UNIVERSITÉ CÔTE D'AZUR

Ifremer

Example imagery of the input data used in this project

Fish-bait-efficiency and benthic stock assessments using deep learning

Kilian Bürgi

kilian.burgi@etu.univ-cotedazur.fr

Mémoire de MSc. MARRES

Université Côte d'Azur

Soutenu le 12 décembre 2021

Directeur de stage

Robin Faillettaz, Ifremer HALGO, Lorient

robin.faillettaz@ifremer.fr

Acknowledgement

I would like to thank my supervisor Dr. Robin Faillettaz for helping and supporting me during the time of the project and always being there when I had questions – as annoying as they might have been. A special thanks goes out to IFREMER Lorient – especially Sonia Méhault, Julien Simon and Dorothé Kopp for providing me with expertise, data, publications, and invaluable knowledge. I would also like to thank Matthew Dawkins the software developer of VIAME with his relentless support for this project.

Abstract

The rise of new technology is continuously generating gigantic datasets, known as “big data”. Imagery to tackle biological and ecological questions, is no exception. Observing and learning from these data is crucial but remains a tedious and labour-intensive process. This project aims to address to what extent computer vision based on deep learning can solve ecological questions while minimizing – if not removing human validation. For this purpose, a convolutional neural network (CNN) was trained on two types of data representing different sampling conditions and species communities. The first aimed at detecting the attraction levels of different types of biodegradable baits using baited remote underwater videos (BRUV). The BRUV footage analysis showed promising results with an average precision (AP), the standard metrics to assess the performance of deep learning models, of 0.827 for fish for the best performing model. An Interest index was introduced to assess each of the different bait types and a cockle bait functioned as the control. The resulting analysis – manual and automated - showed that the biodegradable plastic bait C17 has the greatest potential of replacing an old-fashioned cockle’s bait. The UWTV footage had more diverse classes (17 species, genus, or other taxa) and showed more mitigated results. The fish *Callionymus spp.*, the crustacean *Munida spp.* and the Pennatulaceidae classes were accurately detected with AP values of 0.86, 0.82 and 0.80 respectively. In comparison, the main focus class *Nephrops norvegicus* slightly underperformed, with an AP value of 0.69. Other classes were more difficult to identify as such as “hydrozoa” and “crustacean” (AP of 0.23 and 0.24), due to their high diversity of shapes, colours and sizes. Nevertheless, in regard to other studies and given the challenging nature of marine-derived data, these values are satisfying. This project highlights the promising potential of replacing the labour-intensive human-validated analysis, while identifying the gaps that still need to be overcome. The generated models will help moving toward non-invasive methods with direct applications in marine conservation and fisheries management.

Keywords: Deep learning, BRUV, UWTV, ecology, artificial intelligence, analysis

Table of contents

1. General Introduction

I. Ecological background

Obtaining robust observations has always been a key aspect of biological studies, especially in the field of ecology, to assess behavioural patterns (Underwood *et al.*, 2000). Back in 1831 to 1836, Charles Darwin sailed the HMS Beagle to the Galapagos archipelago to observe the native finks on the different islands and their differences in beak sizes and forms due to different environmental surroundings. His observations heavily contributed in many parts to the theory of evolution widely accepted today. This is a rather simple example compared to the amount and accuracy of recent observations, for example in (Faillettaz *et al.*, 2015) in which simple observations would not have led to the observation of sun-orientation of different fish larvae's, which enabled to decipher more complex theories and hypotheses. In some cases, one observer is not enough, and multiple scientists look at the same data to verify the results. This is labour and time intensive, time which could be spent in aspects of a project not able to be conducted by a computer such as out-of-the-ordinary scenarios or the discussion of the results. Furthermore, there is a potential to be biased towards the proposed hypothesis. Manual validation is more prove to human biases (*i.e.* experience, fatigue, etc.) compared to computer vision which is consistent and more neutral. This could generate unreliable assessments and lead to results that are influenced by one's personal experience. This can lead to under or over estimations of populations in a conservatory light or lead to wrong results which - in this project - can have negative ecological and societal consequences. That is why a need for a reliable and universally applicable tool emerged. This project aims to participate in determining to what extent the current state of the art artificial intelligence could answer ecologically relevant questions.

II. Artificial intelligence & deep learning

In an era with rapidly changing technology with better, stronger and faster computers there is a demand for a universally applicable, low cost tool to analyse vast growing amounts of digital data such as video or photo captures (Parida, 2018). The process of manual analysis is time consuming and the need to speed up this crucial part of the analysis by letting super-computers doing the job is the next step in the history of data assessment, with the help of artificial intelligence (AI). AI is the attempt to project the human process of thinking into a machine or computer (Nilsson, 2009). This task is highly demanding since the human brain consists of a large number of neurons. A total of $2.28 * 10^6$ neurons are approximately found in a human brain (Pakkenberg & Gundersen, 1988) and to reach an efficiency of these neurons. Currently impossible - this exactly is the ultimate goal of AI.

Within the field of AI, machine learning is a well-developed discipline. Machine learning is the training of a computer as a tool for detection of a target based on features extracted from the input data such as shapes, colours, or other statistical values. For example in the study of (Dezecache *et al.*, 2020) the distress calls of infant chimpanzees were processed with the help of a machine learning approach. Recorded calls were labelled as stress calls or not (stressed or not-stressed) depending on a feature manually extracted during the step of labelling. The algorithm was then able to distinguish between the different types of calls and was able to predict if future calls are assigned to the stressed or not-stressed category.

A subgenre or extension of machine learning is the so called deep learning (Miele *et al.*, 2021). Deep learning takes this approach one-step further since it automates the step of feature extraction, which minimizes human interference during this step. In machine and deep learning, depending on the amount of data available, a portion of the dataset is taken to be shown to the computer as a training and is considered the training dataset. The standard split is 70% of the data used as the training dataset

(Huang *et al*., 2019a; Villon *et al*., 2018) but this can vary. If there is more data available, only 10% can be sufficient to train the algorithm and on the other hand in small datasets up to 90% are used for the training (Ovchinnikova *et al.*, 2021). The volume of the training dataset is an important factor on how the performance of the model will change (Joulin *et al.*, 2016; Sun *et al.*, 2017). Through creation of different layers of features, called a convolutional neural network CNN (**Fig. 1** - Kroodsma *et al.*, 2018), which function as a decisional tree, the identified objects are categorized into predefined or non-predefined categories. When the categories are not given, the deep learning approach is unsupervised, if given it is called supervised. In this report, the discussed process is supervised, since we provide the names of the classes.

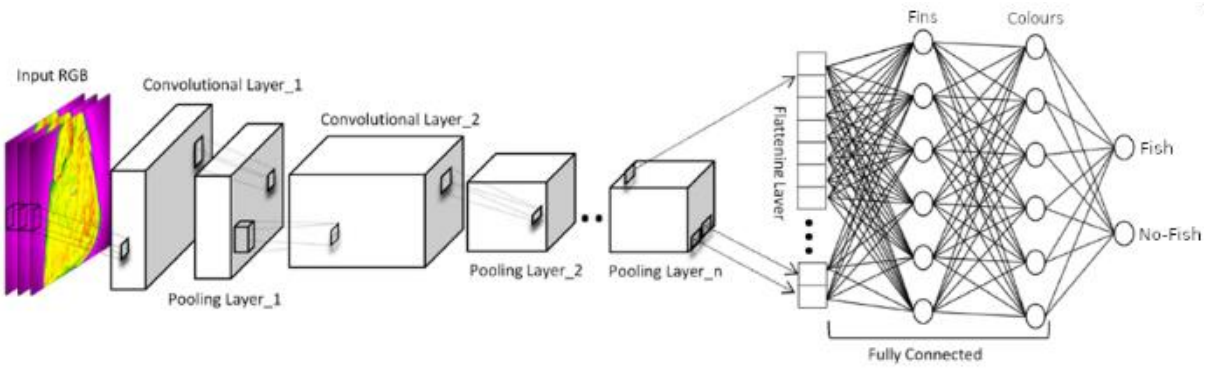

Figure 1, adapted from: Kroodsma et al., 2018. Creation of the CNN. *From left to right: input of an image (colours RGB) and the construction of the different feature-layers (white dots) which resemble certain features extracted from the image. It is referred as "convolutional neural network" due to the creation of convolutional layers that break down the picture into smaller layers to increase accuracy.*

This trained algorithm can then be used to evaluate the remaining data, called the test data, which was not seen by the algorithm, to verify if the training was successful and the algorithm could work and be applied. Whilst nearly completely being freed from human interaction, deep learning is heavily dependent on how good, in terms of diversity and quality, the input data is and requires heavy computational efforts (Panda *et al.*, 2016).

Over the past few years, the development of deep learning has shown massive improvements in scientific fields such as engineering (Krishna Chaitanya & Maragatham, 2021) or the medical field (Maier, 2019). But there are not only applications in these fields but also in ecological fields (Christin *et al.*, 2019, Schofield *et al.*, 2019). Schofield et al, 2019 allowed the recognition of facial structures of wild living chimpanzees by a deep learning algorithm and use this technique to automatically monitor a population of chimpanzees with a fully non-invasive strategy.

The marine environment has been given attention in the past decade too (Xu & Matzner, 2018). However, acquiring of reliable data is more challenging in marine environments (and aquatic environments in general) due to lower light availability, differences in lighting, blurriness and turbulence due to the always moving environment (Sun *et al.*, 2017). Furthermore, the acquisition of these types of data is costly and labour-intensive. Costs that can or will not be spent which lead to training datasets that are not satisfactory. Sun *et al*., 2017 states that no matter how good your model is, if your training dataset is insufficient the results are not satisfying.

A study in 2019 showed the potential of deep learning detecting three genera of marine bottom living animals – sea cucumbers, sea urchins and scallops (Huang *et al.,* 2019a). Average precision numbers averaging at 59% show that 3 out of 5 animals were correctly located and classified. Best detected were sea cucumbers with an average precision value of 0.7979, slightly worse are the scallops detected

with an average precision of 0.6339 and worst detected are sea urchin with an average precision of 0.4682. This shows the prospect of deep learning in aquatic and marine environments. Deep learning has thus shown to be efficient with certain classes of immobile species. Here, we want to determine the potential of deep learning regarding motile organisms, fast-moving fishes or slow-moving organisms facing a moving camera.

III. Short project descriptions

Two different types of data were evaluated in this project, with different recording techniques, species of interest and environments (benthic & demersal) but both having the same aim to assess to what extent the human interference can be lowered or avoided. The two different projects are evaluated and explained in their specific section of the report.

2. General Material & Methods

Project-specific methods are details in each project sections (Project LangolfTV and Project BAITFISH). Yet, the parts that are common to both projects (LangolfTV and BAITFISH) are described below only once to avoid repetitions (*i.e.* 1. Hardware resources, 2. Software, 3. Annotation procedure and 4. model evaluation).

I. Computer, Graphics Processing Unit (GPU) & Operating System (OS)

For the training and object detection a DELL computer running Ubuntu 18.04.5 LTS 64-Bit with 62 GB of RAM, an Intel Xeon Silver 4114 CPU (central processing unit). To bear the graphical and computational effort an NVIDIA® GeForce RTX 2080 Ti (11 GB of GDDR6 memory) Graphics Processing Unit (GPU) was used.

II. VIAME & CFRNN

Most of the work presented here was conducted using the software Video and Image Analytics in Marine Environments (VIAME; v0.15.1). VIAME is an open-source computer vision software platform created to support object detection, object classification and other processes involved in artificial intelligence tangible for a non-expert audience (Dawkins *et al.*, 2018). Started in February 2018, it is still under development and a rapid implementation of fixes and updates lead to an adaptable tool for different project scopes.

III. Annotation

Annotation is the manual marking of an Area of Interest (AoI) in images and videos by creating bounding boxes (BB) around the target object (*i.e.* a fish, a scampi, etc.; **Fig. 2**).

Figure 2, different types of annotation. *In A depicted is a fish object in the footage. Annotated in a pink BB is the S. cantharus and on the bottom the bait is visible. In B a N. norvegicus (pink BB) class and an actiniaria (red BB) class.*

IV. Post processing – model evaluation

IV.i Intersection over Union

Each model was filtered for the category of interest "fish" and then the Intersection over Union (IoU) was calculated. The IoU is a numeric value of the intersection of the groundtruth bounding box and the detected bounding box and evaluates the representativeness of the bounding box (**Figure 3**).

