frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 23 November 2022
DOI 10.3389/fmars.2022.918504

OPEN ACCESS

EDITED BY
Ahmad Salman,
National University of Sciences and Technology (NUST), Pakistan

REVIEWED BY
Ahsan Jalal,
National University of Sciences and Technology (NUST), Pakistan
Miguel Pessanha Pais,
Center for Marine and Environmental Sciences (MARE), Portugal

*CORRESPONDENCE
Rod M. Connolly
r.connolly@griffith.edu.au

SPECIALTY SECTION
This article was submitted to Marine Conservation and Sustainability, a section of the journal Frontiers in Marine Science

RECEIVED 12 April 2022
ACCEPTED 10 November 2022
PUBLISHED 23 November 2022

CITATION
Connolly RM, Jinks KI, Herrera C and Lopez-Marcano S (2022) Fish surveys on the move: Adapting automated fish detection and classification frameworks for videos on a remotely operated vehicle in shallow marine waters. *Front. Mar. Sci.* 9:918504. doi: 10.3389/fmars.2022.918504

Fish surveys on the move: Adapting automated fish detection and classification frameworks for videos on a remotely operated vehicle in shallow marine waters

Rod M. Connolly[1]*, Kristin I. Jinks[1], Cesar Herrera[1] and Sebastian Lopez-Marcano[1,2]

[1]Coastal and Marine Research Centre, Australian Rivers Institute, School of Environment and Science, Griffith University, Gold Coast, QLD, Australia, [2]Quantitative Imaging Research Team, CSIRO, Epping, NSW, Australia

Mobile underwater cameras, diver-operated or on underwater vehicles, have become popular for monitoring fisheries. Replacing divers with cameras has clear advantages, such as creating permanent records and accessing waters unavailable to divers. The use of cameras, however, typically produces large quantities of video that are time-consuming to process manually. Automated analysis of underwater videos from stationary cameras using deep learning techniques has advanced considerably in recent years, but the use of mobile cameras potentially raises new challenges for existing methods. We tested how well three automation procedures for stationary underwater cameras, taking an object-centric rather than background-centric approach, performed on surveys of fish using a mobile camera. We analyzed underwear drone videos from reef and seagrass habitat to detect and count two marine fisheries species, luderick (*Girella tricuspidata*) and yellowfin bream (*Acanthopagrus australis*). Three convolutional neural network (CNN) frameworks were compared: Detectron Faster R-CNN, Detectron2 Faster R-CNN (using a Regional Proposal Network, RPN), and YOLOv5 (a single-stage detector, SSD). Models performed well overall. Per frame, overall F1 scores ranged 81.4 - 87.3%, precision 88.2 – 96.0%, and recall 73.2 - 88.2%. For quantifying MaxN per video, overall F1 ranged 85.9 – 91.4%, precision 81.9 – 95.3%, and recall 87.1 – 91.1%. For luderick, F1 was > 80% for all frameworks per frame and 89% or higher for MaxN. For yellowfin bream, F1 scores were lower (35.0 - 73.8% for frames, 43.4 - 73.0% for MaxN). Detectron2 performed poorly, and YOLOv5 and Detectron performed similarly with advantages depending on metrics and species. For these two frameworks, performance was as good as in videos from stationary cameras. Our findings show that object detection technology is very useful for extracting fish data from mobile underwater cameras for the system

tested here. There is a need now to test performance over a wider range of environments to produce generalizable models. The key steps required area to test and enhance performance: 1. for suites of species in the same habitats with different water clarity, 2. in other coastal environments, 3. trialing cameras moving at different speeds, and 4. using different frame-rates.

KEYWORDS

underwater drone, object detection, diver-operated video (DOV), remotely operated vehicle (ROV), deep learning, ego-motion, fish recognition

Introduction

The use of cameras to survey fish populations is becoming increasingly widespread for fisheries assessments and fish ecology (Sward et al., 2019). Mobile underwater video cameras, either diver operated or on underwater drones, remotely operated vehicles (ROVs) or autonomous underwater vehicles, are beginning to replace traditional underwater visual census (UVC) techniques for habitats such as coral reefs, rocky reefs, deep waters, and offshore gas platforms (Andaloro et al., 2013; Goetze et al., 2015; Sward et al., 2019). Cameras have clear advantages over UVC: 1) cameras produce permanent records during a survey; 2) remote cameras on drones or ROVs can be used in deep waters inaccessible to divers, and in waters with dangerous animals; 3) remote cameras reduce costs associated with diving (Andaloro et al., 2013; Sward et al., 2019; Garner et al., 2021). Cameras can also avoid the known biases of UVC for different types of fish (Bernard et al., 2013; Sheaves et al., 2020). Regardless of the technique utilized, the use of cameras generates large volumes of imagery that can be time-consuming and costly to process manually (Sheaves et al., 2020). Automating the processing and analysis of underwater videos therefore becomes a valuable step for fisheries management.

Deep learning, a form of machine learning, has been used to automate the analysis of imagery from a wide range of environments, including aquatic ecosystems (Dawkins et al., 2017; Jalal et al., 2020; Salman et al., 2020). Deep learning techniques such as convolutional neural networks (CNNs) have proven successful for fish recognition and tracking from stationary cameras such as baited/unbaited remote underwater video systems (BRUVs/RUVs) (Mandal et al., 2018; Villon et al., 2018; Ditria et al., 2020a; Coro and Walsh, 2021; Ditria et al., 2021; Lopez-Marcano et al., 2021). Object recognition can be conventionally achieved by object- and background-centric methods (Heo et al., 2017). In object-centric methods objects of interest are detected by training a model capable of localizing and identifying object features in the image. Background-centric methods, by contrast, first attempt to differentiate background and foreground elements in the image, so moving objects of any type can then be localized and identified. While both methods can accomplish object recognition, the first one is solely concerned with the object detection task, while the second involves image segmentation and object detection tasks. Furthermore, recent advances on instance, semantic and panoptic segmentation combine several computer vision tasks to achieve an holistic understanding of the scene (Kirillov et al., 2019; Kim et al., 2020). However, complex dynamic scenes, such as those created by mobile cameras, pose a different set of challenges to object recognition and segmentation methods. The main challenge is that constantly changing backgrounds can make it difficult to detect objects of interest that are also in motion (Wei et al., 2021). Complementary methods and additional developments and modelling are required for background-centric methods to be reliable on moving cameras at the required accuracy (for instance, Cutter et al., 2015; Heo et al., 2017; Diamantas and Alexis, 2020; Wei et al., 2021); and a holistic understanding of underwater scenes (e.g. panoptic segmentation) requires a considerable effort of annotating and training segmentation models beyond the scope and capabilities of most organizations (but note O'Byrne et al., 2018; Arain et al., 2019; Islam et al., 2020; Liu and Fang, 2020). Given these difficulties and complexities, for single-camera surveys it will be beneficial if existing object-centric methods already developed for underwater imagery prove to be reliable in complex dynamic scenes. Additionally, increased occurrence of blurred images from mobile compared with stationary cameras, as reported in automated image analyses from unmanned aerial vehicles (Sieberth et al., 2013), represents another challenge influencing the applicability of object-centric methods to underwater imagery. It is therefore important to test whether recent successes in automation for underwater surveys using stationary cameras can be achieved for methods using mobile cameras.

