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

Automated Fish Counting in DOV

Kilian Bürgi[a,b], 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 6, 2025

Acknowledgements

This work was only made possible thanks to the collaboration with the projects RECIF (Réseau d'Evaluation des Cantonnements et ZSC en Interface Fonctionnelle) and FEAMPA (Fonds européen pour les affaires 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. 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 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.

Abstract

1 - Underwater video transects are crucial to assess marine biodiversity. The counting of fish individuals in these videos is labour- and time-intensive. An automation of said counting would create non-biased biodiversity data.

2 - For this purpose, we explored traditional methods of counting animals as well as introduced three new methods to count fish from computer vision derived data (single frame detections) resulting in a holistic and fully automated pipeline for fish abundance extraction. The different methods 1) traditional N_{max}, 2) 1d k-means clustering method, 3) an intuitive clustering approach $N_{Heuristic}$ and 4) a Temporal Convolutional Neural Networks (TCN) counting method are proposed on transect data of three Mediterranean species with different ecological niches.

3 - Our results shows evidence of underestimation by the traditional N_{max} while the other methods showed better overall results with the proposed $N_{Heuristic}$ and TCN methods representing the reality the most. With an absolute variation comparable to inter-observer variation, we demonstrated reliable methods for quantifying fish counts within the framework of three different species.

4 - For future projects, incorporating a stereo system could provide more detailed insights into species recovery, and the analysis should be expanded to encompass a broader range of species, including both marine and terrestrial ecosystems.

1 Introduction

The marine environment is facing different critically endangering factors to its inhabitants. Factors such as climate change (Pörtner and Peck (2010)), (mass-)tourism (Weng et al. (2023)) and fishing (Bell et al. (2017)) - especially overfishing (Yan et al. (2021)) - are bringing marine species (*i.e.* mammals, fish, reptiles and invertebrates) populations to a critical low (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 combat the diminishing populations (Hutchings and Reynolds (2004)). Marine protected areas (MPAs) that 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. Among these methods, underwater fish counts play a critical role.

A very important indication for the health of an ecosystem is the count of individual fish, as these measurements provide valuable insights into population dynamics, species diversity, and the overall balance of the aquatic environment (REF). To count fish in the marine environment, today's state of the art techniques rely on divers retrieving biological data in different regions of interest. In these areas, specifically trained experts perform different biodiversity assessments. There are different means to record this diversity - direct methods such as underwater visual census (UVC) or indirect methods which rely on camera deployment.

The traditional way of camera deployment is the Baited Remote Underwater Video (BRUV) approach (Fig. 1-B). To avoid double counting of fish in a stationary setup, the analysis of the biodiversity uses only the frame with the highest number of individuals and is theorised to describe the relative abundance of the specific replicate in that area. The number of fish in this maximised 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 DOV on the other hand uses SCUBA divers or remote operated vehicles (ROV) as a camera-holding vessel recording its view of the appearing and disappearing fish (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, through empirical observations, it is theorised 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)) and would allow more precise abundance data to be collected (Dickens et al. (2011)).

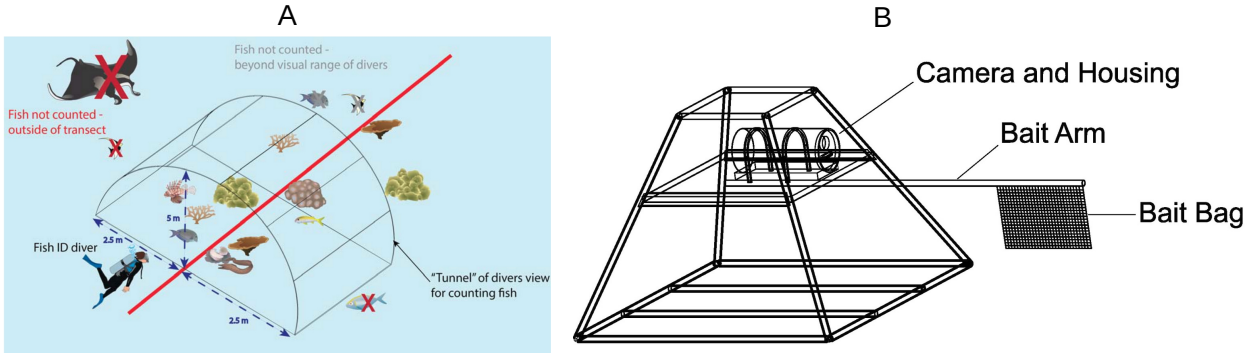

Figure 1: A is an example of a transect setup used in this study with a diver that holds a camera filming his point of view (adapted from Roelfsema et al. (2018)). B explains the concept and construction of a BRUV setup (adapted from Zhang et al. (2024)).

Besides being widely used, both of these methods share the disadvantage of requiring long video analysis times (Schramm et al. (2020)). With advances in technologies in the 21^{st} century, there is a potential to automatise or at least semi-automatise the process of video analysis using machine learning methods (Hoekendijk et al. (2021)). The efforts of

Atlas et al. (2023) presented a deep learning multi object tracker for wild salmon. The salmon swim through a one-directional river fence, getting tracked and counted successfully automatising this procedure. For moving cameras, studies on the automation are scarce. The study of Connolly et al. (2022) was able to accurately predict the frame with the most individuals in a sequence nine out of ten videos. These results are comparable to stationary setups (*i.e.* BRUV). Automatically counting these frames was not stated.

In this study we highlighted the flaws of N_{max} and propose three alternative methods that outperform the N_{max} metric in counting the actual fish abundance in an automated manner. The methods explored are, besides N_{max}, a 1-dimensional (1D) clustering approach, an intuitive clustering approach termed $N_{Heuristic}$ and a Temporal Convolutional Network (TCN) approach. We used them 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 is to find a reliable procedure to count objects in single frame detections of a moving camera and to reveal the possibilities of different methodologies providing this task. This will allow the creation of fast and non-biased data that can be used for further ecological, economical or conservation analyses.

In this study we have three main key *contributions*:
- We present the first fully automated pipeline for Diver Operated Video (DOV) systems, integrating all steps from video recording to extracting the fish abundance.
- We identify critical weaknesses in the widely used N_{max} approach, specifically its tendency to underestimate fish abundance. To overcome these limitations, we propose three novel methods that are significantly reducing underestimation.
- To challenge our methods, we establish two experimental conditions - theoretical and practical. Our approach provides a robust framework for future studies in assessing automated fish abundance extraction.

2 Material 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 will discuss the study area and data collection specifications (see Sec. 2.1), species of interest (see Sec. 2.2), how we used the videos (see Sec. 2.3) and finally give more insights in the different methods (see Sec. 2.4), to be able to reproduce the study for more locations and species.

2.1 Study area and data collection

To cover a great area and wide variety of conditions, we collected videos in eight different locations of the French Riviera in the Mediterranean Sea in standardized fashion (Harmelin-Vivien et al. (1985)). The depth ranged from 1-37m and was executed during the whole year in 2022 (cold- and warm season). Camera-equipped divers did 3 transects of 125 m^2 surface per dive over the period of the year. 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 a full high definition resolution (1920x1080 px). Frames were extracted from these recordings with FPS of 1.

2.2 Species of Interest

In our videos we saw a wide variety of fish species from which we chose three for our study. We chose them because of their distinct ecological niches, that are different enough to challenge these new methods and show the stability of them as well as allow a more broad applicability of these methods in other environments.

The most emblematic species of the French Mediterranean Sea is the endemic dusky grouper (*Epinephelus 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 but a fishing ban in 2003 (Pollard et al. (2018)) shows indication of recovery. Since this species

has only been recently protected, knowing the evolution of this species in a temporal and spatial manner is extremely important.

