Automated Counting of Fish in moving Diver Operated Videos (DOV) for Biodiversity Assessments

Kilian Bürgi, Rémy Sun, Charles Bouveyron, Diane Lingrand, Benoit Dérijard, Frédéric Precioso, Cécile Sabourault

Automated Counting of Fish in moving Diver Operated Videos (DOV) for Biodiversity Assessments

Kilian Bürgi[a,c], Rémy Sun[c], Charles Bouveyron[b], Diane Lingrand[c], Benoit Dérijard[a], Frédéric Precioso[c], and Cécile Sabourault[a,*]

[*]Corresponding author, Cecile.SABOURAULT@univ-cotedazur.fr

[a]Université Côte d'Azur, CNRS, ECOSEAS, Nice, France

[b]Université Côte d'Azur, Inria, CNRS, Laboratoire J.A.Dieudonné, Maasai team, Nice, France

[c]Université Côte d'Azur, Inria, CNRS, I3S, Maasai team, Nice, France

{Kilian.BURGI,Remy.SUN,Charles.BOUVEYRON,Diane.LINGRAND}@univ-cotedazur.fr

{Benoit.DERIJARD,Frederic.PRECIOSO,Cecile.SABOURAULT}@univ-cotedazur.fr

January 28, 2025

Acknowledgements

This work was only made possible thanks to collaborations with the projects RECIF and FishHealth funded by FEAMPA (Fonds européen pour les aires maritimes, la pêche et l'aquaculture) and the divers involved who provided the diver- and video data from the corresponding field campaigns. The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support. We would also like to thank Catherine for the proofreading of this manuscript and her feedback. This project was funded through the UCAJEDI Investments in the Future project managed by

the National Research Agency (ANR) with the reference number ANR-15-IDEX-01 and through 3IA@cote d'azur - ANR-19-P3IA-0002.

Data Availability

Codes, Scripts and numeric datasets (CSV files) are made available on GitHub:
https://github.com/PiSuMp/fishCount_in_DOV
The training data (images and labels) are made available here:
https://doi.org/10.5061/dryad.f7m0cfz6f

Conflict of Interest

The authors declare that they have no conflicts of interest.

Author Contribution Statements

K. Bürgi, R. Sun, C. Bouveyron, D. Lingrand, F. Precioso, B. Dérijard and C. Sabourault conceived the ideas; K. Bürgi, R. Sun, C. Bouveyron, D. Lingrand and F. Precioso designed the machine learning methodology; C. Sabourault and B. Dérijard collected the data; K.Bürgi, C. Sabourault and B. Dérijard analysed the ecological data; K. Bürgi and R. Sun led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Statement on inclusion

Our study brings together authors from a number of different fields of research, including scientists from the field of computer and data science and marine ecology and conservation. The regional data collection was a collaboration with local stakeholders and NGOs. The data and results are shared in different outreach programs (*i.e.* UNOC 2025) in the native language French as well as English to raise awareness.

Abstract

1 - Underwater video transects are crucial to assess marine biodiversity and biomass. The counting of individual fish in these videos is labour- and time-intensive and operator-dependent. Automating this step would create non-biased biodiversity data and decouple the data collection campaign from any field constraints.

2 - To begin developing an automated process, we assessed the commonly used method for counting objects in videos and compared it to three new methods for counting fish using computer vision derived data (single frame detections) that resulte in a holistic and fully automated pipeline for fish abundance measurements. In addition to the commonly used method N_{max}, we included (i) a 1D k-means clustering method, (ii) an intuitive clustering approach, $N_{Heuristic}$, and (iii) a Temporal Convolutional Neural Network (TCN) counting method. We tested these methods on three Mediterranean species from different ecological niches.

3 - We first assessed the methods using manually labelled detections (groundtruth detections) and then incorporated a detector into the pipeline for a more realistic scenario. The results in these two configurations showed evidence of underestimation by N_{max}. The other methods showed better overall results. The proposed $N_{Heuristic}$ and TCN methods are the closest to manual evaluation. With an absolute variation comparable to inter-operator variation, we demonstrated that these are reliable methods for quantifying fish counts for these three different Mediterranean species.

4 - The parameters of these automated methods could be adjusted to suit other species and then be used in monitoring programs, for example to assess biodiversity and biomass in marine protected areas over time.

Keywords: Artificial intelligence; Diver operated videos; Fish count prediction; Fish monitoring; Machine learning; Marine conservation; Object detection; Temporal Convolutional Network

1 Introduction

The marine environment is facing various factors that pose a critical threat to its inhabitants. Factors such as climate change (Pörtner and Peck (2010)), (over-)tourism (Weng et al. (2023)) and fishing (Bell et al. (2017)) - especially overfishing (Yan et al. (2021)) - are threatening marine species populations (*i.e.* mammals, fish, reptiles and invertebrates) (Dìaz et al. (2019)). To counteract these factors different conservation tools (Hilborn et al. (2020); Calò et al. (2022); Ranganathan et al. (2023)) have been implemented to help preserve populations (Hutchings and Reynolds (2004)). Marine protected areas (MPAs), which function as a safe haven for marine species are among those tools. Inside these areas anthropogenic actions (*i.e.* anchoring or fishing) are limited or prohibited. Assessing the effectiveness of Marine Protected Areas (MPAs) requires efficient, unbiased, and reliable data collection methods to monitor species populations and track their changes over time. Underwater evaluations of fish counts and biomass are two measurements that play a critical role.

A very important indication for the health of an ecosystem is the count of individuals of different species, as these measurements provide valuable insights into population dynamics, species diversity, and the overall balance of the aquatic environment. To count fish in the marine environment, traditional techniques rely on divers collecting biological data in different regions of interest. In these areas, specifically trained experts perform different biodiversity assessments. There are different ways to record this diversity - direct methods such as underwater visual census (UVC) or indirect methods which rely on camera deployment (Stobart et al. (2007); Harmelin-Vivien and Hermalin (2013)).

The traditional way of camera deployment is the Baited Remote Underwater Video (BRUV) approach (Fig. 1-B), involving a stationary camera. To avoid double counting of fish in this setup, the analysis uses only the frame with the highest number of individuals, which is hypothesised to present the relative abundance of the specific replicate in that area. The number of fish in this maximal frame is termed N_{max} or MaxN (Ellis and

DeMartini (1995)) and is the most used metric when it comes to analysis of the BRUV (Schobernd et al. (2014); Haberstroh et al. (2022); Villon et al. (2024)).

The Diver Operated Video (DOV) data collection involves scuba divers or remote operated vehicles (ROV) holding a camera and recording fish that appear and diasppear int he field of view (Fig. 1-A). To evaluate DOV, the metrics are predominantly measured manually by an expert and result in abundance (FishAbundance - Schramm et al. (2020); Maslin et al. (2021); Jessop et al. (2022)) and richness data (Langlois et al. (2010); Grane-Feliu et al. (2019); Raoult et al. (2020)). In DOV, empirical observations suggest that the movement of the diver and the fish are antagonistic and therefore the fish move out of the way and do not re-enter the transect at a later stage of the survey, making the N_{max} metric prone to underestimation (Kilfoil et al. (2017); Sherman et al. (2018)). There is the potential, however, for other methods to result in more precise abundance data (Dickens et al. (2011)).

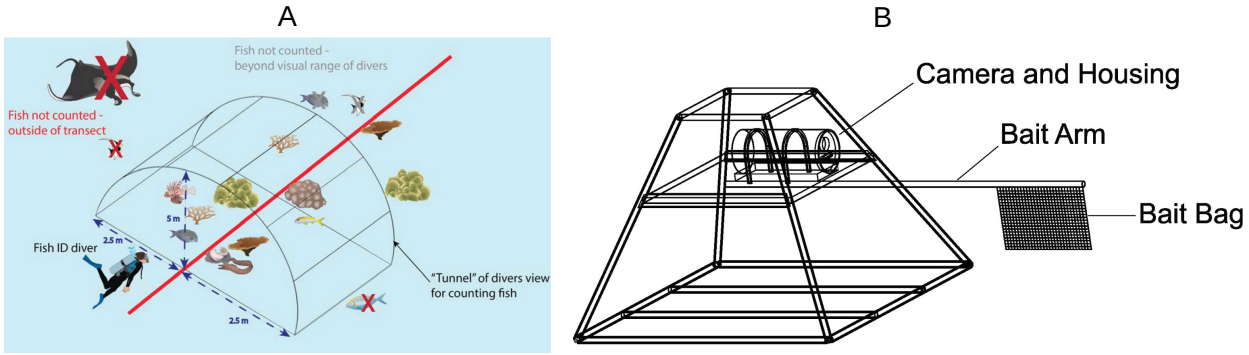

Figure 1: A is an example of a transect setup used in this study with a diver that holding a camera and filming the point of view straight ahead (adapted from Roelfsema et al. (2018)). B explains the concept and construction of a BRUV setup (adapted from Zhang et al. (2024)).

While BRUV and DOV are both widely used, both of these methods require long video analysis times (Schramm et al. (2020)). With advances in technologies in the 21^{st} century,

there is a potential to automate or at least semi-automate the process of video analysis using machine learning methods (Hoekendijk et al. (2021)). The study of Atlas et al. (2023), for example, presented a deep learning multi object tracker for wild salmon. The salmon were tracked and counted as they swam through a one-directional river fence in this successfully automated procedure. Studies combining moving cameras and automation are still scarce. In the study of Connolly et al. (2022), the automated procedure was able to accurately predict the frame with the most individuals in a sequence in nine out of ten videos. These results were comparable to stationary setups (*i.e.*, BRUV). In an example with benthic fish species, Esselman et al. (2025) showed a proof of concept, counting round goby (*Neogobius melanostomus*) in images with a nadir view setup by skipping frames, making it possible to avoid double counting. However, there are still few studies that have focused on the evaluation of moving cameras with pelagic species and the automation of this process.

This study highlights the flaws in N_{max} estimates and proposes three alternative methods that outperform the N_{max} metric in counting the actual fish abundance in an automated manner.

In addition to N_{max} the methods explored were: (i) a 1-dimensional (1D) clustering approach, (ii) an intuitive clustering approach termed $N_{Heuristic}$, and (iii) a Temporal Convolutional Network (TCN) approach. We used the four methods to predict the abundance in 55 videos for three distinctly different Mediterranean species: *Epinephelus marginatus*, *Sciaena umbra* and *Diplodus vulgaris*. The aim of this study was to find a reliable procedure to count objects in single-frame detections with a moving camera and to show the potential of different methodologies to accomplish this task. This will make it possible to rapidly produce fast and non-biased data that can be used for further ecological, economical or conservation analyses.

In this study we provide three main key *contributions*:

- We present the first fully automated pipeline for DOV systems, integrating all steps from video recording to extraction of the fish abundance data.

- We identify critical weaknesses in the widely used N_{max} approach, most importantly its tendency to underestimate fish abundance. To overcome its limitations, we propose three novel methods that significantly reduce underestimation.
- To assess the methods, we established two experimental conditions - theoretical (ideal case) and practical (more realistic automated scenario). The proposed approach provides a robust framework for future studies to assess automated fish abundance measurements.

2 Materials and Methods

In this section we show how we automated the counting of three Mediterranean fish species in underwater videos with three novel methods that have not been explored before. We first present the study area and data collection specifications (see Sec. 2.1), species of interest (see Sec. 2.2), how we processed the videos (see Sec. 2.3) and then give more insights into the different methods (see Sec. 2.4), to enable this study to be reproduced for more locations and species.

2.1 Study area and data collection

To cover a great area and wide variety of habitats, we collected videos in eight different locations of the French Riviera in the Mediterranean Sea in standardized conditions (Harmelin-Vivien et al. (1985)). The depth ranged from 1 to 37m and data was collected in 2022 at different times of the year (cold- and warm season). Camera-equipped divers did 3 transects of 125 m^2 surface per dive. For the recording of the videos, clipboard-mounted GoPro HERO 9 cameras were used. These videos were recorded with a framerate of 24 frames per second (FPS) and full high definition resolution (1920x1080 px). Frames were extracted from these recordings with a framerate of 1 FPS.

2.2 Species of Interest

We observed a wide variety of fish species in the videos, from which we chose three for this study: *E. marginatus*, *S. umbra* and *D. vulgaris*. We chose them because they occupy

distinct ecological niches that are different enough to test these new methods, and ensure they will have a broader applicability to other environments.

