belongie S *et al.* Microsoft COCO: common objects in context. In: *European conference on computer vision , Part V 13*. p. 740–55. Zurich, Switzerland, 2014. https://doi.org/10.1007/978-3-319-10602-1_48

Liu X, Faes L, Kale AU *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019;**1**:e271–97. https://doi.org/10.1016/S2589-7500(19)30123-2

Lomeli MJM, Wakefield WW. The effect of artificial illumination on Chinook salmon behavior and their escapement out of a midwater trawl bycatch reduction device. *Fish Res* 2019;**218**:112–9. https://doi.org/10.1016/j.fishres.2019.04.013

Loshchilov I, Hutter F. 2017 Decoupled weight decay regularization *arXiv preprint*. https://doi.org/10.48550/arXiv.1711.05101

Malherbe C, Vayatis N. Global optimization of Lipschitz functions. In: *Proceedings of the 34thInternational Conference on Machine Learning* . Vol. **70**, p. 2314–23. Sydney, Australia, 2017.

NPFMC 2024 *Fishery Management Plan for Groundfish of the Bering Sea and Aleutian Islands Management Area*. North Pacific Fishery Management Council. Anchorage, AK, USA. https://www.npfmc.org/wp-content/PDFdocuments/fmp/BSAI/BSAIfmp.pdf (18 November 2024, date last accessed).

Ovchinnikova K, James MA, Mendo T *et al.* Exploring the potential to use low cost imaging and an open source convolutional neural network detector to support stock assessment of the king scallop (*Pecten maximus*). *Ecol Inform* 2021;**62**:101233. https://doi.org/10.1016/j.ecoinf.2021.101233

Polyak BT. Some methods of speeding up the convergence of iteration methods. *USSR Comput Math Math Phys* 1964;**4**:1–17. https://doi.org/10.1016/0041-5553(64)90137-5

PyPI 2018 pycocotools. https://pypi.org/project/pycocotools/ (06 October 2023, date last accessed).

Qin H, Li X, Liang J *et al.* DeepFish: accurate underwater live fish recognition with a deep architecture. *Neurocomputing* 2016;**187**:49–58. https://doi.org/10.1016/j.neucom.2015.10.122

Redmon J, Divvala S, Girshick R., *et al.*. You only look once: unified, real-time object detection. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition*. p. 779–88. Las Vegas, NV, USA. https://doi.org/10.1109/CVPR.2016.91

Ren S, He K, Girshick R *et al.* Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2017;**39**:1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Rose CS, Barbee D. Developing and testing a novel active-selection (Act-Sel) bycatch reduction device to quickly alternate trawls between capture and release configurations with real-time triggering. *Fish Res* 2022;**254**:106380. https://doi.org/10.1016/j.fishres.2022.106380

Rosen S, Jörgensen T, Hammersland-White D *et al.* DeepVision: a stereo camera system provides highly accurate counts and lengths of fish passing inside a trawl. *Can J Fish AquatSci* 2013;**70**:1456–67. https://doi.org/10.1139/cjfas-2013-0124

Salman A, Jalal A, Shafait F *et al.* Fish species classification in unconstrained underwater environments based on deep learning. *Limnol Oceanogr Methods* 2016;**14** :570–85. https://doi.org/10.1002/lom3.10113

Sokolova M, Mompó Alepuz A, Thompson F *et al.* A Deep Learning Approach to Assist Sustainability of Demersal Trawling Operations. *Sustainability* 2021;**13**:12362. https://doi.org/10.3390/su132212362

Sreenu G., Saleem Durai M.A. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. *Journal of Big Data* 2019;**6**:48. https://doi.org/10.1186/s40537-019-0212-5

Stram DL, Ianelli JN. Evaluating the efficacy of salmon bycatch measures using fishery-dependent data. *ICES J Mar Sci*, 2015;**72**:1173–80. https://doi.org/10.1093/icesjms/fsu168