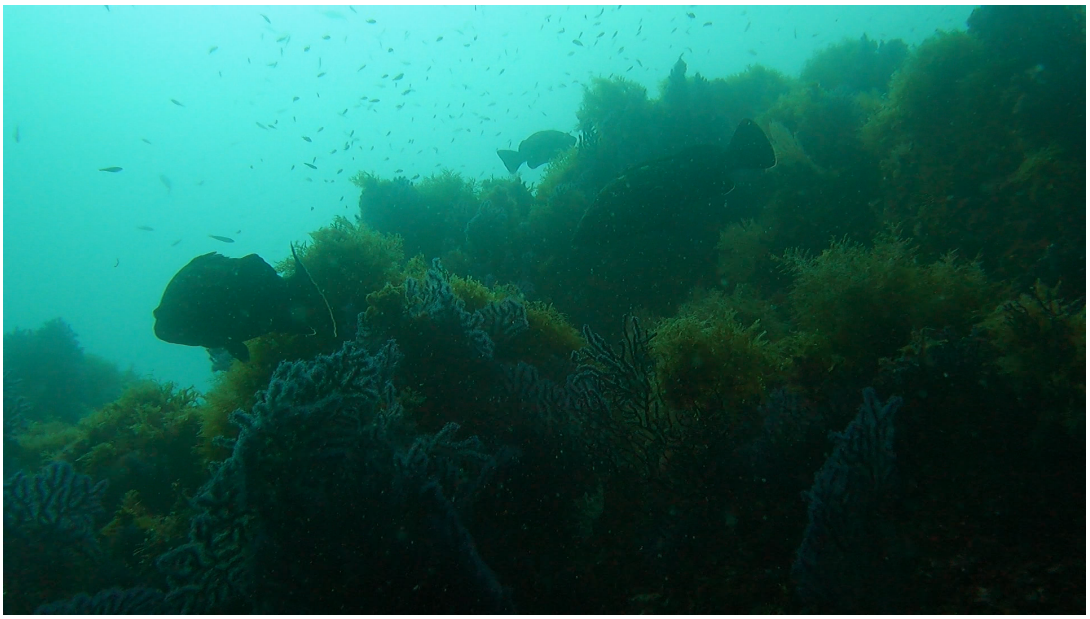

Figure 2: Non-edited example imagery from a transect of three *E. marginatus* individuals in the center of the image. Conditions are variable in the frames and make the detection more difficult.

Besides *E. marginatus* also the brown meagre (*Scianea umbra* - Fig. 3) is protected (Prefectoral orders number 2013357-0002 for Corsica and number 2013357-0007 for continental coast) in French waters. 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 will fill a different ecological niche, challenging our methods.

As a third species we shift from the low occurrence species and look at more abundance species that are present in more videos and increasing high occurrence videos. For this ecological niche, we chose the common two-banded sea bream or *Diplodus 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 are therefore more challenging for the methods. The abundance varies from one to two individuals up to 50 upwards. This new scenario will greatly show the applicability of the different methods to a different ecological niche.

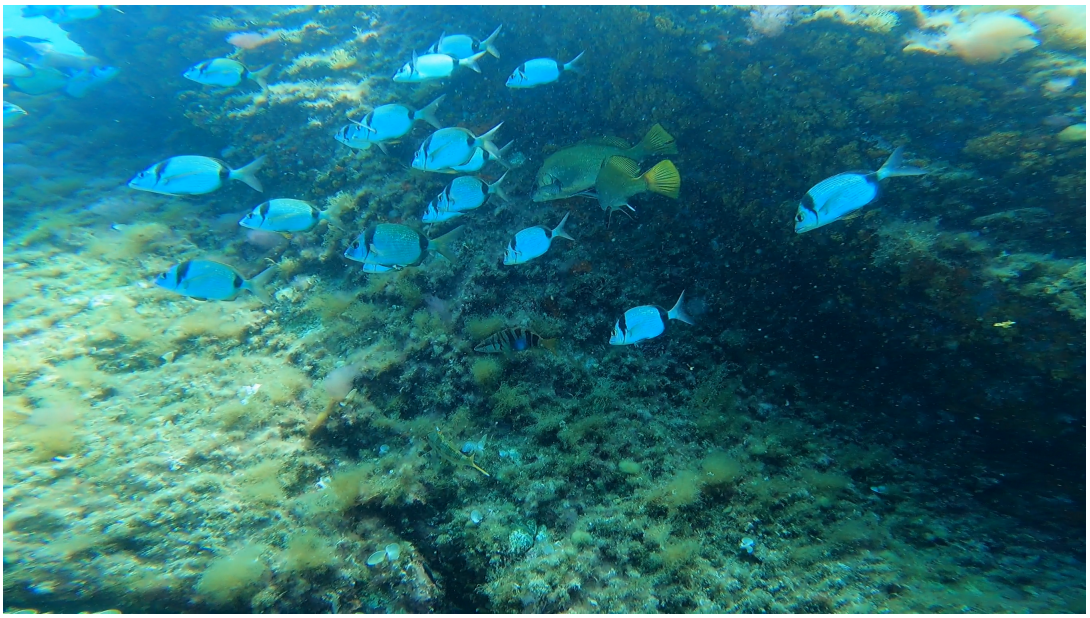

Figure 3: Non-edited example imagery of the transect with over 20 *D. vulgaris* individuals and two *S. umbra* in the middle of the *D. vulgaris* school.

2.3 Obtaining data from the videos

Our videos provided us with sequential frames, forming a 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 our 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 strength of our methods, we defined two types of data as the method input, describing a perfect and a real-world scenario. In the perfect case (see Sec. 3.1), where 100 % of the detections were made correctly for which we used the groundtruth detections to verify the feasibility of the methods proposed, without the interference of a potentially faulty detector. In the fully automated case (see Sec. 3.2), we used the predictions of the detector to see the impact of using a detector in the pipeline. For the output of our methods we wanted to approximate the True FishAbundance. Our method estimated counts are called Estimated FishAbundance from hereby on.

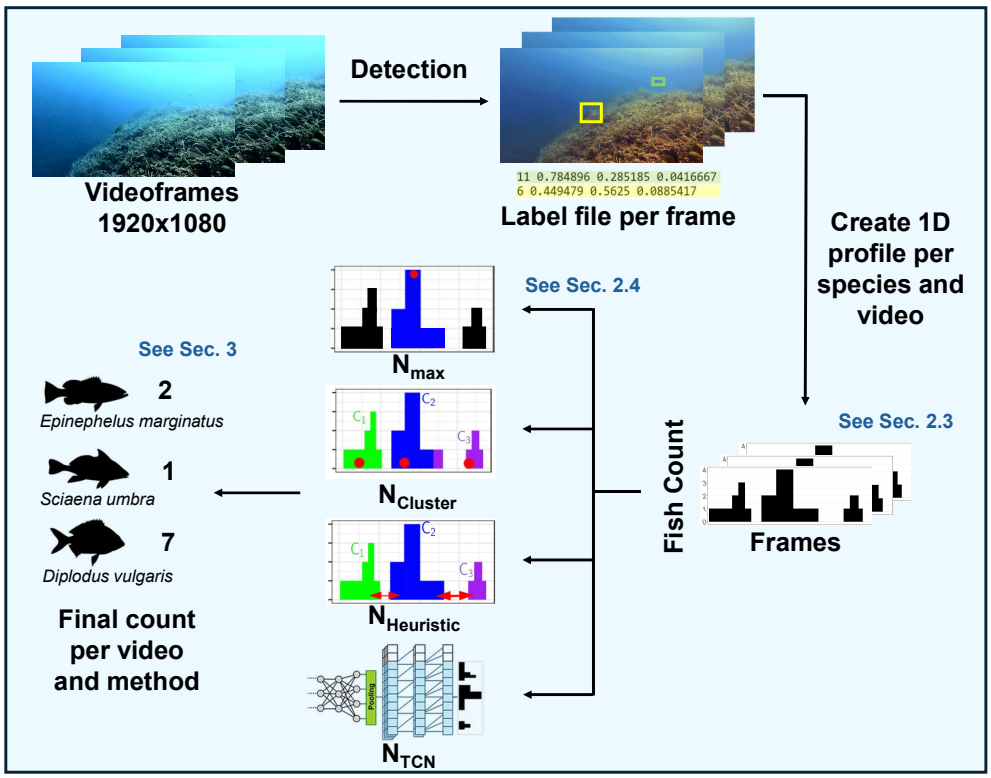

Figure 4: Workflow of the automated pipeline.

2.3.1 Detector training and input data

To find which species are present in which videos automatically, we used a deep learning approach to make predictions on the data. As the detector, we used a slightly varied model described in a previous study (Bürgi et al. (2024)). We kept hyperparameters constant but moved seven videos from the training to the validation set for the detector. We used this validation set to find the f1 score per species for the fully automated case. We excluded five high occurrence videos to enrich the test data set and challenge the methods with abundant videos. 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).

