Canadian Science Publishing

ARTICLE

Early lessons in deploying cameras and artificial intelligence technology for fisheries catch monitoring: where machine learning meets commercial fishing

M. Rizwan Khokher, L. Richard Little, Geoffrey N. Tuck, Daniel V. Smith, Maoying Qiao, Carlie Devine, Helen O'Neill, John J. Pogonoski, Rhys Arangio, and Dadong Wang

Abstract: Electronic monitoring (EM) is increasingly used to monitor catch and bycatch in wild capture fisheries. EM video data are still manually reviewed and adds to ongoing management costs. Computer vision, machine learning, and artificial intelligence-based systems are seen to be the next step in automating EM data workflows. Here we show some of the obstacles we have confronted and approaches taken as we develop a system to automatically identify and count target and bycatch species using cameras deployed to an industry vessel. A Convolutional Neural Network was trained to detect and classify target and bycatch species groups, and a visual tracking system was developed to produce counts. The multiclass detector achieved a mean average precision of 53.42%. Based on the detection results, the visual tracking system provided automatic fish counts for the test video data. Automatic counts were within two standard deviations of the manual counts for the target species and most times for the bycatch species. Unlike other recent attempts, weather and lighting conditions were largely controlled by mounting cameras under cover.

Résumé : La surveillance électronique (SE) est de plus en plus utilisée pour surveiller les prises et prises accessoires dans les pêches de capture. Les données vidéo de SE sont toujours traitées manuellement, rehaussant les coûts de gestion. Des systèmes basés sur la vision artificielle, l'apprentissage automatique et l'intelligence artificielle devraient constituer la prochaine étape de l'automatisation des flux de travail associés aux données de SE. Nous décrivons certains des obstacles que nous avons rencontrés et des approches empruntées dans la mise au point d'un système d'identification automatique et de dénombrement d'individus d'espèces cibles et d'espèces accessoires qui fait appel à des caméras déployées sur un navire commercial. Un réseau neuronal à convolution a été formé pour détecter et classer des groupes d'espèces cibles et accessoires, et un système de suivi visuel a été mis au point pour produire des décomptes. La valeur moyenne de la précision moyenne produite par le détecteur à classes multiples est de 53,42 %. À la lumière des résultats de détection, le système de suivi visuel a produit des dénombrements automatiques de poissons pour les données vidéo expérimentales. Les valeurs produites par ces dénombrements automatiques sont dans la fourchette de deux écarts-types des valeurs obtenues manuellement pour les espèces cibles et, la plupart du temps, pour les espèces accessoires. Contrairement à d'autres tentatives récentes, l'installation des caméras sous couvert a permis en bonne partie de contrôler les conditions météorologiques et d'éclairage. [Traduit par la Rédaction]

Introduction

On-vessel cameras that record fishing operations and catch, known as electronic monitoring (EM), has expanded as costs have declined (van Helmond et al. 2020). EM has the potential to provide full coverage of a fishing trip, a prospect that would be impractical or expensive for human observers, but attractive for managers who require accurate estimates of target species catch, and incidental interactions, particularly with threatened species. Indeed, because of their low relative abundance, threatened species interactions are often rare and can be easily missed if reliant on a sampling program with less than full coverage.

EM has several shortcomings. Without an on-board observer sampling program, biological samples are unable to be captured. Storing, managing and analysing the large amounts of data collected by EM also present challenges as more data are often collected than analysed because of ongoing attendant costs. These costs can be significant enough that in some cases only around 10% of available video is evaluated (Emery et al. 2019*a*). This has motivated the development of automated video analysis using machine-learning and artificial intelligence (MLAI) techniques. MLAI has been used in similar contexts to count seals on rocks from drones (McIntosh et al. 2018) and identify fish species in baited remote underwater video (BRUVs; Siddiqui et al. 2018). A

Received 25 November 2020. Accepted 2 June 2021.

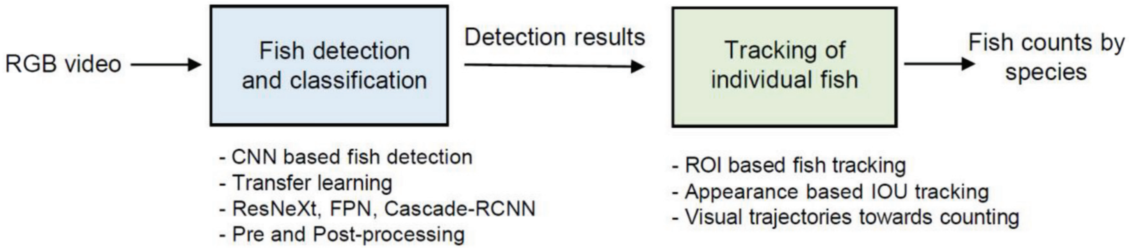
M.R. Khokher, M. Qiao, and D. Wang. CSIRO Data61, P.O. Box 76, Epping, NSW 1710, Australia.
L.R. Little,* G.N. Tuck, and C. Devine. CSIRO Oceans & Atmosphere, G.P.O. Box 1538, Hobart, TAS 7001, Australia.
D.V. Smith. CSIRO Data61, Private Bag 12, Hobart TAS 7001, Australia.
H. O'Neill and J. J. Pogonoski. CSIRO National Research Collections Australia, P.O. Box 1538, Hobart, TAS 7001, Australia.
R. Arangio. Austral Fisheries, Level 4, 50 Oxford Close, West Leederville, WA 6007, Australia.

Corresponding author: L. Richard Little (email: Rich.Little@csiro.au).

*L. Richard Little served as an Associate Editor at the time of manuscript review and acceptance; peer review and editorial decisions regarding this manuscript were handled by J. Jech.

Published at www.cdnsciencepub.com/cjfas on 5 July 2021.

Fig. 1. Steps involved in enumerating on-vessel catch using MLAI techniques in the detector (left box, blue) and tracker (right box, green).


RGB video
Fish detection and classification
Detection results
Tracking of individual fish
Fish counts by species
- CNN based fish detection
- Transfer learning
- ResNeXt, FPN, Cascade-RCNN
- Pre and Post-processing
- ROI based fish tracking
- Appearance based IOU tracking
- Visual trajectories towards counting

large selection of applications of MLAI to fisheries have been shown recently in Beyan and Browman (2020). Lu et al. (2020) for example applied deep Convolutional Neural Networks (CNNs) to digital photos collected by observers on pelagic longliners over a 10-year period. Tseng and Kuo (2020) and then Qiao et al. (2021) have recently proposed solutions for a similar application to this paper by developing deep CNNs to estimate fish counts and identify catch events respectively, from EM video collected from commercially deployed deck cameras.

While the application of MLAI to EM may seem simple, these studies reveal that there are challenges and obstacles to using MLAI operationally for management purposes. To start, the conditions and context of EM footage is often highly variable in terms of light and weather conditions. Camera quality and set-up, encompassing factors such as resolution, frame rate, angle to the region of interest, and infrastructure occlusion all influence the ability of MLAI to provide reasonable object detection rates. Applying a trained algorithm in a new context, such as a fishery or even a new vessel within a fishery, will likely compromise accuracy. Additionally, object detection and classification within an image are only part of the process since counting requires a detected object to be tracked across multiple frames of the video (Tseng and Kuo 2020). This can be particularly challenging for computer vision approaches because fishing vessels are busy places: crew members are often moving and performing several activities at once, with EM capture not necessarily their primary concern. The result is that detected objects can become occluded or have their appearance greatly altered as they are handled. This often leads to fish detections being missed across portions of the video and the need for their trajectories to be interpolated, if avoidance of double counting is desired. Previous work in Tseng and Kuo (2020) and Qiao et al. (2021) utilised fixed spatial and temporal constraints (thresholds) to determine when detected objects are the same fish across video frames. These approaches, however, are highly susceptible to fish being counted multiple times when detections are missed. Here we propose a solution to address this issue by employing a correlation tracker (Lukežic et al. 2017) to interpolate the trajectories of fish across portions of the video where the detector has missed them.