The most emblematic species of the French Mediterranean Sea is the endemic dusky grouper (*E. marginatus* - Fig. 2). It falls into the ecological niche of a solitary predator species. This species is interesting since it has been overfished for decades and subjected to protection efforts since 2003 (Pollard et al. (2018)), now showing signs of recovery. Since this species has only been recently protected, knowing the evolution of populations of this species in a temporal and spatial manner is extremely important.

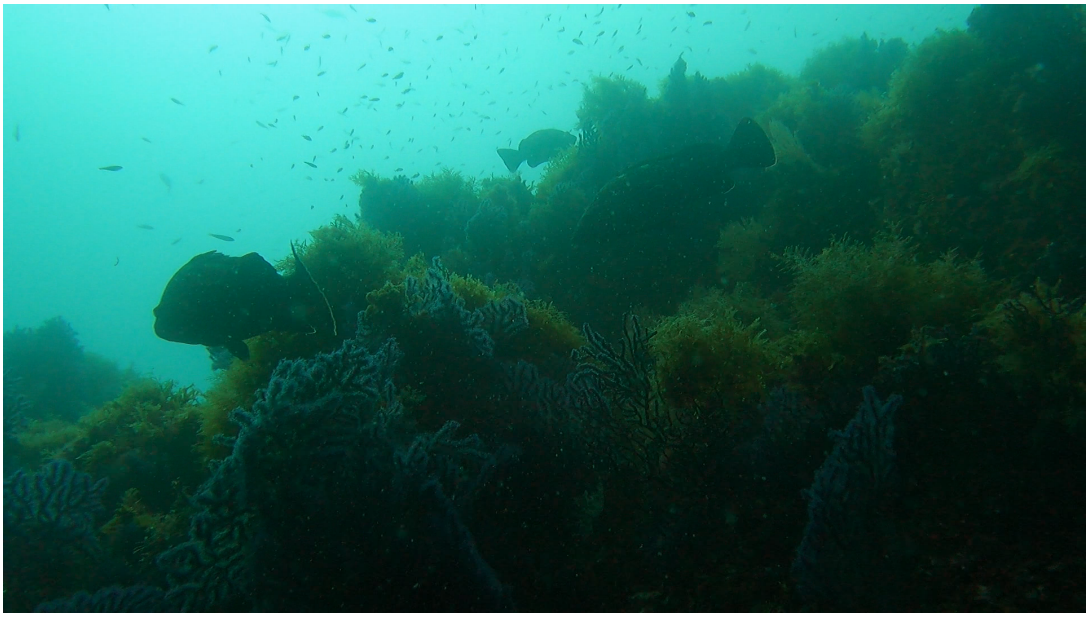

Figure 2: Unedited example image from a transect of three *E. marginatus* individuals in the centre of the image. Conditions are variable in the frames and make the detection more difficult.

The second species, the brown meagre (*S. umbra* - Fig. 3) is also protected in French waters (Prefectoral orders number 2013357-0002 for Corsica and number 2013357-0007 for continental coast). The population is in decline (Harmelin-Vivien et al. (2015)) and therefore it is important to keep track of these fish. They hunt in schools of multiple individuals and thus fill a different ecological niche, enabling us to assess our methods.

For the third species we chose a more abundant species, which is present in more videos and in a greater number of high occurrence videos, the common two-banded sea bream or *D. vulgaris* (Fig. 3). This species lives in large schools above the seabed, scavenging for food. They were found in many of the transects evaluated and therefore present a greater challenge for the methods. The abundance varies from one or two individuals to 50 or more. This different ecological niche will demonstrate the strength and weaknesses of the different methods.

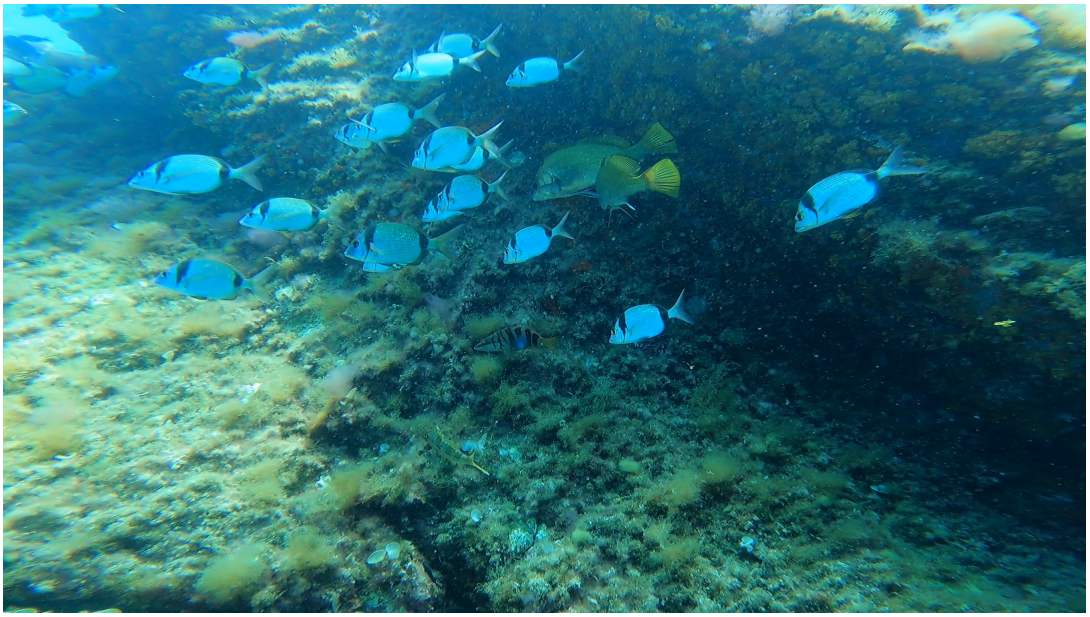

Figure 3: Unedited example image of the transect with around 20 *D. vulgaris* individuals and two *S. umbra* among the *D. vulgaris* school.

2.3 Obtaining data from the videos

The videos provided sequential frames, forming temporal time series. This chronological arrangement allowed us to create 1-dimensional histograms of each video and species (see Sec. 2.3.1). These histograms subsequently served as inputs for the analytical methods (see Sec. 2.4), ultimately giving species-specific counts for each video (see Sec. 3) as an output. The inference pipeline is shown in Figure 4.

To test the robustness of our methods, we used two types of data as the input to each method, the first a theoretical perfect case and the second a more realistic practical

scenario. In the perfect case (see Sec. 3.1), we used the groundtruth detections to verify the feasibility of the methods proposed, thus ensuring 100 % of the detections were correct and no potential error was introduced by a faulty detector. In the practical or fully automated case (see Sec. 3.2), we used the predictions of the detector to see the impact on each method of using a detector in the pipeline. For the output of the methods we wanted to estimate the True FishAbundance. We refer to the method-estimated counts as 'Estimated FishAbundance' henceforth.

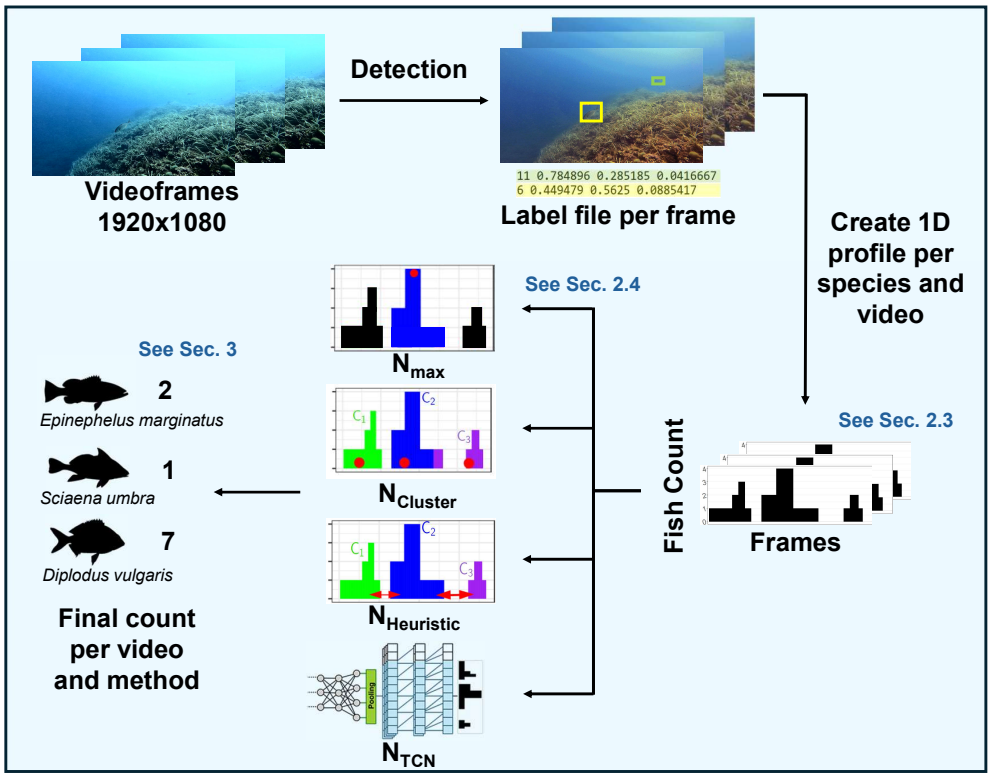

Figure 4: Workflow of the automated pipeline.

2.3.1 Detector training and input data

To detect automatically which species are present in which videos, we used a deep learning approach to make predictions on the data. For the detector, we used a slight variation of the model described in a previous study (Bürgi et al. (2024)). We kept the hyperparameters constant but moved seven videos from the training to the validation set for the detector. We used the validation set to find the f1 score per species for the fully automated case. To enrich the test dataset and assess the methods on high abundance videos, we excluded five

high occurrence videos from training and added them to the test dataset. To analyse these detections, fish counts were aggregated by species and frame, resulting in a one-dimensional time series representing species abundance throughout each video (Fig. 5).

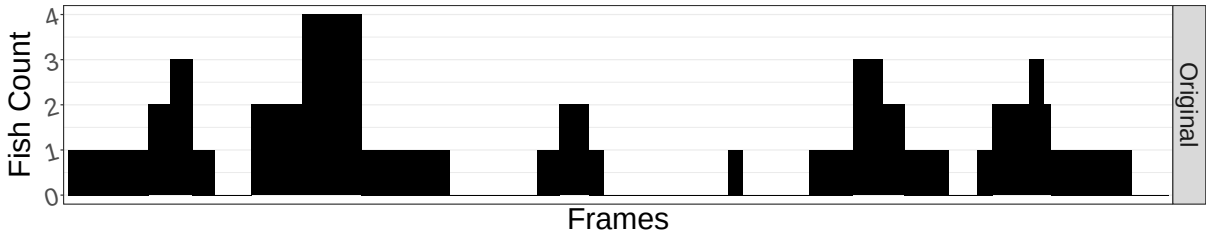

Figure 5: An example of the representation of one fish species in one video: the number of fish from this species is counted per frame, manually or automatically. Each species in each video is represented as this 1D series of values.

The one-dimensional time series (Fig. 5) contain different abundances per species. The training and validation detections for the detector model also form the training and calibration dataset for the counting methods (ii) and (iii). The test set videos were held constant across the methods to have a fair comparison. Table 1 provides more information on the videos used in the TCN training and the $N_{Heuristic}$ calibration.

Table 1: The dataset used for the training of the TCN and the $N_{Heuristic}$ calibration. The training and testing videos are the same for all species to enable a fair comparison. The occurrences differ between the species enabling us to test the methods. The train (train and val combined) and test split used in the detector training are held constant to evaluate the count methods. The number of videos in which there were no occurrences and at least one occurrence are given in columns four and five.

Species	**Category**	**Total Videos**	**0 Videos**	**≥1 Videos**	**Occurrences**
Sciaena umbra	Training	119	112	7	49
	Testing	55	46	9	33
Epinephelus marginatus	Training	119	97	22	55
	Testing	55	36	19	56
Diplodus vulgaris	Training	119	76	43	259
	Testing	55	27	28	334

2.3.2 True FishAbundance

To evaluate the methods, we compared the output of each method with the actual count per video. For this purpose, a marine biology expert counted the actual fish abundance (True FishAbundance) per video, resulting in a groundtruth count per video. When in doubt a second count was made by a different expert and discussed.