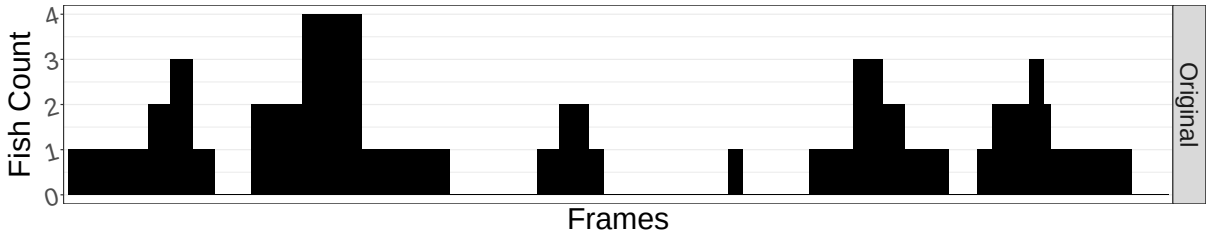

Figure 5: An example of the representation of one fish species in one video: the number of fishes from this species is counted per frame, manually or automatically. Each species in each video is represented as this 1D series of values.

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 3 and 4. The test set videos were held constant across the methods to have a fair comparison. We provide Table 1 for 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 make a fair comparison. The occurrences differ between the species to challenge 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 zero abundance videos and higher than zero abundance videos 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 needed the actual counts 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.

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 gives 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 the different methods not evaluated in this study as well as different species.

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

To have an idea of 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 (Corr_{wo0}), (3) excluding videos with counts of zero and one (Corr_{wo01}), and (4) excluding videos with counts in the range of zero to ten ($\text{Corr}_{wo0:10}$).

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. Our three methods are - 1) 1D clustering termed $N_{Cluster}$, 2)

the manual $N_{Heuristic}$ and 3) 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 traditional method to find the abundance in videos - N_{max}. 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 $N_{Cluster}$

Since N_{max} is only incorporating the peak of one of the schools, 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 using a different approach. The different groups in the 1-dimensional profile (Fig. 5) are hypothesised to be different schools and taking the maximum of each of these cluster is refining the count per video. The 1-dimensional profile deriving from the detections will work well for a clustering approach.

Generally speaking, a k-means clustering approach groups our sequences into k-cluster so that a cost is minimized. The challenge with k-means clustering is finding 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 needs to be adjusted to each individual problem. For each sequence or video the ideal k was found. The peaks of all clusters are then summarized forming 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 $\mathrm{N}_{Cluster}$ 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 $\mathrm{N}_{Heuristic}$ (Eq. 5), a method that employs inter-school distances as a species-specific differentiator.

This method uses the relatively consistent distance characteristic observed for each species, allowing more precise school differentiation based on this distance. The different clusters are differentiated by two variables that are calibrated on the training data set. The variable *threshold* refers to the minimum count for a school to be valid, this was introduced to counteract always occurring species. On the other hand, *n_frames* refers to a delay between schools before a new school is identified. The maxima of each school were then summarised to get an improved count of the fish individuals in the video corresponding to a transect.

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

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 $\mathbf{N}_{TCN}$

Clustering methods typically assume a constant number of individuals within a fish school. 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, there is the possibility 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) 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 with the 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 Figure S1 + S2. The architecture can be seen in Table S1 with 3,713 trainable parameters. The predicted video counts are called 'N_{TCN}' hereby on.

3 Results

In this section, we are going to show the different methods outlined in the Material and Methods section. The methods follow the same order as well as the species to maintain a reader flow. We commence with the perfect case (see Sec. 3.1) and then use the fully automated case (see Sec. 3.2) to challenge our methods.

3.1 Perfect Case on groundtruth test labels

In this first case we test the fish counting impacted solely by our methods 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 investigated first the species of *E. marginatus*. It is an uncommon species and high occurrence videos are rare. In all test videos, we have seen a total of 56 individuals with the majority being in multiple one occurrence videos. In Table 2 we can see that all methods out compete N_{max} in all metrics provided. Best performing is the method of $N_{Heuristic}$ with an absolute error (AE) of 13 or 23 % over- or under-estimation. The correlation decreases if we exclude the 0 and 1 occurrence videos below 0.60 for N_{max} while the others stay constant above. The exclusion results in a reduction of correlation for N_{max} from 0.897 to 0.544, whereas $N_{Heuristic}$ also decreases, but to a lesser extent, from 0.957 to 0.820.

Table 2: The different methods with the different metrics are presented in this table for the species *E. marginatus*. Correlation with the actual counts on the test set are indicated with all points included (All), 0 excluded (wo0) and 0 and 1 excluded (wo01) to show the strength 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**$_{wo0}$	**Corr**$_{wo01}$
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) show a clear underestimation of the count with N_{max} while it is much more stable with the other three methods. We can see that with an increase in occurrence in the videos, our methods handle this 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 our methods.