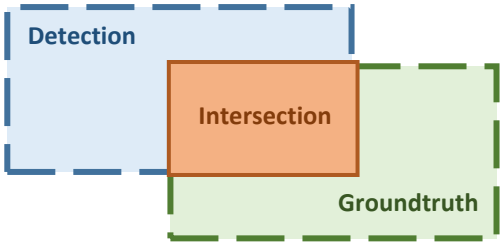

Figure 3, IoU calculation. *The IoU is calculated as the area that is described by both the detection (predicted) and the groundtruth bounding box value and give a value that ranges between 0 and 1 and gives insight on how well the prediction represents the groundtruth bounding box*

$$IoU = \frac{A_{groundtruth} \cap A_{detection}}{A_{groundtruth} \cup A_{detection}} \quad \textbf{(1)}$$

The calculation of the IoU for each detection allows the final step to assess if a detection is considered correct (True Positive) or false (False Positive). With different IoU thresholds - 0.25, 0.5 and 0.75 – different AP were calculated and represented. This allows the evaluation of what the impact of the bounding box accuracy means to the overall model performance.

IV.ii Metrics

The metrics used to assess the results were precision, recall, F1 score and the (mean) average precision (Everingham *et al.*, 2010). The precision is the portion of objects correctly identified (**Equation 2**) when looked at all detections – correctly or incorrectly identified. The recall is defined as the fraction of correctly detected objects when looked at all the objects that needed to be detected – the groundtruth (**Equation 3**).

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad \textbf{(2)}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad \textbf{(3)}$$

With True Positives being the correctly located and classified objects, False Positives being the wrongly classified objects (*i.e.* bait identified as fish or nothing identified as bait) and False Negatives being the missing objects, which were not detected. True Negatives, which are another class to measure the performance, would represent all the areas that were correctly not detected as a AoI. However, in object detection, this metrics is not useful and is ignored since it would artificially inflate the result.

Depending on the question to be answered, these standalone metrics can be used to evaluate object detection models. When the interest lies in the highest possible value for both the metrics, then the F1 score can be considered in addition, since it combines the precision and the recall (**Equation 4**) and is the harmonic mean of the two. The F1 score will be used to determine the confidence threshold in the step of post-processing – bait attraction levels.

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad \textbf{(4)}$$

Before the performance of object detection may vary depending on the amount of data considered, the most widely used object detector metric is the average precision (AP). Like the F1-score, the AP uses the both precision and the recall but sequentially plots them against each other – with the x-axis as the recall and the y-axis as the precision – to generate a precision-recall curve (PR curve). The area under the curve (AUC) corresponds to the AP, and the AP can thus be calculated for each class and the mean of all classes is called the mean average precision (mAP) (Hui, 2018).

--- ---

The following sections 3. and 4. detail the two case studies considered in this report, each with project-specific Introduction, M&M, Results and Discussion.

--- ---

3. Project: LangolfTV

Introduction

The LangolfTV missions are international seabed stock assessment expeditions (**Fig. 4**). These expeditions are conducted to assess the abundance of *Nephrops norvegicus* and other seabed dwelling organisms. The goal is to estimate the stock size of these economically relevant organisms and to ensure that the socio-economic important harvest of the free living *N. norvegicus* in the area is ensured for future generations. (Ifremer, 2017)

Figure 4, depiction of the expedition vessel "The Celtic Explorer" *during of the missions. The camera that is lowered and described in the next paragraph is visible.*

Here, in line with the BAITFISH project, this project aims to determine to what extent human validation can be replaced by a computer doing the same work in order to automatize the stock assessment process and subsequently increase the amount surface and amount of data that can be evaluated. Yet here, since we focus on detecting the occurrence of a specific species, we postulate that if the computer can detect and locate more than 75% (AP > 0.75) (Knausgård *et al.*, 2021) of the *N. norvegicus*, then the computer has the potential of replacing a physical human being counting the animals manually.

The Norway lobster called *Nephrops norvegicus* - referred to as Scampi—are a species of Nephropidae or lobsters and are distributed on the continental shelf and slope in the North Atlantic (M. P. Johnson *et al.*, 2013) all the way down to the Canary Islands, the western Mediterranean Sea in particular in the Adriatic Sea and the Aegean Sea (Lolas & Vafidis, 2021). Due to their benthic lifestyle, the species is exposed to a variety of different stressors such as low oxygen levels (Hagerman & Baden, 1988), predators such as shore crabs and squat lobsters (Albalat *et al.*, 2016) humans and others (Canli & Furness, 1993, 1995). Another stress is applied to these animals: the fishing industry is highly interested in harvesting this species with trawling techniques (Leocádio *et al.*, 2012) for its high economical value for food consumption all over the world (FAO, 2020). An estimated 300 million euros per year is the revenue is generated from catching these organisms (*Landings of Fishery Products.*, n.d.). Of interest here, the species prefer muddy soil to build protective burrows in the seabed to endure the stress of predation, competition and fishing stressor (M. L. Johnson & Johnson, 2013). As of now, the abundance is estimated by counting the number of burrows and not actual individuals. This study aims at automatically determining the abundance of Scampi that are either outside of their burrow or, at least, partially visible. Since the data collection is occurring at the same sites every year since 2017, providing new indices of abundance difference between years would be useful to managers for highlighting ups and downs in local abundances and to define where the fishing effort should be focused on. This is essential for to reduce the risk of depletion of Scampi populations.

Material & Methods

I. Data collection

The footage was captured by towing a 2048 x 1152 px camera, light and laser-equipped sledge above the sea ground (**Fig. 5**) by a trawler—this technique is known as an underwater television (UWTV) survey (Campbell *et al.*, 2009). In total, 44 stations (*i.e.* locations) where captured and analysed. All species beneath the camera were recorded and the video files were manually reviewed by marine scientists.

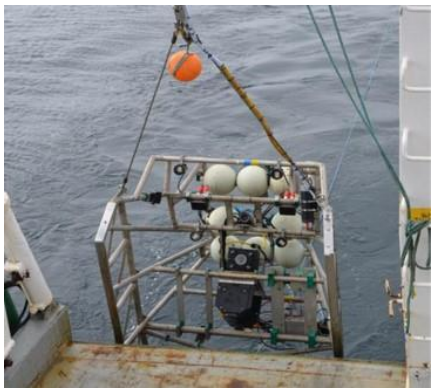

***Figure 5, the sledge** used in the assessments of the seabed in the LangolfTV project.*

II. Annotation conversion

Some LangolfTV data had been annotated with a different software called Labelix, and thus had to be pre-processed and converted into a format compatible with VIAME (see above). Labelix creates an xml file as output, whilst VIAME uses a csv format as input for the annotations. To convert the Labelix derived annotations and exploit them in VIAME, the different image attributes where extracted, filled into a data frame and saved in .csv format using RStudio v1.4.1106 and packages "XML" and "tidyverse" – especially the function grep.

III. Training Data

The project collected data in 2019 at 44 different stations in the Gulf of Biscay. 39 of these stations showed low to high-moderate diversity of classes and functioned as the training dataset. This led to 79% or 31'736 annotations of the complete data collected in 2019 being treated as the training data. For a full overview of the 18 classes and number of annotations, see **Table 1**. These annotations were marked on 20'364 images, corresponding to roughly one to two annotations in each frame. It is important to note that not all the original annotations were used, since there were classes with only two annotations available. The low numbered classes were fused with bigger classes to have better overall class levels (**Appendix IV**). Class imbalance, a recurrent issue with machine learning, are regulated by the usage of the metrics mAP, which already accounts for this type of issues.

***Table 1, depiction of the split of the train and the test annotation dataset used for the LangolfTV subproject.** Classes were fused to allow a better overview and training process.*

class	total	train	test	percentage of test
actiniaria	4401	3607	794	18%
actinopterygii	1082	912	170	16%
callionymus	400	356	44	11%
cephalopoda	327	297	30	9%
crevette	2655	2382	273	10%
crinoidea	2500	2386	114	5%
crustacean	1096	995	101	9%
flatfish	730	547	183	25%
gadiforme	530	425	105	20%
hydrozoa	1494	1355	139	9%
munida	6183	5001	1182	19%
nephrops_norvegicus	2874	2474	401	14%

pennatulacea	7299	4790	2509	34%
polychaeta	792	670	122	15%
scyliorhinus	38	18	20	53%
terrier	7663	5341	2322	30%
starfish	261	180	81	31%
	40325	**31736**	**8589**	**21 %**

IV. Test data

Stations 2, 15, 18, 52 and 68 were held back due to their high diversity levels and inclusion of rare species the algorithm was trained on (**Fig. 6**). In total, the 8'589 remaining annotations from 4'643 images were used to evaluate the models' performance. The stations evaluated have all the classes included, which allows the calculation of average precisions (AP) for each class and a mean average precision (mAP) for all the classes together - see paragraph below VII. Since the focus lies on detecting Scampi and their burrows, these classes' AP are of particular interest for this study.

Figure 6, example species observed during the LangolfTV 2019 mission *used in the project of 2021. A shows the species of interest N. norevgeicus, B shows a crustacean (unidentifiable but not a Scampi nor a Munida) and C shows an example of a Munida spp.*

V. Post processing – Confusion matrix

After the creation of the model and the evaluation of its performance, a confusion matrix could be computed. The confusion matrix is an important tool to see how exactly the training efforts can be focused to generate a new model with better results on underperforming classes when numerous classes are considered. A confusion matrix enables to highlight strengths and flaws in the proposed model and give an overview on how the model has performed when compared to all classes and gives more insights on which class performed well and which are underperforming.

Results

I. Model evaluation

The performance of the model created with the training and test data presented above are presented with the Precision-Recall curves at IoU levels of 0.25, 0.5 and 0.75 (**Figure 7**).

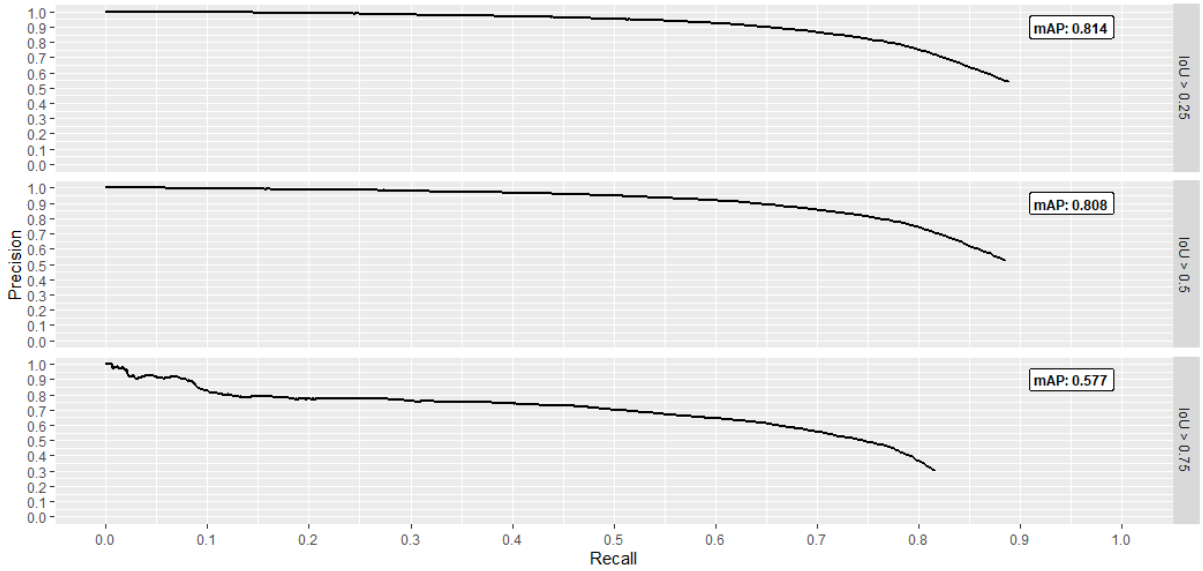

***Figure 7, Precision-Recall curves depicting the performance of the LangolfTV model**. Three levels of IoU are presented to have an overview on how this affects a model's performance. The mAP values are depicted in the rectangle boxes in the top right corner. (Abbreviations: mAP = mean average precision, IoU = Intersection over Union)*

In **Figure 7**, different IoU levels show moderate to high difference in the model's performance. The difference between IoU 0.25 and 0.5 is lower than the difference between either the IoU 0.25 or 0.5 and the IoU 0.75. The mAP of all the classes is 0.814 for IoU > 0.25, 0.808 for IoU > 0.5 and 0.577 for IoU > 0.75. The two lower IoU thresholds show a gradual decrease ending with a steep decline at higher recall values. For the highest IoU threshold, the detections with a high precision already show a pattern of decline and start to drop at 0.02 recall value. False Negatives (FN) range from ~ 12% for 0.25 and 0.5 to 17% for 0.75.