The ability to control and test different camera configurations, angles, and set-ups, could also significantly improve the accuracy of MLAI techniques. Here we also address the challenge of improving MLAI accuracy by testing different camera set-ups and configurations. Our ability to address this challenge benefited from a science–industry collaboration interested in monitoring the catch and bycatch of operations.

While EM is being used increasingly by fisheries managers, this collaboration was motivated by the realisation of the importance of advanced analytics, and methods to collect, process, and interpret data by a seafood producer. The potential cost savings, environmental outcomes, safety and waste reduction are substantial, not only through monitoring fishing activities, but also through increased understanding and management of the supply chain (Christiani et al. 2019). Improved supply chain management practices can provide publicly verified sustainability practices, optimise fishing operations, and document logistical processes in transporting and storing product at temperature (cold chain).

The Heard Island and McDonald Islands (HIMI) Patagonian toothfish fishery is a Marine Stewardship Council certified fishery, targeting Patagonian toothfish (*Dissostichus eleginoides*), and incidentally catching grenadiers (Macrouridae), skates and rays (Arhynchobatidae and Rajidae), and to a lesser degree morid cods (*Antimora rostrata*). There is considerable interest in rapidly and efficiently monitoring the catch interactions with these species from a regulatory and operator point of view. Advantages include real-time catch recording for quota reconciliation, logbook recording, early market knowledge, bycatch minimisation and compliance, and reduced monitoring costs.

The correlation tracker, and different camera angles and configuration setups were tested on a longline vessel in this fishery using GoPro cameras. The goal was to produce accurate counts, show the effect of changing camera configurations on the accuracy of the algorithm, and address some of the limitations of MLAI, particularly in operations and procedures. We offer suggestions for how these can be overcome, and lessons for practitioners in the future.

Materials and methods

We developed a model for on-vessel catch counting, called the enumerator, that incorporates MLAI techniques, and deployed cameras to a commercial sub-Antarctic longliner to collect data, which we labelled for training and testing purposes. The experimental and reporting setup includes parameter settings for detection and tracking modules, and methods to evaluate them.

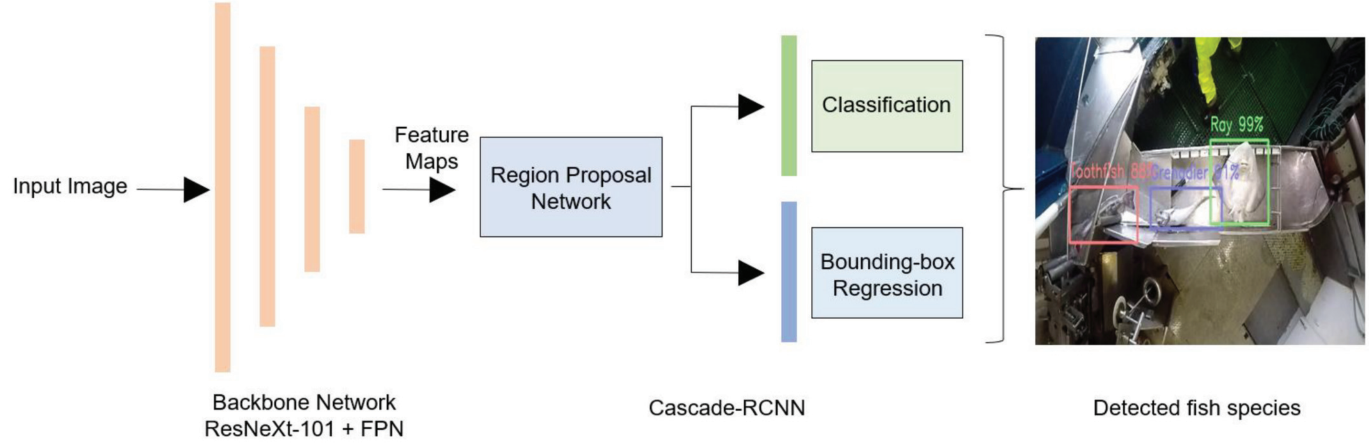
On-vessel fish counting based on MLAI: the enumerator

Automated fish counting involved a two-step process (Fig. 1). First, a CNN model was trained and then used to detect and classify target objects in each video frame (the detector). Detections were tracked by matching detections in a frame to the detections in the next frame based upon their overlapping region (the tracker). Frames with missing or occluded detections were interpolated between subsequent frames. Trajectories formed in tracking resulted in counts.

Multiclass fish species detection and classification: the detector

We used a deep learning-based object detection framework to localize and detect fish in the video imagery. Deep learning approaches have been widely and successfully used for various object detection tasks (Rawat and Wang 2017; Voulodimos et al. 2018). The multiclass fish detection and species classification model we used is based on a CNN backbone (Fig. 2), which consisted of a series of convolution and pooling layers that take an image as input to extract image features. A ResNet backbone network (He et al. 2016) uses short-cut connections in the form of a residual block in the network. Here we used an upgrade of

Fig. 2. Fish species detection and classification model based on convolutional neural networks. (FPN: feature pyramid network; RCNN: region-based convolution neural network.)



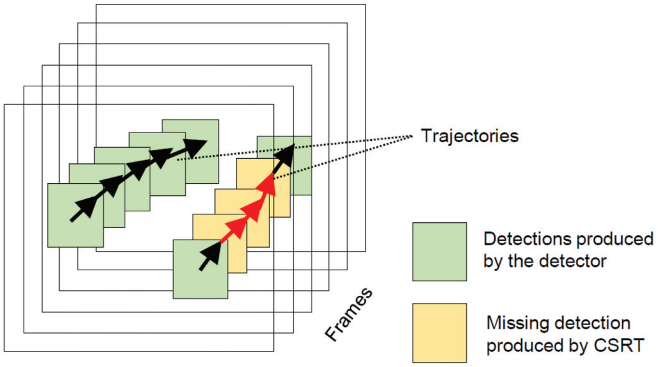
ResNet, ResNeXt (Xie et al. 2017) as the backbone network, with 101 layers (convolution and pooling) along with a Feature Pyramid Network (FPN; Lin et al. 2017). ResNeXt uses group convolution to reduce the number of parameters and increase accuracy. FPN exploits the natural pyramid structure of the CNNs and uses features from the deeper layers.

The image features extracted from the backbone network were used to detect and classify objects in the video data. A multistage object detector referred to as a cascaded-RCNN (Cai and Vasconcelos 2018) was adapted to perform multiclass fish detection. Cascade-RCNN is a sequence of detectors which are trained with increasing overlapping thresholds between objects and ground-truth regions at each stage. This solves problems like overfitting because cascade-RCNN is more selective against close false positives. Each stage is based on a Faster-Region-based Convolution Neural Network (Faster-RCNN; Ren et al. 2017) detector. Faster-RCNN have three components: Region Proposal Network (RPN), classification layer, and regression layer (Fig. 2), with the RPN generating locations of potential objects in the image (i.e., region proposals). A predefined set of bounding-boxes or anchors were used to identify the objects from the background by exploiting the image features produced by the backbone network. The region proposals were refined and forwarded to the classification layer to assign confidence scores to the objects, and then to the regression layer to determine the best coordinates of the bounding-boxes around the objects.

Fish tracking and counting: the tracker

Detections alone are unable to produce fish counts from video, given a single object will appear in subsequent frames. Given the detection results for each video frame obtained from the detector (Fig. 3), a tracking-by-detection (Bochinski et al. 2017, 2018) approach was adapted to track individual fish in the videos to produce counts. Detections were obtained in the form of a bounding-box for each fish detected in a frame and matched with the bounding-box detections in the next frame based upon their overlap. The overlap between two bounding-box detections (x, y) was determined using Intersection-Over-Union (IOU) calculated as $\mathrm{IOU}(x, y) = (\mathrm{Area}(x) \cap \mathrm{Area}(y))/(\mathrm{Area}(x) \cup \mathrm{Area}(y))$. Trajectories were formed by connecting pairs of detections across subsequent frames possessing the highest IOU (Fig. 3). The detections that possessed lower IOU values with existing trajectories were assigned to be a new trajectory. Trajectories were terminated if there were no new detections assigned to it for ttl (time-to-live) frames. Trajectories were discarded if their length was shorter than t frames or they did not possess a detection with a confidence score higher than a threshold σ. This approach is known as V-IOU tracking (Bochinski et al. 2018).