2.3.3 Evaluation metrics

To evaluate the accuracy of the different count methods and their ability to grasp the actual biodiversity, we introduced different metrics. The first metric is the absolute error (AE - Eq. 1) which enables a direct comparison of our proposed methods to the N_{max} method.

$$\text{AE} = |\text{True FishAbundance} - \text{Estimated FishAbundance}| \tag{1}$$

The absolute percentage error (APE - Eq. 2) allows a relative comparison between

other methods not evaluated in this study.

$$\text{APE} = \left(\frac{|\text{True FishAbundance} - \text{Estimated FishAbundance}|}{|\text{True FishAbundance}|} \right) \times 100 \tag{2}$$

To have an idea of hte linear relationship between the true FishAbundance (manual) and the estimated FishAbundance (automated), we calculated the Pearson correlation coefficient under different exclusion criteria: (1) all videos included (Corr_{All}), (2) excluding videos with zero counts ($\text{Corr}_{w/o0}$), (3) excluding videos with counts of zero or one ($\text{Corr}_{w/o01}$), and (4) excluding videos with counts in the range of zero to ten ($\text{Corr}_{w/o0:10}$).

2.4 Counting Methods

We wanted to show the risks and flaws of using N_{max} in a DOV setup and use N_{max} as the baseline for our three improved methods. Previous studies have shown underestimation of true fish abundance in videos when utilising the N_{max} metric (Schobernd et al. (2014); Campbell et al. (2015); Sherman et al. (2018)). We introduce 3 novel methods besides the commonly used N_{max}, to find the most suitable count method for the different ecological niches of fish. These three methods are - (i) 1D clustering termed $\text{N}_{Cluster}$, (ii) the manual $\text{N}_{Heuristic}$ and (iii) a Temporal Convolutional Network (TCN) approach termed N_{TCN} to evaluate the fish abundance.

2.4.1 N_{max}

As a baseline we used the commonly used N_{max} to find the abundance in the videos. N_{max} uses a snapshot of the sequence with the highest count of individuals and uses this count as the sequence abundance (Eq. 3).

$$N_{max} = max\{N_f\}, \quad f = 1, 2, ..., F \tag{3}$$

Where:

- N_f: Number of individuals counted in frame f.
- F: Total number of frames in the video.

2.4.2 $\mathbf{N}_{Cluster}$

Since N_{max} only counts one peak per video, information before and after this peak is lost and not incorporated into the count. Using one value per video is not ideal and we thought of a potentially better approach. The different groups in the 1-dimensional profile (Fig. 5) are hypothesised to be different schools so summing the maximum of each of these clusters refines the count per video. A clustering approach will work well to determine clusters from the 1-d profile of detections derived from each video.

Generally speaking, a k-means clustering approach groups the sequences into k clusters so that a cost is minimized. The challenge with k-means clustering is to find the correct value of k. For this purpose we used the R package *Ckmeans.1d.dp* (Wang and Song (2011)) that clusters 1-dimensional data dynamically into different clusters. We provided a range of k (1 to 10) since never more than 10 schools of fish were observed - this range needs to be adjusted according to each individual problem. For each sequence or video, the ideal k was found. The peaks of all clusters were then summed to form a better representation of the fish count over time (Eq. 4).

$$N_{Cluster} = \sum_{j=1}^{N_{Clus}} max\{C_j\} \tag{4}$$

Where:

- N_{Clus}: Total number of clusters identified in the video.
- C_j: Cluster j
- j: Cluster index (1,2,...,j)

2.4.3 $\mathbf{N}_{Heuristic}$

The k-means clustering method used for $N_{Cluster}$ relies on statistical principles that may not align with how a human would intuitively approach the problem. Therefore, we simplified the problem and we were able to adopt a natural and intuitive solution to differentiate

between the various fish groups in the videos. We introduced $N_{Heuristic}$ (Eq. 5), a method that employs inter-school distances as a species-specific differentiator.

This method used the relatively consistent characteristic distance between schools observed for each species, allowing more precise school differentiation based on this distance. The different clusters were differentiated by two variables that were calibrated on the training dataset. The variable *threshold* referred to the minimum count for a cluster to be considered a school, this was introduced to counteract always occurring species. The second variable, *n_frames* referred to a delay between schools before a new school was identified. The maxima of each school were then summed to get an improved count of the fish individuals in each transect video.

$$N_{Heuristic} = \sum_{j=1}^{N_{Schools}(n_{frames},threshold)} max\{C_j\} \tag{5}$$

Where:

- $N_{Schools}$: Total number of clusters identified in the video.
- n_{frames}: Frame delay between two clusters
- $threshold$: Minimum individual count for a cluster to be valid
- C_j: Cluster j
- j: Cluster index (1,2,...,j)

2.4.4 N_{TCN}

Clustering methods typically that the number of individuals within a fish school remains constant. However, fish schools are dynamic systems where individuals frequently join or leave. The proposed clustering methods do not account for the dynamic nature of this group composition, which may affect the accuracy of fish counts. With the rise of neural networks (NN) in recent years, it is possible to use an NN to account for this more dynamic and complex behaviour of the fish. This is why we introduced a Temporal Convolutional Network (TCN, Bai et al. (2018)) as a third method.

A TCN is a Convolutional Neural Network (CNN) but excels in utilising temporal data (*i.e.* time series). The two main advantages of TCN are 1) the property to keep temporal information between the datapoints (*i.e.*, timepoint_0, timepoint_1 and timepoint_n) and 2) that it is parameter-efficient, making it well-suited for scenarios where data is limited. These advantages led to the decision to utilise a TCN for this study. The sequences of counts were prepared to fit the input format of the TCN (predictor = sequence of counts, target = $\text{N}_{TCNSpecies1}$, $\text{N}_{TCNSpecies2}$, $\text{N}_{TCNSpecies3}$). We trained the TCN model on batches of size 64 using a stochastic gradient descent (SGD) optimisation function, a learning rate of 0.01 and trained for a total of 1,250 epochs. Five independent trainings were conducted and the average is presented with the corresponding standard deviation. For graphical representation, we chose the model that had the lowest absolute error on the test set. The training and validation loss curve can be seen in Figures S1 and S2. The architecture can be seen in Table S1 with 3,713 trainable parameters. We called the predicted video counts 'N_{TCN}' henceforth.

3 Results

In this section, we show the outputs of the different methods. We commence with the perfect case (see Sec. 3.1) and then use the fully automated case (see Sec. 3.2) to test our methods. The methods and species follow the same order to aid readability in this section.

3.1 Perfect case testing methods on groundtruth test labels

In this first case we tested how the fish count was impacted solely by the methods used and not by the object detection task. We used the groundtruth labels on the test set to assess the performance without the impact of the detector performance.

3.1.1 *Epinephelus marginatus*

We first investigated the species *E. marginatus*. It is a relatively rare species and videos with a high occurrence of this species are scarce. In all test videos, we observed a total of

56 individuals with the majority being in multiple single occurrence videos. In Table 2 we can see that all methods perform better than N_{max} in all metrics provided. The best performing is the $N_{Heuristic}$ method with an absolute error (AE) of 13 or percentage error of 23 % over- or under-estimation. If we exclude the videos with 0 or 1 occurrences, the correlation decteases below 0.60 for N_{max} while it stays constant above 0.6 for the other methods. The exclusion results in a reduction of the correlation coefficient for N_{max} from 0.897 to 0.544, whereas for $N_{Heuristic}$ also it decreases, but to a much lesser extent, from 0.957 to 0.820.

Table 2: The different metrics for the different methods are presented for the species *E. marginatus*. The correlations between the estimated counts and the actual counts on the test set are indicated with all points included (All), 0 excluded (w/o0) and 0 and 1 excluded (w/o01) to show the robustness of the methods in high occurrence videos. Percentage values were rounded to have 0 decimals. For N_{TCN} the standard deviation was calculated for the 5 replicates we trained.

Method	**AE**$_{All}$	**APE**$_{All}$	**Corr**$_{All}$	**Corr**$_{w/o0}$	**Corr**$_{w/o01}$
N_{max}	18	32%	0.905	0.752	0.544
$N_{Cluster}$	**13**	**23%**	0.942	0.899	0.751
$N_{Heuristic}$	**13**	**23%**	**0.957**	**0.901**	**0.820**
N_{TCN}	**15±2**	**27±4%**	0.932±0.012	0.841±0.030	0.701±0.034

The visual representation of the counts (Fig. 6) showed a clear underestimation of the counts with N_{max} while it is much more stable with the other three methods. We can see that with an increase in occurrence of the species in the videos, our methods handle these cases much better than the more commonly used N_{max}. The difference to the ideal line shows that none of the methods shows a perfect result but the trend is towards less miscounting with the new methods.

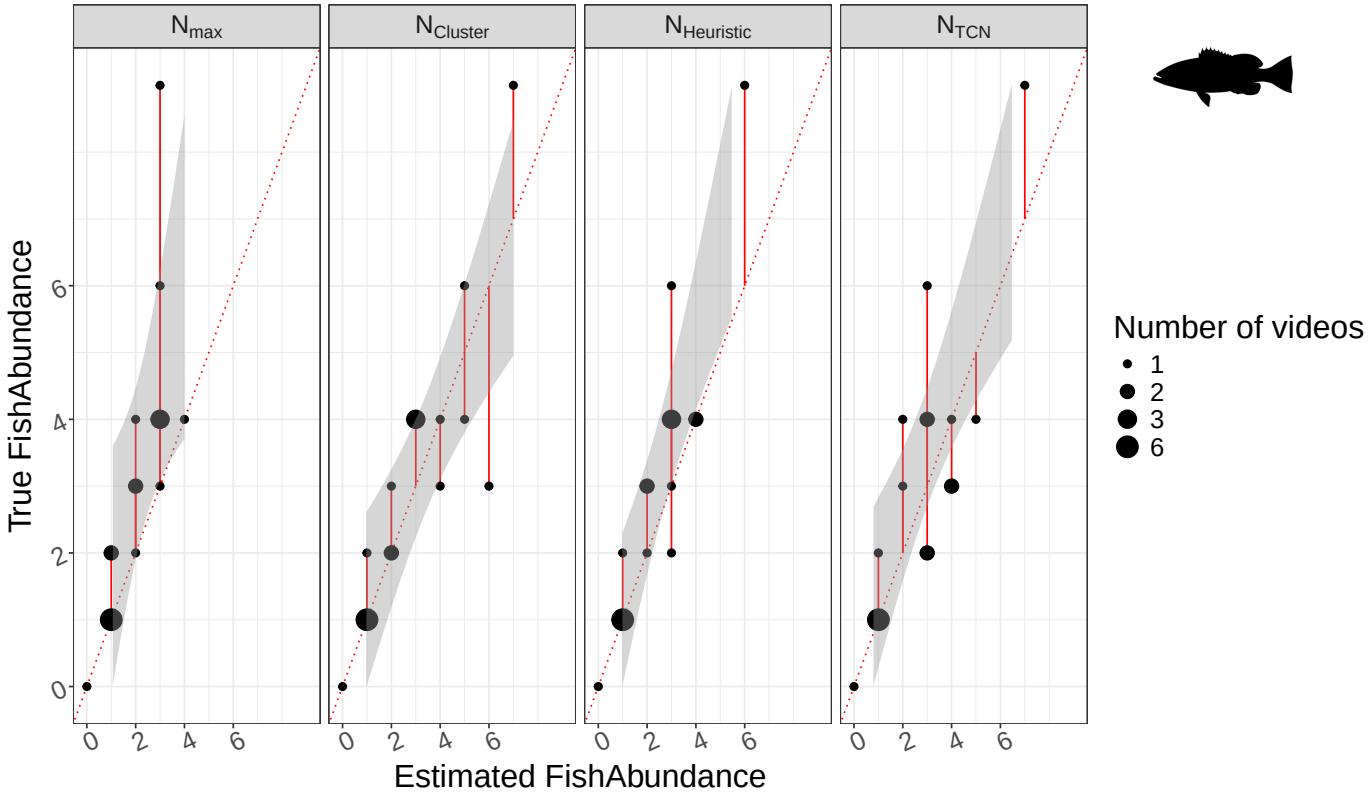

Figure 6: Linear regression (grey area indicates the 95% confidence interval) between the actual number of individuals of *E. marginatus* in test videos (y-axis) and the estimated count by the different methods (x-axis). The size of the points indicates more videos overlapping with the corresponding methods estimated count and the actual count. The dashed red line indicates a perfect result. Underestimations are seen above the line and overestimations are seen below the line. The point size of 0,0 was reduced to 1 for the graphical representation. A total of 19 videos had 56 *E. marginatus* present.