II. Individual average precisions

Individual APs are depicted in **Table 4**. Average precision ranges from 0.24 to 0.86 for the different classes. The most influencing classes are "pennatulacea" and "munida" with numerous detections (2'509 and 1'182) and high AP (0.8 and 0.82).

The "callionymus" has the highest AP value (AP=0.86), while the "hydrozoa" and "crustacean" classes have the two lowest level of correctness, each with AP values of 0.24.

***Table 2, individual AP per class.** (Abbreviations: AP = average precision)*

class	AP
terrier	0.36
starfish	0.74
scyliorhinus	0.38
polychaeta	0.23
pennatulacea	0.80
nephrops_norvegicus	0.69
munida	0.82
hydrozoa	0.24
gadiforme	0.71
flatfish	0.52
crustacean	0.24
crinoidea	0.63
crevette	0.68
cephalopoda	0.44
callionymus	0.86
actinopterygii	0.52
actinaria	0.60

III. Confusion matrix

To get a better overview on the results of the precision-recall curves, the IoU threshold of 0.5 (*i.e.* the most commonly used threshold in the literature (Huang *et al.*, 2019a)) was chosen for the confusion matrix (**Fig. 8**). For a confusion matrix, the confidence level to consider the detections as correct also need to be defined. A confidence threshold of 0.33 was selected by calculating the highest F1 score. This, if the predicted confidence in the classification of one object is below 0.33, it will be considered a false negative (FN), while if it is higher than 0.33, it will be a true positive. Within the 17 classes consider, five classes (crustacean, nephrops_norevigucs, munida pennatulacean and terrier) were chosen for closer inspection since these classes either performed well, poorly or are of interest for the economy in the area (**Fig. 8 – yellow, green, red, white, and blue rectangle**).

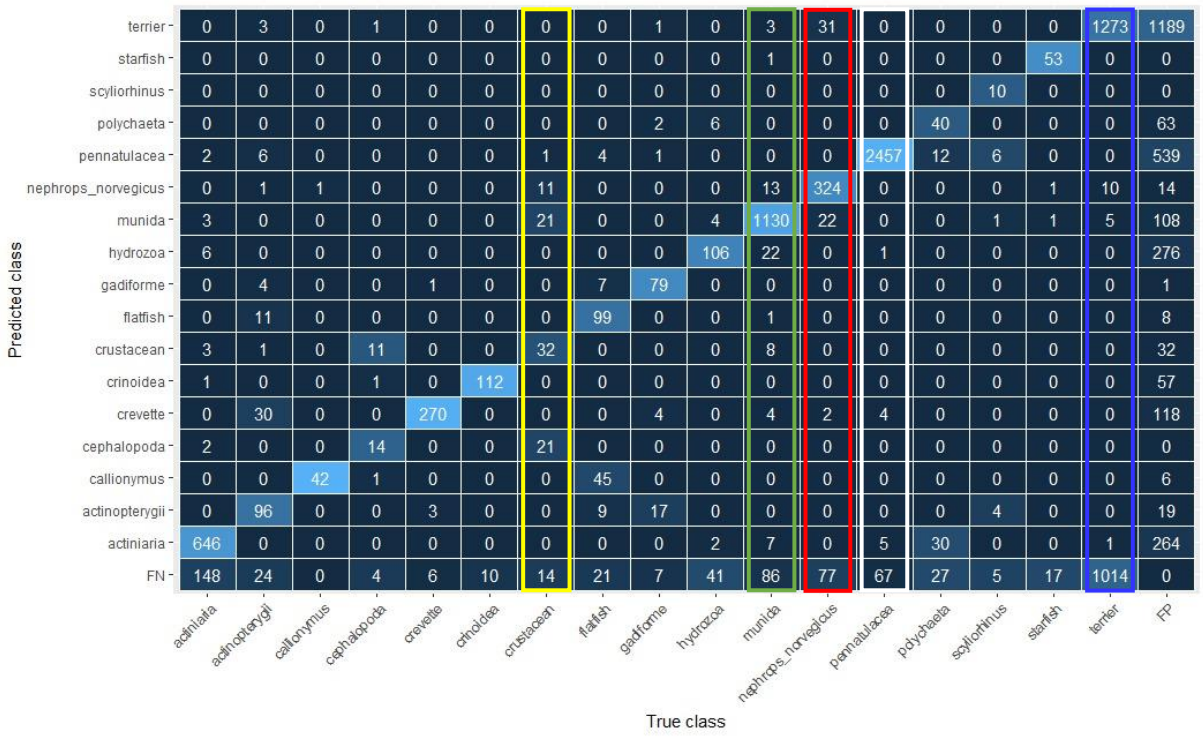

Figure 8, confusion matrix for all the classes *implemented in the model within this project. On the x-axis the actual class in the groundtruth is reported and on the y-axis the models predicted class. The lighter the colour per cell, the better the performance of the model for this class. Colour code for the rectangles is yellow for the crustacean, green for the Munida spp., red for the Scampi, white for the pennatulacean and blue for the burrow (in French terrier) class. False negatives and false positives are bounding boxes that are drawn in the background and do not represent any class. (Abbreviations: FN = background false negative, FP = background false positive).*

Overall, there is a clear diagonal line observable, which indicates that the predicted classes are mostly represented as the true classes. There is one clear exception with the "terrier" (French burrow) class (**Fig. 8 – blue rectangle**): 1'273 detections are counted as true positives (TP) but another 1'014 background false negatives (bFN) and 1'189 background false positives (bFP) are also detected. This means that a large proportion of burrows are not detected or located correctly (only around 36% are correctly detected). Another underperforming class is the "crustacean" class (**Fig. 8 – yellow rectangle**) which showed confusions with other more abundant classes such as "munida" and "nephrops_norvegicus". In total 32 true predictions and 100 wrong predictions make up this class, this only 24% are predicted correctly.

The Scampi trigger moderate results (class "nephrops_norvegicus", **Fig. 8 – red rectangle**): a total of 324 of the Scampi are correctly detected but non-background false positives (FP) remain frequent (55 detections, 14%). In addition, there are 77 bFN and 14 bFP detections, which adds up to a final mean

of 69% of correct predictions for Scampi (correct = correct class and correct location of the object on the frame).

One of the classes, the "pennatulacea" was the most accurately detected (**Fig. 8 – white rectangle**), with 2'457 TP, 10 FP, 67 bFN and 539 bFP. More than 80% of all the detections made were thus correctly detected. Another easy-to-detect class was the "munida", having 1'130 TP, 59 FP, 86 bFN and 108 bFP (**Fig. 8 – green rectangle**). It overpasses the pennatulacea class with a mean correct prediction of 82%. These two classes make up most of the data and therefore have a big impact on the mean average precision.

Discussion

This project presents a new method to estimate the benthic organism's abundance when dealing with Underwater Television (UWTV) surveillance data. The great advantage of this method is that the organism found on the seabed floor are in general less fast-moving organisms and have less year-to year diversity. On the other hand, the seabed is more divers when it comes to its inhabitants – as seen in these projects. Nevertheless, the advantage of more sessile and slower organisms allows to use the same model each year to assess the abundance of these communities, which, over longer time scales, could save a lot of labour and time.

Contributing most to the high mAP value are the two classes "pennatulacean" and "munida", which both combined are over 3'500 annotations of a total of 8'500 annotations worth. This means if these classes perform well (AP > 0.5), the whole model would be considered as "good"—that is one of the flaws of the metric, even though it is accounted for it during the calculation.

Special attention is given to the Scampi class and the performance of the model on this class. The overall AP value of 0.69 is considerably lower than expected and does not reach the level proposed in the hypothesis. This may result from the morphology of other crustacean species such as *Munida spp.* and the general class of crustaceans, which can easily be confounded with *N. norvegicus* (**Figure 6**) and are frequently falsely identified as such (Figure confusion matrix). Another error-prone class is the terrier class. Its underperformance could be due to the inconsistency of the training dataset containing un-annotated burrows. Furthermore, it could be because of the training data containing burrowed Scampis which are not 100% visible and then the model detects a Scampi instead of a burrow. There is potential by improving the detection of scampi in their burrow by adding diverse data (angle, parts of the scampi that is visible, etc.) into the model training.

The created model is thus able to predict several, uniquely shaped and formed species with an AP greater 0.75, showing a solid potential to complement the assessment of these species or genus. Yet, comparing the results to other publications suggests that the model can be considered as efficient as others (eg. Huang *et al.*, 2019b). In the publication of Huang *et al.*, 2019b obtained AP values 0.47 to 0.80 and an overall mAP of 0.6, which are in agreement of the results of this projects. For some classes, "crevette", "munida" and "pennatulacea" with mAP values over 0.8, this method could indeed be applicable for multi-year variability in abundances. An AP over 0.5 is commonly assumed as "good" (Li *et al.*, 2015). However, our results indicate that considering this threshold alone can lead to under- or overestimation of wild living populations, therefore misleading conservation efforts in this area. Here of particular interest, the key class "nephrops_norvegicus" has an AP of 0.69 (ie. considered "good to excellent" based on (Li *et al.*, 2015)), which translates the 69% of correct detections and 31% of incorrect ones on this class. Further fine-tuning is thus mandatory to obtain a model that reaches satisfactory performances to feed stock assessment data, and one should not stick to numerical values without understanding what is behind and the potential consequences of error propagation if the data are exploited for modelling or stock management.

Next steps for this project could be the implementation of a class-dependent fine tuning and further analysis of the underperforming classes. This could give a greater insight on why exactly certain classes did not perform well and to what extent this performance could be improved. Adding more data from more stations and years could be a great improvement.

4. Project: BAITFISH - Behaviour, performAnce, Impacts of Trap FISH

Introduction

This project aims to develop more selective fishing gear targeting economically relevant fish species in the Golf of Biscay whilst decrease bycatch or avoid trapping unwanted or even protected species (Ifremer, 2018). During summer seasons the species *Spondyliosoma cantharus* or commonly known as the black seabream is very abundant in the area around the Golf of Biscay. Non-selective fish traps were chosen by numerous fishermen and led to the increase of targeted bycatch of lobster and other crustaceans without quota and therefore the need for a newer more selective fish traps emerged. Furthermore, economically a great interest in this fish exists in the region around the Gulf of Biscay (Perodou & Nedelec, 1980). Due to their low fat content, low liver lipid content and high Hepatosomatic Index (Rizzo & Bazzoli, 2020) they became a popular target for the artisanal and commercial fishing industry (Future Market Insights, 2020; Kora *et al.*, 2000) and food industry.

To increase the time of attraction for the fish, three biodegradable plastics were introduced and tested. The smell of these baits' dissolves slowly into the water column attracting nearby fish downwards of the current and potentially allows the trap to be longer in water with attracting only fish as interest.

Here, we focus on the spatial, temporal, and behavioural differences between the fish approaching different types of baits. Patterns were first analysed manually, then automatically with the deep learning networks trained on the data. Both methods were then compared to determine the potential of a computer replacing human validation when dealing with this kind of sea trials.

Material & Methods

I. Data collection

In a preliminary project, different baits were tested to evaluate the most efficient bait type between several organic and inorganic matters (lures, light, cuttlefish, cockles, etc.) (Fluhr *et al.*, in prep). The most efficient in attraction and withhold were the cockles which functioned as the control in the experiment evaluated in this report. For the current project, Baited Remote Underwater Videos (BRUVs) were recorded by lowering dropcams (**Fig. 9**) in different locations recording the area of and around the bait structure.