Fig. 3. Fish tracking adapted from V-IOU-based tracking using visual information (Bochinski et al. 2018).


Trajectories
Detections produced by the detector
Missing detection produced by CSRT
Frames

Missing detections, which can cause the trajectories to jump and produce incorrect and multiple counts, were addressed with a discriminative correlation filter with Channel and Spatial Reliability Tracker (CSRT; Lukežic et al. 2017) to track trajectories, find missing detections and smooth the trajectory (Fig. 3). This provides a means to fill in the gaps of the spatial trajectory to reduce fragmentation and avoid multiple counting of fish. Once trajectories were generated, class labels and their associated detection confidence scores were used to identify the species of each tracked fish. Trajectories were classified to species with the most frequent assignment of confidence scores greater than the threshold σ.

Dataset

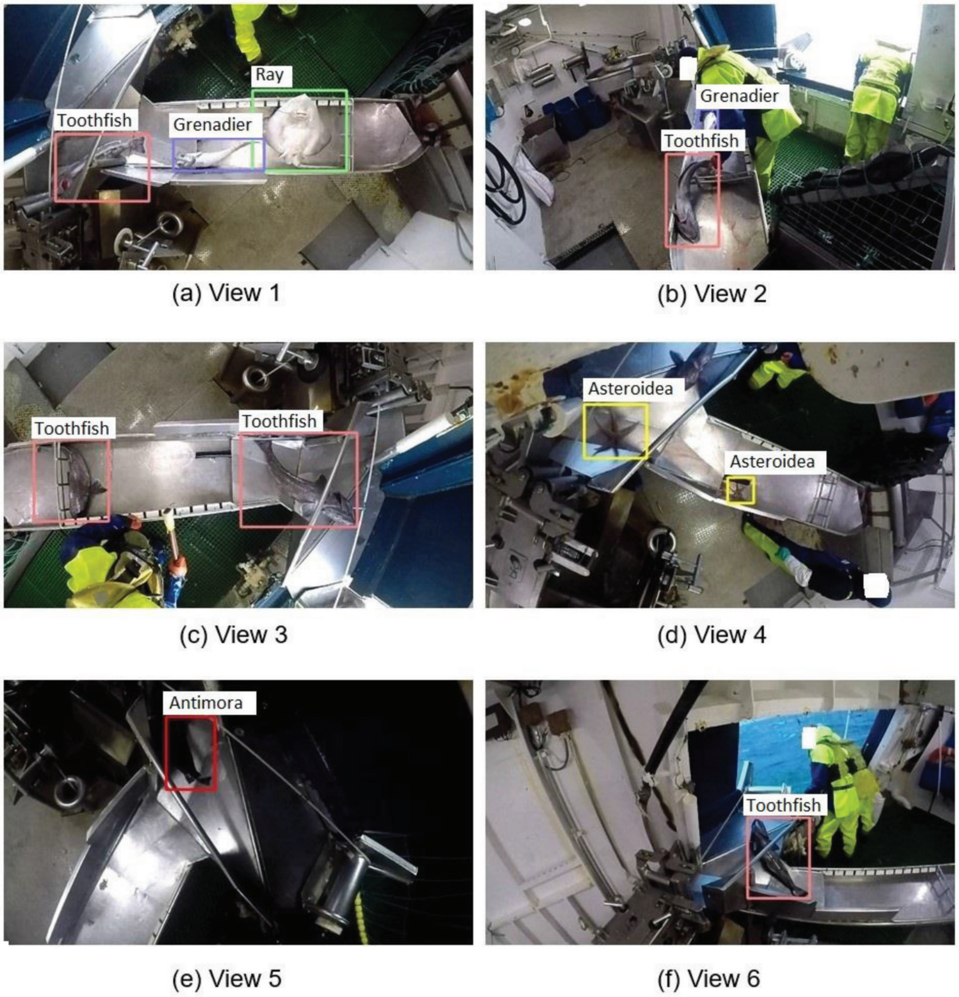
Video data acquisition

GoPro Hero5 Black cameras were used to collect 23 RGB videos from a commercial longline fishing vessel ranging between 12 to 18 min in duration. The videos were captured at a resolution of 848 × 480 with 240 frames per second (fps). Six different camera views were captured from the vessel with each having a different perspective of the line, catch, and processing components (Fig. 4).

Ground-truth for image and video data

Labelled image data from 15 videos were used to train the detector and validate performance. A total of 1200 images containing five species groups: toothfish, grenadier, skate or ray, *Antimora*, and

Fig. 4. Sample video frames from six camera views with ground-truth labels for different species groups.



(a) View 1 (b) View 2

(c) View 3 (d) View 4

(e) View 5 (f) View 6

Table 1. Number of labelled instances for each species.

Species	Count
Toothfish	890
Grenadier	529
Skates and rays	214
Antimora	106
Asteroidea	231

Asteroidea (sea star) were manually labelled by five human experts to capture observer variability (Table 1).

Experimental setup

Training and analysis were performed using a Tesla P100-SXM2 GPU with 16GB of memory. The implementation of the multiclass fish species detection was adapted from the MMDetection toolbox (Chen et al. 2019). This toolbox makes use of different libraries including Python 3.6, PyTorch 1.6, OpenCV 3.4.3, and MMCV. The 1200 labelled images were randomly divided into a training set (70%) and a testing set (30%).

The extracted video frames from the test videos were then used for frame-by-frame tracking to produce counts. Once an animal was caught and de-hooked from the fishing line, it moved onto a sorting tray where it became difficult to detect and track due to occlusions from other fish. Consequently, to minimise these effects, a region of interest (ROI) that contained the de-hooking area, was used to initially detect and track animals on the test video.

Parameter settings

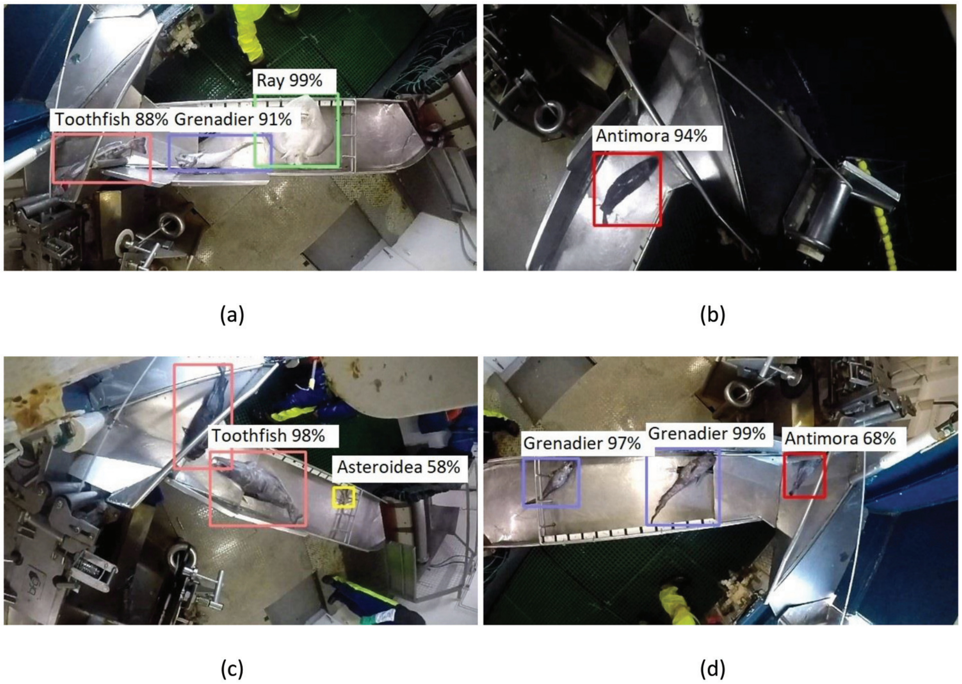
To handle the imbalance in the dataset, we performed data augmentation by horizontally flipping images of *Antimora*, skate or rays, and Asteroidea. Since our dataset is small, we employed a transfer learning approach for training. That is, the detection network was initialized with a pretrained detection model trained on a benchmark object detection dataset called COCO (Lin et al. 2014). The network was then fine-tuned on our training set.