3.1.2 *Sciaena umbra*

The second species we investigated was *S. umbra*. This species is also rare but appears in larger schools of up to 20 individuals. We observed this species in 9 test videos. All three methods showed high correlation values and low miscounts of 21 % or lower (Table 3), making any of them suitable to count predatory schooling species. N_{max} fails to count the absolute fish abundance and 33 % of the individuals are miscounted. Correlation values significantly drop from 0.766 to 0.382 when the lower occurrence videos are removed. The best performing method is N_{TCN} with only 15% of the fish being miscounted and

correlation values of 0.975 even with the low occurrence videos excluded.

Table 3: The different metrics for the different methods are presented for the species *S. umbra*. The correlations with the actual counts on the test set are indicated with all points included (All), 0 excluded (w/o0) and 0 and 1 excluded (w/o01) to show the robustness of the methods in high occurrence videos. Percentage values were rounded to have 0 decimals. For N_{TCN} the standard deviation was calculated for the 5 replicates we trained.

Method	**AE**$_{All}$	**APE**$_{All}$	**Corr**$_{All}$	**Corr**$_{w/o0}$	**Corr**$_{w/o01}$
N_{max}	11	33%	0.766	0.450	0.382
$N_{Cluster}$	7	21%	0.966	0.934	0.929
$N_{Heuristic}$	**6**	**18%**	0.976	0.966	0.965
N_{TCN}	**5±1**	**15±3%**	**0.987±0.006**	**0.978±0.010**	**0.977±0.011**

We finally looked at the visual representation of the count data for the different methods(Fig. 7). First, the low correlation values generated by N_{max} depend on only one video that has more than six occurrences. This video is better counted with the other methods and therefore leads to the higher correlation values for these methods. This gives an indication how the different methods can outperform N_{max} on high occurrence videos, which N_{max} struggles with. While N_{TCN} has an absolute error of zero in this specific video, the other methods struggle with errors ranging from 1 for $N_{Cluster}$ to 4 $N_{Heuristic}$.

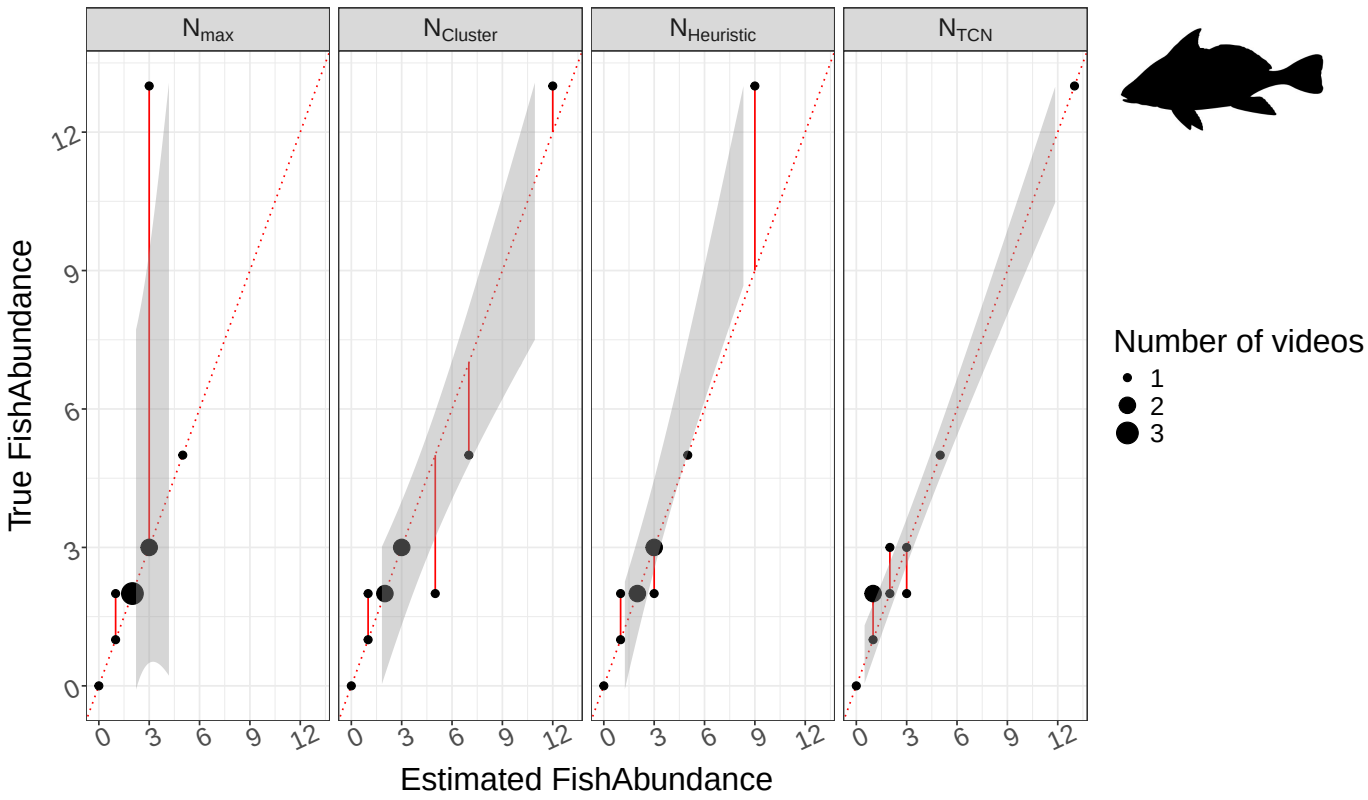

Figure 7: Linear regression (grey area indicates the 95% confidence interval) between the actual number of individuals of *S. umbra* in test videos (y-axis) and the counts of the different methods (x-axis). The size of the points indicates more videos overlapping with the corresponding methods estimated count and the actual count. The dashed red line indicates a perfect result. Underestimations are seen above the line and overestimations are seen below the line. The majority of the videos (n=45) were at point 0,0. The point size of 0,0 was reduced to 1 for the graphical representation. A total of 9 videos had 33 *S. umbra* present.

3.1.3 *Diplodus vulgaris*

The last species that we looked at was the schooling fish and commonly seen *D. vulgaris* (Table 4). The expected increase in number of individuals was thereby used to test the methods. This increase in numbers had the biggest impact on two methods for this species, namely N_{max} and $N_{Cluster}$. With error rates of 40 % the counting of this species was inadequate. However, for the other two methods the error rate is halved and is around 20 % for $N_{Heuristic}$ and N_{TCN}. All of the proposed methods have a correlation over 0.90.

When we excluded the videos with 10 or less individuals, the correlation for $N_{Heuristic}$ and N_{TCN} stayed above 0.90, which further underscores the broad applicability of these methods for different ecological niches.

Table 4: The different metrics for the different methods are presented for the species *D. vulgaris*. The correlations with the actual counts on the test set are indicated with all points included (All), 0 excluded (w/o0), 0 and 1 excluded (w/o01) and videos with less than 10 individuals excluded (w/o0:10) to show the robustness of the methods in high occurrence videos. Percentage values were rounded to have 0 decimals. For N_{TCN} the standard deviation was calculated for the 5 replicates we trained.

Method	**AE**$_{All}$	**APE**$_{All}$	**Corr**$_{All}$	**Corr**$_{w/o0}$	**Corr**$_{w/o01}$	**Corr**$_{w/o0:10}$
N_{max}	135	40%	0.907	0.882	0.859	0.718
$N_{Cluster}$	130	39%	0.94	0.925	0.91	0.822
$N_{Heuristic}$	**64**	**19%**	**0.991**	**0.988**	**0.986**	**0.980**
N_{TCN}	**67±12**	**20±4%**	0.980±0.004	0.975±0.004	0.968±0.006	0.937±0.009

The smaller number of no occurrence videos made data more available and favoured the two methods that needed training or calibration. This is clearly visible in the graphical representation (Fig. 8) of the FishAbundance. Both better performing methods seem to underestimate the count a little but keep the distance to the perfect dashed red line as small as possible. $N_{Cluster}$ overestimates the majority of the videos that contain 20 or more fish, which seems to be a limit to this method. On the other hand, N_{max} underestimates the count in all videos and the majority of the miscounting occurs in the videos that contain more than 15 individuals, which seems to be the limit of this method.

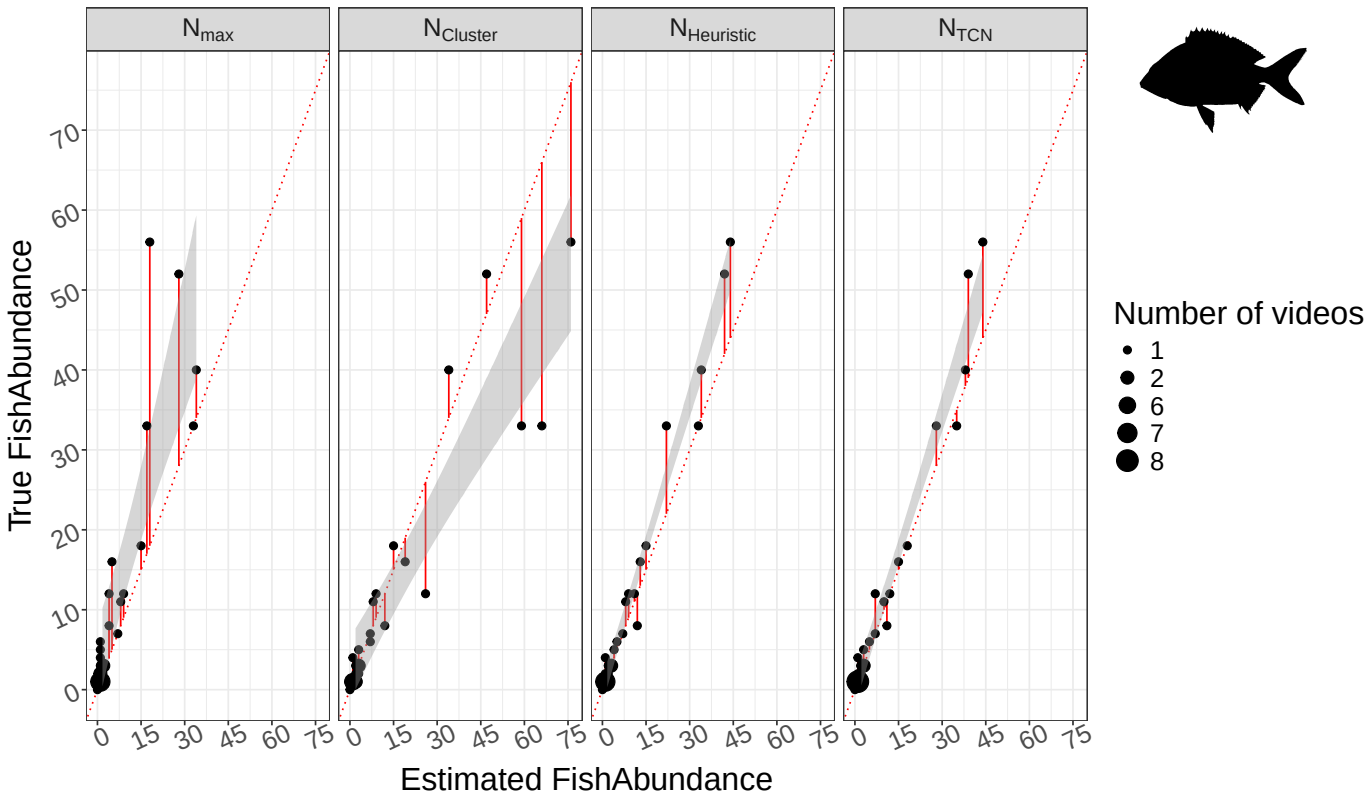

Figure 8: Linear regression (grey area indicates the 95% confidence interval) between the actual number of individuals of *D. vulgaris* in test videos (y-axis) and the counts of the different methods (x-axis). The size of the points indicates more videos overlapping with the corresponding methods estimated count and the actual count. The dashed red line indicates a perfect result. Underestimations are seen above the line and overestimations are seen below the line. The videos at point 0,0 (n=25) were reduced in their point size to 1 for the graphical representation. A total of 28 videos had 334 *D. vulgaris* present.

3.2 Fully automated case testing methods on detections

In this section we explored the impact of utilising a detector and its detections instead of the groundtruth labels. It is important to assess real world applications of the problem and see the feasibility with an imperfect detector with potential for improvement. For each species we found the best performing confidence threshold based on the respective f1 score on the validation set. We determined the confidence thresholds as follows, 0.55 for *E. marginatus*, 0.60 for *S. umbra* and 0.45 for *D. vulgaris*.