Automation, in the form of object detection on underwater videos from mobile cameras, has been used in certain situations. For example, Francisco et al. (2020) used a multi-device camera system to track fish against a complex background, one a diver-operated stereo-video system, the other a single camera fixed on

the seabed. They used CNNs to detect the position of fish within a 3D re-creation of the substrate background. The success of the algorithms used by Francisco et al. (2020) is promising but is difficult to apply to the more typical situation of a diver or drone operating a single camera device. Machine learning algorithms have also been used to guide underwater ROVs, for example in tracking marine animals such as jellyfish (Katija et al., 2021). Walther et al. (2004) used selective attention algorithms on ROVs to automatically detect and track a range of animals, including jellyfish. However, although these studies used automated object detection, none of them reported on the effectiveness of automated data extraction on videos from mobile cameras. The need remains, therefore, to test and report the performance of deep learning algorithms for identifying and counting fish in videos from mobile cameras.

We tested how object-centric automation procedures established for stationary underwater cameras performed on videos from a mobile camera. Our aim was to establish whether videos obtained from a mobile camera could be reliably trained and analyzed using deep learning to achieve high accuracy of detecting and counting fish. To achieve this, we developed multispecies fish detection models trained and tested on videos from an underwater drone. We compared performance of object detection per frame, and on extraction of MaxN values, the maximum number of fish of a particular species in any single frame of a video (Langlois et al., 2020). Performance was compared with that in comparable previous studies on stationary cameras. Given the increasingly widespread use of mobile cameras in fisheries science, our finding that an existing object-detection method developed for stationary cameras is suitably reliable for identifying and counting fish in videos from mobile cameras offers a potentially important step-change in image processing efficiency.

Materials and methods

Data collection and preparation

To achieve our aim of testing the effectiveness of deep learning techniques on videos from mobile cameras, we used CNNs to detect and count two common species of fish in surveys using an underwater drone. We surveyed a mixed reef and seagrass habitat in 3 m water depth, in Tallebudgera estuary, southeast Queensland, Australia (28°05'54.0"S, 153°27'20.5"E). The two target species, luderick (*Girella tricuspidata*) and yellowfin bream (hereafter bream, *Acanthopagrus australis*), are commercially and recreationally harvested fisheries species that also have important ecological roles in coastal waters (Ferguson et al., 2013; Gilby et al., 2017).

Three surveys using a QYSea FiFish V6 drone (Figure 1) filming at 25 fps were conducted on each of three days during November 2021. Each survey ran for 20 min, haphazardly criss-crossing the ~ 1 ha site, with a minimum of 20 min between consecutive surveys. Videos from each survey were clipped into 2.5 min segments to obtain multiple replicates per survey per day. This resulted in 8 segments per survey (except for one survey cut short by an increase in turbidity that reduced visibility). Video snippets were created from these segments, focusing on periods with one or more fish, for a total of 237 snippets of varying length (approximately evenly spread across surveys and days).

We created two completely independent datasets: 1) the training dataset, using snippets from surveys on two of the three days, and 2) the testing dataset, using snippets from surveys on the remaining day. All individuals of the two target species were annotated using bounding boxes in all snippets in both datasets, on 5 frames per second of video (i.e., every fifth frame), for a

FIGURE 1
Images showing underwater drone with camera during fish surveys (left), and an example image from a video of fish being automatically identified and counted using Detectron model (TP = true positive identification) (right).

total of 11,714 annotations (8,830 luderick and 2,884 bream; Table 1). Although luderick and bream were the most common species, two other species occurred occasionally, common silverbiddy (*Gerres subfasciatus*) and estuary glassfish (*Ambassis marianus*), and a small number of annotations of these other species were also included in the training dataset (total of 460 annotations); the objective was not to count these non-target species in testing performance, but to optimize the model for the target species (a technique used previously on stationary videos (Connolly et al., 2021).

Object detection, MaxN, and performance metrics

For object detection, the tasks of localizing and classifying fish into the two pre-determined labels, we used three CNN frameworks, the first being particularly widely used recently for fish detection in videos from fixed cameras (e.g. Ditria et al., 2020a; Ditria et al., 2021; Lopez-Marcano et al., 2021): (1) Detectron Faster R-CNN (Ren et al., 2015) with a RestNet50 configuration pre-trained using the ImageNet1k dataset (Deng et al., 2009), (2) Detectron2 (Wu et al., 2019) Faster R-CNN (Ren et al., 2015) with ResNet50 and Feature Pyramid Network (FPN) configuration pre-trained using the Common Objects in Context (COCO v2017) dataset (Lin et al., 2014) and a Regional Proposal Network (RPN), and (3) YOLOv5 (Jocher et al., 2020) a single shot detector (SDD) with new CSP-Darknet53 backbone (Wang et al., 2020), Spatial Pyramid Pooling-Fast (SPP) neck, and YOLOv3 head (Redmon and Farhadi, 2018) trained on the COCO dataset (Lin et al., 2014) with pre-trained weights from YOLOv5x. Model training, evaluation and testing were conducted in two cloud platforms: Detectron was run on a Microsoft Azure Data Science virtual machine equipped with an NVIDIA V100 GPU, while Detectron2 and YOLOv5 were ran on a virtual machine provided by the Australian Research Data Commons Nectar Research Cloud equipped with an NVIDIA A100 GPU. All models were developed using Python PyTorch framework (Paszke et al., 2019). We used base model pre-trained weights during model initialization to transfer CNN general purpose detection knowledge to the fish detection domain, an effective and common procedure used in the literature (Ditria et al., 2020a; Zhuang et al., 2020; Saleh et al., 2022a). This procedure also shortens training time in new datasets. The hyper-parameters used for each model can be found online (https://github.com/globalwetlands/fish-on-mobile-cameras). Overfitting was mitigated by: using the early-stopping technique (Prechelt, 2012), performing data augmentation during training, and by assessing model performance (i.e. loss) in an evaluation set (a subset of the training set not used during model training).