Figure 9, basic principle of the baited remote underwater setup used in BAITFISH.

This experiment wanted to evaluate if the hooking of animal biomass can be reduced by using biodegradable plastic pallets induced with cuttlefish powder as bait. The footage was generated during one-day boat trips from 20. to 23. of July 2020 with GoPro Hero 4 cameras and in the format of 1920x1080px and 24 frames per second (fps). The videos were manually analysed by a student of the UBS (Université Bretagne-Sud), then revised and re-formatted to fit the output of the detector and to acquire comparable results.

II. Annotation

The annotation for the BAITFISH data did not need a conversion since the videos were not processed before and were annotated directly in VIAME. In addition, the fish class, the plant and bait class were introduced to reduce the amount of falsely positive predicted particles, plants and the bait-apparatus.

III. Training Data

Training data are the images that are presented to the computer from which the algorithm learns the different classes – in this example the fish, plant, and bait classes. The baseline model 0 is an already existing, pre-trained fish detector available in VIAME. It was trained using CFRNN on four National Oceanic and Atmospheric Administration (NOAA) datasets containing various footage of different sites. Dropcams in the Gulf of Mexico, the MOUSS protocol in the Pacific Ocean (Miller-Greene *et al.*, 2020) and the HabCam (**HAB**itat mapping **CAM**era) system in the Atlantic Ocean (Cojanu & Hugus, 2016). This detector is constantly updated, and the detector used in this project was downloaded the 10th of March 2021.

The model 0 is thus performant to identify fishes in general, yet this baseline model performed poorly on the species considered here, since the conditions and species characteristics are unknown to the model. Thus, going from model 0, BRUV footage of older expeditions in the area Gulf of Biscay was used to train the different models. The images and annotations were gradually increased to evaluate the effect on the model's performance by doing so (**Table 3**). Different levels of complexity due to turbidity and levels of fish were chosen to have a great variety of input data. To save time spent annotating, the frame rate was reduced to five fps from 24. Therefore, 9,277 images - used for model 11 & 14 – correspond for approximately 30mins of video footage. The difference between the model 11 and model 14 are different only by their parameterization of the neural network (indicated by the * in **Table 3**). Augmentation is the process of increasing the number of images by rotation, cutting and other image alterations is done automatically in VIAME and optimized for the deep learning network architecture used (here, CFRNN).

Table 3, model names and training efforts (annotations) used. *The *-symbol stands for a different validation set size and was corrected only for model 14.*

Model	**Images**	**Annotations**	*Fish*	*Plants*	Bait
0 – The Basic	**0**	**0**	*0*	*0*	0
5 – The Middle	**1830**	**4002**	*1875*	*222*	1905
7 – Only Fish	**1830**	**1875**	*1875*	*0*	0
8 – 7K	**7141**	**24017**	*19214*	*879*	3942
11 – Close Second	**9277**	**30433**	*19214*	*6073*	5146
14 – Top Model*	**9277**	**30433**	*19214*	*6073*	5146

IV. Extracting the groundtruth dataset

To evaluate the performance of each BAITFISH model, a 17-min video (5,100 frames) with varying environmental factors such as turbidity and camera angle (**Fig. 10**) was concatenated and fully annotated, totalizing 6,660 fish and 5,100 baits. These handmade annotations are the baseline dataset,

or *groundtruth,* used in the following sections to compare the object detections performance of the different models.

Important note: There was no plant annotated in this video, since the interest of the project lies on the ability to detect the fish object.

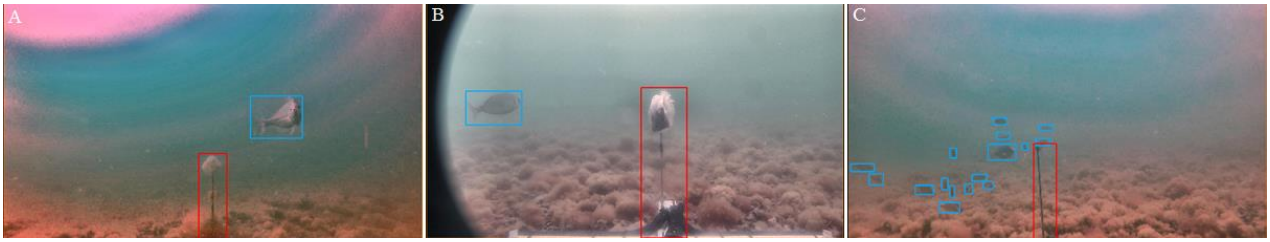

Figure 10, Snapshots of the test video used for the groundtruth in the BAITFISH project*. A: low light condition, B: clear conditions but with camera frame visible, and C: school of fish passing by the bait. In blue are the annotated fish and in red the different baits. Note that these images are already enhanced by VIAME as they were during the training. Important note that the colour correction is not applied for the training only when the video is loaded into VIAME directly.*

V. Manual processing

The videos were looked through by hand and the time of the fish on screen was noted with different additional information such as behaviour. A general model was applied to the video and allowed the identification of potential fish and their corresponding timestamps. All the fish positions were evaluated and looked at and allowed the manual analysis of the videos.

VI. Post processing – bait attraction levels

The ecological question to be addressed here is to determine the attraction levels of one control and three biodegradable plastic baits developed within the BAITFISH and Indigo projects, using the presence and behaviour of *S. cantharus* around the baits. Since all data need to be validated to reach this objective, this will later enable the comparison between the attraction levels obtained manually and automatically with the trained models presented above.

VI.i Confidence threshold

Every detection receives a certain confidence of correctness ranging from 0 to 1 (equivalent to a probability) by the model. The higher this value, the more certain that the chosen class is correct. When several classes are probable, the model will provide the confidence level of all possible classes (from 1 to *n* classes, *n* being the number of different classes used for training the model). Here, we assume that the class with was attributed the highest confidence is the most likely, and all secondary classifications were discarded.

To reduce the False Positives, a confidence threshold was calculated for the best performing model based on the highest F1 score by plotting the confidence (on the x-axis) in 0.01 steps against the F1 score (y-axis). To assess further the effect of the confidence threshold, a lower and higher (± 0.1) threshold are also included into the evaluation.

VI.ii Area of Interest (AoI)

The Area of Interest (AoI) is the portion of the screen surrounding the bait in a radius of 450px - see the red circle in **Figure 11**. The area has been defined in an ongoing study (Fluhr *et al.*, *in prep*), which shows promising results.

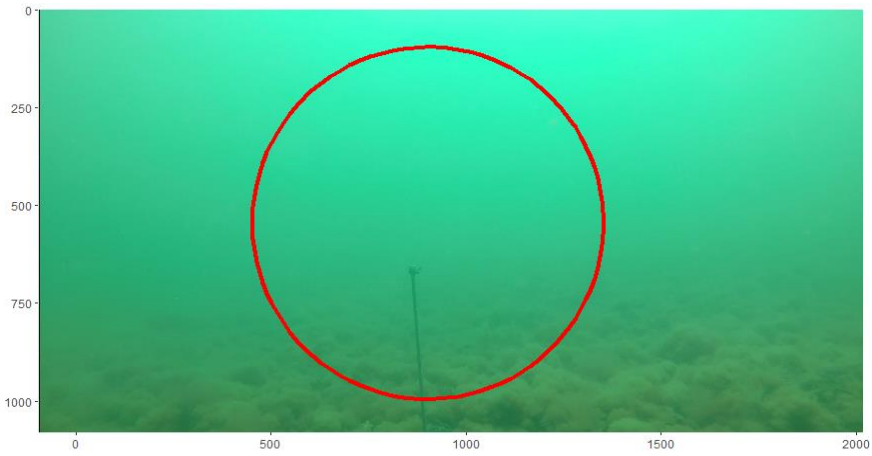

Figure 11, Area of Interest (AoI) *as described above.*

The centre of the bounding box (mean of x_1 and x_2 and mean of y_1 and y_2) functioned as the point to which the distance to the bait is calculated (**Fig. 12**) and to decide if an object is in the AoI or not.

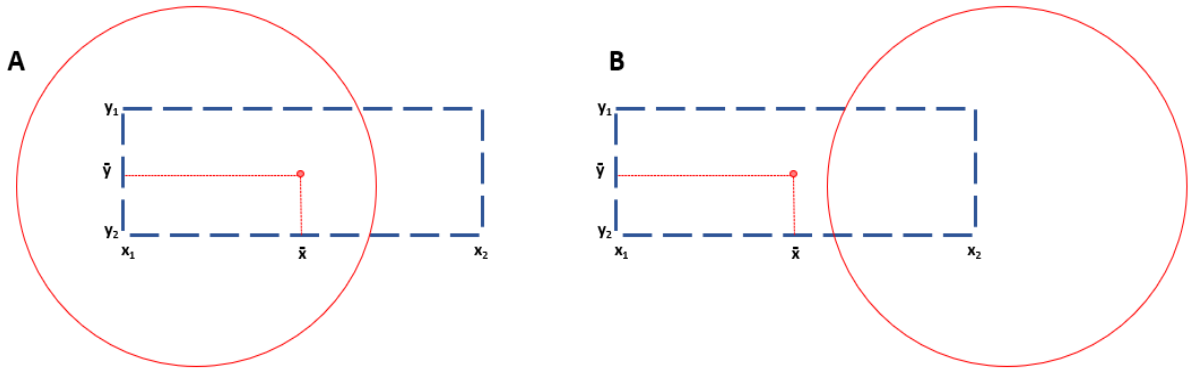

Figure 12, shows the definition of a fish inside (A) and outside (B) of the AoI. *Blue corresponds to the bounding box of a fish and the red circle indicates the AoI.*

Fish-tracks that lasted less than 2 frames were systematically false-positives, and were considered as such in the analyses, and not counted as fish nor considered for the evaluation of fish counts within the AoI.

VI.iii Interest Index (Ii)

After this cleaning process (*i.e.* discarding fish with low confidence or tracks too shorts), each detected fish can be considered as a fish and is categorised into one of five categories (**Table 4**) depending on if and how long the fish was on the screen and in the AoI.

Table 4, categories of interest *and the time spent in AoI or one the screen.*

Category	Times spent in AoI (s)	Category
1	0 (only on screen not in AoI)	On the screen
2	< 1	Passing
3	1-2	Mildly interested
4	2-3	Interested
5	> 3	Eating

The five categories are then corrected for the total number of fish in the video (**Equation 5**), since this factor plays a crucial role and needs to be incorporated into the equation. The category of each fish is divided by the number of fish and summed up to have a corrected mean of all the categorised fish, as:

$$Ii = \sum \frac{category_i}{n_{fish_j}} \qquad \textbf{(5)}$$

VI.iv Statistical test

A Kruskall-Wallis test was conducted to test significancy between the bait types of each method. A pairwise Wilcoxon-test was used to check the differences in the bait types and to further evaluate the result. As a software RStudio v2021.09.0 Build 351 and R 4.1.1 with the packages *ggplot2* 3.3.5 and *ggrepel* 0.9.1 for graphics and the package *stats* 4.1.1 for statistics. The methods – automated and manual - were not compared in-between since there was an observable significance difference.

Results

I. Model performance

To assess the models' performance, two different sets of curves were dawn. One set for the mean average precision (mAP) for the two classes bait and fish and one set for the average precision (AP) of the fish class alone (**Table A1**).