The multiclass detector was implemented through the PyTorch framework using the MMDetection toolbox. A three-stage cascade-RCNN detector was trained with increasing IOU thresholds of 0.5, 0.6, and 0.7. A smoothed L_1 loss function for bounding-box regression and a classic cross-entropy loss function for classification were implemented. A stochastic gradient descent optimizer was used in the training. Hyperparameters in the optimizer were set as follows: (1) the learning rate, which was used to control the model response to the estimated error, was set to 0.02; (2) the batch size was set to 3, meaning that 3 images were processed by the GPU simultaneously before the network was updated. During testing, the IOU threshold was set to 0.5, a standard value for object detection tasks in computer vision, meaning that a detected bounding-box sharing 50% overlap with the corresponding ground-

Table 2. Tracker parameter settings tuned through a grid search procedure.

Parameter	Description	Value
t	Minimum number of frames considered to retain a trajectory	15
ttl	Time-to-live representing number of frames with no new detections before a trajectory is considered terminated	30
σ	Confidence threshold: a trajectory must contain at least one detection with confidence score higher than σ threshold	0.3

Fig. 5. Detection results produced by the multiclass species groups detection model. The rectangles or bounding-boxes represent the region containing the detected objects with species group identification labels and associated detection confidence. Species groups are represented by the coloured bounding-boxes: blue, green, red, yellow, and peach, for grenadier, ray, *Antimora*, Asteroidea, and toothfish, respectively: (*a*) skate and ray (confidence 85%) and two grenadiers (confidence 45%, 53%); (*b*) toothfish (confidence 89%) and *Antimora* (confidence 92%); (*c*) two toothfish (confidence 98%, 96%) and Asteroidea (confidence 59%); (*d*) two grenadiers (confidence 59%, 94%) and *Antimora* (confidence 31%).



(a)

(b)

(c)

(d)

truthed object was considered to be a true object detection (a true-positive: TP), otherwise it was considered to be a false-positive (FP).

The frame rate was set to 30 fps. To find missing detections between frames, the CSRT tracker waited for ttl = 30 frames (Table 2) for a detection to reappear and then filled in the missing detections. An animal was counted if the detection confidence score was at least 80% and assigned to the species class with the highest predicted probability.

Evaluation measures

The detection network output was compared to ground-truthed data by calculating the precision and recall values as $p = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$ and $r = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, respectively. The precision gives a percentage that shows how accurately the network detects an object and the recall gives a percentage that shows how many actual targets are detected out of all true targets. The average precision (AP; Zhang and Su 2012) was calculated for each species group and then averaged across species groups (mAP; Manning et al. 2008). For species counts, the results from the tracking model were compared to manual counts from human experts.

Results

Multiclass fish species detection

Despite the complex background (e.g., Fig. 5*a*), the detector was able to provide reliable results (Fig. 5), achieving more than 80% confidence in most detections. Detection was difficult in circumstances associated with the interclass similarities between the toothfish, grenadier, and *Antimora* species groups, due to low resolution images and variable lighting conditions. These issues could be managed by changing camera position, increasing camera resolution and improving the lighting conditions.

Figure 6 indicates that roughly 20 epochs were sufficient to train the detector, which took 98 min (4.9 min per epoch). The epoch is a hyperparameter defining the number of passes of the entire training dataset the machine learning algorithm has completed. The behaviour of the bounding-box regression loss, classification loss, and overall training loss varied during the training process (Fig. 6). Classification loss decreased significantly through the first three training epochs (Fig. 6*b*), while the most dramatic decrease in the overall training error occurred between the epochs 8 and 12 (Fig. 6*c*). Bounding-box regression error declined steadily over the first 9 epochs (Fig. 6*a*). Both bounding-box regression loss and classification loss converged around epoch 11, while the overall training loss converged around epoch 17.

The detector was evaluated on the test dataset using the trained detector model after every epoch. Figure 7 shows the mAP achieved on the test dataset after each epoch. The highest mAP of 56.69% was achieved at epochs 9 to 11, after that the mAP stabilized from epochs 13 to 20. We chose the detector that was trained for 20 epochs (when mAP stabilized) instead of 11 epochs. Although the detector trained for 11 epochs had a higher mAP, it detected many false positives which eventually created trouble during tracking.

Fig. 6. Training convergence of the multiclass fish species group detector. The graphs correspond to (*a*) bounding-box regression loss vs. number of epochs, (*b*) classification loss vs. number of epochs, and (*c*) overall training loss vs. number of epochs.

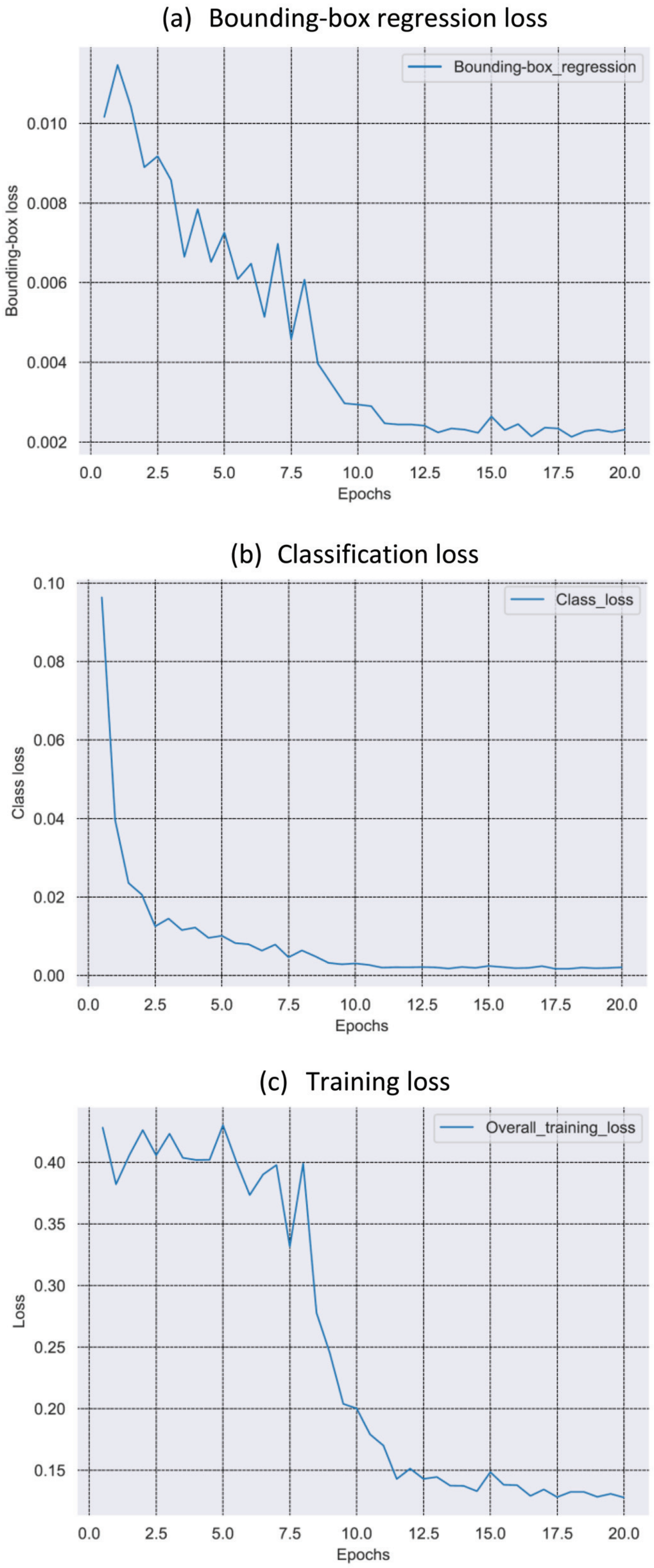


Fig. 7. Test accuracy in terms of mAP (%) achieved versus training epochs.