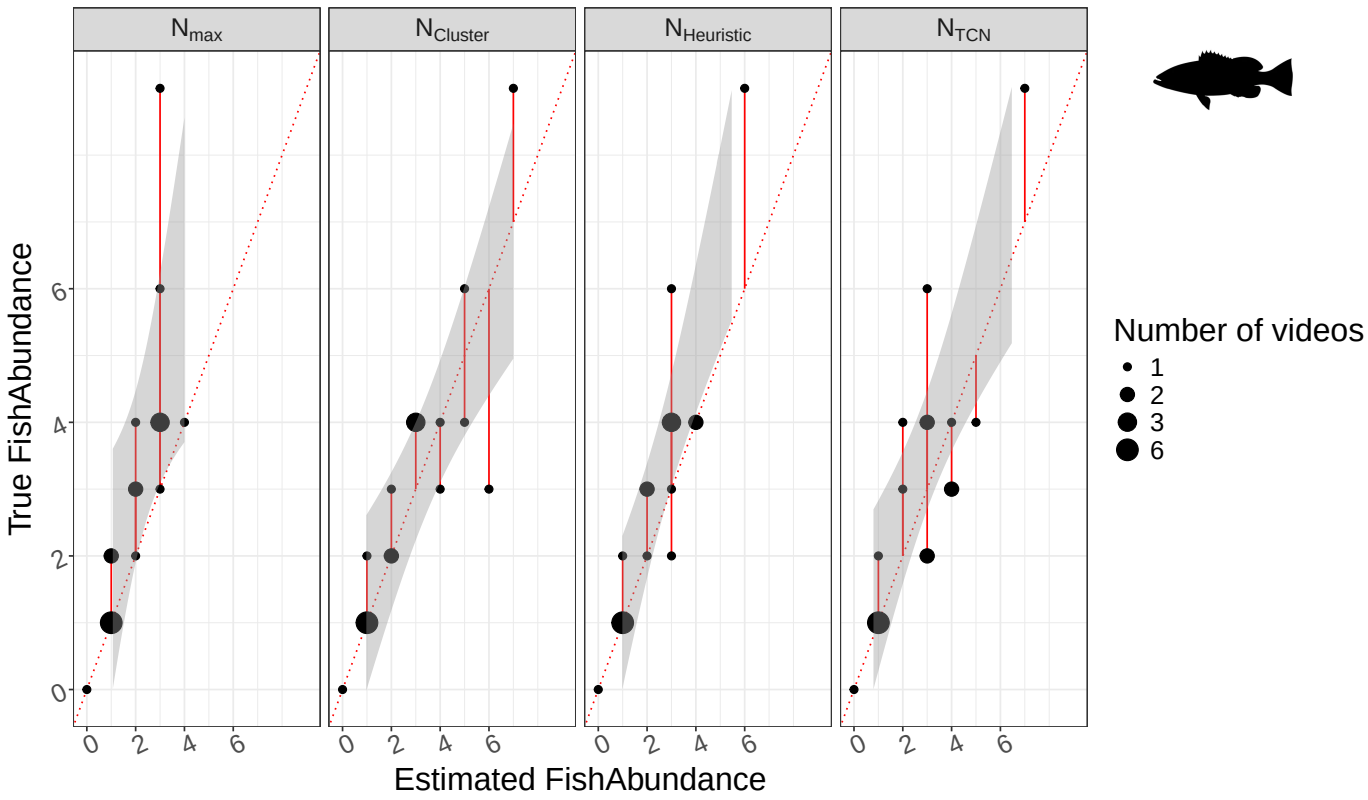

Figure 6: Linear regression (grey area indicates the 95% confidence interval) between he actual number of individuals of *E. marginatus* in test videos (y-axis) and the estimated count by the different methods (x-axis). The red line indicates a perfect prediction of the count. Size of the circular shapes present how many videos fall in this count-category. The majority of the videos (n=36) were at point 0,0. 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 is *S. umbra* in Table 3. Also this species is rare but appears in larger schools of up to 20 individuals. We observed this species in 9 test videos. All our three methods have high correlation values and low miscount of 21% or lower, making any of them suitable to count the ecological functional group of schooling predatory species. N_{max} fails to count the absolute fish abundance and 33% of the individuals are miscounted. Correlation values significantly drop from 0.771 to 0.382 when removing the lower occurrence videos. 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 methods with the different metrics are presented in this table for the species *S. umbra*. Correlation with the actual counts on the test set are indicated with all points included (All), 0 excluded (wo0) and 0 and 1 excluded (wo01) to show the strength 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**$_{wo0}$	**Corr**$_{wo01}$
N_{max}	11	33%	0.766	0.45	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 looked visually into how the different methods were presenting the count data (Fig. 7). The first thing that can be seen is that the insufficient correlation values generated by N_{max} depends on only one video that has more than six occurrences. This video is better counted with the other methods and therefore leads to the better correlation values for these methods. This gives an indication how the different methods can outperform N_{max} on high occurrence videos while N_{max} struggles with that. None of the methods receive a perfect result on this particular video but $N_{Cluster}$ present the best result with only one individuals missed.

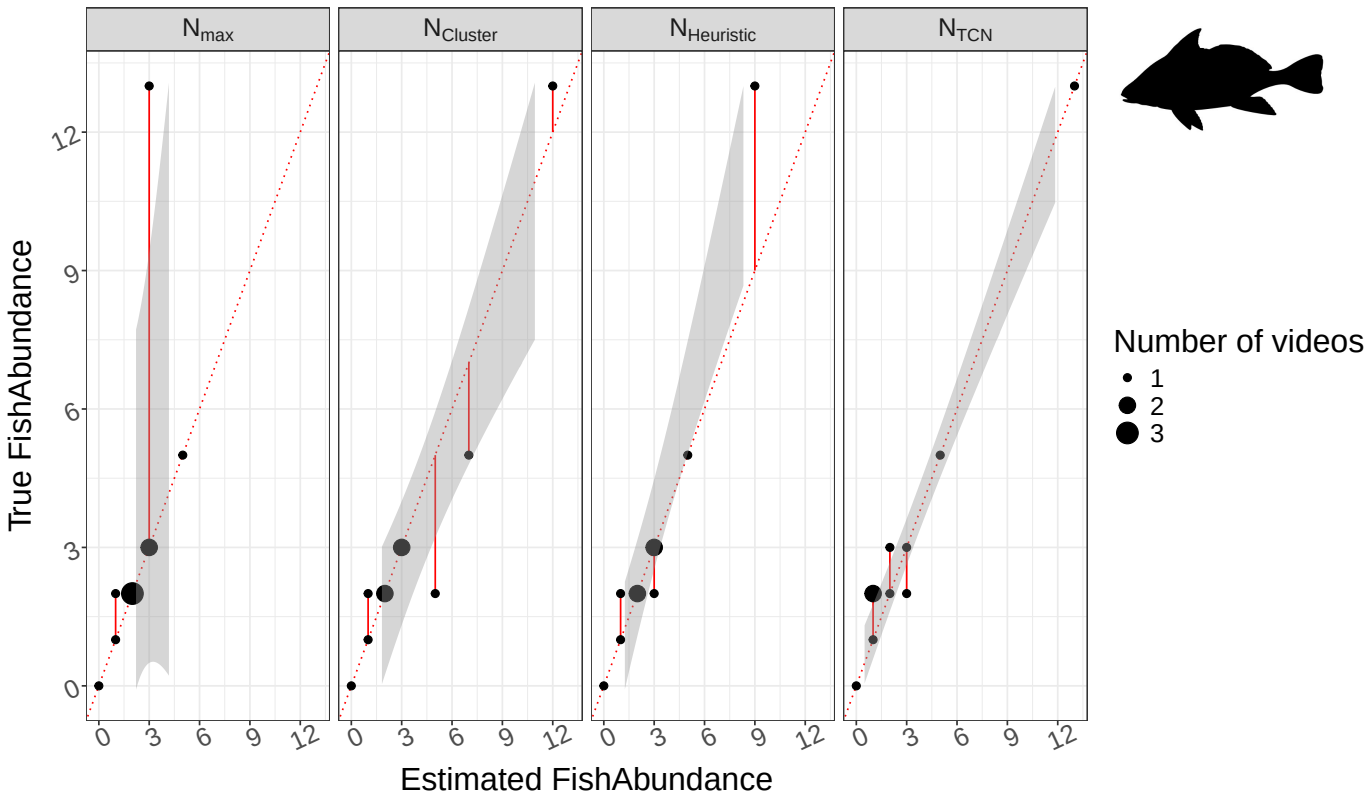

Figure 7: Linear regression (grey area indicates the 95% confidence interval) between he actual number of individuals of *S. umbra* in test videos (y-axis) and the counts of the different methods (x-axis). Size of the points indicate more videos overlapping with the corresponding methods and the actual count. The dashed red line indicates a perfect result, above the line is an underestimation and under the line indicates an overestimation. 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 and commonly seen *D. vulgaris* (Table 4). This gives a new scenario for the methods and with the expected increase in number of individuals, we also expected the challenge for the methods to be higher. This difficulty can be seen for two methods for this species - N_{max} and $N_{Cluster}$. With error rates of 40% the counting of this species is insufficient. However, for the other two species the error rate is halved and is around 20% for $N_{Heuristic}$ and N_{TCN}. All of our 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 over 0.90, which further underscores the broadened applicability of these methods for different ecological niches.

Table 4: The different methods with the different metrics are presented in this table for the species *D. vulgaris*. Correlation with the actual counts on the test set are indicated with all points included (All), 0 excluded (wo0), 0 and 1 excluded (wo01) and videos with less than 10 individuals excluded (wo0:10) to show the strength 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**$_{wo0}$	**Corr**$_{wo01}$	**Corr**$_{wo0:10}$
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 decrease in no occurrence videos made data more available and favoured the two methods that need a training or a calibration. This is clearly visible in the graphical representation (Fig. 8) of the FishAbundance. Both better performing methods seem to underestimate the count a bit but keep the distance to the perfect dashed red line as minimal 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} is underestimating the count in all videos and the majority of the miscounting occurs in the videos that contain more than 15 individuals seeming to be the limit of this method.