3.2.1 *Epinephelus marginatus*

Accurately determining the counts of *E. marginatus* is crucial, even when using a detector system. This ensures that newly recorded data can be reliably evaluated and closely reflects actual population dynamics and distribution. We see an increased error for all the methods (Table 5) in comparison with the perfect case (Table 2). The effect of this imperfection is greater on the correlation of N_{max} than the other methods that continue to present values above 0.750 while N_{max} drops to 0.444 for $Corr_{w/o01}$. Most of these errors derived from false positive counts in no and one occurrence videos (except N_{max}). This is observable for both the rarer species since the contribution of the low occurrence videos is bigger than for the more common *D. vulgaris*.

Table 5: The different methods tested on the detector predictions are presented for the species *E. marginatus*. The correlations with the actual counts on the test set are indicated with all points included (All), 0 excluded (w/o0) and 0 and 1 excluded (w/o01) to show the robustness of the methods in high occurrence videos. Percentage values were rounded to have 0 decimals. For N_{TCN} the standard deviation was calculated for the 5 replicates we trained.

Method	**AE**$_{All}$	**APE**$_{All}$	**Corr**$_{All}$	**Corr**$_{w/o0}$	**Corr**$_{w/o01}$
N_{max}	34	61%	0.800	0.710	0.444
$N_{Cluster}$	29	52%	0.876	**0.928**	**0.882**
$N_{Heuristic}$	**19**	**34%**	**0.906**	**0.894**	**0.790**
N_{TCN}	**21±6**	**38±11%**	**0.913±0.034**	**0.899±0.045**	**0.826±0.078**

In Figure 9 the over- or under- estimation is presented. We can see that N_{max} and $N_{Heuristic}$ both tend to underestimate (with varying effect) the count. The biggest error is observable here with the false positives on the horizontal line of $y = 0$. Trends of $N_{Cluster}$ and N_{TCN} show a clear indication that the performance is better than N_{max}. $N_{Heuristic}$ has lower error rates due to fewer false positives being counted towards the abundance. This can be seen numerically in Table 5.

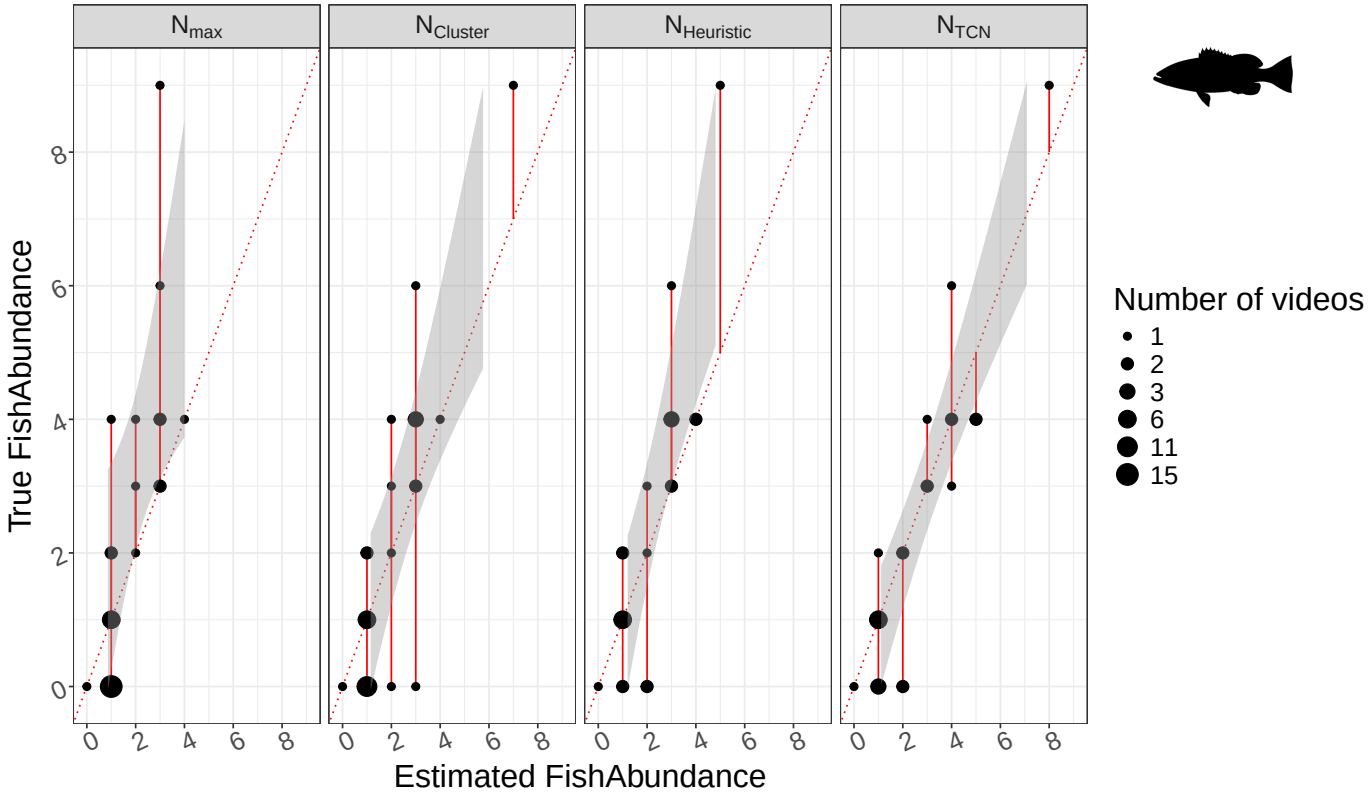

Figure 9: Linear regression (grey area indicates the 95% confidence interval) between the actual number of individuals of *E. marginatus* in test videos (y-axis) and the estimated count by the different methods (x-axis) on the detector predictions. The size of the points indicates more videos overlapping with the corresponding methods estimated count and the actual count. The dashed red line indicates a perfect result. Underestimations are seen above the line and overestimations are seen below the line. The point size of 0,0 was reduced to 1 for the graphical representation. A total of 19 videos had 56 *E. marginatus* present.

3.2.2 *Sciaena umbra*

The biggest difference between the perfect and the fully automated case can be seen for *S. umbra* (Tables 3 and 6). The results for the perfect case can be considered very good with low error rates, while using the detector increased the error by up to 55% for N_{max} and up to 49% for the other methods. The most stable results were obtained with the $N_{Heuristic}$ method, with an increase of 37% from 18% to 55%. This can be explained by an insufficient detection capability for this species in the test dataset. Correlation values

remained above 0.9 for the proposed methods, even when low occurrence videos were excluded. In contrast, for N_{max}, the correlation reached 0.812 under the same exclusion conditions. These results are subjected to caution since the sample dataset is low with only 9 videos for this species.

Table 6: The different methods on the detector predictions are presented for the species *S. umbra*. The correlations with the actual counts on the test set are indicated with all points included (All), 0 excluded (w/o0) and 0 and 1 excluded (w/o01) to show the robustness of the methods in high occurrence videos. Percentage values were rounded to have 0 decimals. For N_{TCN} the standard deviation was calculated for the 5 replicates we trained.

Method	**AE**$_{All}$	**APE**$_{All}$	**Corr**$_{All}$	**Corr**$_{w/o0}$	**Corr**$_{w/o01}$
N_{max}	29	88%	0.727	0.784	0.812
$N_{Cluster}$	23	70%	0.903	0.924	0.918
$N_{Heuristic}$	**18**	**55%**	**0.931**	**0.966**	**0.963**
N_{TCN}	**18±4**	**55±12%**	**0.930±0.017**	**0.963±0.014**	**0.961±0.015**

For *S. umbra*, the false positive rate is the highest, as clearly illustrated in the graphical representation (Fig. 10). The false positives on y = 0 (equivalent to $w/o0$) ranged from 11 individuals for $N_{Cluster}$ to 6 for $N_{Heuristic}$ (N_{max} = 9, N_{TCN} = 7). This shows that the N_{TCN} and $N_{Heuristic}$ are more robust against false positives but are still affected by the inclusion of a detector in the process. The single video containing more than 10 individuals contributed significantly to the error in N_{max}, favouring the new proposed methods.

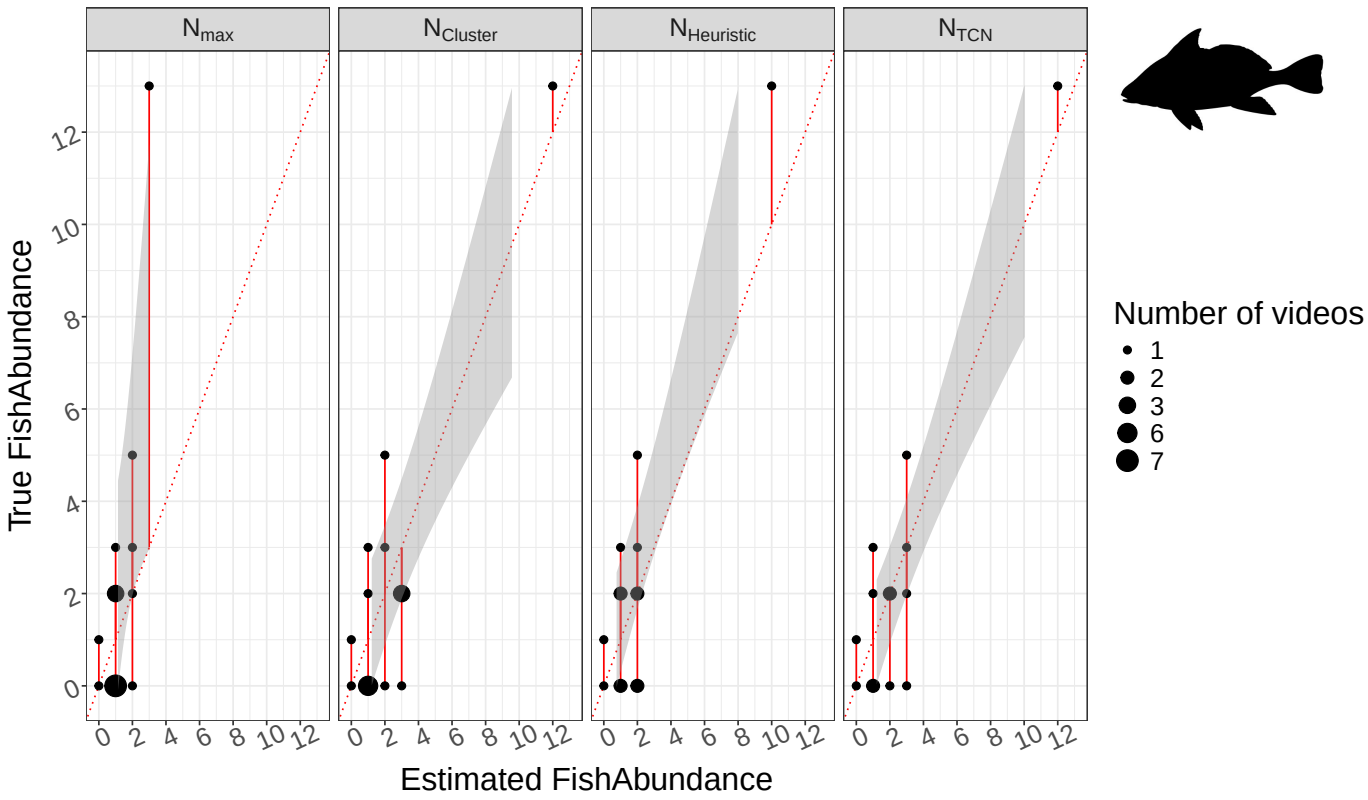

Figure 10: Linear regression (grey area indicates the 95% confidence interval) between the actual number of individuals of *S. umbra* in test videos (y-axis) and the counts of the different methods (x-axis) on the detector predictions. The size of the points indicates more videos overlapping with the corresponding methods estimated count and the actual count. The dashed red line indicates a perfect result. Underestimations are seen above the line and overestimations are seen below the line. The majority of the videos (n=45) were at point 0,0. The point size of 0,0 was reduced to 1 for the graphical representation. A total of 9 videos had 33 *S. umbra* present.

3.2.3 *Diplodus vulgaris*

The results obtained for the third species showed the least change in error rates for the three species between the perfect and fully automated scenario (Table 7), ranging from -4% (false negatives decreasing the count to a better result) for $N_{Cluster}$ to 14% for $N_{Heuristic}$. This stability may be attributed to the increased number of individuals, which not only enhanced counting accuracy but also improved the training effectiveness of the deep learning model. The error percentage remains below 40% for all our proposed methods

while for N_{max} it is 50%. In most cases, correlations remain above 0.9, both with and without exclusions. However, when the occurrence range of 0 to 10 is excluded, correlations for $N_{Cluster}$ and N_{max} drop below 0.9 while $N_{Heuristic}$ and N_{TCN} stay above this value.