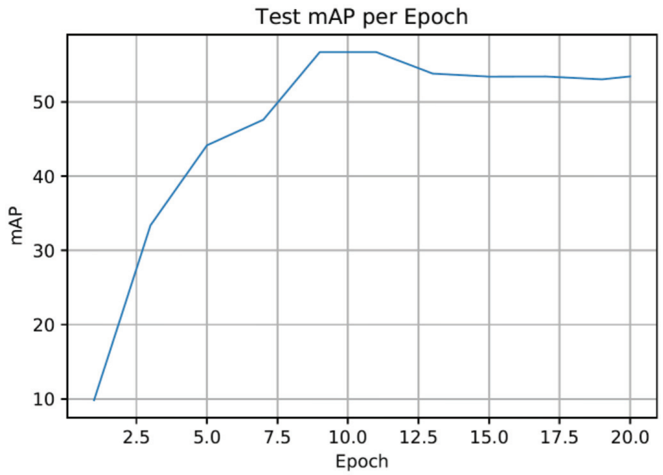
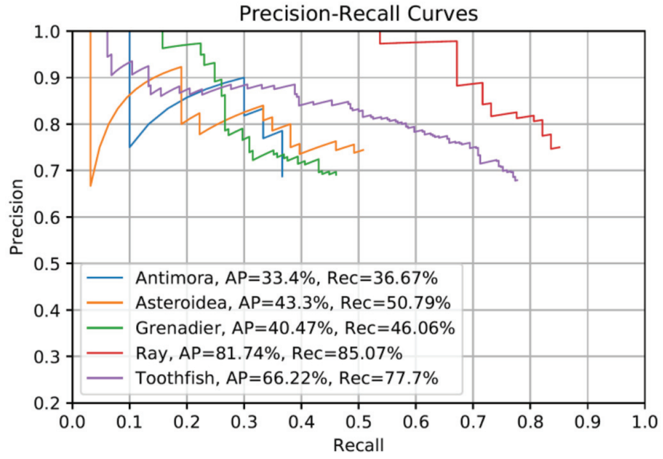


Fig. 8. Precision–recall curves of species groups achieved during detection on test data.



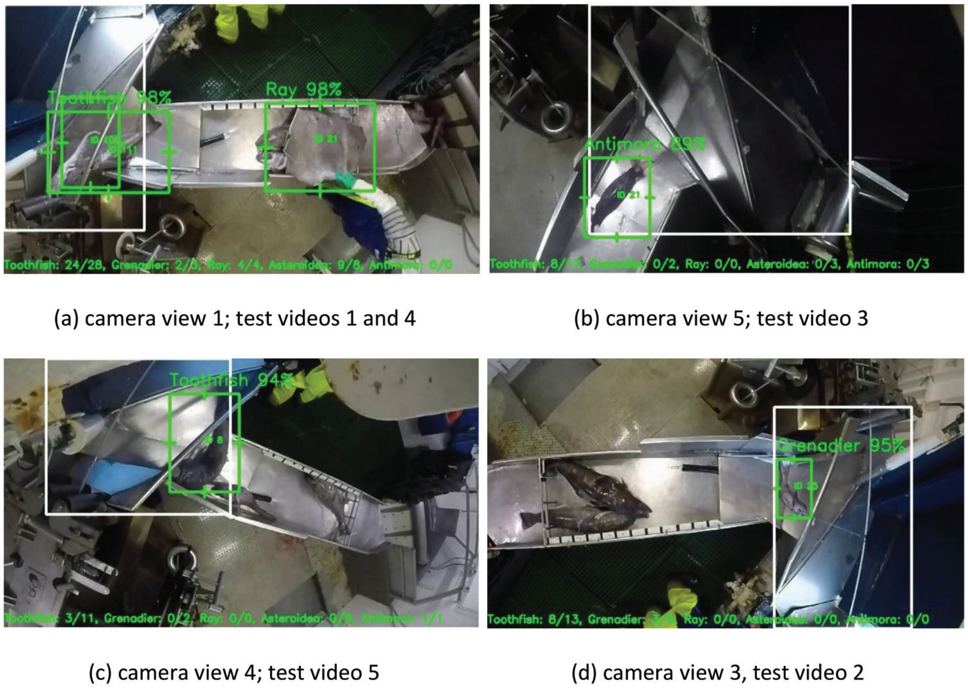
The detector was trained on 840 images. Class (i.e., species groups) imbalance within the training set made the detector biased towards the classes with more training images (i.e., toothfish; Table 1). Figure 8 shows that in the precision–recall curves for species groups during detection, the class "Ray" achieved the highest AP of 81.74%. This was likely due to their distinctive appearance compared to the other species groups. The target species (i.e., "Toothfish") achieved an AP of 66.22%. The bycatch species, like "*Antimora*" and "Grenadier", presented interclass similarities with "Toothfish" and, having fewer samples in the training set, produced lower APs. The AP of "*Antimora*" and "Grenadier" species were 33.40% and 40.47%, respectively. These results indicate that the AP can be potentially improved with more training data.

The multiclass fish species detector was evaluated using 3-fold cross validation. The detector achieved mAP of 52.74% (mean) ±1.56% (standard deviation) among all folds (Molinaro et al. 2005). Despite these results, the subsequent tracking process improved accuracy and the ability to capture missing detections, which is reflected in the counting process (next section).

Fish tracking and counting in videos

The multiclass detector was applied to five test videos from the four camera views displayed in Fig. 9. The frame rate (30 fps) was enough to capture and track an animal on the fishing line. For all

Fig. 9. Tracking and counting of fish in the test videos. The white rectangles represent the ROIs where the fish were tracked. The green bounding-boxes represent the tracked fish. The fish counts for individual fish species are given at the bottom of the video frames.


Toothfish 98%
Ray 98%
ID 21
Toothfish: 24/28, Grenadier: 2/3, Ray: 4/4, Asteroidea: 9/8, Antimora: 0/0
Antimora 83%
ID 21
Toothfish: 8/17, Grenadier: 0/2, Ray: 0/0, Asteroidea: 0/3, Antimora: 0/3
Toothfish 94%
Toothfish: 3/11, Grenadier: 0/2, Ray: 0/0, Asteroidea: 0/9, Antimora: 1/1
Grenadier 95%
Toothfish: 8/13, Grenadier:
Ray: 0/0, Asteroidea: 0/0, Antimora: 0/0

(a) camera view 1; test videos 1 and 4

(b) camera view 5; test video 3

(c) camera view 4; test video 5

(d) camera view 3, test video 2

Table 3. Trajectory filtering towards counting based on bounding-box confidence and trajectory length.

Video	No. of tracked IDs by the tracking system	No. of filtered IDs based on bounding-box confidence	No. of filtered IDs based on trajectory length
1	75	62	42
2	42	36	19
3	69	52	23
4	90	73	47
5	26	18	15

species groups except skates and rays, tracking occurred in the ROI. The skates and rays species group was tracked on the sorting tray (Fig. 9*a*) because they often were removed from the line directly at the sea door and placed directly on the sorting tray.

The number of trajectories produced to track animals within the test videos are given in Table 3. These trajectories need to be filtered to produce final fish counts. The trajectories were first filtered based on the confidence scores of the bounding-box detections present in the trajectories. That is, if a trajectory did not contain any bounding-box with a detection confidence score of more than 80%, it was discarded. This process resulted in a 20.8% decrease in trajectories on average (Table 3). Trajectories with a length of less than 15 frames, were discarded as well. This further decreased the number of trajectories down to 48.34% of the original number. The final number of trajectories that contributed towards producing the fish in each test video is given in the last column of Table 3.