TABLE 1 Numbers of annotations for training and testing of models on underwater drone videos (relative abundances of the two species varied between training and testing because of natural fluctuations among survey days).

Annotations	Total	Luderick	Bream
Training	8,282	5,660	2,622
Testing	3,432	3,170	262

We had two strategies for optimizing model performance: adjusting confidence thresholds (CTs) and using a spatio-temporal filter for each target species. The varying CTs change how many predictions of a species occurrence are confirmed in each frame. We varied CTs between 0-95% in 5% increments. For the spatio-temporal filter, we used sequential non-maximum suppression (Seq-NMS), which links detections in neighboring frames (Han et al., 2016). We had Seq-NMS turned on and off, both before and after CT selection, and varied the number of frames for Seq-NMS from 1 to 15 frames.

For each of the three frameworks, we developed one model trained to identify the two species classes. We tested the performance of models using two key metrics of fish abundance, separately for the two species: 1) count per frame (object detection), and 2) MaxN per video snippet. Count per frame was calculated over a total of 1,939 frames in the testing dataset, and MaxN was calculated across 84 video snippets. Two performance criteria, precision and recall, were determined for each confidence threshold and Seq-NMS frame. Precision measures the fraction of fish detections that were correct, and recall measures the fraction of fish actually present that were detected.

$$Precision = \frac{True\ \ Positive}{True\ \ Positive + False\ \ Positive}$$

$$Recall = \frac{True\ \ Positive}{True\ \ Positive + False\ \ Negative}$$

Overall performance for count per image and MaxN was determined by the F1 score, which represents the balance between precision (P) and recall (R):

$$F1 = 2 \times \frac{P \times R}{P + R}$$

At the optimal CT, we proceeded to run two additional replicate models, giving a total of three models generated from the same training dataset. F1 performance could therefore be reported as an average with an estimate of confidence (standard error).

We assessed two factors potentially contributing to prediction errors, fish bounding box blurriness and fish relative size (the size of the bounding box). Blurriness was estimated using a Laplacian-based operator (LAP family as per Pertuz et al., 2013), which assesses the 'amount' of edges using the second derivative of the image. The underlying assumption is that fish in focus present more conspicuous edges than blurred fish. For calculating Laplacian values, we firstly computed the

Sobel operator using a kernel of 3-pixel size convolved over the bounding box, with scale factor of 1. We then applied the Laplacian method to return a value for each pixel within the bounding box. We reported the variance of Laplacian values as our blurriness estimate, with lower Laplacian values indicating high blurriness, and vice versa. Fish relative size was reported directly as pixel area, the product of the bounding box width and height.

Results

The models generally performed well. Detectron2 performed more poorly than other frameworks, whereas YOLOv5 and Detectron performed similarly, sometimes better, sometimes worse than each other, depending on key metrics and species. On a count per frame basis (Table 2), the average overall model F1 scores for the three frameworks ranged from 81.4 - 87.3%, precision 88.2 – 96.0%, and recall 73.2 - 88.2%. Of the two better frameworks, Detectron had fewer False Negatives than YOLOv5 (16% vs 20%), but more False Positives (in all 382 vs 115). The F1 score on a count per frame basis for luderick specifically was > 80% for all frameworks; and lower for bream ranging 35.0 – 73.8%.

Performances based on extraction of MaxN values (Table 3) were solid overall; for the different frameworks, F1 ranged from 85.9 – 91.4%, precision 81.9 – 95.3%, and recall 87.1 – 91.1%. For extraction of MaxN, F1 values were high for luderick (89% or higher) and lower for bream (43.4 - 73.0%). Detectron outperformed both other frameworks for bream, with higher precision *and* recall and thus superior overall performance (Tables 2, 3).

The best performance metrics occurred with Seq-NMS turned on prior to applying confidence thresholds and using three frames. A CT of 40% for both species maximised F1 and balanced precision and recall (shown for Detectron: Figure 2).

Further analysis to determine potential factors contributing to prediction errors, especially false negatives and particularly for bream, revealed that false negatives (FNs) were consistently labelled as such across the models, with 74% of luderick FNs and 67% of bream FNs classified as such in all three models. So these errors are consistent and therefore worth assessing for other causative factors. Blurriness did not differ among classes (TP, FN, FP) for either species (Figure 3). If anything, there was a bias

TABLE 2 Count per frame results.

	Detectron			Detectron2			YOLOv5		
	Overall model	**Luderick**	**Bream**	**Overall model**	**Luderick**	**Bream**	**Overall model**	**Luderick**	**Bream**
F1 score	85.8	86.7	73.8	81.4	80.8	35.0	87.3	87.6	69.3
Precision	88.2	88.5	84.3	91.7	95.7	28.3	96.0	95.8	82.2
Recall	83.5	85	65.6	73.2	69.9	45.8	80.0	80.7	59.9
Ground-truths (GT)	3432	3170	262	3432	3170	262	3432	3170	262
True positives	2867	2695	172	2512	2215	120	2745	2557	157
False negatives (FN)	565	475	90	920	955	142	687	613	105
FN proportion of GT	0.16	0.15	0.34	0.27	0.30	0.54	0.20	0.19	0.40
False positives	382	350	35	227	100	304	115	112	34

Overall and per species model performance for three CNN frameworks on the fish count per frame task (confidence thresholds of 40% for both species).

TABLE 3 MaxN per video results.