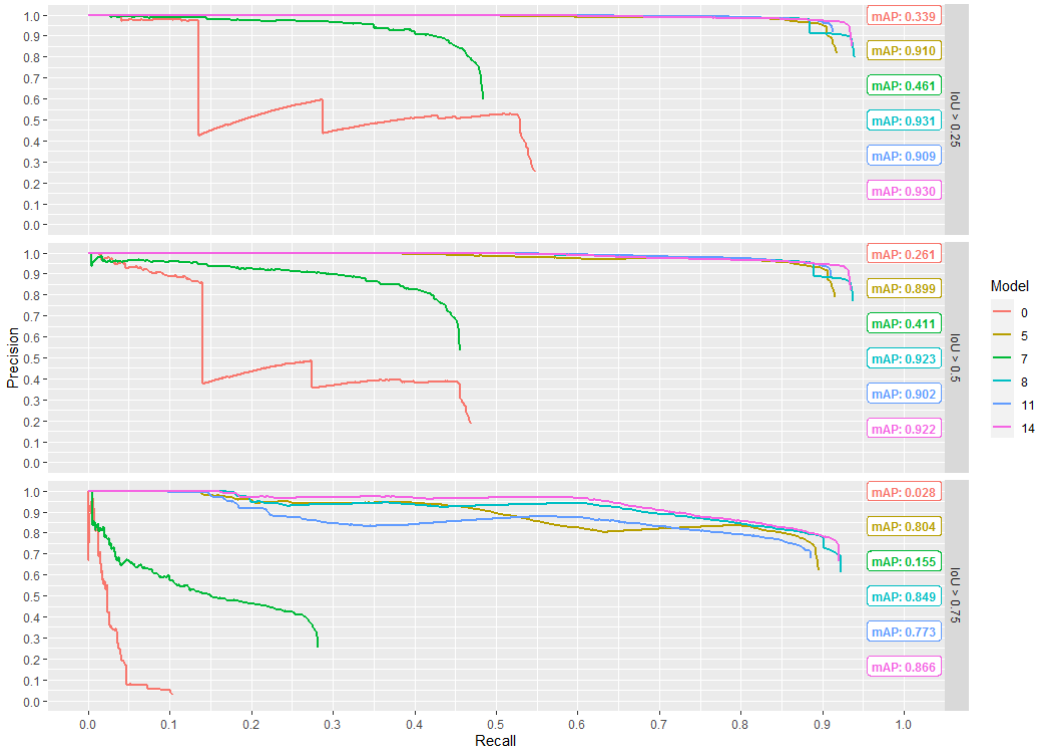

Figure 13, Precision-Recall curves of the two classes fish and bait *– indicated with colours are the different models. Labels indicate the mAP for each model for the two classes. The different plots/facets explain the PR curves at different IoU thresholds 0.25, 0.5 and 0.75. (Abbreviations: IoU = Intersection over Union, mAP = mean average precision, PR = Precision-Recall)*

The mAP values for the first set of curves - bait and fish - are presented on **Figure 13**. Model 0 is outperformed by every other model since there is no bait class trained, which leads to inflation of false negatives and therefore severe smaller mAP values (IoU > 0.75 = 0.028, IoU > 0.5 = 0.261, IoU > 0.25 = 0.339). The second model is model 7 trained only on the fish class and for the same reason as model 0 it does not perform well with values ranging from 0.155 for IoU greater than 0.75 up to 0.461 with an

IoU threshold of 0.25. That means roughly every second detection is correctly located. Model 5 has the least annotation compared to the remaining models but is with values 0.804 (IoU > 0.75), 0.899 (IoU > 0.5) and 0.910 (IoU > 0.25) above 0.5, which can be considered good to excellent. The next best model is model 8 and differentiates from the other models by having less bait and plant annotations. MAP values of 0.849, 0.923 and 0.931 were achieved for the different thresholds. The increase of 0.5 and 0.25 compared to 0.75 is explained by the difficulty for the model to locate and mark the correct location of the objects in the video. Model 11 and model 14 are different in terms of parameters chosen and achieved values of 0.723, 0.824 and 0.831 for model 11 and 0.816, 0.859 and 0.867 for model 14.

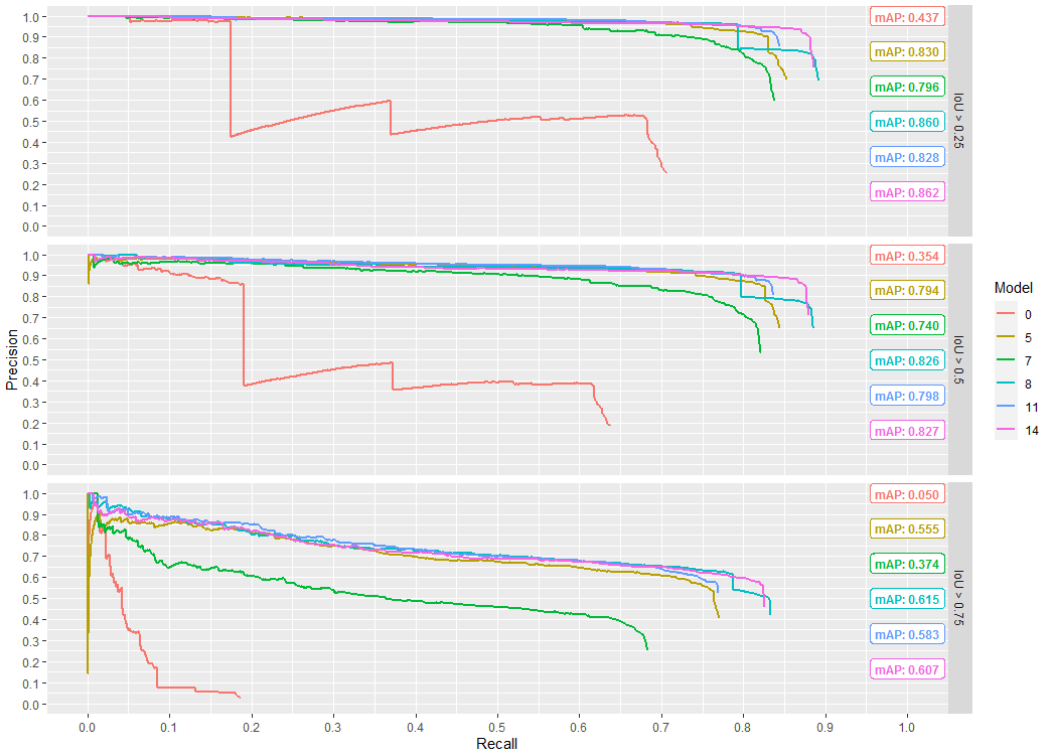

Figure 14, Precision-Recall curves of the class fish *– indicated with colours are the different models. Labels indicate the mAP for each model. The different plots/facets explain the PR curves at different IoU thresholds. (Abbreviations: IoU = Intersection over Union, AP = average precision)*

Overall, the IoU thresholds, a consistent increase of the AP is observable (**Fig. 14**) when going from model 0 to model 8 except for model 7, which was a test model trained only on fish and no other classes (plants and baits) and to see the effect of the exclusion of non-fish classes on the average precision on the class of fish. The models 8, 11 and 14 seem to have reached a plateau since the AP is in the range of ± 0.03 with 0.7 to 0.72 for IoU > 0.25, 0.67-0.7 for IoU > 0.5 and 0.5-0.52 for IoU > 0.75.

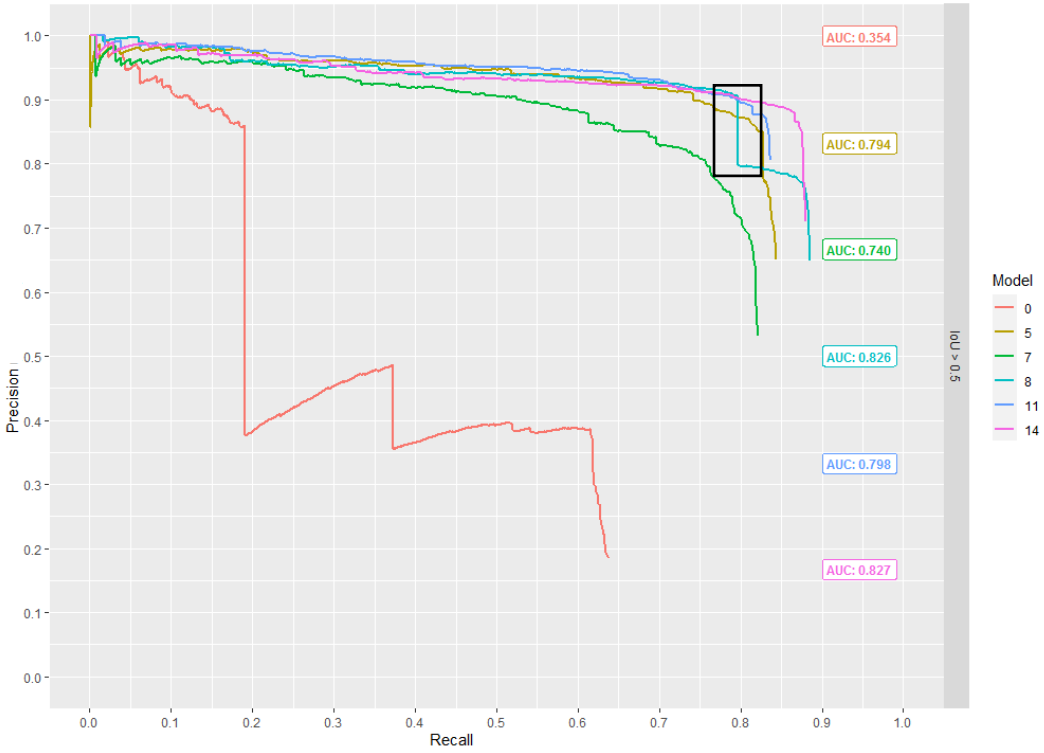

Figure 15, Precision-Recall curves of the class fish *– indicated with colours are the different models. Labels indicate the AP for each model. The different plots/facets explain the PR curves at different IoU thresholds. Black box indicates a special type of error of object detection explained in* ***Figure 13****. (Abbreviations: IoU = Intersection over Union, AP = average precision)*

For the following results, an IoU > 0.5 is considered since it is the most commonly used threshold. The **Figure 15** and following figures only show the result for AP of the models at the selected threshold of IoU > 0.5. As expected, the generic fish model 0 performs poorly when it is compared to the other five models. With an AP of 0.307, the model 0 generates approximately one third of correct detections. By providing more images and annotations, we could fine-tune the models and largely improve the detections. Better performant was the model 7, which was trained on only the fish annotations. The AP doubled when compared to the model 0 with a value of 0.614. An AP of around 0.68 ± 0.02 (models 8, 11 and 14) means that approximately for every three fish detected, two of them are correctly identified and located. Model 14 has the highest AP with 0.695 for the class fish and was the model chosen for the detection & evaluation of the biodegradable-plastic baits videos.

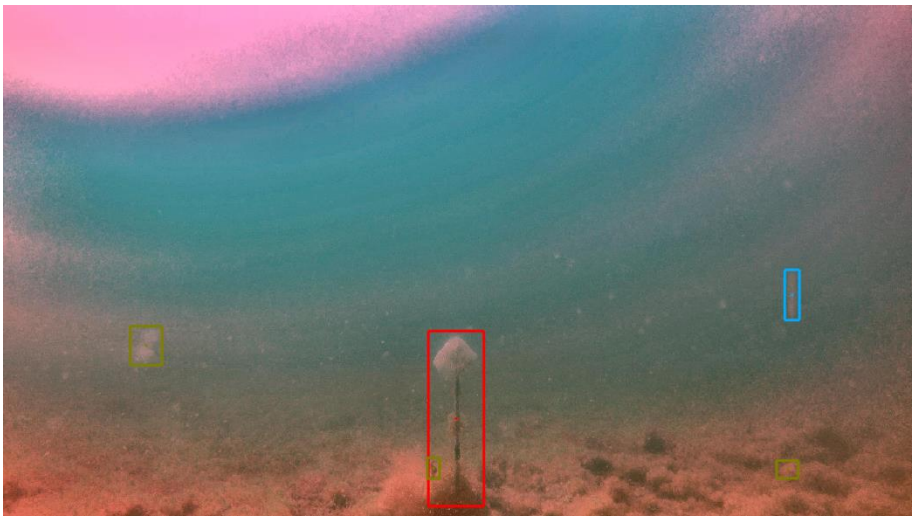

Figure 16, error of model 8 *detecting the GoPro light as a fish (blue BB). Correctly identified bait (red BB) and plants (green BB). This was later avoided by adding annotations of the light as the plant class for models 11 and 14. Important note that the colour correction is not applied for the training only when the video is loaded into VIAME directly.*

Special attention was given to re-occurring and avoidable errors. In **Figure 16** provides an example of such re-occurring detection with model 8. On the right-hand corner there is an incorrect detection since the GoPro light is detected as a fish. To counteract this mistake, the models 11 and 14 contain images with this light annotated. This error hinders the model 8 from being one of the best models in this experiment (cf. the black rectangle on **Figure 15**).