Tan M, Pang R, Le QV. EfficientDet: scalable and efficient object detection. In: *2020 IEEE/CVF Conference on Computer vision and Pattern Recognition*. p. 10781–90. Seattle, WA, USA, 2020. https://doi.org/10.1109/CVPR42600.2020.01079

Tseng CH, Kuo YF. Detecting and counting harvested fish and identifying fish types in electronic monitoring system videos using deep convolutional neural networks. *ICES J Mar Sci* 2020;**77**:1367–78. https://doi.org/10.1093/icesjms/fsaa076

Tucker S, Trudel M, Welch DW *et al.* Life history and seasonal stock-specific ocean migration of Juvenile Chinook Salmon. *Trans Am Fish Soc* 2011;**140**:1101–19. https://doi.org/10.1080/00028487.2011.60703

Underwood MJ, Rosen S, Engås A *et al.* Deep vision: an in-trawl stereo camera makes a step forward in monitoring the pelagic community. *PLoS One* 2014;**9**:e112304. https://doi.org/10.1371/journal.pone.0112304

Wang J, Fu P, Gao RX. Machine vision intelligence for product defect inspection based on deep learning and Hough transform. *J Manuf Syst* 2019;**51**:52–60. https://doi.org/10.1016/j.jmsy.2019.03.002

Wilchek M, Hanley W, Lim J *et al.* Human-in-the-loop for computer vision assurance: a survey. *Eng Appl Artif Intell* 2023;**123**:106376. https://doi.org/10.1016/j.engappai.2023.106376

Williams K, Towler R, Wilson C. Cam-trawl: a combination trawl and stereo-camera system. *Sea Technology* 2010;**51**:45–50.

Witherell D, Ackley D, Coon C. An overview of salmon bycatch in Alaska groundfish fisheries. *Alaska Fishery Research Bulletin* 2002;**9**:53–64.

Xie J, Hu K, Zhu M *et al.* Investigation of different CNN-based models for improved bird sound classification. *IEEE Access* 2019;**7**:175353–61. https://doi.org/10.1109/ACCESS.2019.2957572

Yochum N, Stone M, Breddermann K *et al.* Evaluating the role of bycatch reduction device design and fish behavior on Pacific salmon (*Oncorhynchus* spp.) escapement rates from a pelagic trawl. *Fish Res* 2021;**236**:105830. https://doi.org/10.1016/j.fishres.2020.105830

Yu H, Chen C, Du X *et al.* 2020 Tensor Flow Model Garden. https://github.com/tensorflow/models (25 July 2022, date last accessed).

Zhang S, Jafari O, Nagarkar P. A survey on machine learning techniques for auto labeling of video, audio, and text data. *arXiv preprint*. 2021. https://doi.org/10.48550/arXiv.2109.03784 (13 May 2025, date last accessed).

Zhang S, Yang X, Wang Y *et al.* Automatic fish population counting by machine vision and a hybrid deep neural network model. *Animals* 2020;**10**:364. https://doi.org/10.3390/ani10020364

Zhao S, Zhang S, Liu J *et al.* Application of machine learning in intelligent fish aquaculture: a review. *Aquaculture* 2021;**540**:736724. https://doi.org/10.1016/j.aquaculture.2021.736724

Zoph B, Cubuk ED, Ghiasi G *et al.* Learning data augmentation strategies for object detection. In: *Computer Vision–ECCV 2020: 16th European Conference Proceedings, Part XXVII* . p. 566–83. Glasgow, UK, 2020. https://doi.org/10.48550/arXiv.1906.11172

Dewei Yi, Hasan Bayarov Ahmedov, Shouyong Jiang, Yiren Li, Sean Joseph Flinn, Paul G Fernandes. Coordinate-Aware Mask R-CNN with Group Normalization: A underwater marine animal instance segmentation framework. *Neurocomputing* 2024; **583**. https://doi.org/10.1016/j.neucom.2024.127488.

Handling editor: Cigdem Beyan