Table 7: The different methods on the detector predictions are presented for the species *D. vulgaris*. The correlations with the actual counts on the test set are indicated with all points included (All), 0 excluded (w/o0), 0 and 1 excluded (w/o01) and videos with less than 10 individuals excluded (w/o0:10) to show the robustness of the methods in high occurrence videos. Percentage values were rounded to have 0 decimals. For N_{TCN} the standard deviation was calculated for the 5 replicates we trained.

Method	**AE**$_{All}$	**APE**$_{All}$	**Corr**$_{All}$	**Corr**$_{w/o0}$	**Corr**$_{w/o01}$	**Corr**$_{w/o0:10}$
N_{max}	166	50%	0.939	0.926	0.910	0.819
$N_{Cluster}$	116	35%	0.937	0.924	0.910	0.838
$N_{Heuristic}$	**111**	**33%**	**0.978**	**0.975**	**0.969**	**0.964**
N_{TCN}	**103±14**	**31±4%**	**0.982±0.007**	**0.979±0.009**	**0.975±0.011**	**0.958±0.023**

We visually assessed the impact of the absolute error and if there was an over- or underestimation (Fig. 11). We can see that N_{max} and $N_{Heuristic}$ underestimate the count while $N_{Cluster}$ overestimates the count but less so than with groundtruth labels, explaining the 4 % decrease in absolute error. For this ecological niche, the best performer is the N_{TCN} method which does not over- or underestimate the count but has a balanced variance around the ideal line. This is also observed numerically with high correlation values.

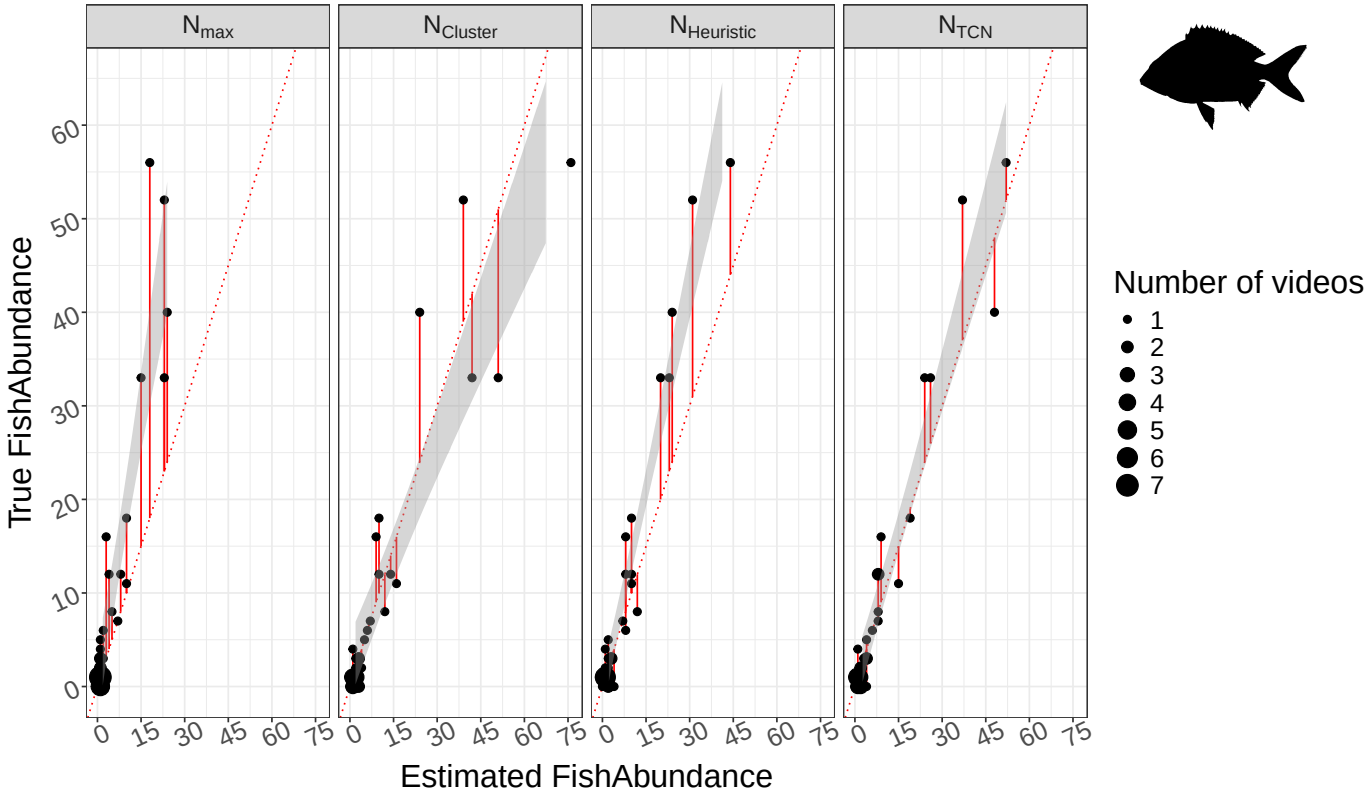

Figure 11: Linear regression (grey area indicates the 95% confidence interval) between the actual number of individuals of *D. vulgaris* in test videos (y-axis) and the counts of the different methods (x-axis) on the detector predictions. The size of the points indicates more videos overlapping with the corresponding methods estimated count and the actual count. The dashed red line indicates a perfect result. Underestimations are seen above the line and overestimations are seen below the line. The videos at point 0,0 (n=25) were reduced in their point size to 1 for the graphical representation. A total of 28 videos had 334 *D. vulgaris* present.

4 Discussion

Key Message To the extent of our knowledge, this was the first study to explore automated FishAbundance counting in a DOV setup. In this proof of concept, we proposed three automated fish counting methods that used detections derived from a deep learning model as an input. This automation process will significantly reduce the analysis time associated with manually calculating FishAbundance (Haberstroh et al. (2022)) or

N_{max} (Raoult et al. (2020)).

The three new proposed methods of fish counting all outperformed the commonly used N_{max} method. Even the simplest method $N_{Cluster}$ outperforms the metric N_{max} widely used in BRUV and DOV. In all cases, N_{max} underestimated the true abundance in the videos, even with perfectly labelled images, by up to 40 % and with varying linear relationships to the true FishAbundance, making it impossible to generalize.

The issue with N_{max} in DOV is that distinct groups of the same fish species within a transect are not counted. This is especially important in the case of *E. marginatus* as this species exhibits solitary and territorial behaviour (Pollard et al. (2018)), characterized by limited mobility. This implies that on a transect, multiple individuals can be spread out, which leads to an underestimation (Sherman et al. (2018)). This is evident when we removed the no and one occurrence videos from the analysis and the correlation dreceased drastically. In contrast, the correlation remained relatively stable across the other methods. For conservation purposes, it is important to rely on count data since the number of individuals is important for an evaluation of biomass, which is related to the health of a local population.

Methods Application of N_{max} gave a frequency rather than a count for the species and can already give valuable insights into species recovery. As Campbell et al. (2015) mentioned, the N_{max} metric works for location-expanding species that appear in low numbers in new areas. However, in these kinds of situations the new methods tested provide even greater insights.

For different scenarios, N_{max} chronically underestimate the counts. For comparison, in the perfect case the N_{TCN} and $N_{Heuristic}$ metrics both have error percentages lower than 30 %, which are in the range of the error rates of divers (Pais and Cabral (2018); Ward-Paige et al. (2010)). The $N_{Cluster}$ method shows evidence of adequate counting capability when the scenario is less complex, such as fewer individuals or further apart schools. The great advantage of $N_{Cluster}$ is that no prior knowledge is needed for calibration or training. The

only adjustable parameter is the choice of the number of clusters, 'k' that should be considered. This is dependent on video duration, species and ecological niche. Empirically, the best trade-off between computational effort and accuracy was to use k=10 for the algorithm as none of the videos had more than 10 peaks. The clear downside of this method is the accuracy, even though still outperforming N_{max}, it is outperformed by the other proposed methods.

The $N_{Heuristic}$ method also groups the different fish schools into clusters and sums the peak of each school to determine the final count. The difference between $N_{Cluster}$ and $N_{Heuristic}$ is that $N_{Heuristic}$ uses an intuitive procedure to differentiate the clusters, dictated by a subset of the data provided. The drawback of this method is that part of the data available is used for calibration and cannot be used in the analysis. However, the increase in the correlation and decrease of error rate between estimated and actual counts makes this approach valuable for instances when there is abundant data available and the task does not exceed a certain complexity.

Taking it a step further we introduced N_{TCN}, which allows the rapid addition of new species into the process, which is not always possible with $N_{Heuristic}$. Furthermore, in complex examples (*i.e.* more individuals, less accurate detector, etc.) the TCN outperforms the other methods and should always be favoured. Overall, when data is available the TCN approach is the most stable and high-performing method.

Impact of data scarcity on counting performance Organism counts and the resulting density numbers are among the most important ecological indicators for evaluating the health of natural ecosystems (Ramos et al. (2012)). Especially for the two protected species *E. marginatus* and *S. umbra*, a head count is of the utmost importance to monitor the evolution of their populations, their population dynamics and, ultimately, potential recovery. Especially for these species a complete detector pipeline is important.

In the tested cases, the detector does not always provide satisfactory results. Hence there is room for improvement on the detection task that can be fixed by adding more

images to the training dataset. Especially with rare species, the image pool is small; this scarcity of the data is observable with *S. umbra*, which had only 9 videos available in the test set and 8 videos in the training set. This data scarcity affects the detector more than the counting methods, as seen by the differences in the count between the fully automated case (Table 6) and the perfect case (Table 3). The error rate increased from 20 % to 60 % for this specific species, which is not sufficient to confidently predict the count for *S. umbra*. For the other species the differences in the error rate between fully automated and perfect cases are smaller. Linear correlation values are less affected by the detector than the to absolute errors, with changes in value typically less than $\pm$ 0.1.

Integrating a computer vision model with one of the proposed methods offers researchers the ability to collect novel data in multiple ways. Firstly, it provides more time for the ecological analysis of the results generated by these methods. Secondly, the framework presents alternatives to the commonly used N_{max}. This will allow more precise fish abundance measurements without operator dependent manipulations such as diving or manual video evaluation. Finally, it enables the use of a remotely operated vehicle (ROV), allowing transects to be conducted from a safe distance. This will enable an increased frequency of biodiversity assessments and a reduction in diver accidents, and will help our understanding of the marine environment and its evolution (Buscher et al. (2020)).

Future applications Fish biomass is another very important indicator for evaluating the health of more ecosystems. Therefore, the size per individual is an important indication for the well-being of a species (Duplisea and Castonguay (2006); Hallett et al. (2012)). The size of a fish can also give insights into the reproductive status of a population (Uusi-Heikkilä (2020)) and the distribution of adults and juveniles, which is an important indicator in recovery in areas of interest (Molloy et al. (2009)). Using the tested methods, a stereo camera system could automatically choose the frames with the highest appearances in both camera videos, detect the fish, extract the size and make an automated sizing of all the fish involved per school and not overall per video with N_{max}.

Furthermore, wherever there is a deep learning model available, labels are already available or can be easily obtained and, therefore, the methods can be calibrated or trained without further effort, which makes the methods applicable to more scientific fields. This approach could facilitate and accelerate the identification and counting of non-indigenous species using a moving camera, which may include a variety of both amateur and professional setups, and can be applied to a wide range of environments, including marine fish (Martìnez-González et al. (2021)) and terrestrial plants (Dyrmann et al. (2021)). Due to different direct and indirect anthropogenic actions, invasion of non-indigenous species has become a threat for the environment and knowing the extent and dynamics of these invasions is crucial for the health of local and endemic ecosystems. While prevention is still the most successful tool (Keller et al. (2008)), early recognition can lead to more efficient management of these invasions (*e.g.*, the black-striped mussel in Darwin Harbor, Australia (Ferguson (1999)), and the algae *Caulerpa taxifolia* in Agua Hedionda Lagoon and Huntington Harbor, USA (Anderson (2005))).

