Impact of deep learning and post-processing algorithms performances on biodiversity metrics assessed on videos

Valentine Fleuré, Kévin Planolles, Thomas Claverie, Baptiste Mulot, Sébastien Villéger

RESEARCH ARTICLE

Impact of deep learning and post-processing algorithms performances on biodiversity metrics assessed on videos

Valentine Fleuré[1,2*], **Kévin Planolles**[1,3], **Thomas Claverie**[4], **Baptiste Mulot**[2], **Sébastien Villéger**[1]

1 MARBEC, University Montpellier, CNRS, Ifremer, IRD, Montpellier, France, **2** ZooParc de Beauval & Beauval Nature, Saint-Aignan, France, **3** Research-team ICAR, LIRMM, University Montpellier, CNRS, Montpellier, France, **4** UMR ENTROPIE, IRD, IFREMER, CNRS, Univ La Réunion, Saint Denis, Réunion, France

* valentine.fleure@gmail.com

OPEN ACCESS

Citation: Fleuré V, Planolles K, Claverie T, Mulot B, Villéger S (2025) Impact of deep learning and post-processing algorithms performances on biodiversity metrics assessed on videos. PLoS One 20(8): e0327577. https://doi.org/10.1371/journal.pone.0327577

Editor: Tzen-Yuh Chiang, National Cheng Kung University, TAIWAN

Received: March 17, 2025

Accepted: June 17, 2025

Published: August 11, 2025

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0327577

Abstract

Assessing the escalating biodiversity crisis, driven by climate change, habitat destruction, and exploitation, necessitates efficient monitoring strategies to assess species presence and abundance across diverse habitats. Video-based surveys using remote cameras are a promising, non-invasive way to collect valuable data in various environments. Yet, the analysis of recorded videos remains challenging due to time and expertise constraints. Recent advances in deep learning models have enhanced image processing capabilities in both object detection and classification. However, the impacts on models' performances and usage on assessment of biodiversity metrics on videos is yet to be assessed. This study evaluates the impacts of video processing rates, detection and identification model performance, and post-processing algorithms on the accuracy of biodiversity metrics, using simulated remote videos of fish communities and 14,406 simulated automated processing pipelines. We found that a processing rate of one image per second minimizes errors while ensuring detection of all species. However, even near-perfect detection (both recall and precision of 0.99) and identification (accuracy of 0.99) models resulted in overestimation of total abundance, species richness and species diversity due to false positives. We reveal that post-processing model outputs using a confidence threshold approach (i.e., to discard most erroneous predictions while also discarding a smaller proportion of correct predictions) is the most efficient method to accurately estimate biodiversity from videos.

Introduction

The escalating biodiversity crisis, exacerbated by climate change, habitat destruction, and exploitation, underscores the urgent need for efficient monitoring strategies of

Data availability statement: All data are fully available. Research data are available at https://doi.org/10.5281/zenodo.15519964. Code is available at https://github.com/valentine-fleure/deep_perf.

Funding: V. F. was supported by the Association Beauval Nature (CIFRE 2022/0127). K. P was supported by the IA-Biodiv ANR project FISH-PREDICT (ANR-21-AAFI-0001-01). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

biodiversity [1]. Such efficient monitoring requires consistently tracking the presence and abundance of species at large spatial scales, in various habitats and frequently to assess ecosystem changes accurately [2]. This challenge is particularly critical in underwater environments, where conducting biodiversity surveys is logistically complex and resource-intensive. Coral reefs, which alone host nearly one-fifth of all marine life on Earth [3], are among the most threatened ecosystems, making it essential to develop efficient monitoring techniques to assess and protect their biodiversity effectively. Various community assessment methods are conventionally used but none of them are yet satisfactory. Sampling-based approaches such as fishing and trawling observations yield valuable data, but they often come with ecological impacts [4]. Among non-invasive methods, eDNA allows censusing a high number of species (including elusive or cryptic ones) but it does not provide estimates of species abundance [5]. Acoustic based approaches allow long term monitoring and development of ecosystem health proxies, but do not census all species and do not accurately estimate abundance [6]. Visual counting methods performed by scuba divers are reference approaches to survey community composition but are notoriously time-consuming and require highly trained experts [7]. In response to these challenges, the use of video-based surveys has emerged as a promising solution [8] with diverse approaches depending on the animals and ecosystems studied with underwater camera traps [9], baited underwater stereo-cameras for marine predators [10,11], long-duration remote underwater cameras for reef fishes [12,13], autonomous underwater vehicles for reef fishes [14], and unmanned aerial vehicle for marine megafauna [15,16]. These image recorders allow for accurate and non-invasive monitoring of marine and terrestrial ecosystems and have thus been increasingly used for the last decade [8]. However, the on-screen analysis of recorded videos by experts remains a significant obstacle to scalability. The detection and identification of all targeted species are indeed demanding tasks even for highly-trained experts [17,18].