The accuracy of the automated species group counts varied across the different camera views (Fig. 10). Observer based species counts also varied across the different camera views. For toothfish, the MLAI counts were within 2 standard deviations (SD) of the mean across all videos (Fig. 10*a*). Grenadier tended to be underestimated, mainly because fishers sometimes manually removed grenadier from the fishing line, or they fell onto the floor and were directly placed into the sorting tray without moving through the ROI. On a few occasions, the landing of a partial fish indicated it had been depredated (i.e., eaten by a predator); such cases were identified to be a problem for the detector. Skates and rays offered few difficulties for identification and counting as a species group, except for a few occurrences in the test videos where an individual was cut directly from the line and almost immediately returned to the water. In such cases we modified the expert observer counts by removing instances where we would not reasonably expect the detector to identify and count a fish (Fig. 10*b*). The Asteroidea count in video 5 (camera view 4; Fig. 4) was highly inaccurate because the camera view provided relatively short tracking trajectories for such small objects. *Antimora* were relatively rare in the training and testing video (only videos 3 and 5), but the identification and counts were still within 2 SD of the human observer estimate.

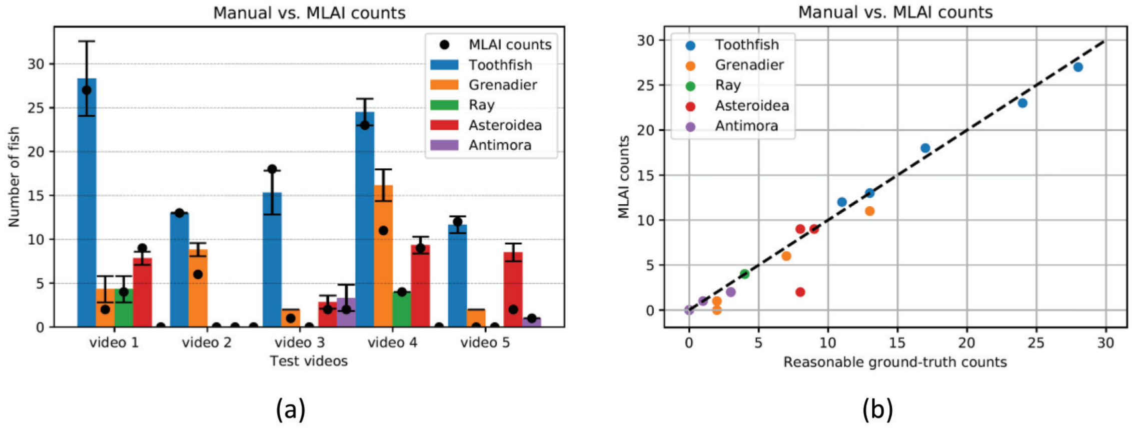
Discussion

The development of MLAI-based approaches to automate fishery catch and bycatch monitoring is a challenging task given the quantity of data and the complexity of the operating environment. With support from industry and fishery management agencies, we are confronting these challenges, and have identified a series of lessons and challenges to the operational deployment of the technology.

Training data acquisition

Many fisheries catch a range of species under a range of conditions, which adds to the work needed to develop annotated training datasets. In the case we presented, data were collected under highly controlled conditions and so weather and lighting conditions had little effect. The main source of variability was camera position. Our case study also had only a handful of relatively distinct classes (i.e., species groups), and as a result, we were able to compile a labelled database relatively easily. Other fisheries have

Fig. 10. Mean fish counts for (*a*) human observers (bars) ±2 SD, across 5 species and MLAI counts (black dots). (*b*) Scatter plot shows under- or overestimates of MLAI species counts in comparison with counts that could be reasonably achieved (excluding edge cases).



greater diversity in catch, and conditions. Qiao et al. (2021) examined fewer frames than examined here out of 720 h of video from more variable conditions in the Australian eastern tuna and billfish fishery (ETBF), and found 46 species identified by EM reviewers. Of these, five were target, eight by-product, and the remainder bycatch species (Emery et al. 2019*a*). Prioritisation of species or groups for MLAI consideration therefore is critical because otherwise, some species or species groups with inadequate training data might still require monitoring through logbooks, human auditing of video, or on-vessel human observation.

Prioritisation can be made based on an economic, ecological, or vulnerability status. The problem, however, is that in a multispecies context EM data are usually unbalanced, reflecting relative abundances. *Antimora* in our data for example, were uncommon with only 106 instances (i.e., 7%) in the training dataset compared to the other species groups. In the two test videos where *Antimora* were present, the enumerator was accurate. This is most likely because *Antimora* is a distinctive species comparing with others in the training dataset. More generally however, acquiring enough data to use MLAI on rare species can be challenging, and several approaches could be used to address the problem. Firstly, images can be synthesized by mimicking capture either electronically or physically in a controlled environment. Alternatively, where limited amounts of data exist, augmentation techniques are commonly used. We used this technique here by applying horizontal transformations to each training image.

Another method for acquiring imagery of rare species would be to use machine learning techniques to generate synthetic training images. EM data may also not be able to capture interactions that occur off-camera, but still might indicate anomalous behaviour of crew in capturing non-target species. Handling of bycatch is markedly different to that of target species, and if the movement patterns of crew to remove bycatch can be uniquely characterised (for example, leaning over the vessel to cut line), then the MLAI algorithm might be trained to detect these behaviours. An algorithm that bookmarks or extracts such an event in a video recording can then be checked by an analyst.

Lastly, crew can be requested either to ensure direct and unhindered vision from the on-vessel cameras, or to hold specimens directly in front of a camera (Gilman et al. 2019). This approach has been trialled in the Australian tuna fishery (ETBF) for seabird bycatch since 1 July 2020.

Obtaining enough data for automated EM analysis is also a challenge. EM data are valuable, both commercially and from a privacy perspective, which can make access difficult (Gilman et al. 2019; van Helmond et al. 2020). Public datasets of annotated fish images to assist in model development are available (Cutter et al. 2015; The Nature Conservancy 2020; Boom et al. 2012). The most relevant is the FishNet open image dataset (The Nature Conservancy 2020), which contains approximately 35 000 annotated images from commercial fishing vessels. The dataset currently consists of 26 species, a majority of which have less than 100 annotated instances, and consequently, is not likely to possess the quantity and diversity of images needed to train models for a vast majority of applications.

Transferring fish knowledge into networks

Even if the public fish image datasets do not provide a complete solution for training, they do provide a useful starting point for transfer learning (Yosinski et al. 2014). Such an approach is commonly used in computer vision by transferring the visual knowledge from networks trained on very large, general image databases, such as ImageNet (Deng et al. 2009) or Open Images (Kuznetsova et al. 2020), to models with specific applications.

Recent fish detection and recognition models have applied transfer learning to general feature extraction (FE) networks as the starting point for model training. Tseng et al. (2020) and French et al. (2020) developed fish species detection models by using a Mask R-CNN (He et al. 2017) architecture that had been pre-trained with the general COCO dataset (Lin et al. 2014); the Mask R-CNN was then retrained upon relatively small datasets of the target species. Siddiqui et al. (2018) combined a Support Vector Machine (SVM) classifier with three different pre-trained FE based CNNs (AlexNet, VGGNet and ResNet) to address the limited data that were available for training an underwater fish recognition system.

There are yet to be any existing models that retrain FE networks on publicly accessible fish databases, such as FishNet. This could potentially improve performance and reduce the quantity of application-specific training images that will be needed.

Classification bias

Training datasets should produce minimal bias in classifiers. It is generally not advisable for instance, to use training datasets that possess a heavy imbalance between the different species of interest. While it is far more convenient to prepare training datasets skewed by species with an abundance of images (i.e., the majority classes), it is likely that any trained classifier will exhibit some bias towards the majority class. This is particularly problematic when rare species with less training instances (i.e., minority classes) are important to detect due to their vulnerability status for instance. Trained detectors with high recall and a greater tolerance for false positives, which can be scrutinised, are clearly desirable when there is a strong need to detect threatened species interactions.