	Detectron			Detectron2			YOLOv5		
	Overall model	**Luderick**	**Bream**	**Overall model**	**Luderick**	**Bream**	**Overall model**	**Luderick**	**Bream**
F1 score	90	93.4	73	85.9	87.1	43.4	91.4	90.3	65.3
Precision	95.3	98.2	78.9	81.88	89.38	29.03	91.8	90.7	57.1
Recall	87.1	89.9	71.4	90.37	84.87	85.71	91.1	89.9	76.2
Ground-truths (GT)	135	119	21	135	119	21	135	119	21
True positives	122	107	15	122	101	18	123	107	16
False negatives (FN)	18	12	6	13	18	3	12	12	5
FN proportion of GT	0.11	0.1	0.28	0.10	0.15	0.14	0.09	0.10	0.24
False positives	6	2	4	27	12	44	11	11	12

Overall and per species model performance for three CNN frameworks on the MaxN fish per video task (confidence thresholds of 40% for both species).

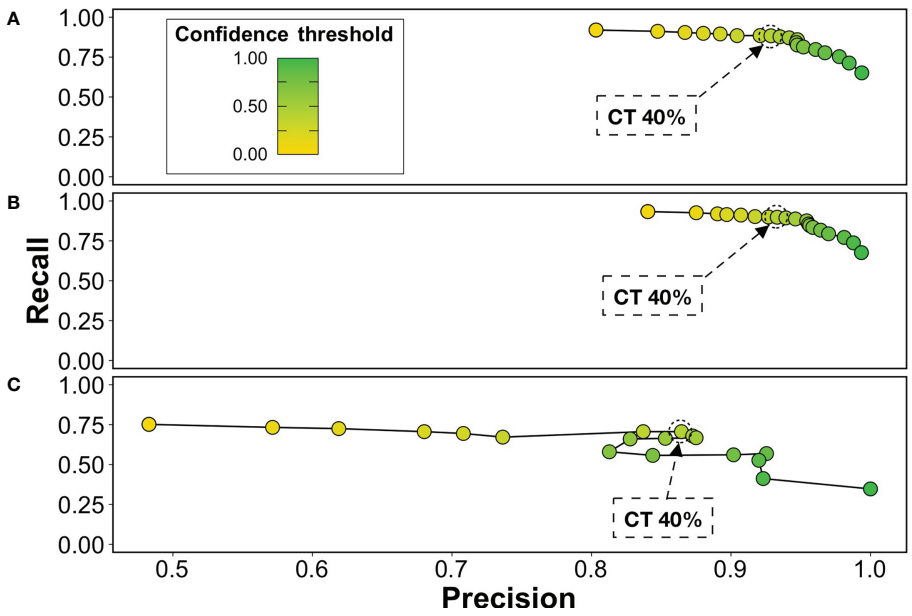

FIGURE 2
Precision and recall values for count per frame results for **(A)** overall model, **(B)** luderick and **(C)** bream, shown for Detectron as example. Confidence intervals are in 5% increments; a confidence threshold of 40% is highlighted as resulting in the best F1 score, for performance as reported in Table 2.

towards lower blurriness in FNs; although this was very slight, it shows clearly that the FNs were not the result of high blurriness. Fish relative size (bounding box area) did differ among classes, with FNs for bream (but not luderick) having a preponderance of smaller areas (Figure 4). Frequencies of image sizes for bream in training and evaluation datasets were similar although there were slightly fewer small images in training (Figure 5).

Discussion

The performance of the deep learning models was suitable for multi-species detection and counting of fish in mobile underwater videos. For the two better performing frameworks, Detectron and YOLOv5, overall F1 values above 85% for object detection (counts per frame) and above 90.0% for MaxN values are comparable with

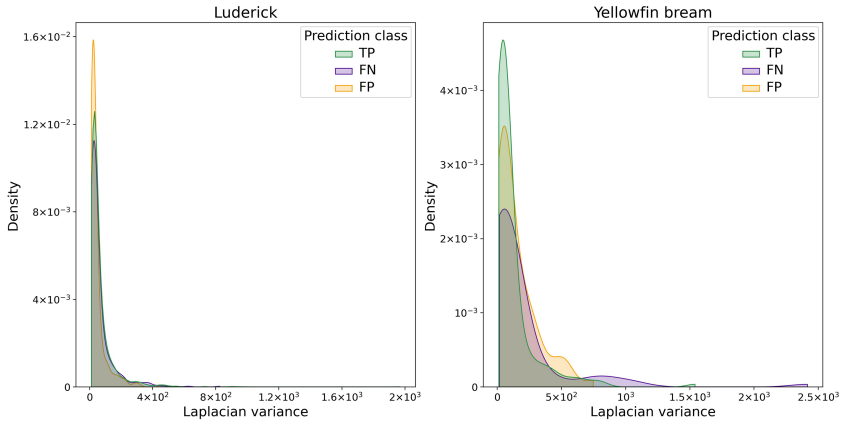

FIGURE 3
Distribution of blurriness values for three classes of predictions (True Positives, False Negatives, False Positives), for luderick and bream, shown for Detectron model. No substantial differences among classes are evident. Low Laplacian variance values are indicative of high blurriness. Results are for one of the three Detectron models, other model results were very similar.

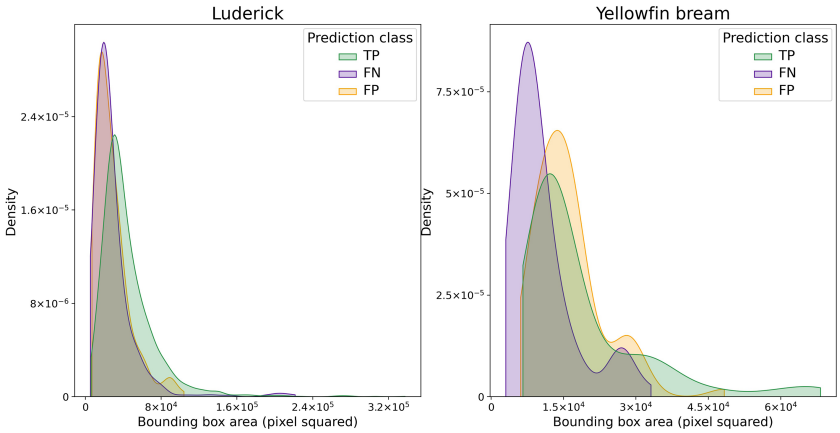

FIGURE 4
Distribution of fish relative sizes (bounding box areas in pixels) for three classes of predictions (True Positives, False Negatives, False Positives), for luderick and bream. No substantial differences among classes are evident for luderick but for bream image size tends to be smaller for FNs). Results are for one of the three models using the Detectron framework, other model results were very similar.

those using deep learning on videos from stationary cameras. CNN models in recent multispecies analyses of videos from underwater stationary cameras have produced overall performances (as % of object detection correct, similar to F1) of: 86.9% average on 18 fish species (Villon et al., 2018), and 78.0% on 20 species (with values for individual species ranging from 63 – 99%, Villon et al., 2021). Single species CNN models for stationary cameras in local waters near the current survey location have produced a range of F1 values in object detection analyses: 87.6% and 92.3% over seagrass habitat (Ditria et al., 2020a; Ditria et al., 2020b); 83.0% and 90.6% over reef (Ditria et al., 2020b; Lopez-Marcano et al., 2022), albeit with much lower values where training did not include the habitat over which test videos were filmed (e.g. 58.4% and 73.3%, Ditria et al., 2020b).