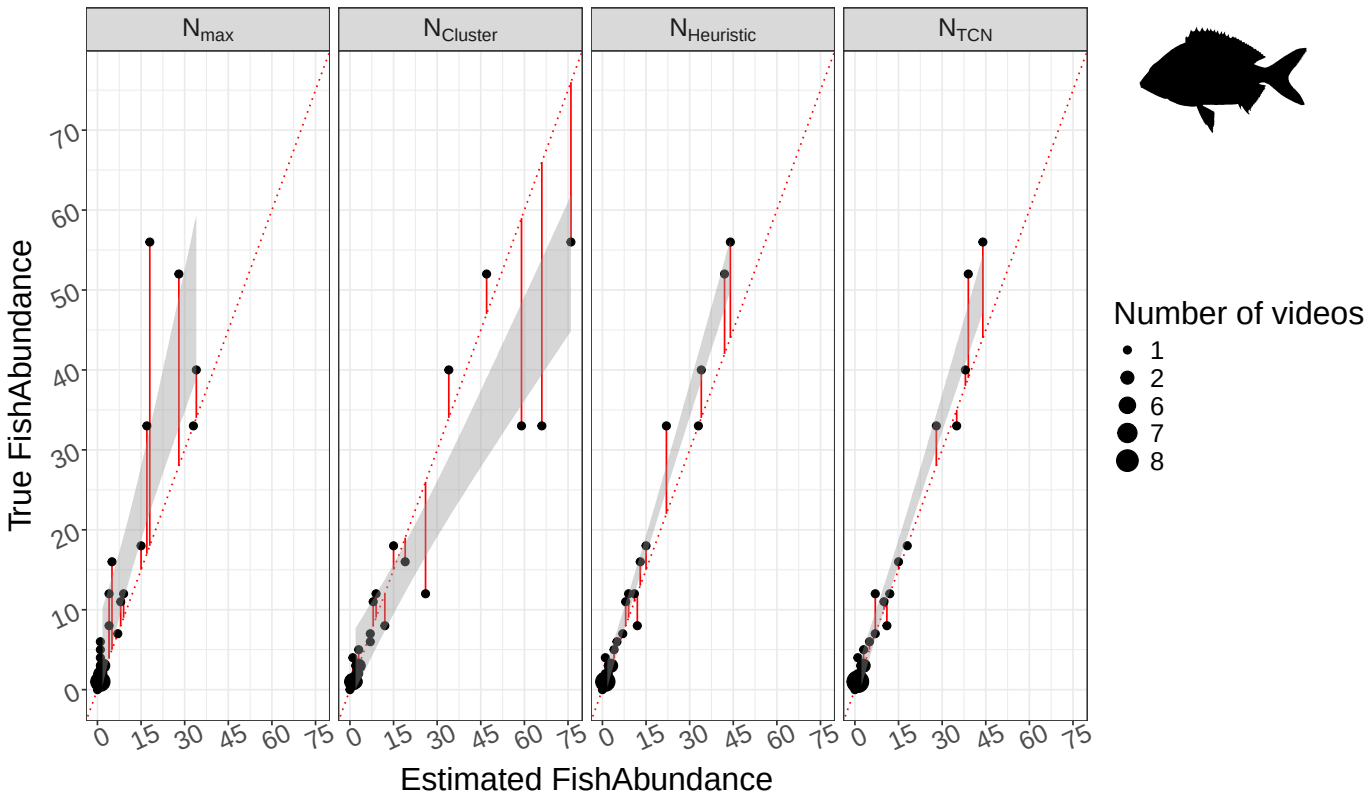

Figure 8: Linear regression (grey area indicates the 95% confidence interval) between he actual number of individuals of *D. vulgaris* in test videos (y-axis) and the counts of the different methods (x-axis). Size of the points indicate more videos overlapping with the corresponding methods and the actual count. The dashed red line indicates a perfect result, above the line is an underestimation and under the line indicates an overestimation. 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 on test detections

In this section we explored the impact of utilizing 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 by the respective f1 score on the validation set of the detector training. We determined the confidence thresholds as followed, 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 increase of error from all the methods (Table 5) when in comparison with the perfect case (Table 2). The effect of this imperfection is heavier on the correlation of N_{max} than the other methods that keep values above 0.750 while N_{max} drops to 0.444 for the $Corr_{wo01}$. Most of these errors derive from false positive counts in zero and one occurrence videos since when removed, the correlation is higher than with the inclusion (except N_{max}). This is observable for both more rarer species since the effect 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 in this table for the species *E. marginatus*. Correlation with the actual counts on the test set are indicated with all points included (All), 0 excluded (wo0) and 0 and 1 excluded (wo01) to show the strength 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**$_{wo0}$	**Corr**$_{wo01}$
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} are showing clear indication that the performance is better than N_{max}. $N_{Heuristic}$ has lower error rates due to less false positives being counted towards the abundance with the fp exclusion mechanism of the method (to be written in M&M). This can be seen

numerically in Table 5.

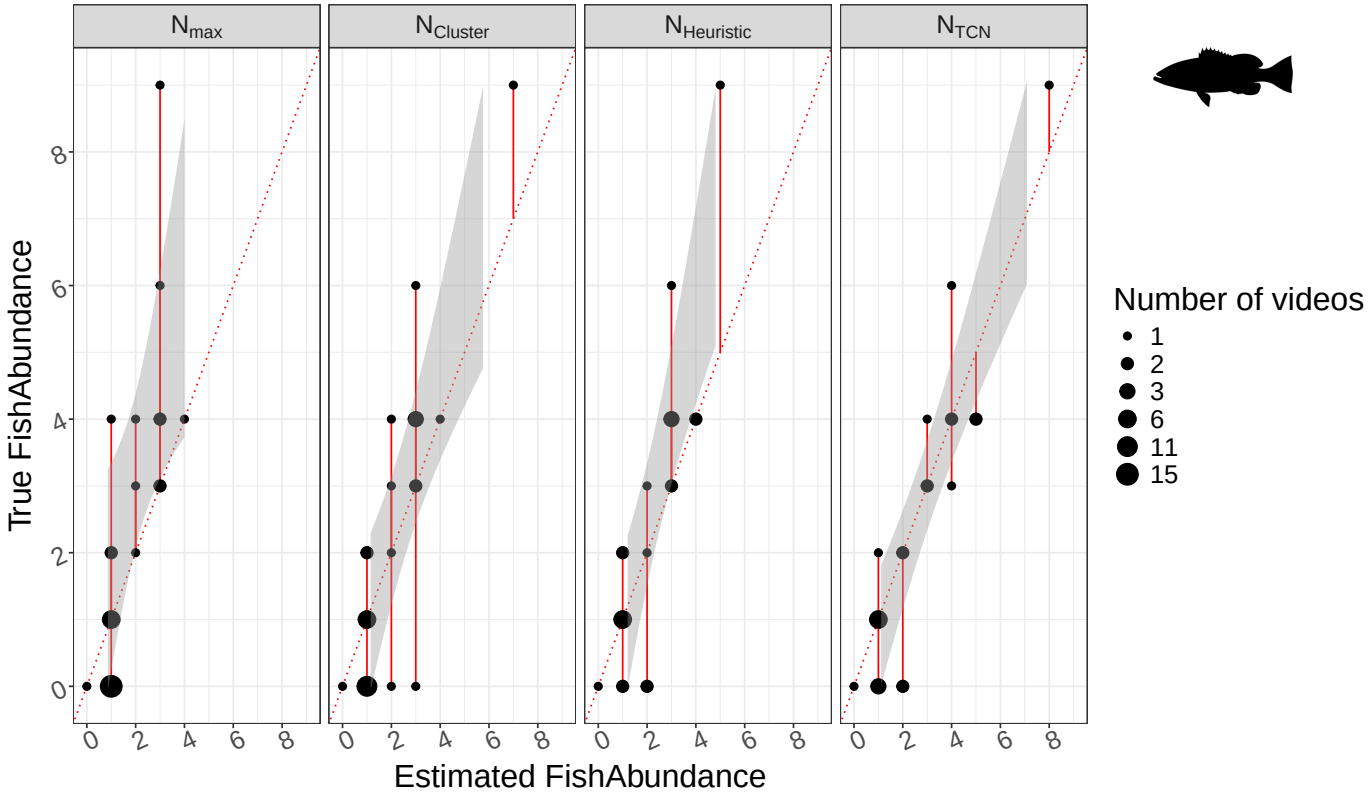

Figure 9: Linear regression (grey area indicates the 95% confidence interval) between he actual number of individuals of *E. marginatus* in test videos (y-axis) and the estimated count by the different methods (x-axis) on the detector predictions. The red line indicates a perfect prediction of the count. Size of the circular shapes present how many videos fall in this count-category. The majority of the videos (n=36) were at point 0,0. 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 the increase in challenge with the utility of the detector saw an increase of error of up to 55% for N_{max} and up to 49% for the other methods. Most stable was $N_{Heuristic}$ with an increase of 37% from 18% to 55%. This can be explained by insufficient detection capability of this species in the test dataset. Correlation values

remain above 0.9 for the proposed methods, even when low-occurrence videos are excluded. In contrast, for N_{max}, correlation reach 0.812 under the same exclusion conditions. These results are to be enjoyed with caution since the sample since is very low with only 9 videos for this species.

Table 6: The different methods on the detector predictions are presented in this table for the species *S. umbra*. Correlation with the actual counts on the test set are indicated with all points included (All), 0 excluded (wo0) and 0 and 1 excluded (wo01) to show the strength 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**$_{wo0}$	**Corr**$_{wo01}$
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 *wo*0) range from 11 individuals for $N_{Cluster}$ and 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 contributes significantly to the error in N_{max}, favoring our methods. This highlights a potential trend within this ecological niche or fish type, suggesting improved counting accuracy in high-occurrence videos from our methods.