Different strategies can be employed to reduce classification bias. Images from the majority classes can be under-sampled to achieve a better balance with minority class species, but this is considered wasteful when the class skew is significant. An alternative option is to augment the minority classes by generating synthetic images (Shorten and Koshgoftaar 2019) using different image transformations, kernels or deep learning (Frid-Adar et al. 2018). Whilst synthetic image generation increases the size and diversity of species or species groups in the minority class, it must also ensure that the data remains representative of them and does not introduce significant noise. Tseng et al. (2020), Allken et al. (2019) and Zheng et al. (2018) augmented training datasets for fish detection and recognition by applying different transformations (i.e., flipping, shifting, blurring, rotating and scaling) to annotated fish from training images.

Taxonomic classification

Ideally, computer vision models should be able to classify all catch and bycatch at the species level, however, we have found that it can be particularly difficult to discriminate between fish species that possess subtle differences in external appearance, a challenging task even for a human expert. In many applications, it may be sufficient to classify the catch at higher taxonomic ranks (genus, family or order), particularly if it provides greater assurance about the classification accuracy. In terms of species collected in the HIMI toothfish fishery, grenadiers of the genus *Macrourus* are particularly difficult to separate even by trained taxonomists. An additional Southern Ocean species was only recently recognised and described using a combination of molecular and traditional taxonomic methods (Smith et al. 2011; McMillan et al. 2012). At least three, but possibly four, *Macrourus* species are known to be present at Heard Island (CSIRO Australian National Fish Collection records) and the complexity of separating the species suggests significant additional work would be needed to train an MLAI system to distinguish them. Additionally, multiple species of skates of the genus *Bathyraja* are present at HIMI and these species are still being investigated in a taxonomic sense (e.g., Last et al. 2016). When the species are more adequately resolved, the MLAI technology could potentially be trained to detect differences in colour and shape that may separate the species.

We suggest MLAI models be designed to provide a hierarchical, multi-label classification (Wehrmann et al. 2018) where the detected catch is classified at multiple taxonomic ranks (i.e., species, genus, family) simultaneously, each with an associated confidence measure to indicate the granularity at which the classification can be trusted. Bayesian neural networks (Gal and Ghahramani 2015) are an important approach to address the issue of classification uncertainty. In addition, the level of hierarchical classification that is satisfactory to a manager should be specified prior to embarking on a potentially expensive MLAI application for a fishery. If it is critical that species level identification is needed, then MLAI may not be the answer.

Automated monitoring in a complex work environment

Whilst models can be developed to accurately detect, classify and track fish using relatively cheap and low-resolution video cameras, additional challenges are introduced by the deployment of such cameras in the complex work environment of a fishing vessel. Camera configuration is critically important, as shown, especially because placement can differ widely across vessels. For example, French et al. (2020) showed that expert human observers performed better than MLAI algorithms when tested on data from a separate vessel. In a study comparing vessel logbook records of catch against EM records in the ETBF, Emery et al. (2019*b*) postulate that inappropriate camera positioning led to lower EM records of protected species than logbooks. In addition, in some circumstances poor lighting also meant that EM analysts were not able to determine whether seabird mitigation devices were being deployed. These issues will also cause problems for any application of MLAI to EM footage.

Fish tracking is complicated by the busy and crowded fishing vessel environment. The interaction between fish and workers often leads to occlusions, and on longliners the consistent motion patterns of fish attached to the fishing line becomes far less predictable once handled by workers. Fish occlusion and unpredictable motion patterns can significantly degrade tracking performance, and hence, lead to duplicated counting of fish.

We also note that not all work practices on fishing vessels are conducive to video monitoring. In our longline application for example, rays normally would be removed from the fishing line prior to entering the vessel and, hence, often were immediately transferred below the deck or returned to the water. Despite the accuracy of the enumerator, skates and rays were only present in the field of view in our footage for a minimal period. This could present difficulties for identifying objects that are less morphologically distinct. Similar drawbacks to using MLAI can occur with fish handled by the crew that bypass the ROI due to needs for expediency or practicality (for example, a fish may come off the hook, fall to the deck floor and then be placed directly into the sorting tray by crew). It is worth considering if and how simple modifications to work practice or the work environment itself can be made to enhance the ability to monitor catch without causing work disruption.

The detector processed the test data at a rate of about 5 to 6 images per second using a GPU. The ability ultimately to capture and process video for near-real time counts is highly desirable for informing fishing operations, managing production (Christiani et al. 2019), and conserving vulnerable bycatch species, especially if coupled with telecommunications technology. Deploying onboard edge computing devices equipped with GPUs, to batch or parallel process videos as they are captured, potentially adds another disruption or imposition on vessel work practices. With regulated work practice, automated identification and counting can become a reality, which will ultimately benefit the fishery, and manage its effect on the wider ecosystem.

Acknowledgements

This project was supported by the CSIRO Machine Learning & Artificial Intelligence Future Science Platform. We thank Martin Exel and Austral Fisheries Pty Ltd. for their ongoing support. Al Graham (CSIRO, Hobart) and Narissa Bax assisted with acquisition and photography of voucher specimens into the CSIRO Australian National Fish Collection, respectively.

References

Allken, V., Handegard, N.O., Rosen, S., Schreyeck, T., Mahiout, T., and Malde, K. 2019. Fish species identification using a convolutional neural network trained on synthetic data. ICES J. Mar. Sci. **76**(1): 342–349. doi:10.1093/icesjms/fsy147.

Beyan, C., and Browman, H.I. 2020. Setting the stage for the machine intelligence era in marine science. ICES J. Mar. Sci. **77**: 1267–1273. doi:10.1093/icesjms/fsaa084.

Bochinski, E., Eiselein, V., and Sikora, T. 2017. High-Speed tracking-by-detection without using image information. *In* Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September, 2017. pp. 1–6. doi:10.1109/AVSS.2017.8078516.

Bochinski, E., Senst, T., and Sikora, T. 2018. Extending IOU based multi-object tracking by visual information. *In* Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 27–30 November, 2018. pp. 1–6. doi:10.1109/AVSS.2018.8639144.

Boom, B.J., Huang, P.X., Spampinato, C., Palazzo, S., He, J., Beyan, C., et al. 2012. Long-term underwater camera surveillance for monitoring and analysis of fish populations. *In* Proceedings of the International Workshop on Visual Observation and Analysis of Animal and Insect Behavior (VAIB), Tsukuba, Japan.

Cai, Z., and Vasconcelos, N. 2018. Cascade R-CNN: delving into high quality object detection. *In* Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 18–23 June, 2018. pp. 6154–6162. doi:10.1109/CVPR.2018.00644.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., et al. 2019. MMDetection: open MMLab detection toolbox and benchmark. arXiv:1906.07155 [cs.CV].

Christiani, P., Claes, J., Sandnes, E., and Stevens, A. 2019. Precision fisheries: navigating a sea of troubles with advanced analytics. McKinsey & Company.

Cutter, G., Stierhoff, K., and Zeng, J. 2015. Automated detection of rockfish in unconstrained underwater videos using Haar cascades and a new image dataset: labeled fishes in the wild. *In* Proceedings of the 2015 IEEE Winter Applications and Computer Vision Workshops, Waikoloa, HI, 6–9 January, 2015. pp. 57–62. doi:10.1109/WACVW.2015.11.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. 2009. ImageNet: a large-scale hierarchical image database. *In* Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, Fla., 20–25 June, 2009. pp. 248–255. doi:10.1109/CVPR.2009.5206848.

Emery, T.J., Noriega, R., Williams, A.J., and Larcombe, J. 2019*a*. Changes in logbook reporting by commercial fishers following the implementation of electronic monitoring in Australian Commonwealth fisheries. Mar. Pol. **104**: 135–145. doi:10.1016/j.marpol.2019.01.018.

Emery, T.J., Noriega, R., Williams, A.J., and Larcombe, J. 2019*b*. Measuring congruence between electronic monitoring and logbook data in Australian Commonwealth longline and gillnet fisheries. Oceans Coastal Manage. **168**: 307–321. doi:10.1016/j.ocecoaman.2018.11.003.