Model performance on individual fish species varied. For luderick, counts per frame showed good accuracy for the two

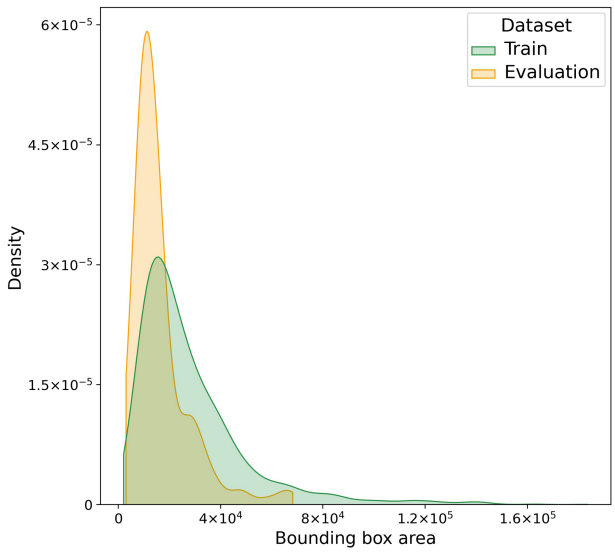

FIGURE 5
Distribution of relative fish sizes (bounding box areas in pixels) in the training and evaluation datasets for bream. Size distributions were similar, although with slightly fewer small images used in training.

better performing models (F1 approx. 87% for both), and for MaxN excellent accuracy (93.4% for Detectron, 90.3% for YOLOv5). This level of accuracy for MaxN extraction for luderick more than matches recent evaluations from stationary cameras ranging from 65.0 – 91.5% (F1 values from Ditria et al., 2020b). We therefore conclude that for this species, automated analysis of videos from mobile cameras is suitable for routine use in surveys. This finding is of immediate relevance given current interest in the significant grazing impacts of luderick in seagrass habitat (Wendländer et al., 2020), and early indications of a range shift for this species under climate change (Pollock, 2017).

Model accuracy for bream was lower than for luderick. Even for the two better performing frameworks, accuracy for counts per frame was below 80% (73.8% for Detectron, 69.3% for YOLOv5, as well as for MaxN (73% for Detectron, 65.3% for YOLOv5). Object detection performance on stationary cameras for this species has had a wide range of outcomes in recent studies, namely: 91% F1 value in a single species model (Lopez-Marcano et al., 2021), but also 75.0% in a three species model (Lopez-Marcano et al., 2022). For bream, we are unable to clearly state whether automation algorithms are performing differently on mobile than on stationary cameras. However there does seem room for substantial improvement in deep learning models for bream, a species that is abundant and important for estuarine and marine ecosystem function in local estuaries (Gilby et al., 2017; Olds et al., 2018; Henderson et al., 2020).Given that juveniles and adults share the same habitats, training datasets become too variable where there are significant intra-specific variabilities in sizes, colorations and behaviors. Employing adaptive deep learning frameworks that account for this large variability may help improve bream models (Qiu et al., 2019).

Mobile cameras can produce some level of additional blurring due to ego-motion, which results in more rapid changes in object perspective and scale (Chuang et al., 2017). To some extent, this effect could be prevented by using higher frame rates. Errors in model predictions do not appear to be related to the movement of the camera. Blurriness values were much the same for true and false predictions. Errors were more prevalent for smaller fish images, either for small fish or fish further from the camera. This could be an issue for mobile cameras if the movement of the drone, vehicle, or diver on which the camera is mounted scare fish to swim away. Several attraction and repulsion effects have been documented during fish surveys using ROVs (e.g., Stoner et al., 2008; Baker et al., 2012; Laidig et al., 2013). The source of bias and behavioral effect on fish taxa are not universal, but standard surveying procedures can minimize sampling variation (Stoner et al., 2008; Sward et al., 2019). Future work can experiment with the speed of movement of the camera, a factor that has been insufficiently assessed in the literature (Sward et al., 2019), potentially balancing efficiency of surveying large areas of the seabed against the proportion of images that are blurred. Notwithstanding the need for further testing and fine-tuning, the object detection method employed here seems effective for mobile cameras, and avoids issues with having to 'know the background' and subtract it prior to conventional deep learning for object detection (Wei et al., 2021).

Other challenges for automated fish identification and counting in videos from mobile cameras are the same as for videos from stationary cameras. Detection of crypto-benthic fish can be unreliable, for example, either when using ROVs on offshore industrial platforms (Andaloro et al., 2013), or when using stationary cameras (Sheaves et al., 2020). However mobile cameras probably work well for pelagic species, but cannot avoid the same challenge faced with stationary cameras of repeatedly counting the same individuals of schooling species.