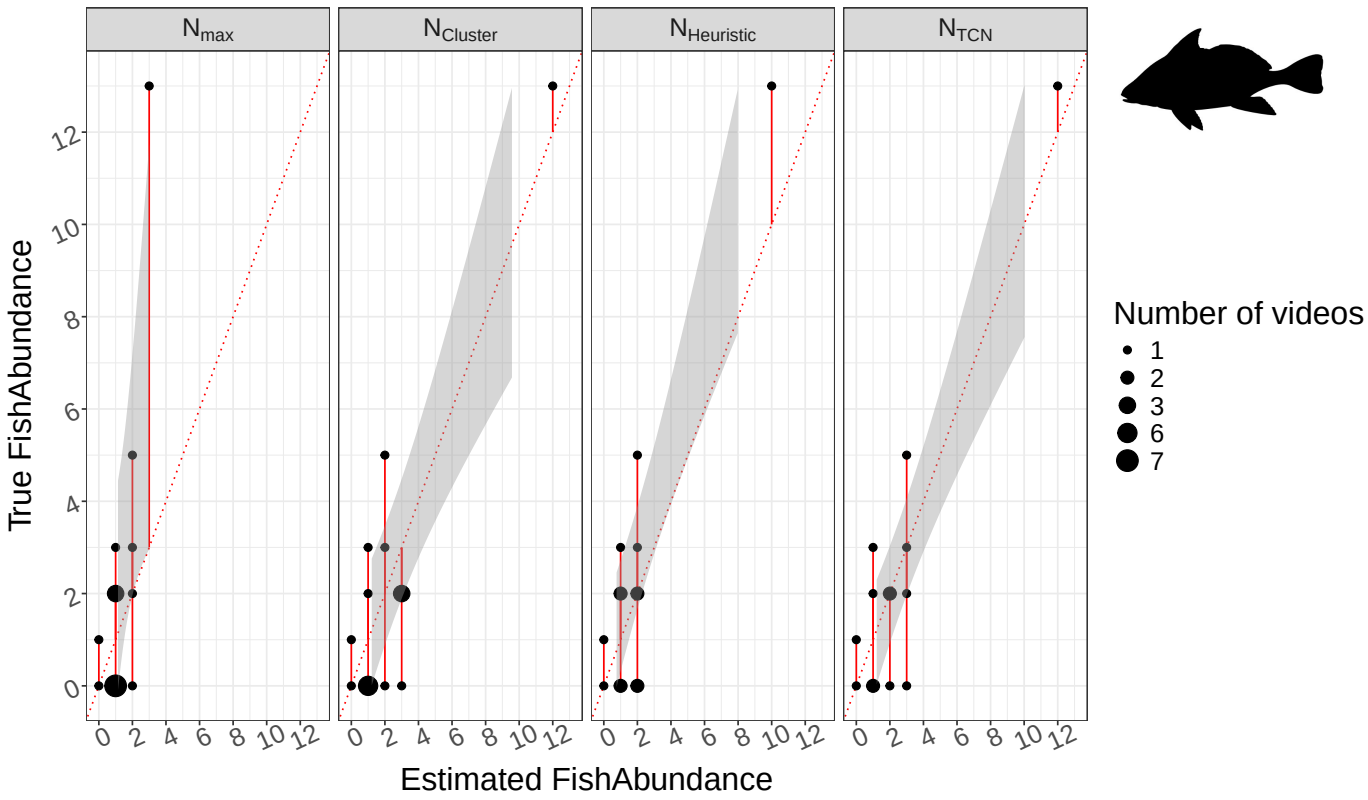

Figure 10: Linear regression (grey area indicates the 95% confidence interval) between he actual number of individuals of *S. umbra* in test videos (y-axis) and the counts of the different methods (x-axis) on the detector predictions. Size of the points indicate more videos overlapping with the corresponding methods and the actual count. The dashed red line indicates a perfect result, above the line is an underestimation and under the line indicates an overestimation. 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*

For our third species with a very different ecological niche than the two before we explored the impact of choosing a detector doing the detections instead of relaying on the groundtruth samples. We see the least change in error rates between all the species (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 enhances counting accuracy but also improves the training effectiveness of the DL

model. The error percentage stay under 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.

Table 7: The different methods on the detector predictions are presented in this table for the species *D. vulgaris*. Correlation with the actual counts on the test set are indicated with all points included (All), 0 excluded (wo0), 0 and 1 excluded (wo01) and videos with less than 10 individuals excluded (wo0:10) to show the strength 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**$_{wo0}$	**Corr**$_{wo01}$	**Corr**$_{wo0:10}$
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 assessed visually the impact of the absolute error and if there is an over- or underestimation (Fig. 11). We can see that N_{max} and $N_{Heuristic}$ underestimate the count while $N_{Cluster}$ is overestimating the count but less 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- nor underestimate the count but has a balanced variance around the ideal line. This is also numerically visible with high correlation values.

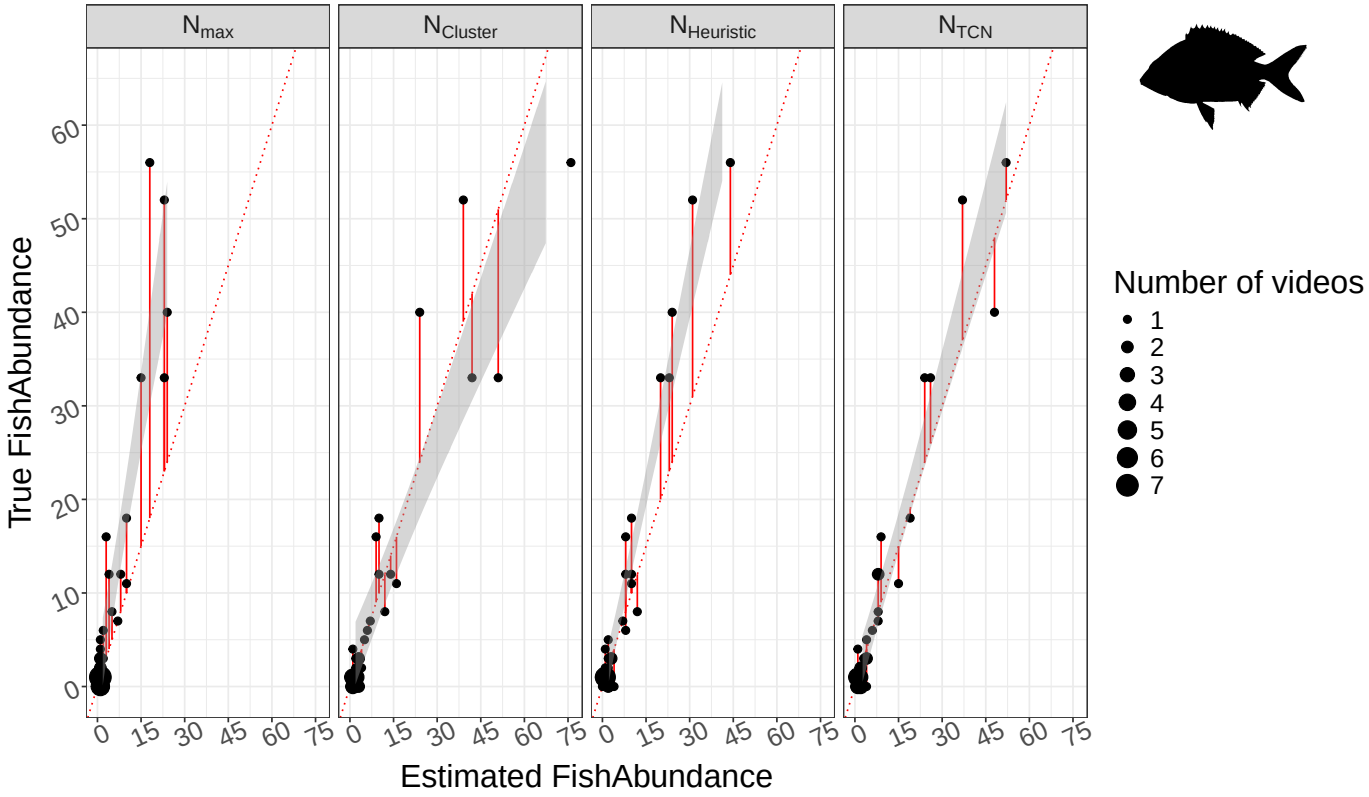

Figure 11: Linear regression (grey area indicates the 95% confidence interval) between he actual number of individuals of *D. vulgaris* in test videos (y-axis) and the counts of the different methods (x-axis) on the detector predictions. Size of the points indicate more videos overlapping with the corresponding methods and the actual count. The dashed red line indicates a perfect result, above the line is an underestimation and under the line indicates an overestimation. 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 is the first study that explores the automated FishAbundance counting in a DOV setup. We proposed three automated fish counting methods that work with detections from a deep learning model as an input in this proof of concept. These automation processes will significantly reduce the analysis time associated with manually calculating FishAbundance (Haberstroh et al. (2022)) or N_{max} (Raoult et al. (2020)).