French, G., Mackiewicz, M., Fisher, M., Holah, H., Kilburn, R., Campbell, N., and Needle, C. 2020. Deep neural networks for analysis of fisheries surveillance video and automated monitoring of fish discards. ICES J Mar. Sci. **77**: 1340–1353. doi:10.1093/icesjms/fsz149.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. 2018. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, **321**: 321–331. doi:10.1016/j.neucom.2018.09.013.

Gal, Y., and Ghahramani, Z. 2015. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. arXiv:1506.02142 [stat.ML].

Gilman, E., Legorburu, G., Fedoruk, A., Heberer, C., Zimring, M., and Barkai, A. 2019. Increasing the functionalities and accuracy of fisheries electronic monitoring systems. Aquat. Conserv Mar. Freshw. Ecosyst. **29**: 901–926. doi:10.1002/aqc.3086.

He, K., Zhang, X., Ren, S., and Sun, J. 2016. Deep residual learning for image recognition. *In* Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 27–30 June, 2016. pp. 770–778. doi:10.1109/CVPR.2016.90.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. 2017. Mask R-CNN. *In* Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October, 2017. pp. 2980–2988. doi:10.1109/ICCV.2017.322.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., et al. 2020. The Open Images Dataset V4: unified image classification, object detection, and visual relationship detection at scale. Int. J. Comput. Vis. **128**: 1956–1981. doi:10.1007/s11263-020-01316-z.

Last, P.R., Stehmann, M.F.W., Séret, B., and Weigmann, W. 2016. Chapter 20: Softnose skates, family Arhynchobatidae. *In* Rays of the World. *Edited by* P.R. Last, W.T. White, M.R. de Carvalho, B. Séret, M.F.W. Stehmann, and G.J.P. Naylor. CSIRO Publishing, Melbourne, Australia. pp. 364–472.

Lin, T.-Y., Maire., M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., and Zitnick, L. 2014. Microsoft COCO: common objects in context. *In* Computer Vision – ECCV 2014, 13th European Conference, Zurich, Switzerland, 6–12 September, 2014. *Edited by* D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars. Lecture Notes in Computer Science, Vol. 8693. Springer, Cham. pp. 740–755. doi:10.1007/978-3-319-10602-1_48.

Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. 2017. Feature pyramid networks for object detection. *In* Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 21–26 July, 2017. pp. 936–944. doi:10.1109/CVPR.2017.106.

Lu, Y.-C., Tung, C., and Kuo, Y.-F. 2020. Identifying the species of harvested tuna and billfish using deep convolutional neural networks. ICES J. Mar. Sci. **77**: 1318–1329. doi:10.1093/icesjms/fsz089.

Lukežic, A., Vojír, T., Zajc, L.C., Matas, J., and Kristan, M. 2017. Discriminative correlation filter with channel and spatial reliability. *In* Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 21–26 July, 2017. pp. 4847–4856. doi:10.1109/CVPR.2017.515.

Manning, C., Raghavan, P., and Schütze, H. 2008. Introduction to information retrieval. Cambridge University Press, Cambridge, UK.

McIntosh, R.R., Holmberg, R., and Dann, P. 2018. Looking without landing-using remote piloted aircraft to monitor fur seal populations without disturbance. Front. Mar. Sci. **5**: 2296–7745. doi:10.3389/fmars.2018.00202.

McMillan, P.J., Iwamoto, T., Stewart, A.L., and Smith, P.J. 2012. A new species of grenadier, genus *Macrourus* (Teleostei, Gadiformes, Macrouridae) from the southern hemisphere and a revision of the genus. Zootaxa, **3165**: 1–24. doi:10.11646/zootaxa.3165.1.1.

Molinaro, A.M., Simon, R., and Pfeiffer, R.M. 2005. Prediction error estimation: a comparison of resampling methods. Bioinformatics, **21**(15): 3301–3307. doi:10.1093/bioinformatics/bti499.

Qiao, M., Wang, D., Tuck, G.N., Little, L.R., Punt, A.E., and Gerner, M. 2021. Deep learning methods applied to electronic monitoring data: automated catch event detection for longline fishing. ICES J. Mar. Sci. **78**: 23–35. doi:10.1093/icesjms/fsaa158.

Rawat, W., and Wang, Z. 2017. Deep convolutional neural networks for image classification: a comprehensive review. Neural Comput. **29**: 2352–2449. doi:10.1162/neco_a_00990. PMID:28599112.

Ren, S., He, K., Girshick, R., and Sun, J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intel. **39** (6): 1137–1149. doi:10.1109/TPAMI.2016.2577031. PMID:27295650.

Shorten, C., and Koshgoftaar, T.M. 2019. A survey on image data augmentation for deep learning. J. Big Data, **6**: 60. doi:10.1186/s40537-019-0197-0.

Siddiqui, S.A., Salman, A., Malik, M.I., Shafait, F., Mian, A., Shortis, M.R., and Harvey, E.S. 2018. Automatic fish species classification in underwater videos: exploiting pre-trained deep neural network models to compensate for limited labelled data. ICES J. Mar. Sci. **75**: 374–389. doi:10.1093/icesjms/fsx109.

Smith, P.J., Steinke, D., McMillan, P.J., Stewart, A.L., McVeagh, S.M., Díaz de Astarloa, J.M., et al. 2011. DNA barcoding highlights a cryptic species of grenadier *Macrourus* in the Southern Ocean. J. Fish Biol. **78**: 355–365. doi:10.1111/j.1095-8649.2010.02846.x. PMID:21235567.

The Nature Conservancy. 2020. FishNet Open Image Dataset. Available from https://www.fishnet.ai/.

Tseng, C.-H., and Kuo, Y.-F. 2020. Detecting and counting harvested fish and identifying fish types in electronic monitoring system videos using deep convolutional neural networks. ICES J. Mar. Sci. **77**: 1367–1378. doi:10.1093/icesjms/fsaa076.

Tseng, C.-H., Hsieh, C.-L., and Kuo, Y.-F. 2020. Automatic measurement of the body length of harvested fish using convolutional neural networks. Biosyst. Eng. **189**: 36–47. doi:10.1016/j.biosystemseng.2019.11.002.

van Helmond, A.T.M., Mortensen, L.O., Plet-Hansen, K.S., Ulrich, C., Needle, C.L., Oesterwind, D., et al. 2020. Electronic monitoring in fisheries: lessons from global experiences and future opportunities. Fish Fish. **21**: 162–189. doi:10.1111/faf.12425.

Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. 2018. Deep learning for computer vision: a brief review. Comp. Intel. Neurosci. **2018**: 7068349. doi:10.1155/2018/7068349.

Wehrmann, J., Cerri, R., and Barros, R. 2018. Hierarchical multi-label classification networks. *In* Proceedings of the International Conference on Machine Learning, Stockholmsmässan, Stockholm, Sweden, 10–15 July, 2018. PMLR Vol. 80. pp. 5075–5084.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. 2017. Aggregated residual transformations for deep neural networks. *In* Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 21–26 July, 2017. pp. 5987–5995. doi:10.1109/CVPR.2017.634.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. 2014. How transferable are features in deep neural networks? *In* NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems, Cambridge, Mass., 8–13 December, 2014. Volume 2. pp. 3320–3328.

Zhang, P., and Su, W. 2012. Statistical inference on recall, precision and average precision under random selection. *In* Proceedings of the 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, Chongqing, China, 29–31 May, 2012. pp. 1348–1352. doi:10.1109/FSKD.2012.6234049.

Zheng, Z., Guo, C., Zheng, X., Yu, Z., Wang, W., Zheng, H., et al. 2018. Fish recognition from a vessel camera using deep convolutional neural network and data augmentation. *In* Proceedings of 2018 OCEANS – MTS/IEEE Kobe Techno-Oceans (OTO), Kobe, Japan, 6 December, 2018. doi:10.1109/OCEANSKOBE.2018.8559314.