4.1 Conclusion

In conclusion, we presented three distinct methods for automatically and accurately estimating fish abundance using diver-operated videos. While N_{max} remains vital for stationary camera setups, moving cameras offer an opportunity to explore alternative counting methods, reducing labour and increasing efficiency. By introducing a comprehensive pipeline based on single-frame detections from a deep learning model, these methods become broadly applicable beyond underwater environments. Overall, this approach enables more frequent and accurate data collection, enhancing ecological research and conservation efforts.

References

Anderson, L. W. (2005). California's reaction to caulerpa taxifolia: a model for invasive species rapid response. *Biological Invasions*, 7:1003–1016. https://doi.org/10.1007/s10530-004-3123-z.

Atlas, W. I., Ma, S., Chou, Y. C., Connors, K., Scurfield, D., Nam, B., Ma, X., Cleveland, M., Doire, J., Moore, J. W., et al. (2023). Wild salmon enumeration and monitoring using deep learning empowered detection and tracking. *Frontiers in Marine Science*, 10:1200408. https://doi.org/10.3389/fmars.2023.1200408.

Bai, S., Kolter, J. Z., and Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint*. https://doi.org/10.48550/arXiv.1803.01271.

Bell, J. D., Watson, R. A., and Ye, Y. (2017). Global fishing capacity and fishing effort from 1950 to 2012. *Fish and Fisheries*, 18(3):489–505. https://doi.org/10.1111/faf.12187.

Bürgi, K., Bouveyron, C., Lingrand, D., Dérijard, B., Precioso, F., and Sabourault, C. (2024). Towards a fully automated underwater census for fish assemblages in the mediterranean sea. *Ecological Informatics*, page 102959. https://doi.org/10.1016/j.ecoinf.2024.102959.

Buscher, E., Mathews, D. L., Bryce, C., Bryce, K., Joseph, D., and Ban, N. C. (2020). Applying a low cost, mini remotely operated vehicle (rov) to assess an ecological baseline of an indigenous seascape in canada. *Frontiers in Marine Science*, 7. https://doi.org/10.3389/fmars.2020.00669.

Calò, A., Pereñiguez, J. M., Hernandez-Andreu, R., and García-Charton, J. A. (2022). Quotas regulation is necessary but not sufficient to mitigate the impact of scuba diving in a highly visited marine protected area. *Journal of Environmental Management*, 302:113997. https://doi.org/10.1016/j.jenvman.2021.113997.

Campbell, M. D., Pollack, A. G., Gledhill, C. T., Switzer, T. S., and DeVries, D. A. (2015). Comparison of relative abundance indices calculated from two methods of generating video count data. *Fisheries Research*, 170:125–133. https://doi.org/10.1016/j.fishres.2015.05.011.

Connolly, R. M., Jinks, K. I., 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. *Frontiers in Marine Science*, 9:918504. https://doi.org/10.3389/fmars.2022.918504.

Dìaz, S., Settele, J., Brondízio, E. S., Ngo, H. T., Agard, J., Arneth, A., Balvanera, P., Brauman, K. A., Butchart, S. H., Chan, K. M., et al. (2019). Pervasive human-driven decline of life on earth points to the need for transformative change. *Science*, 366(6471):eaax3100. https://doi.org/10.1126/science.aax3100.

Dickens, L. C., Goatley, C. H., Tanner, J. K., and Bellwood, D. R. (2011). Quantifying relative diver effects in underwater visual censuses. *PloS one*, 6(4):e18965. https://doi.org/10.1371/journal.pone.0018965.

Duplisea, D. E. and Castonguay, M. (2006). Comparison and utility of different size-based metrics of fish communities for detecting fishery impacts. *Canadian Journal of Fisheries and Aquatic Sciences*, 63(4):810–820. https://doi.org/10.1139/f05-261.

Dyrmann, M., Mortensen, A. K., Linneberg, L., Høye, T. T., and Bjerge, K. (2021). Camera assisted roadside monitoring for invasive alien plant species using deep learning. *Sensors*, 21(18):6126. https://doi.org/10.3390/s21186126.

Ellis, D. and DeMartini, E. (1995). Evaluation of a video camera technique for indexing abundances of juvenile pink snapper, pristipomoides filamentosus, and other hawaiian insular shelf fishes. *Oceanographic Literature Review*, 9(42):786.

Esselman, P. C., Moradi, S., Geisz, J., and Roussi, C. (2025). A transferable approach for quantifying benthic fish sizes and densities in annotated underwater images. *Methods in Ecology and Evolution*, 16(1):145–159. https://doi.org/10.1111/2041-210X.14453.

Ferguson, R. (1999). The effectiveness of australia's response to the black striped mussel incursion in darwin, australia. In *A report of the marine pest incursion management workshop*, pages 27–28. Citeseer.

Grane-Feliu, X., Bennett, S., Hereu, B., Aspillaga, E., and Santana-Garcon, J. (2019). Comparison of diver operated stereo-video and visual census to assess targeted fish species in mediterranean marine protected areas. *Journal of Experimental Marine Biology and Ecology*, 520:151205. https://doi.org/10.1016/j.jembe.2019.151205.

Haberstroh, A. J., McLean, D., Holmes, T. H., and Langlois, T. (2022). Baited video, but not diver video, detects a greater contrast in the abundance of two legal-size target species between no-take and fished zones. *Marine Biology*, 169(6):79. https://doi.org/10.1007/s00227-022-04058-3.

Hallett, C. S., Valesini, F. J., Clarke, K. R., Hesp, S. A., and Hoeksema, S. D. (2012). Development and validation of fish-based, multimetric indices for assessing the ecological health of western australian estuaries. *Estuarine, Coastal and Shelf Science*, 104:102–113. https://doi.org/10.1016/j.ecss.2012.03.006.

Harmelin-Vivien, M., Cottalorda, J.-M., Dominici, J.-M., Harmelin, J.-G., Le Diréach, L., and Ruitton, S. (2015). Effects of reserve protection level on the vulnerable fish species sciaena umbra and implications for fishing management and policy. *Global Ecology and Conservation*, 3:279–287. https://doi.org/10.1016/j.gecco.2014.12.005.

Harmelin-Vivien, M. and Hermalin, J.-G. (2013). How to assess the effects of protection on fish? the port-cros national park and the first underwater visual censuses in the mediterranean sea. *Sci. Rep. Port-Cros natl. Park, Fr*, 27:369–375.

Harmelin-Vivien, M. L., Harmelin, J.-G., Chauvet, C., Duval, C., Galzin, R., Lejeune, P., Barnabé, G., Blanc, F., Chevalier, R., Duclerc, J., et al. (1985). Evaluation visuelle des peuplements et populations de poissons méthodes et problèmes. *Revue d Écologie (La Terre et La Vie)*, 40(4):467–539. https://doi.org/10.3406/revec.1985.5297.

Hilborn, R., Amoroso, R. O., Anderson, C. M., Baum, J. K., Branch, T. A., Costello, C., De Moor, C. L., Faraj, A., Hively, D., Jensen, O. P., et al. (2020). Effective fisheries management instrumental in improving fish stock status. *Proceedings of the National Academy of Sciences*, 117(4):2218–2224. https://doi.org/10.1073/pnas.1909726116.

Hoekendijk, J. P., Kellenberger, B., Aarts, G., Brasseur, S., Poiesz, S. S., and Tuia, D. (2021). Counting using deep learning regression gives value to ecological surveys. *Scientific reports*, 11(1):23209. https://doi.org/10.1038/s41598-021-02387-9.

Hutchings, J. A. and Reynolds, J. D. (2004). Marine fish population collapses: consequences for recovery and extinction risk. *BioScience*, 54(4):297–309. https://doi.org/10.1641/0006-3568(2004)054[0297:MFPCCF]2.0.CO;2.

Jessop, S. A., Saunders, B. J., Goetze, J. S., and Harvey, E. S. (2022). A comparison of underwater visual census, baited, diver operated and remotely operated stereo-video for sampling shallow water reef fishes. *Estuarine, coastal and shelf science*, 276:108017. https://doi.org/10.1016/j.ecss.2022.108017.

Keller, R. P., Frang, K., and Lodge, D. M. (2008). Preventing the spread of invasive species: economic benefits of intervention guided by ecological predictions. *Conservation Biology*, 22(1):80–88. https://doi.org/10.1111/j.1523-1739.2007.00811.x.

Kilfoil, J. P., Wirsing, A. J., Campbell, M. D., Kiszka, J. J., Gastrich, K. R., Heithaus, M. R., Zhang, Y., and Bond, M. E. (2017). Baited remote underwater video surveys undercount sharks at high densities: insights from full-spherical camera technologies. *Marine Ecology Progress Series*, 585:113–121. https://doi.org/10.3354/meps12395.

Langlois, T. J., Harvey, E. S., Fitzpatrick, B., Meeuwig, J. J., Shedrawi, G., and Watson, D. L. (2010). Cost-efficient sampling of fish assemblages: comparison of baited video stations and diver video transects. *Aquatic biology*, 9(2):155–168. https://doi.org/10.3354/ab00235.

Martìnez-González, Á. T., Ramìrez-Rivera, V. M., Caballero-Vázquez, J. A., and Jáuregui, D. A. G. (2021). Deep learning algorithm as a strategy for detection an invasive species in uncontrolled environment. *Reviews in Fish Biology and Fisheries*, 31(4):909–922. https://doi.org/10.1007/s11160-021-09667-7.

Maslin, M., Louis, S., Godary Dejean, K., Lapierre, L., Villéger, S., and Claverie, T. (2021). Underwater robots provide similar fish biodiversity assessments as divers on coral reefs. *Remote Sensing in Ecology and Conservation*, 7(4):567–578. https://doi.org/10.1002/rse2.209.

Molloy, P. P., McLean, I. B., and Côté, I. M. (2009). Effects of marine reserve age on fish populations: a global meta-analysis. *Journal of Applied Ecology*, 46(4):743–751. https://doi.org/10.1111/j.1365-2664.2009.01662.x.

Pais, M. P. and Cabral, H. N. (2018). Effect of underwater visual survey methodology on bias and precision of fish counts: a simulation approach. *PeerJ*, 6:e5378. https://doi.org/10.7717/peerj.5378.

Pollard, D., Afonso, P., Bertoncini, A., Fennessy, S., Francour, P., and Barreiros, J. (2018). Epinephelus marginatus. *The IUCN Red List of Threatened Species*, 2018:e–T7859A100467602.

Pörtner, H. O. and Peck, M. A. (2010). Climate change effects on fishes and fisheries: towards a cause-and-effect understanding. *Journal of fish biology*, 77(8):1745–1779. https://doi.org/10.1111/j.1095-8649.2010.02783.x.

Ramos, S., Amorim, E., Elliott, M., Cabral, H., and Bordalo, A. A. (2012). Early life stages of fishes as indicators of estuarine ecosystem health. *Ecological Indicators*, 19:172–183. https://doi.org/10.1016/j.ecolind.2011.08.024.

Ranganathan, C. S., Raman, R., Parikh, S., Rajesh, S., Meenakshi, R., and Muthulekshmi, M. (2023). Iot applications in marine monitoring: Protecting ocean health and biodiversity. In *2023 International Conference on Sustainable Communication Networks*

and Application (ICSCNA), pages 305–310. https://doi.org/10.1109/ICSCNA58489.2023.10370314.

Raoult, V., Tosetto, L., Harvey, C., Nelson, T. M., Reed, J., Parikh, A., Chan, A. J., Smith, T. M., and Williamson, J. E. (2020). Remotely operated vehicles as alternatives to snorkellers for video-based marine research. *Journal of Experimental Marine Biology and Ecology*, 522:151253. https://doi.org/10.1016/j.jembe.2019.151253.

Roelfsema, C., Bayraktarov, E., van den Berg, C., Breeze, S., Grol, M., Kenyon, T., de Kleermaeker, S., Loder, J., Mihaljevic, M., Passenger, J., et al. (2018). Ecological assessment of the flora and fauna of flinders reef, north moreton island, queensland.

Schobernd, Z. H., Bacheler, N. M., and Conn, P. B. (2014). Examining the utility of alternative video monitoring metrics for indexing reef fish abundance. *Canadian Journal of Fisheries and Aquatic Sciences*, 71(3):464–471. https://doi.org/10.1139/cjfas-2013-0086.

Schramm, K. D., Harvey, E. S., Goetze, J. S., Travers, M. J., Warnock, B., and Saunders, B. J. (2020). A comparison of stereo-bruv, diver operated and remote stereo-video transects for assessing reef fish assemblages. *Journal of Experimental Marine Biology and Ecology*, 524:151273. https://doi.org/10.1016/j.jembe.2019.151273.