II. Biodegradable-plastic baits analysis

II.i Confidence threshold

To define an objective cut of the F1 score, the score was plotted against the confidence threshold and the highest value for the F1 score was chosen to represent the threshold needed (**Figure 17**).

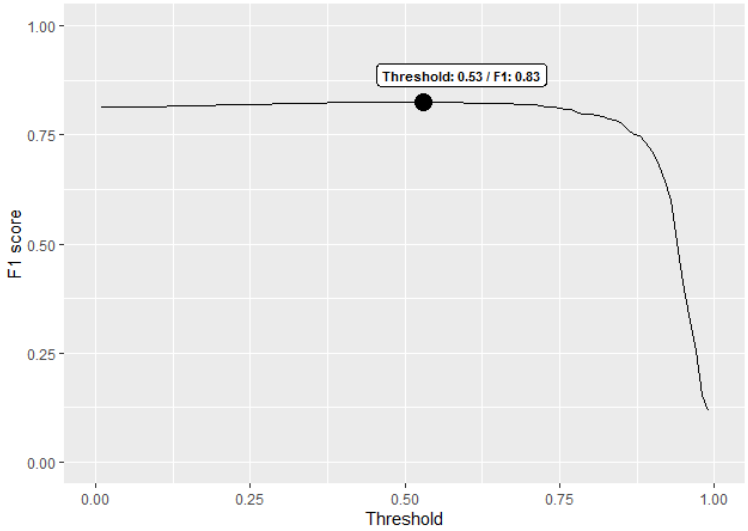

Figure 17, evaluation of the confidence threshold for model 14. *The F1 score is plotted against the threshold to get the highest possible value for the F1 score.*

The curve here does not follow the expected pattern (see **Appendix I** for the expected pattern). As explained in Czakon (2021), this is due to the low detection number at lower confidences and those detections being valid. This definition of threshold is used when all classes are balanced and of the same importance. Since the interest only lies on the fish, this technique is used to find the best

threshold only for the fish class. At a confidence threshold of 0.53 the F1 score is highest with a value of 0.83. This means that the further analysis will be done with a threshold of 0.53, a lower (0.43) and a higher (0.63) variant for comparison.

II.ii Number of fish

Looked at first were the number of fish predicted for each biodegradable bait type. A Kruskal-Wallis test and a pairwise Wilcox-test was then done to see if the data is different from each other. A boxplot graph was created to present a visual output of the data (**Figure 18**).

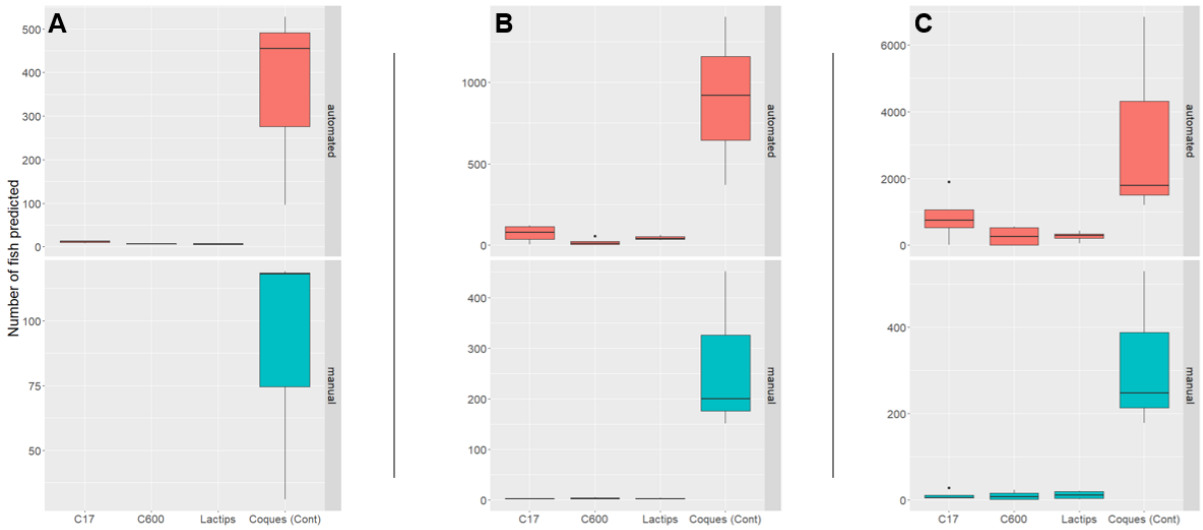

Figure 18, number of fish predicted by analysis method – automated or manual*. A presents the data after 15 minutes of the videos have passed, B shows after 59 minutes and in C described is the whole video. This was done to have the temporal factor also included. Red colour indicates the automated data, whilst the blue colour indicates the manually acquired data.*

The conditions of application are violated for the manual analysed data (presence of heteroscedasticity, Bartlett test, $p < 0.05$), and the Kruskal-Wallis test was thus applied.

Overall, the Kruskal-Wallis test showed significant differences in the automated data (p-value for 15mins = 0.05, p-value for 59mins = 0.04, p-value for the whole video = 0.05) as well as significant differences in the manually acquired data (p-value for 15min = NA, p-value for 59min = 0.16, p-value for the whole video = 0.066). The Lactips and C600 show no statistical differences in attraction level (pairwise Wilcoxon test, $p > 0.05$). The pairwise Wilcoxon test shows highest confidence of difference for C17 and C600 a p-value of 0.34. The shape of the boxplot follows similar patterns with an outlier for the C17 baits and low values for C600. Yet, there is a difference in the total number of fish observed for the two methods. Nevertheless, similar patterns are observable with high numbers for the control bait and lower numbers for the biodegradable baits. The high number of fish detected by the model is an indicator that the model is overestimating the number of fish in the video, which is expected since if the track of a fish is lost, the same fish may be counted multiple times.

II.iii Interest index (Ii)

To counteract this overestimation, an Interest index (Ii) was calculated. The calculation is explained in **Equation 5**.

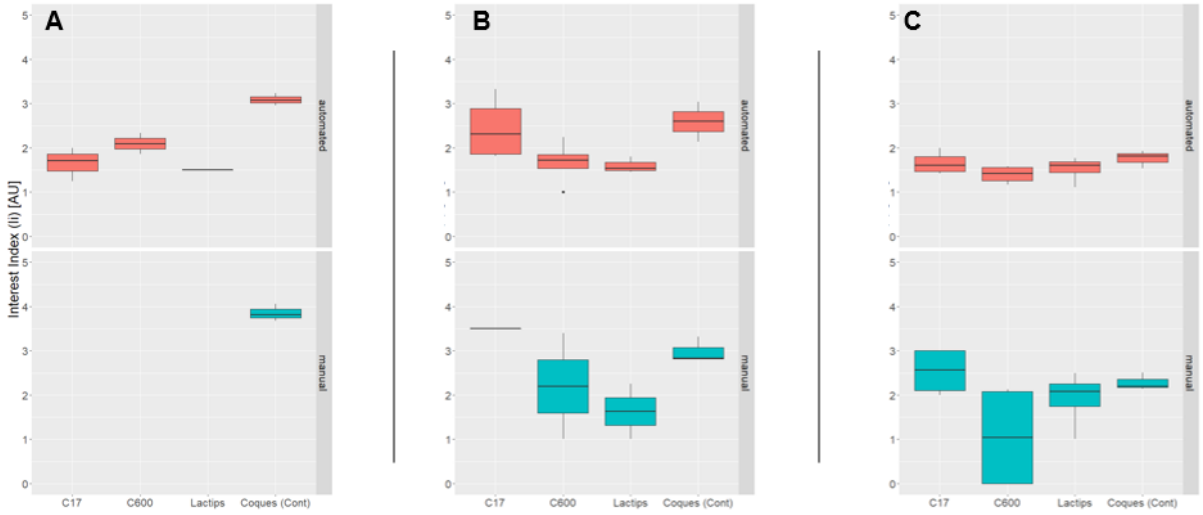

Figure 19, boxplot of the Interest index (Ii) *for the different bait types. A presents the data after 15 minutes of the videos have passed, B shows after 59 minutes and in C described is the whole video. The colour code is red for automatically analysed and blue for manually analysed data.*

The interest index Ii returns a value for each video, which are presented on **Figure 19**. Clear differences are observable in the 15min group since there was fish detected even though there were no fish observed by manual analysis. Interestingly, for both the automated and manual dataset the index is high for the control bait (automated = 3.089451, manual 3.850622) which indicates an early presence of fish staying in the area of interest for a long time (**Figure 19 – A**). A Kruskal-Wallis test showed no significant difference in the bait's attraction levels for the automated data with a p-value of 0.06, no test was conducted for the manual data since there is only one group. **Figure 19 – B** indicates the Interest index after 59min. Automated data shows lower levels of interest levels in comparison to the manually evaluated data (Ii = 2.590784 vs 2.98497). The boxplots do follow similar patterns within the two methods conducted. For the whole video analysis (**Fig. 19 – C**), the boxplots follow similar patterns, forming a u-shape. Overall, the interest index is similar when looked at the mean between the two methods, with no significant difference is observable when investigating each method separately – automated p-value = 0.50, manual p-value = 0.25).

II.ii Heatmapping the number of fish

Without great extra effort, heatmaps (**Fig. 20**) were created and visually explored to detect if there is a dense concentration of fish in the AoI.

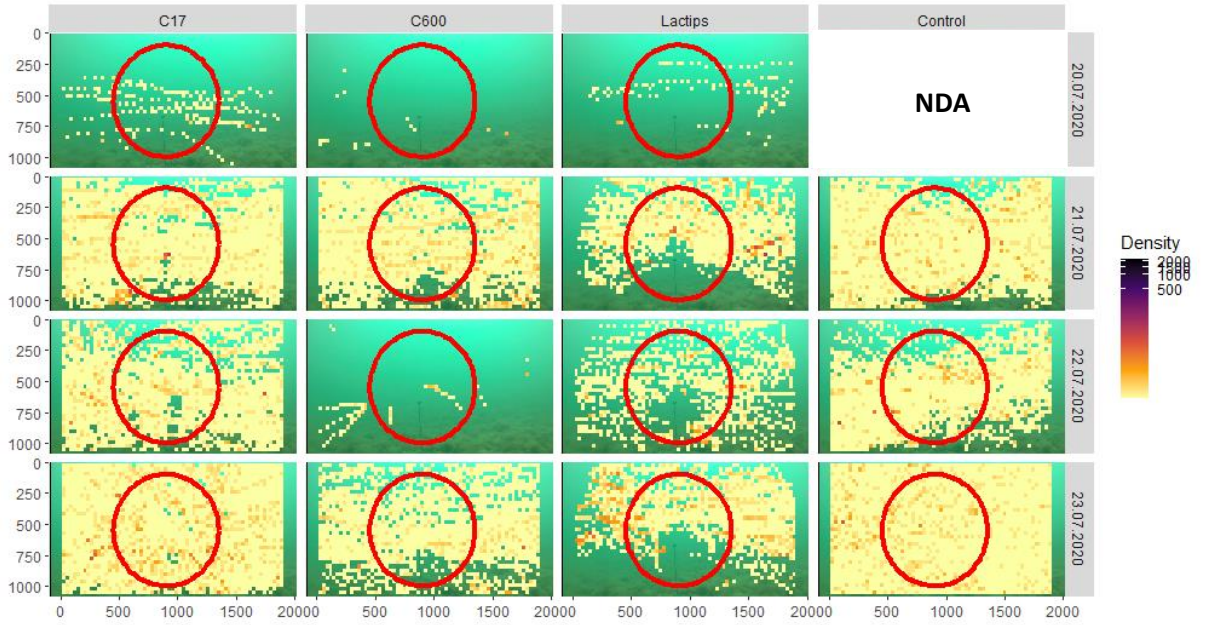

Figure 20, heatmaps describing the fish in AoI. *Each plot represents a video and each 30px by 30px field is used to present the concentration/density of the detection of fish at this exact 30x30-area. The plots are divided by baits (C17, C600, Lactips, Control (Cockles)) and by dates (20-23.07.2020). The colours of the 30x30 fields represent the density from dark (= high density) to light (= low density). (Abbreviations: NDA = No Data Available)*

The bait's level of attraction of the C17 bait shows the strongest attraction levels of the three bioplastic baits, as indicated by the larger numbers of fish inside of the AoI, followed by the C600 and Lactips. The control, with raw cockle, remains the bait type with the highest attraction potential overall. Which aligns with the results of **Figure 18**. Another aspect is that there were temporal differences between the same baits but different dates – C600 20.07.2020 and 21.07.2020 are an example.