Another challenge for developing deep learning models to underwater environments and fish detection is that datasets are often small relative to those for common terrestrial topics (Jin & Liang, 2017; Saleh et al., 2022b). Overfitting is therefore difficult to avoid because just a small sample from the population of fish and underwater conditions is used during model training and optimization. In our study, keeping imagery from one of the three days for testing, independently of videos from other days used for training, was the best way to reduce overfitting. Nevertheless, given the modest size of our dataset, with imagery from consecutive days, and from a single estuary, the variability within and among training and testing datasets is less than would be expected if other estuaries and times of year were included. Our intention is not to provide a model to be used out-the-box, in fact our model would likely underperform in a novel dataset. To our knowledge, there is not a publicly available dataset from mobile cameras for fish detections that we could use for testing, thus restricting us from providing an unbiased estimate of performance. We are making our dataset public so others can use it as part of the training and optimization process, or for an unbiased assessments of model performance. As fish monitoring is anticipated to be enhanced by new technologies including remotely operated and autonomous underwater vehicles, upcoming fish detections models would likely have access to a larger bank of fish images captured from mobile cameras. Furthermore, aside from ego-motion and behavioral effects from mobile cameras on the quality of fish images, which can create a distinctive type of dataset, fish detection models for mobile cameras could take advantage of images and annotations captured from stationary ones. Further research, using a mix of images from stationary and mobile cameras for fish detection model is needed to elucidate if high prediction performance for both types of images can be accomplished, and/or if low-dimensional features in those types of images are similar.

In summary, we find that existing deep learning procedures developed for stationary cameras can be used to reliably identify and count fish in videos from mobile cameras. Extraction of MaxN values, the most commonly used indicator of fish abundances (Harvey et al., 2021), were automated for target

species with some confidence in the current study. There is variability in performance among species, even for the two target species in the current study, and further testing is warranted on mobile cameras used in surveys of other species and habitats. Cameras on underwater drones and remotely operated and automated vehicles offer advantages over traditional UVC, for example avoiding observer bias (Sheaves et al., 2020; Ditria et al., 2022). Video surveys can also circumvent the weakness of UVCs in situations with high fish abundances, particularly when automation is used in conjunction with additional post-processing mathematical solutions to increase the rigor of automated fish counting (Connolly et al., 2021). We encourage analysis of imagery from mobile cameras to be included in ongoing refinement of deep learning procedures for automated fish counts in underwater videos.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

Ethics statement

The animal study was reviewed and approved by Griffith University Animal Ethics Committee.

Author contributions

Conceptualization: all authors. Methodology: all authors led by RC. Interpretation of results: all authors. Writing: all authors led by RC. Funding and resources: RC, SL-M. All authors contributed to the article and approved the submitted version.

Funding

The work was supported by the Australian Research Data Commons (ARDC), the Microsoft AI for Earth program, and the Global Wetland Project (GLOW), which draws support from a philanthropic trust that neither seeks nor permits publicity.

Acknowledgments

We greatly appreciate assistance in the field and laboratory from M. Kitchingman, C. McAneney, and A. Shand, and with software support from E. Jinks. The paper benefitted from discussions with staff and students in the Global Wetland Project.

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.

Publisher's note

References

Andaloro, F., Ferraro, M., Mostarda, E., Romeo, T., and Consoli, P. (2013). Assessing the suitability of a remotely operated vehicle (ROV) to study the fish community associated with offshore gas platforms in the Ionian Sea: a comparative analysis with underwater visual censuses (UVCs). *Helgol. Mar. Res.* 67, 241–250. doi: 10.1007/s10152-012-0319-y

Arain, B., Mccool, C., Rigby, P., Cagara, D., and Dunbabin, M. (2019). "Improving underwater obstacle detection using semantic image segmentation," in *2019 International Conference on Robotics and Automation (ICRA): IEEE)* (Montreal, Canada: IEEE). 9271–9277.

Baker, K. D., Haedrich, R. L., Snelgrove, P. V., Wareham, V. E., Edinger, E. N., and Gilkinson, K. D. (2012). Small-scale patterns of deep-sea fish distributions and assemblages of the grand banks, Newfoundland continental slope. *Deep Sea Res. Part I: Oceanographic Res. Papers* 65, 171–188. doi: 10.1016/j.dsr.2012.03.012

Bernard, A. T. F., Götz, A., Kerwath, S. E., and Wilke, C. G. (2013). Observer bias and detection probability in underwater visual census of fish assemblages measured with independent double-observers. *J. Exp. Mar. Biol. Ecol.* 443, 75–84. doi: 10.1016/j.jembe.2013.02.039

Chuang, M.-C., Hwang, J.-N., Ye, J.-H., Huang, S.-C., and Williams, K. (2017). Underwater fish tracking for moving cameras based on deformable multiple kernels. *IEEE Trans. Syst. Man Cybern.* 47, 2467–2477. doi: 10.1109/tsmc.2016.2523943

Connolly, R. M., Fairclough, D. V., Jinks, E. L., Ditria, E. M., Jackson, G., Lopez-Marcano, S., et al. (2021). Improved accuracy for automated counting of a fish in baited underwater videos for stock assessment. *Front. Mar. Sci.* 8. doi: 10.3389/fmars.2021.658135

Coro, G., and Walsh, M. B. (2021). An intelligent and cost-effective remote underwater video device for fish size monitoring. *Ecol. Inform.* 63, 101311. doi: 10.1016/j.ecoinf.2021.101311

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 - 2015 IEEE Winter Conference on Applications of Computer Vision Workshops, WACVW 2015)* (Washington DC, USA: IEEE). 57–62.

Dawkins, M., Sherrill, L., Fieldhouse, K., Hoogs, A., Richards, B., Zhang, D., et al. (2017). "An open-source platform for underwater image and video analytics," in

2017 IEEE Winter Conference on Applications of Computer Vision (WACV): IEEE) (Washington DC, USA: IEEE). 898–906.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., and Fei-Fei, L. (2009). "Imagenet: A large-scale hierarchical image database," in *IEEE conference on computer vision and pattern recognition* (Miami, USA: IEEE). 248–255.

Diamantas, S., and Alexis, K. (2020). "Optical flow based background subtraction with a moving camera: application to autonomous driving," in *International symposium on visual computing*. Eds. G. Bebis, Z. Yin, E. Kim, J. Bender, K. Subr, B. C. Kwon, J. Zhao, D. Kalkofen and G. Baciu (Cham: Springer), 398–409.