Sherman, C. S., Chin, A., Heupel, M. R., and Simpfendorfer, C. A. (2018). Are we underestimating elasmobranch abundances on baited remote underwater video systems (bruvs) using traditional metrics? *Journal of Experimental Marine Biology and Ecology*, 503:80–85. https://doi.org/10.1016/j.jembe.2018.03.002.

Stobart, B., García-Charton, J. A., Espejo, C., Rochel, E., Goñi, R., Reñones, O., Herrero, A., Crec'hriou, R., Polti, S., Marcos, C., et al. (2007). A baited underwater video technique to assess shallow-water mediterranean fish assemblages: Methodological evaluation. *Journal of Experimental Marine Biology and Ecology*, 345(2):158–174. https://doi.org/10.1016/j.jembe.2007.02.009.

Uusi-Heikkilä, S. (2020). Implications of size-selective fisheries on sexual selection. *Evolutionary Applications*, 13(6):1487–1500. https://doi.org/10.1111/eva.12988.

Villon, S., Iovan, C., Mangeas, M., and Vigliola, L. (2024). Toward an artificial intelligence-assisted counting of sharks on baited video. *Ecological Informatics*, 80:102499. https://doi.org/10.1016/j.ecoinf.2024.102499.

Wang, H. and Song, M. (2011). Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming. *The R journal*, 3(2):29.

Ward-Paige, C., Mills Flemming, J., and Lotze, H. K. (2010). Overestimating fish counts by non-instantaneous visual censuses: consequences for population and community descriptions. *PLoS One*, 5(7):e11722. https://doi.org/10.1371/journal.pone.0011722.

Weng, K. C., Friedlander, A. M., Gajdzik, L., Goodell, W., and Sparks, R. T. (2023). Decreased tourism during the covid-19 pandemic positively affects reef fish in a high use marine protected area. *Plos one*, 18(4):e0283683. https://doi.org/10.1371/journal.pone.0283683.

Yan, H. F., Kyne, P. M., Jabado, R. W., Leeney, R. H., Davidson, L. N., Derrick, D. H., Finucci, B., Freckleton, R. P., Fordham, S. V., and Dulvy, N. K. (2021). Overfishing and habitat loss drive range contraction of iconic marine fishes to near extinction. *Science Advances*, 7(7):eabb6026. https://doi.org/10.1126/sciadv.abb6026.

Zhang, Y., Ou, Z., Tweedley, J. R., Loneragan, N. R., Zhang, X., Tian, T., and Wu, Z. (2024). Evaluating the effectiveness of baited video and traps for quantifying the mobile fauna on artificial reefs in northern china. *Journal of experimental marine biology and ecology*, 573:152001. https://doi.org/10.1016/j.jembe.2024.152001.

Supplementary Material

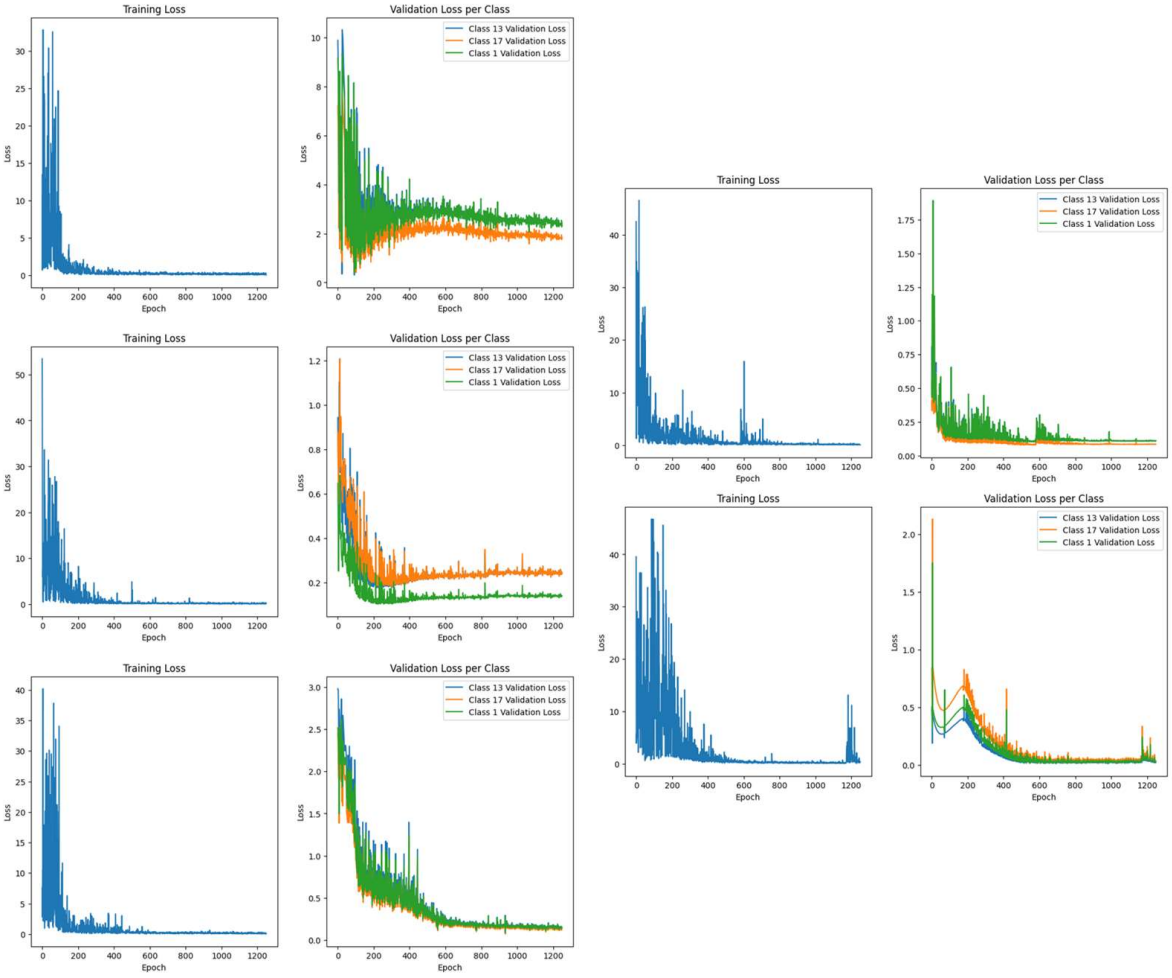

Figure S1: The 5 training runs for the TCN model for the perfect case used in the study.

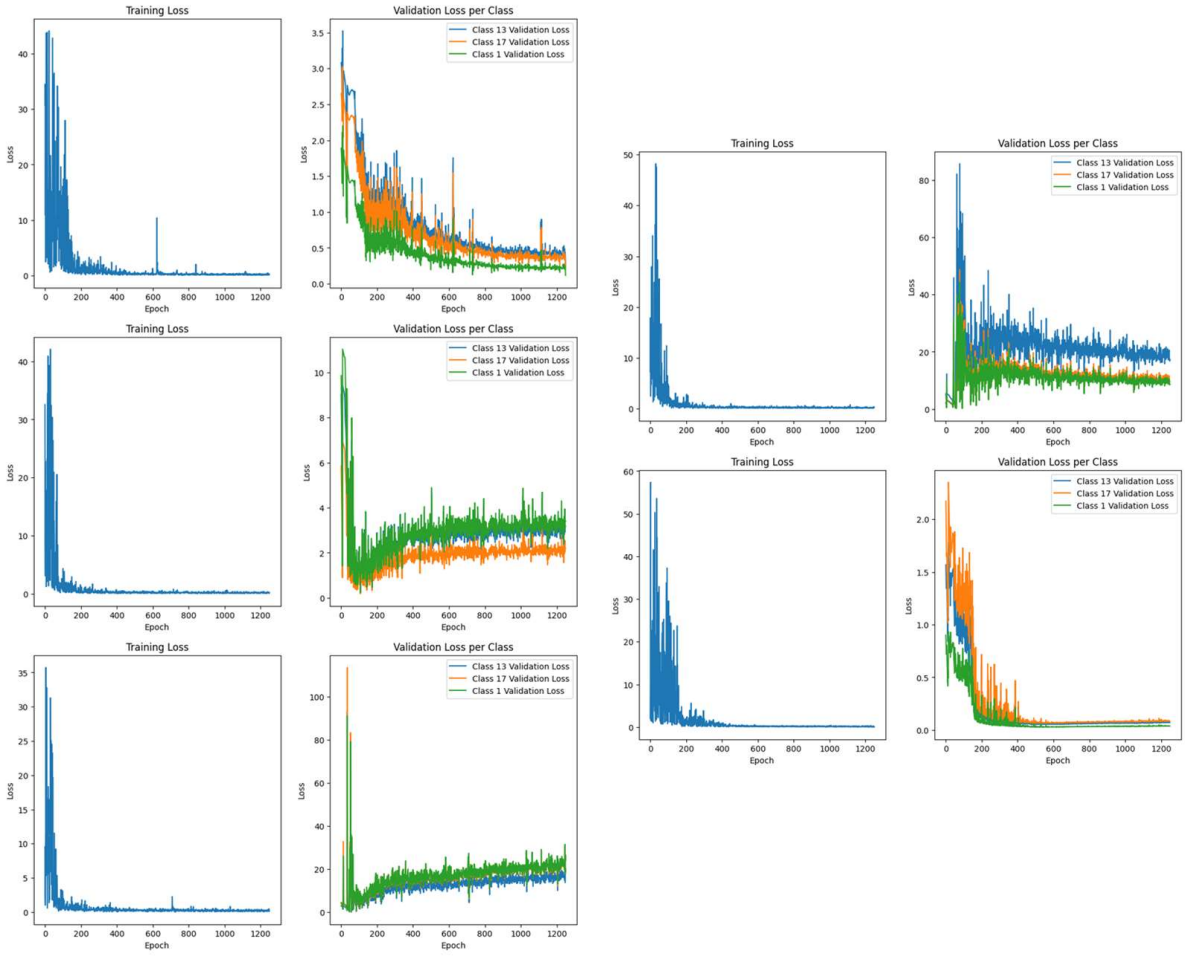

Figure S2: The 5 training runs for the TCN model for the fully automated case used in the study.

Table S1: The TCN model used in the study.

Layer (type)	Output Shape	Param #
CausalConv1d-1	[-1, 20, 709]	100
BatchNorm1d-2	[-1, 20, 709]	40
ReLU-3	[-1, 20, 709]	0
Dropout-4	[-1, 20, 709]	0
CausalConv1d-5	[-1, 20, 709]	1,620
BatchNorm1d-6	[-1, 20, 709]	40
ReLU-7	[-1, 20, 709]	0
Dropout-8	[-1, 20, 709]	0
Conv1d-9	[-1, 20, 709]	40
ReLU-10	[-1, 20, 709]	0
TemporalBlock-11	[[-1, 20, 709], [-1, 20, 709]]	0
CausalConv1d-12	[-1, 10, 709]	810
BatchNorm1d-13	[-1, 10, 709]	20
ReLU-14	[-1, 10, 709]	0
Dropout-15	[-1, 10, 709]	0
CausalConv1d-16	[-1, 10, 709]	410
BatchNorm1d-17	[-1, 10, 709]	20
ReLU-18	[-1, 10, 709]	0
Dropout-19	[-1, 10, 709]	0
Conv1d-20	[-1, 10, 709]	210
ReLU-21	[-1, 10, 709]	0
TemporalBlock-22	[[-1, 10, 709], [-1, 10, 709]]	0
CausalConv1d-23	[-1, 5, 709]	205
BatchNorm1d-24	[-1, 5, 709]	10
ReLU-25	[-1, 5, 709]	0
Dropout-26	[-1, 5, 709]	0
CausalConv1d-27	[-1, 5, 709]	105
BatchNorm1d-28	[-1, 5, 709]	10
ReLU-29	[-1, 5, 709]	0
Dropout-30	[-1, 5, 709]	0
Conv1d-31	[-1, 5, 709]	55
ReLU-32	[-1, 5, 709]	0
TemporalBlock-33	[[-1, 5, 709], [-1, 5, 709]]	0
TCN-34	[-1, 5, 709]	0
AvgPool1d-35	[-1, 5, 1]	0
Flatten-36	[-1, 5]	0
Linear-37	[-1, 3]	18

Total params: 3,713
Trainable params: 3,713
Non-trainable params: 0

Input size (MB): 0.00
Forward/backward pass size (MB): 2011.53
Params size (MB): 0.01
Estimated Total Size (MB): 2011.55