With the three proposed methods of fish counting, we were able to show that even the simplest method in $N_{Cluster}$ outperforms the metric N_{max} widely used in BRUV and less used in DOV. N_{max} in all cases was underestimating 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 problem.

The issue with N_{max} in DOV is that distinct groups of the same fish species within a transect may go uncounted. This is especially important in the case of *E. marginatus* as this species exhibits solitary and territorial behavior (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 zero and one occurrence videos from the analysis and the correlation drastically dropped. In contrast, the correlation remained relatively stable across the other methods. For protection efforts and justifications, it is important to incorporate the count since the number of individuals is important for biomass calculations and overall health of a local population.

Methods The product of the N_{max} method is rather a frequency than a count for this species and can already give valuable insights on the species recovery. As Campbell et al. (2015) correctly mentioned, the N_{max} metric works for location-expanding species that appear in low numbers in new areas. In situations like this MeanCount is not ideal, while our methods also cover this type of scenario and can give even greater insights into this.

For a different scenario, N_{max} is chronically underestimating the count. On the other hand, our top performers in N_{TCN} and $N_{Heuristic}$ both have an error percentage of lower than 30% in a perfect case which are comparable to the error rate of divers (Pais and Cabral (2018); Ward-Paige et al. (2010)). $N_{Cluster}$ shows evidence of sufficient counting capability when the scenario is less complex and data is rare. The great advantage of $N_{Cluster}$ that no prior knowledge is needed for calibration nor training. The only influenceable parameter is the choice of how many clusters 'k' should be considered. This is dependent on video

length, 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.

$N_{Heuristic}$ also groups the different fish schools into clusters and uses the peak of each school to summarize the final count. The difference between $N_{Cluster}$ and $N_{Heuristic}$ is, that $N_{Heuristic}$ uses an intuitive procedure to justify the cluster differentiation that is dictated by a subset of the data provided, closely resembling each species by two parameters. 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 linear relation and decrease of error rate makes this approach valuable for instances when there is data available and the task does not exceed a certain complexity.

Taking it a step further we introduced N_{TCN}, that allows the fast addition of new species into the method pool that $N_{Heuristic}$ does not always allow. Furthermore, in complex examples (*i.e.* more individuals, less performant 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 performant method.

Impact of data scarcity on counting performance Organism counts and the resulting density numbers are one of the most important ecological indicators for health and state of natural systems (Ramos et al. (2012)). Especially for the two species *E. marginatus* and *S. umbra* who were protected just in recent time, a head count is of utmost importance to follow their evolution and potential recovery. Especially for these species a complete detector pipeline is important.

In our case, 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 training images. Especially with rare species, the image pool is small and this scarcity of the data is observable with the *S. umbra* that only had 9 videos available in the test set and 8

videos in the training set. This data scarcity affects more the detector than our method as seen by the differences in the count between the fully automated case (Table 6) and the perfect case (Table 3). The error rate increase 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 difference in error rate between fully automated and perfect case are less prominent. Linear correlation values are less affected by the detector compared 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 analyzing the results generated by these methods. Secondly, it enables the use of a remotely operated vehicle (ROV), allowing transects to be conducted from a safe distance. This will lead to increased frequencies of biodiversity assessments, helping our understanding of the marine environment and its evolution (Buscher et al. (2020)).

Future applications But not only the count but also the size per individual is an important indication for the well-being of a species (Duplisea and Castonguay (2006); Hallett et al. (2012)). With these methods a stereo system could automatically chose 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 made and therefore, the methods can be calibrated or trained without a more-effort, which makes the methods applicable to more scientific fields. This approach could facilitate and accelerate the identification and counting of invasive species using a moving camera, which may vary in origin from amateur to professional setups, and can be applied to a 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 alien species has become a threat for the environment and knowing the extent

of these invasions is crucial for healthy local and endemic ecosystems. While prevention is still the most successful tool (Keller et al. (2008)), an early recognition can lead to a more efficient battle against these invasions (*i.e.* 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 in diver-operated videos. While N_{max} remains vital for stationary camera setups, moving cameras offer an opportunity to explore alternative counting methods, reducing labor 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.

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.

Bai, S., Kolter, J. Z., and Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint 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.

513 Bürgi, K., Bouveyron, C., Lingrand, D., Dérijard, B., Precioso, F., and Sabourault,
514 C. (2024). Towards a fully automated underwater census for fish assemblages in the
515 mediterranean sea. *Ecological Informatics*, page 102959.

516 Buscher, E., Mathews, D. L., Bryce, C., Bryce, K., Joseph, D., and Ban, N. C. (2020).
517 Applying a low cost, mini remotely operated vehicle (rov) to assess an ecological baseline
518 of an indigenous seascape in canada. *Frontiers in Marine Science*, 7.

519 Calò, A., Pereñiguez, J. M., Hernandez-Andreu, R., and García-Charton, J. A. (2022).
520 Quotas regulation is necessary but not sufficient to mitigate the impact of scuba diving
521 in a highly visited marine protected area. *Journal of Environmental Management*,
522 302:113997.

523 Campbell, M. D., Pollack, A. G., Gledhill, C. T., Switzer, T. S., and DeVries, D. A. (2015).
524 Comparison of relative abundance indices calculated from two methods of generating
525 video count data. *Fisheries Research*, 170:125–133.

526 Connolly, R. M., Jinks, K. I., Herrera, C., and Lopez-Marcano, S. (2022). Fish surveys on
527 the move: Adapting automated fish detection and classification frameworks for videos
528 on a remotely operated vehicle in shallow marine waters. *Frontiers in Marine Science*,
529 9:918504.

530 Dìaz, S., Settele, J., Brondízio, E. S., Ngo, H. T., Agard, J., Arneth, A., Balvanera,
531 P., Brauman, K. A., Butchart, S. H., Chan, K. M., et al. (2019). Pervasive human-
532 driven decline of life on earth points to the need for transformative change. *Science*,
533 366(6471):eaax3100.

534 Dickens, L. C., Goatley, C. H., Tanner, J. K., and Bellwood, D. R. (2011). Quantifying
535 relative diver effects in underwater visual censuses. *PloS one*, 6(4):e18965.

536 Duplisea, D. E. and Castonguay, M. (2006). Comparison and utility of different size-based
537 metrics of fish communities for detecting fishery impacts. *Canadian Journal of Fisheries*
538 *and Aquatic Sciences*, 63(4):810–820.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Hutchings, J. A. and Reynolds, J. D. (2004). Marine fish population collapses: consequences for recovery and extinction risk. *BioScience*, 54(4):297–309.

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.

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.

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.

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.

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.

Maslin, M., Louis, S., Godary Dejean, K., Lapierre, L., Villéger, S., and Claverie, T. (2021).

592 Underwater robots provide similar fish biodiversity assessments as divers on coral reefs.
593 *Remote Sensing in Ecology and Conservation*, 7(4):567–578.

594 Pais, M. P. and Cabral, H. N. (2018). Effect of underwater visual survey methodology on
595 bias and precision of fish counts: a simulation approach. *PeerJ*, 6:e5378.

596 Pollard, D., Afonso, P., Bertoncini, A., Fennessy, S., Francour, P., and Barreiros, J.
597 (2018). Epinephelus marginatus. *The IUCN Red List of Threatened Species*, 2018:e–
598 T7859A100467602.

599 Pörtner, H. O. and Peck, M. A. (2010). Climate change effects on fishes and fisheries:
600 towards a cause-and-effect understanding. *Journal of fish biology*, 77(8):1745–1779.

601 Ramos, S., Amorim, E., Elliott, M., Cabral, H., and Bordalo, A. A. (2012). Early life stages
602 of fishes as indicators of estuarine ecosystem health. *Ecological Indicators*, 19:172–183.
603 Assessing ecological quality in estuarine and coastal ecosystems.

604 Ranganathan, C. S., Raman, R., Parikh, S., Rajesh, S., Meenakshi, R., and Muthulek-
605 shmi, M. (2023). Iot applications in marine monitoring: Protecting ocean health and
606 biodiversity. In *2023 International Conference on Sustainable Communication Networks*
607 *and Application (ICSCNA)*, pages 305–310.