Ditria, E. M., Buelow, C. A., Gonzalez-Rivero, M., and Connolly, R. M. (2022). Artificial intelligence and automated monitoring for conservation of marine ecosystems: a perspective. *Front. Mar. Sci* 9, 918104. doi: 10.3389/fmars.2022.918104

Ditria, E. M., Jinks, E. L., and Connolly, R. M. (2021). Automating the analysis of fish grazing behaviour from videos using image classification and optical flow. *Anim. Behav.* 177, 31–37. doi: 10.1016/j.anbehav.2021.04.018

Ditria, E. M., Lopez-Marcano, S., Sievers, M., Jinks, E. L., Brown, C. J., and Connolly, R. M. (2020a). Automating the analysis of fish abundance using object detection: optimizing animal ecology with deep learning. *Front. Mar. Sci.* 7. doi: 10.3389/fmars.2020.00429

Ditria, E. M., Sievers, M., Lopez-Marcano, S., Jinks, E. L., and Connolly, R. M. (2020b). Deep learning for automated analysis of fish abundance: the benefits of training across multiple habitats. *Environ. Monit. Assess.* 192, 698. doi: 10.1007/s10661-020-08653-z

Ferguson, A. M., Harvey, E. S., Taylor, M. D., and Knott, N. A. (2013). A herbivore knows its patch: luderick, *Girella tricuspidata*, exhibit strong site fidelity on shallow subtidal reefs in a temperate marine park. *PloS One* 8, e65838. doi: 10.1371/journal.pone.0065838

Francisco, F. A., Nuhrenberg, P., and Jordan, A. (2020). High-resolution, non-invasive animal tracking and reconstruction of local environment in aquatic ecosystems. *Mov. Ecol.* 8, 27. doi: 10.1186/s40462-020-00214-w

Garner, S. B., Olsen, A. M., Caillouet, R., Campbell, M. D., and Patterson, W. F. (2021). Estimating reef fish size distributions with a mini remotely operated vehicle-integrated stereo camera system. *PloS One* 16, e0247985. doi: 10.1371/journal.pone.0247985

Gilby, B. L., Olds, A. D., Yabsley, N. A., Connolly, R. M., Maxwell, P. S., and Schlacher, T. A. (2017). Enhancing the performance of marine reserves in estuaries: Just add water. *Biol. Conserv.* 210, 1–7. doi: 10.1016/j.biocon.2017.03.027

Goetze, J. S., Jupiter, S. D., Langlois, T. J., Wilson, S. K., Harvey, E. S., Bond, T., et al. (2015). Diver operated video most accurately detects the impacts of fishing within periodically harvested closures. *J. Exp. Mar. Biol. Ecol.* 462, 74–82. doi: 10.1016/j.jembe.2014.10.004

Han, W., Khorrami, P., Paine, T. L., Ramachandran, P., Babaeizadeh, M., Shi, H, et al. (2016). Seq-nms for video object detection. *arXiv preprint* 1–9. doi: 10.48550/arXiv.1602.08465

Harvey, E. S., Mclean, D. L., Goetze, J. S., Saunders, B. J., Langlois, T. J., Monk, J., et al. (2021). The BRUVs workshop-an Australia-wide synthesis of baited remote underwater video data to answer broad-scale ecological questions about fish, sharks and rays. *Mar. Policy* 127, 104430. doi: 10.1016/j.marpol.2021.104430

Henderson, C. J., Gilby, B. L., Schlacher, T. A., Connolly, R. M., Sheaves, M., Maxwell, P. S., et al. (2020). Low redundancy and complementarity shape ecosystem functioning in a low-diversity ecosystem. *J. Anim. Ecol.* 89, 784–794. doi: 10.1111/1365-2656.13148

Heo, B., Yun, K., and Choi, J. Y. (2017). "Appearance and motion based deep learning architecture for moving object detection in moving camera," in *2017 IEEE International Conference on Image Processing (ICIP): IEEE)* (Washington DC, USA: IEEE) 1827–1831.

Islam, M. J., Edge, C., Xiao, Y., Luo, P., Mehtaz, M., Morse, C., et al. (2020). "Semantic segmentation of underwater imagery: Dataset and benchmark," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS): IEEE)* (Washington DC, USA: IEEE).

Jalal, A., Salman, A., Mian, A., Shortis, M., and Shafait, F. (2020). Fish detection and species classification in underwater environments using deep learning with temporal information. *Ecol. Inform.* 57, 101088. doi: 10.1016/j.ecoinf.2020.101088

Jin, L., and Liang, H. (2017). "Deep learning for underwater image recognition in small sample size situations," in *Aberdeen, UK: IEEE* (Washington DC, USA: IEEE), 1–4. doi: 10.1109/OCEANSE.2017.8084645

Jocher, G., Stoken, A., and Borovec, J. (2020). *ultralytics/Yolov5: v3.1 - bug fixes and performance improvements* (Berne, Switzerland: Zenodo). doi: 10.5281/zenodo.4154370

Katija, K., Roberts, P. L. D., Daniels, J., Lapides, A., Barnard, K., Risi, M., et al. (2021). "Visual tracking of deepwater animals using machine learning-controlled robotic underwater vehicles," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA: IEEE (Washington DC, USA: IEEE). doi: 10.1109/wacv48630.2021.00090

Kim, D., Woo, S., Lee, J.-Y., and Kweon, I. S. (2020). "Video panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Washington DC, USA: IEEE). 9859–9868.

Kirillov, A., He, K., Girshick, R., Rother, C., and Dollar, P. (2019). "Panoptic segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Washington DC, USA: IEEE). doi: 10.1109/CVPR42600.2020.00988

Laidig, T. E., Krigsman, L. M., and Yoklavich, M. M. (2013). Reactions of fishes to two underwater survey tools, a manned submersible and a remotely operated vehicle. *Fish. Bull.* 111, 54–67. doi: 10.7755/Fb.111.1.5

Langlois, T., Goetze, J., Bond, T., Monk, J., Abesamis, R. A., Asher, J., et al. (2020). A field and video annotation guide for baited remote underwater stereo-video surveys of demersal fish assemblages. *Methods Ecol. Evol.* 11, 1401–1409. doi: 10.1111/2041-210x.13470

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014) in *European conference on computer vision*. 740–755 (Cham: Springer).

Liu, F. F., and Fang, M. (2020). Semantic segmentation of underwater images based on improved deeplab. *J. Mar. Sci. Eng.* 8, 188. doi: 10.3390/jmse8030188

Lopez-Marcano, S., E, L. J., Buelow, C. A., Brown, C. J., Wang, D., Kusy, B., et al. (2021). Automatic detection of fish and tracking of movement for ecology. *Ecol. Evol.* 11, 8254–8263. doi: 10.1002/ece3.7656

Lopez-Marcano, S., Turschwell, M. P., Brown, C. J., Jinks, E. L., Wang, D., and Connolly, R. M. (2022). Computer vision reveals fish behaviour through structural equation modelling of movement patterns *J Res. Square* 1–24 . doi: 10.21203/rs.3.rs-1371027/v1

Mandal, R., Connolly, R. M., Schlacher, T. A., and Stantic, B. (2018). "Assessing fish abundance from underwater video using deep neural networks," in *2018 International Joint Conference on Neural Networks (IJCNN): IEEE)* (Washington DC, USA: IEEE) 1–6.