Discussion

Automatic predictions and models' performance

Overall, increasing the size of the training effort improved the performance of the models. This is particularly striking when comparing the mAP from the model 0 to model 14. The poor performance of model 0 highlights the importance of case-specific datasets, which allows the algorithm to be specialised in the species of interest and the conditions of the area. Our results pinpoint how species-specific training data control the sensitivity in the performance of a deep learning model.

The increasement of the aP for fish from 0.354 to 0.827 allows us to assume, that with increasing efforts, the models get more performant. This is goes hand in hand with the observations done by Sun *et al.*, 2017 and Joulin *et al.*, 2016 who state that the increase in volume of the data is linked to better visual detection. The drawback is a loss in efficiency, since more complex models require longer processing for the detections process; here, it reached up to several days for one video. Nevertheless, there is a plateau observable for models 8, 11 and 14 and the increase in data is not enough to improve the models since the AP are in similar ranges to each other. This is surprising since the effort to train these models, namely more annotations and more time, was heavily increased, especially for model 11 and 14. With increased training volume, the change in mAP decreased and more annotations and images were needed to further improve the models to new mAP levels.

Another interesting thing about the choice of the classes is that an easy to detect class – in this example the bait – can heavily increase mAP values from 0.827 (without the bait) to 0.922 (with the bait). This

shows that even a metrics routinely used in deep learning like the mAP has flaws and should be interpreted with caution. Indeed, easy or hard to detect classes can inflate or deflate the result and lead to misinterpretation, that is why the individual AP need to be considered too and looked at. Especially when the class of interest is not the one that is easily detectable as presented in Jalal *et al.*, 2020, where *Abudefduf vaigiensis* (45% average) severely underperformed compared to the other classes – *i.e. Chaetodon speculum* (100% average) – and the overall performance of < 90% average.

The effect of adding bait and other particles that increased the noise in fish detections into the training process was an approach to compare how much this differs the AP of the fish. This effect is visible in the OnlyFish model 7 (AP_{fish} = 0.740), which was trained on the same amount of fish annotations as model 5 (AP_{fish} = 0.794) but severally underperformed when compared to each other (**Figure 14**). This is due to the bait and other particles being detected as a fish and this leads to an increase in false positives for the class of fish.

Considering the challenging conditions in the created test video, the AP values can be considered good, when compared with other projects with the same intention. Two examples (Huang *et al.*, 2019b; Politikos *et al.*, 2021) show that AP values ranging from 0.4682 for sea urchins, 0.7979 for sea cucumbers, 0.6339 for scallops, and 0.62 for marine litter can be considered the current state of the art, which is on track with (yet even lower) the results that have been accomplished in this project.

Adding to that, a software bug led to inconsistent validation sets, which, at first, decreased the performance of the models and confused the interpretation of the models' performance. This bug has been fixed in interaction with the developer of the VIAME software, Matt Dawkins (NOAA). Models 1-4 encountered a bug in the conversion of the videos into single-frame pictures, because the extraction led to a small lag in timestamps. The resulting lag was visible in the trained detections by always being a head of the fish (**Figure A1**). This behaviour has also been fixed with the help of Matt Dawkins.

Fish bait attraction levels

While promising, considering only the model's prediction to answer an ecological question remains limited on several aspects. The main issue was that the overall number of fish predicted was inflated compared to the manual counts. This likely results from fish leaving the camera field of view, by remaining in an ambiguous positioning over several frames, or by hiding behind an object and then re-emerging. In this scenario, the model will necessarily predict multiple fish instead of one, which can add up quickly to many fish being double or triple counted. Another aspect leading to overestimation could be the "noise". This is a well-known issue when considering using an artificial intelligence approach (Stekolshchik, 2020). As of today, this issue can only be counteracted with the development of more performant trackers or the implementation of a dynamical error correction factor that changes the number of predicted fish. This overestimation could also result from the lower numbers of the Interest index of the automated data, since the cut-off of certain fish individuals creates multiple fish, which are considered less long in the area of interest and therefore alter the mean index towards lower values.

Another aspect was the aspect of time, since there were great differences between the two analysis methods after only 15min and 59min but if we look at the whole video, the difference gets smaller. This indicates that the model's prediction gets more confident and more precise when looked at longer videos. This indicates that the potential of the created model lies more in the evaluation of larger datasets or longer videos compared and decreases with short duration experiments or essays.

The Interest index showed clear differences in numeric values between bait types for the two methods evaluated but there is a more qualitative approach, which allows to make certain assumptions. The

results indicate that some behaviours can be more easily assessed than others using deep learning. For example, there was a high number of fish for 15min with both methods in the control group, followed by a decline for 59min since the bait was eaten and the fish started only passing through and for the whole video it stayed the same. This indicates that the attraction level for the control falls rapidly after deploying the bait into the water and does not need to stay in the water for several hours. This pattern was accurately depicted by the automated method. Whilst the biodegradable baits had stable and mediocre attraction levels for the first hour, but their attraction potential rose over time until the end of the sampling period. Interestingly, the C17 bait is the one outperforming the control towards the end – this could be due to the mentioned depletion of the physical bait in the control group or due to the effectivity of this type of bait. These biodegradable baits have the advantages of staying in the water longer since it will not be eaten by the fish immediately and could lead to a longer deployment time whilst also safe resources.

5. General Discussion

Thus, even though different in their nature, camera type, environmental conditions and number of classes considered, both projects had similar mean average precision values – 0.827 for BAITFISH and 0.808 for LangolfTV. Both mAP values are considered "good to excellent" (Li *et al.*, 2015). The small difference of these values can be explained by the differences in the training effort and training data quality, as well as in the number of classes predicted. The more classes a model must predict the harder it is to classify it correctly and not as a wrong class (Shahinfar *et al.*, 2020). With this in mind, it is likely that the LangolfTV, which based on the mAP slightly underperformed when compared to the BAITFISH project, is actually the more performant of the two models overall.

Although performing well in terms of mAP values, both projects highlighted the difficulty of answering an ecological question. Immobile and distinctive classes were detected excellent whilst other faster moving, burrowing or similar classes were underperforming due to misclassification. Fish not moving and staying close to the bait structure were detected very well whilst fish that entered the screen rapidly were misclassified or missed. Nevertheless, both projects showed patterns and evidence of being able to answer such questions, which indicate the possibilities to be utilized as a reproducible and constant tool in this area of expertise. For example, in fisheries management, marine policy engagement or conservation of the ocean.

Both these models thus have the potential to be implemented into different aspects and tools to help in marine conservation, fisheries, or scientific research. For example, the LangolfTV model could be used to minimize the bycatch of the bottom trawling technic and move fishing efforts away from areas known for sensitive species and large biodiversity.

6. Conclusion

These two projects show the progress and opportunities arising when using an artificial intelligence approach when looking at ecologically relevant data. It always has been important to look at data to find evidence for a hypothesis proposed. For these two projects a similar approach was chosen to conduct the feasibility of deep learning replacing the human behind the keyboard and it was found that there is a possibility to do so but not for every project.

As the BAITFISH project showed, the algorithm had good prediction numbers with an AP of 0.826 but the number of fish were overestimated due to noise and other factors discussed above. Nevertheless, the results on the interest index showed potential to answer the bait attraction levels in both methods described.

For the LangolfTV project certain classes were predicted good and others were not. There is room for improvement on this model too with more data but the project highlights that the utility of deep learning is within the realm of possibility for continental shelf stock assessments.

7. Appendix

Appendix I – expected curvature of F1 score x threshold

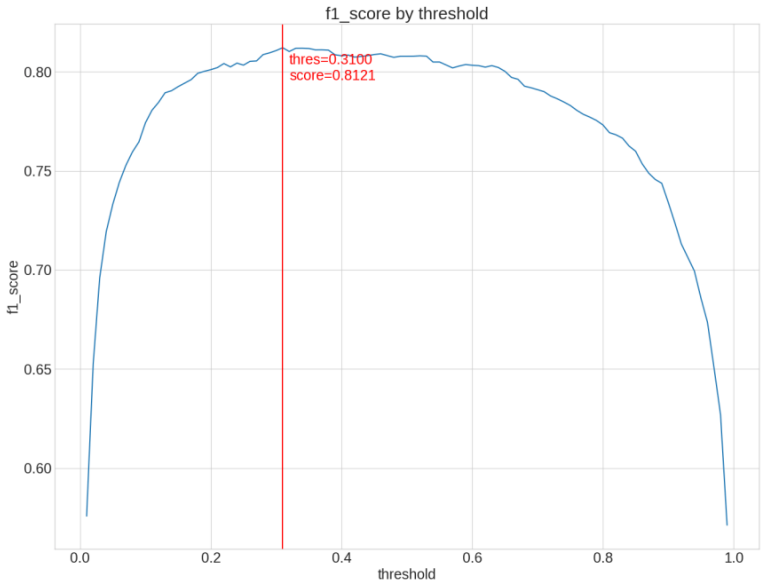

Figure A1, from (Czakon, 2021)

An example of the F1 score/threshold curve applied in this project.

Appendix II – (Mean) Average precision values for all the models of two class analysis

Table A1, values of (mean) average precisions of all the models and IoU thresholds

Model	IoU threshold	aP (fish)	mAP (fish + bait)
0	0.25	0.437	0.339
0	0.5	0.354	0.261
0	0.75	0.050	0.028
5	0.25	0.830	0.910
5	0.5	0.749	0.899
5	0.75	0.555	0.804
7	0.25	0.796	0.461
7	0.5	0.740	0.411
7	0.75	0.374	0.155
8	0.25	0.860	0.931
8	0.5	0.826	0.923
8	0.75	0.615	0.849
11	0.25	0.828	0.909
11	0.5	0.798	0.902
11	0.75	0.583	0.773
14	0.25	0.862	0.930
14	0.5	0.827	0.922
14	0.75	0.607	0.866

Appendix III – VIAME error frame change

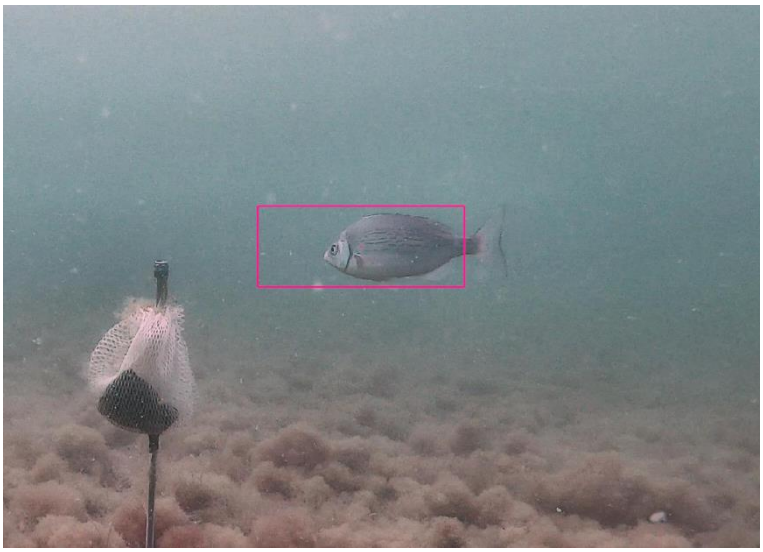

Figure A2, Error encountered when training the data on extracted frames, *the algorithm was trained lagging in front of the fish for most of the time. This was discovered after already training 4 of the models.*

Error encountered during the training of the data because of the frame extraction delay.