608 Raoult, V., Tosetto, L., Harvey, C., Nelson, T. M., Reed, J., Parikh, A., Chan, A. J.,
609 Smith, T. M., and Williamson, J. E. (2020). Remotely operated vehicles as alternatives
610 to snorkellers for video-based marine research. *Journal of Experimental Marine Biology*
611 *and Ecology*, 522:151253.

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

615 Schobernd, Z. H., Bacheler, N. M., and Conn, P. B. (2014). Examining the utility of
616 alternative video monitoring metrics for indexing reef fish abundance. *Canadian Journal*
617 *of Fisheries and Aquatic Sciences*, 71(3):464–471.

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.

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.

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.

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.

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.

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.

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.

Supplementary Material

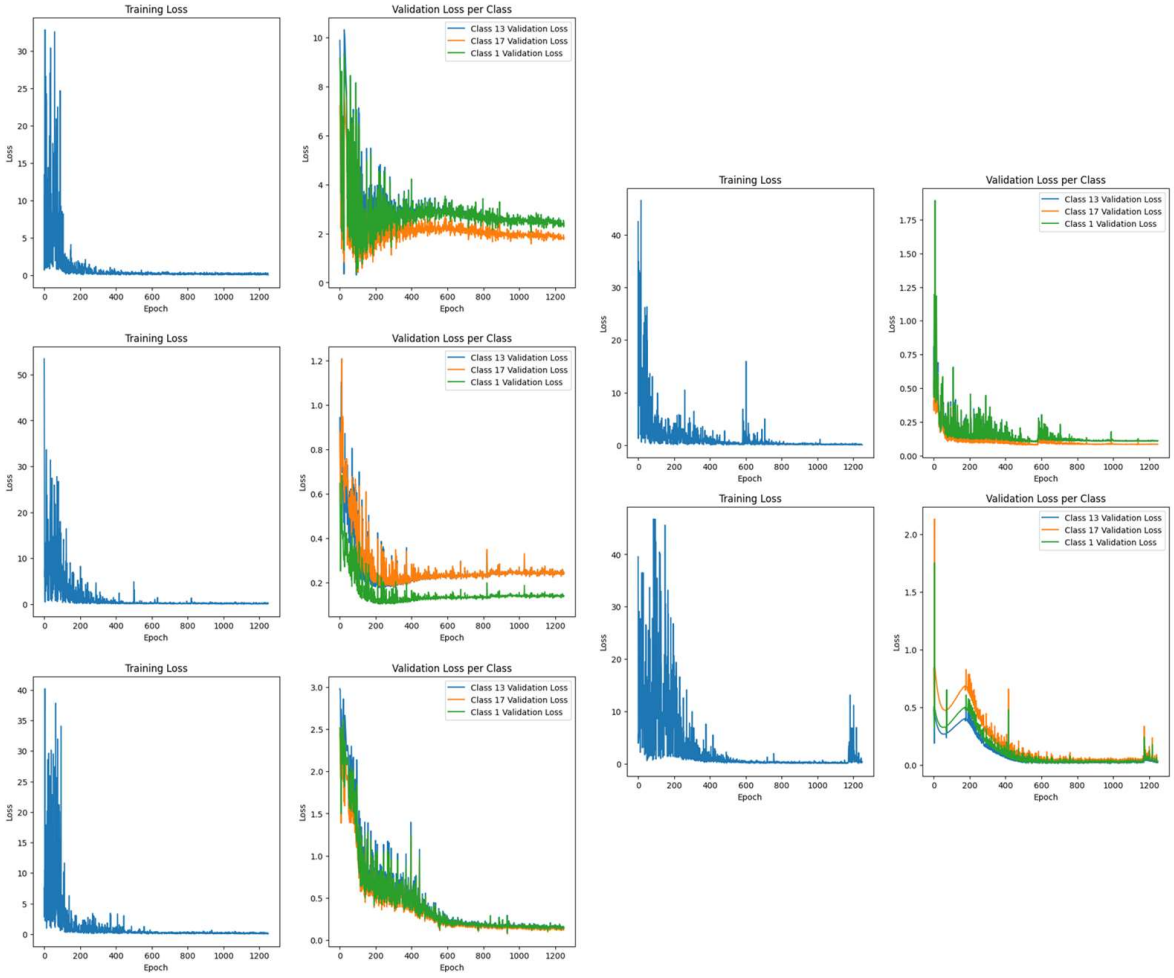

Figure S1: The 5 training runs for the TCN model for the perfect case used in the study.

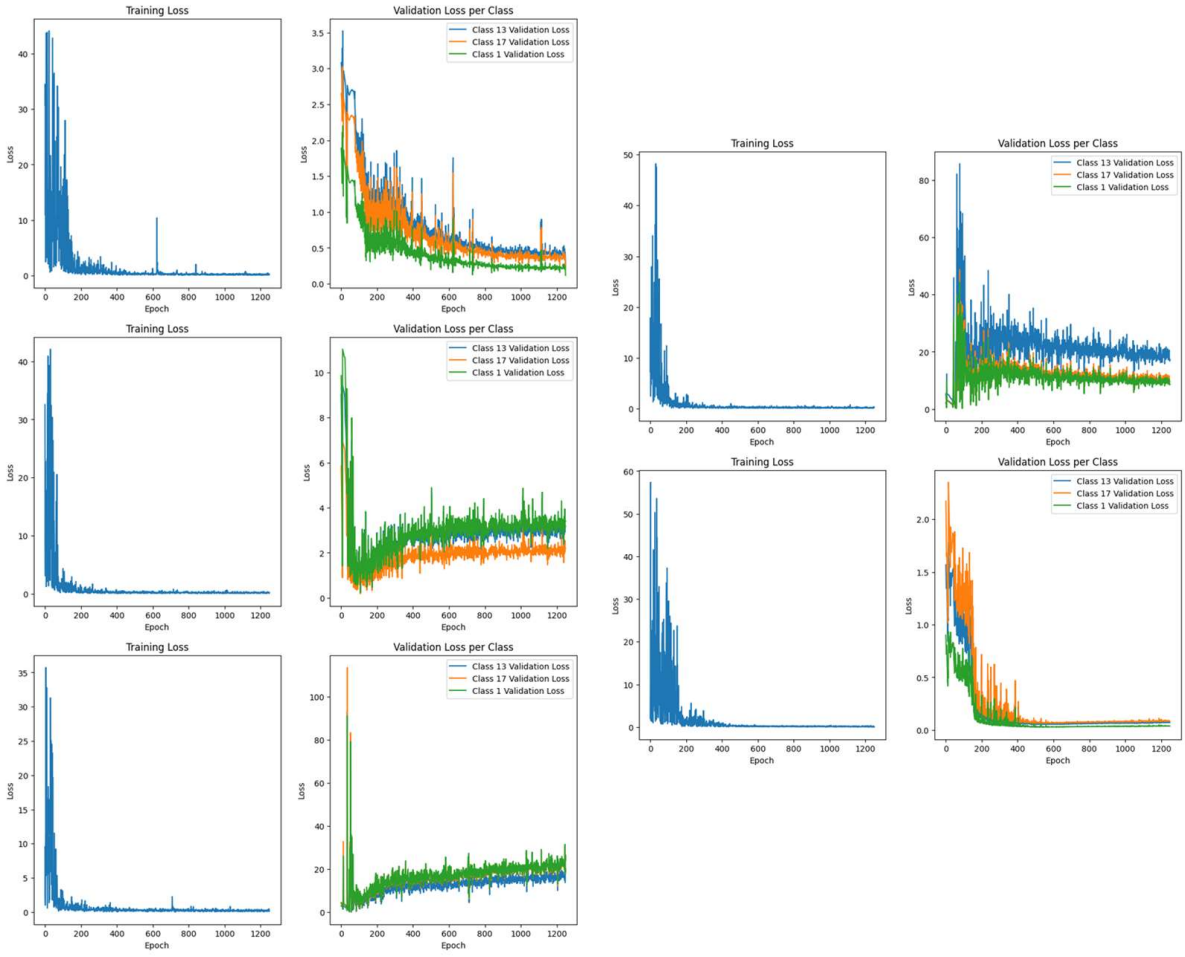

Figure S2: The 5 training runs for the TCN model for the fully automated case used in the study.

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