O'Byrne, M., Pakrashi, V., Schoefs, F., and Ghosh, B. (2018). Semantic segmentation of underwater imagery using deep networks trained on synthetic imagery. *J. Mar. Sci. Eng.* 6, 93. doi: 10.3390/jmse6030093

Olds, A. D., Frohloff, B. A., Gilby, B. L., Connolly, R. M., Yabsley, N. A., Maxwell, P. S., et al. (2018). Urbanisation supplements ecosystem functioning in disturbed estuaries. *Ecography* 41, 2104–2113. doi: 10.1111/ecog.03551

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* 32, 1–12. doi: 10.48550/arXiv.1912.01703

Pertuz, S., Puig, D., and Garcia, M. A. (2013). Analysis of focus measure operators for shape-from-focus. *Pattern. Recogn.* 46, 1415–1432. doi: 10.1016/j.patcog.2012.11.011

Pollock, B. R. (2017). Latitudinal change in the distribution of luderick *Girella tricuspidata* (Pisces: Girellidae) associated with increasing coastal water temperature in eastern Australia. *Mar. Freshw. Res.* 68, 1187–1192. doi: 10.1071/Mf16070

Prechelt, L. (2012). "Early stopping-but when?," in *Neural networks: Tricks of the trade*. Eds. G. Montavon, G. Orr and K. Müller (Berlin: Springer) 53–67.

Qiu, H., Li, H., Wu, Q., Meng, F., Ngan, K. N., and Shi, H. (2019). A2RMNet: Adaptively aspect ratio multi-scale network for object detection in remote sensing images. *Remote Sens* 11, 1594. doi: 10.3390/rs11131594

Redmon, J., and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv* 1–6. doi: 10.48550/arXiv.1804.02767

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* 28, 1–14. doi: 10.48550/arXiv.1506.01497

Saleh, A., Sheaves, M., Jerry, D., and Azghadi, M. R. (2022b) *Applications of deep learning in fish habitat monitoring: A tutorial and survey*. Available at: http://arxiv.org/abs/2206.05394.

Saleh, A., Sheaves, M., and Rahimi Azghadi, M. (2022a). Computer vision and deep learning for fish classification in underwater habitats: A survey. *Fish Fisheries* 23, 977–999. doi: 10.1111/faf.12666

Salman, A., Siddiqui, S. A., Shafait, F., Mian, A., Shortis, M. R., Khurshid, K., et al. (2020). Automatic fish detection in underwater videos by a deep neural network-based hybrid motion learning system. *ICES J. Mar. Sci.* 77, 1295–1307. doi: 10.1093/icesjms/fsz025

Sheaves, M., Bradley, M., Herrera, C., Mattone, C., Lennard, C., Sheaves, J., et al. (2020). Optimizing video sampling for juvenile fish surveys: Using deep learning and evaluation of assumptions to produce critical fisheries parameters. *Fish. Fish.* 21, 1259–1276. doi: 10.1111/faf.12501

Sieberth, T., Wackrow, R., and Chandler, J. H. (2013). Automatic isolation of blurred images from UAV image sequences. *Int. Arch. Photogramm. Remote Sens* XL-1/W2, 361–366. doi: 10.5194/isprsarchives-XL-1-W2-361-2013

Stoner, A. W., Ryer, C. H., Parker, S. J., Auster, P. J., and Wakefield, W. W. (2008). Evaluating the role of fish behavior in surveys conducted with underwater vehicles. *Can. J. Fish. Aquat. Sci.* 65, 1230–1243. doi: 10.1139/F08-032

Sward, D., Monk, J., and Barrett, N. (2019). A systematic review of remotely operated vehicle surveys for visually assessing fish assemblages. *Front. Mar. Sci.* 6. doi: 10.3389/fmars.2019.00134

Villon, S., Iovan, C., Mangeas, M., Claverie, T., Mouillot, D., Villéger, S., et al. (2021). Automatic underwater fish species classification with limited data using few-shot learning. *Ecol. Inform.* 63, 101320. doi: 10.1016/j.ecoinf.2021.101320

Villon, S., Mouillot, D., Chaumont, M., Darling, E. S., Subsol, G., Claverie, T., et al. (2018). And fast identification of coral reef fishes in underwater images. *Ecol. Inform.* 48, 238–244. doi: 10.1016/j.ecoinf.2018.09.007

Walther, D., Edgington, D. R., and Koch, C. (2004). "Detection and tracking of objects in underwater video," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004: IEEE)* (Washington DC, USA: IEEE).

Wang, C. Y., Liao, H. Y., Wu, Y. H., Chen, P. Y., Hsieh, J. W., and Yeh, I. H. (2020). "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (Washington DC, USA: IEEE) 390–391.

Wei, J., Jiang, J., and Yilmaz, A. (2021). MBS-net: A moving-camera background subtraction network for autonomous driving. *Int. Arch. Photogramm. Remote Sens* 43, 69–76. doi: 10.5194/isprs-archives-XLIII-B2-2021-69-2021

Wendländer, N. S., Lange, T., Connolly, R. M., Kristensen, E., Pearson, R. M., Valdemarsen, T., et al. (2020). Assessing methods for restoring seagrass (*Zostera muelleri*) in australia's subtropical waters. *Mar. Freshw. Res.* 71, 996–1005. doi: 10.1071/MF19237

Wu, Y., Kirillov, A., Massa, F., and Lo W.Y and Girshick, R. (2019) *Detectron2*. Available at: https://github.com/facebookresearch/detectron2.

Zhuang, P., Wang, Y., and Qiao, Y. (2020). "Wildfish++: A comprehensive fish benchmark for multimedia research," in *IEEE Transactions on Multimedia*, (Washington DC, USA: IEEE). 23. 3603–3617. doi: 10.1109/TMM.2020.3028482