Appendix IV - Classes of LangolfTV before fusion of classes

terrier=7663
actiniaria=4228
nephrops_norvegicus=2874
munida=6183
microchirus_variegatus=350
hydrozoa=01494
actinauge=173
gadiforme=318
crevette=2655
brachyura=1086
pennatula_phosphorea=6514
callionymus=400
poisson=59
actinopterygii=952
pennatulacea=626
pleuronectiforme=10
sabellidae=615
ophiuroidea=74
lepidorhombus_whiffiagonis=356
octopodiforme=294
polychaeta=177
bryozoa=13
autre=8
solea_solea=14
Conger=036
echinoidea=15
crinoidea=2485
trisopterus=163
laser=66
virgularia_mirabilis=159
asteroidea=187

paguroidea=10
micromesitius_poutassou=7
scyliorhinus=38
triglidae=35
merluccius_merluccius=42
sepidae=33
sepiidae=1

8. References

Albalat, A., Collard, A., Brucem, C., Coates, C. J., & Fox, C. J. (2016). Physiological Condition, Short-Term Survival, and Predator Avoidance Behavior of Discarded Norway Lobsters (Nephrops norvegicus). *Journal of Shellfish Research*, *35*(4), 1053–1065. https://doi.org/10.2983/035.035.0428

Campbell, N., Dobby, H., & Bailey, N. (2009). Investigating and mitigating uncertainties in the assessment of Scottish Nephrops norvegicus populations using simulated underwater television data. *ICES Journal of Marine Science*, *66*(4), 646–655. https://doi.org/10.1093/icesjms/fsp046

Canli, M., & Furness, R. W. (1993). Toxicity of heavy metals dissolved in sea water and influences of sex and size on metal accumulation and tissue distribution in the norway lobster Nephrops norvegicus. *Marine Environmental Research*, *36*(4), 217–236. https://doi.org/10.1016/0141-1136(93)90090-M

Canli, M., & Furness, R. W. (1995). Mercury and cadmium uptake from seawater and from food by the Norway lobster Nephrops norvegicus. *Environmental Toxicology and Chemistry*, *14*(5), 819–828. https://doi.org/10.1002/etc.5620140512

Christin, S., Hervet, É., & Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution*, *10*(10), 1632–1644. https://doi.org/10.1111/2041-210X.13256

Cojanu, D., & Hugus, E. (2016). HABCAM - HABitat mapping CAMera system is a window to the sea floor. *Https://Www.Whoi.Edu/*. https://www.whoi.edu/oceanus/feature/habcam/

Czakon, J. (2021). *F1 Score vs ROC AUC vs Accuracy vs PR AUC: Which Evaluation Metric Should You Choose?* Neptune.Ai. https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc

Dawkins, M. D., Sherrill, L., Crall, J., Hoogs, A., Zhang, D., Richards, B., Prasad, L., & Williams, K. (2018, February 12). *VIAME: Video and Image Analytics in Marine Environments*. 2018 Ocean Sciences Meeting. https://agu.confex.com/agu/os18/meetingapp.cgi/Paper/321598

Dezecache, G., Zuberbühler, K., Davila-Ross, M., & Dahl, C. D. (2020). A machine learning approach to infant distress calls and maternal behaviour of wild chimpanzees. *Animal Cognition*. https://doi.org/10.1007/s10071-020-01437-5

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, *88*(2), 303–338. https://doi.org/10.1007/s11263-009-0275-4

Faillettaz, R., Blandin, A., Paris, C. B., Koubbi, P., & Irisson, J.-O. (2015). Sun-Compass Orientation in Mediterranean Fish Larvae. *PLOS ONE*, *10*(8), e0135213. https://doi.org/10.1371/journal.pone.0135213

FAO. (2020). *The State of World Fisheries and Aquaculture 2020*. FAO. https://doi.org/10.4060/ca9229en

Fluhr, J., Kopp, D., Robert, M., Morandeau, F., Simon, J., Baudry, J., Ledreau, N., & Méhault, S. (in prep). *How does black sea bream reach the attractive bait? A behavior-based approach*.

Future Market Insights. (2020). *Sea Bream Market Analysis and Review 2019—2029 | Future Market Insights (FMI)*. https://www.futuremarketinsights.com/reports/sea-bream-market

Hagerman, L., & Baden, S. P. (1988). Nephrops norvegicus: Field study of effects of oxygen deficiency on haemocyanin concentration. *Journal of Experimental Marine Biology and Ecology*, *116*(2), 135–142. https://doi.org/10.1016/0022-0981(88)90051-2

Huang, H., Zhou, H., Yang, X., Zhang, L., Qi, L., & Zang, A.-Y. (2019a). Faster R-CNN for marine organisms detection and recognition using data augmentation. *Neurocomputing*, *337*, 372–384. https://doi.org/10.1016/j.neucom.2019.01.084

Huang, H., Zhou, H., Yang, X., Zhang, L., Qi, L., & Zang, A.-Y. (2019b). Faster R-CNN for marine organisms detection and recognition using data augmentation. *Neurocomputing*, *337*, 372–384. https://doi.org/10.1016/j.neucom.2019.01.084

Hui, J. (2018). *MAP (mean Average Precision) for Object Detection*. https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173

Ifremer. (2017). *LangolfTV* [Website]. https://wwz.ifremer.fr/peche/Le-role-de-l-Ifremer/Recherche/Projets/Description-projets/Langolf-TV

Ifremer. (2018). *BAITFISH* [Website]. https://wwz.ifremer.fr/peche_eng/Le-role-de-l-Ifremer/Recherche/Projets/Description-projets/BaitFISH

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

Johnson, M. L., & Johnson, M. P. (2013). *Advances in Marine Biology—The Ecology and Biology of Nephrops norvegicus* (64th ed.). https://books.google.ch/books?id=Efu1M-adVxAC&hl=de&source=gbs_navlinks_s

Johnson, M. P., Lordan, C., & Power, A. M. (2013). Chapter Two—Habitat and Ecology of Nephrops norvegicus. In M. L. Johnson & M. P. Johnson (Eds.), *Advances in Marine Biology* (Vol. 64, pp. 27–63). Academic Press. https://doi.org/10.1016/B978-0-12-410466-2.00002-9

Joulin, A., van der Maaten, L., Jabri, A., & Vasilache, N. (2016). Learning Visual Features from Large Weakly Supervised Data. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 67–84). Springer International Publishing. https://doi.org/10.1007/978-3-319-46478-7_5

Knausgård, K. M., Wiklund, A., Sørdalen, T. K., Halvorsen, K. T., Kleiven, A. R., Jiao, L., & Goodwin, M. (2021). Temperate fish detection and classification: A deep learning based approach. *Applied Intelligence*. https://doi.org/10.1007/s10489-020-02154-9

Kora, H., Tsuchimoto, M., Miyata, K., Osato, S., Wang, Q., Apablaza, P. A. G., Mishima, T., & Tachibana, K. (2000). Estimation of body fat content from standard body length and body weight on cultured red sea bream. *Fisheries Science*, *66*(2), 365–371. https://doi.org/10.1046/j.1444-2906.2000.00056.x

Krishna Chaitanya, G., & Maragatham, G. (2021). Object and Obstacle Detection for Self-Driving Cars Using GoogLeNet and Deep Learning. In D. J. Hemanth, G. Vadivu, M. Sangeetha, & V. E. Balas (Eds.), *Artificial Intelligence Techniques for Advanced Computing Applications* (pp. 315–322). Springer. https://doi.org/10.1007/978-981-15-5329-5_30

Kroodsma, D. A., Mayorga, J., Hochberg, T., Miller, N. A., Boerder, K., Ferretti, F., Wilson, A., Bergman, B., White, T. D., Block, B. A., Woods, P., Sullivan, B., Costello, C., & Worm, B. (2018). Tracking the global footprint of fisheries. *Science*, *359*(6378), 904–908. https://doi.org/10.1126/science.aao5646

Landings of fishery products. (n.d.). http://ec.europa.eu/eurostat.

Leocádio, A. M., Whitmarsh, D., & Castro, M. (2012). Comparing Trawl and Creel Fishing for Norway Lobster (Nephrops norvegicus): Biological and Economic Considerations. *PLOS ONE*, *7*(7), e39567. https://doi.org/10.1371/journal.pone.0039567

Li, X., Shang, M., Qin, H., & Chen, L. (2015). Fast accurate fish detection and recognition of underwater images with Fast R-CNN. *OCEANS 2015 - MTS/IEEE Washington*, 1–5. https://doi.org/10.23919/OCEANS.2015.7404464

Lolas, A., & Vafidis, D. (2021). Population Dynamics, Fishery, and Exploitation Status of Norway Lobster (Nephrops norvegicus) in Eastern Mediterranean. *Water*, *13*(3), 289. https://doi.org/10.3390/w13030289

Maier, A. (2019). A gentle introduction to deep learning in medical image processing. *Z Med Phys*, 16.

Miele, V., Dray, S., & Gimenez, O. O. (2021). Images, écologie et deep learning. *Regards Sur La Biodiversité*. https://hal.archives-ouvertes.fr/hal-03142486

Miller-Greene, D. R., Amin, R., & Taylor, J. C. (2020). *MOUSS protocol for the Pacific Islands Fisheries Science Center*. https://doi.org/10.25923/7Q3T-YK14

Nilsson, N. J. (2009). *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819346

Ovchinnikova, K., James, M. A., Mendo, T., Dawkins, M., Crall, J., & Boswarva, K. (2021). Exploring the potential to use low cost imaging and an open source convolutional neural network detector

to support stock assessment of the king scallop (Pecten maximus). *Ecological Informatics*, *62*, 101233. https://doi.org/10.1016/j.ecoinf.2021.101233

Pakkenberg, B., & Gundersen, H. J. G. (1988). Total number of neurons and glial cells in human brain nuclei estimated by the disector and the fractionator. *Journal of Microscopy*. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2818.1988.tb04582.x

Panda, P., Sengupta, A., & Roy, K. (2016). Conditional Deep Learning for energy-efficient and enhanced pattern recognition. *2016 Design, Automation Test in Europe Conference Exhibition (DATE)*, 475–480.

Parida, V. (2018). *Digitalization. In Addressing Societal Challenges* (pp. 23–38). Luleå University of Technology. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1191622&dswid=181

Perodou, J.-B., & Nedelec, D. (1980). Bilan d'exploitation du stock de Dorade grise. *Science et Pêche*, *308*, 1–7.

Politikos, D. V., Fakiris, E., Davvetas, A., Klampanos, I. A., & Papatheodorou, G. (2021). Automatic detection of seafloor marine litter using towed camera images and deep learning. *Marine Pollution Bulletin*, *164*, 111974. https://doi.org/10.1016/j.marpolbul.2021.111974

Rizzo, E., & Bazzoli, N. (2020). *Hepatosomatic Index—An overview | ScienceDirect Topics*. https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/hepatosomatic-index

Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D., & Carvalho, S. (2019). Chimpanzee face recognition from videos in the wild using deep learning. *Science Advances*, *5*(9), eaaw0736. https://doi.org/10.1126/sciadv.aaw0736

Shahinfar, S., Meek, P., & Falzon, G. (2020). "How many images do I need?" Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring. *Ecological Informatics*, *57*, 101085. https://doi.org/10.1016/j.ecoinf.2020.101085

Stekolshchik, R. (2020). Noise, overestimation and exploration in Deep Reinforcement Learning. *ArXiv:2006.14167 [Cs, Stat]*. http://arxiv.org/abs/2006.14167

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *2017 IEEE International Conference on Computer Vision (ICCV)*, 843–852.

Underwood, A. J., Chapman, M. G., & Connell, S. D. (2000). Observations in ecology: You can't make progress on processes without understanding the patterns. *Journal of Experimental Marine Biology and Ecology*, *250*(1), 97–115. https://doi.org/10.1016/S0022-0981(00)00181-7

Villon, S., Mouillot, D., Chaumont, M., Darling, E. S., Subsol, G., Claverie, T., & Villéger, S. (2018). A Deep learning method for accurate and fast identification of coral reef fishes in underwater images. *Ecological Informatics*, *48*, 238–244. https://doi.org/10.1016/j.ecoinf.2018.09.007

Xu, W., & Matzner, S. (2018). Underwater Fish Detection Using Deep Learning for Water Power Applications. *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, 313–318. https://doi.org/10.1109/CSCI46756.2018.00067