Meanwhile, the rise of the use of image recording and artificial intelligence applied to image processing have made significant progress during the last decade, mostly through the progress of deep learning algorithms [19]. Today, two of the main families of deep learning models applied to images are used: detection models and classification models. Detection models are used to find all objects of interest in an image [20]. Classification models are used to associate each image with a label [21]. While multi-class detection models do both tasks simultaneously [22,23], training them efficiently remains a challenge when most classes are rare, which is the case for most species assemblages [24–26]. Deep learning algorithms can identify all animals on images considerably faster than humans [27]. However, detection algorithms produce both false negatives (individuals that are not detected) and false positives (regions of image without any individuals but detected as containing one). In addition, identification algorithms may assign the wrong species to the correctly detected individuals. Finally, all false positives from the detection step, yield a misidentification as classifiers always attribute one of the learned classes to an input object. One approach to limit model errors consists in thresholding [28], i.e., setting aside raw outputs

according to a user-defined criterion. This criterion can be based on a confidence score returned by the models [29] or on the size of the bounding box [30].

Computer vision models are designed to process images [31]. Hence, to process videos the first step is to extract a subset of the video frames after choosing a frame rate (e.g., 1 image per second among the 25 or 30 images per second recorded). Then, models will analyze these images one by one. The selected frame processing rate affects the number of images analyzed hence the processing time as well as the potential to accurately assess biodiversity. Increasing the processing frame rate is expected to reduce false negatives from the detection model (more frames will increase the chance to catch an individual) and hence allow censusing more species and more individuals. Meanwhile, an increasing processing rate is also expected to increase the number of false positives from both detection and classification models resulting in biased species richness and species abundances. However, the impact of frame processing rate on quality of detection and identification outputs is yet to be assessed.

Furthermore, the ultimate goal of such automated video processing is to compute biodiversity metrics, such as total abundance, species richness, and species diversity, based on outputs from the models on the sequence of images from the video. In more detail, the species richness will be accurately estimated as long as at least one individual of each species is correctly detected and classified in at least one frame of the video, with the condition that none of the detections are classified as a species that does not appear in the video. Total abundance estimated using the maximum number of individuals present simultaneously in a single image (i.e. maxN) can be accurately estimated only if for each species all individuals are detected and identified on at least one of the images where the maximum number is visible [32]. This metric is highly sensitive to detection, as only a few frames typically contain the maximum number of individuals, hence false negatives prevent from censusing all individuals. It is also sensitive to classification, since misclassifications (i.e., false negative) in those key frames decrease the number of individuals from a given species seen simultaneously. Hence, some biodiversity metrics are by definition more sensitive to model errors, yet such differences remain unevaluated to date.

In the present study, we evaluate the influence of video processing rates, the performance of detection and identification models, and the use of post-processing thresholds on the accuracy of biodiversity metrics estimation using simulations of remotely censused communities. We achieve this using simulated remote videos of fish communities and simulated 14,406 automated processing pipelines, and calculating error rates and biodiversity metrics.

Materials and methods

Simulating video-based community surveys

We considered a hypothetical case study with a regional pool of 30 species (close to the richness observed for vertebrates in terrestrial and aquatic temperate ecoregions) [24,33,34].

The 30 species were divided into 8 groups with contrasting abundance and temporal abundance variation in a site (Table 1). More precisely, groups were characterized by varying: (i) abundance, defined as the maximum number of individuals seen simultaneously (ranging from 1 to 20); (ii) relative temporal occupancy, or the proportion of time they were present (ranging from 0.5% to 90%); and (iii) the clustering of their presence across time, represented by the number of time slots (from 1 to 15) when a species was visible.

These groups were designed to represent the variability of abundance, mobility (i.e., average moving speed and proportion of time spent stationary for feeding or resting), and gregariousness observed in most animal communities surveyed with remote cameras (e.g., solitary fox vs gregarious mouflon, gregarious surgeonfishes vs solitary barracuda). We note that the scarcest and most elusive species is represented by a single individual recorded for only half a second over 10 minutes while the most abundant species is represented by 20 individuals visible for 2 minutes over 10 minutes and the most present species is represented by 2 individuals visible for 9 minutes over 10 minutes (S1 Fig). The number of species in each group varied from 3 to 5 species (Table 1).

Table 1. Values used to describe the groups of species considered in this study.

group	abundance	relative temporal occupancy (%)	number of time slots	number of species in the regional pool	number of species in each 10-min video
1	20	20	5	3	1
2	2	90	15	3	1
3	2	30	15	3	2
4	8	10	1	3	2
5	8	5	1	3	2
6	1	5	1	5	2
7	1	1	1	5	3
8	1	0.5	1	5	2

https://doi.org/10.1371/journal.pone.0327577.t001

We simulated fixed remote video surveys with a total duration of 10 minutes. This length was chosen as a trade-off between the ability to simulate contrasted temporal abundance of species and the size of simulated datasets.

We considered that each video survey allows recording 15 species (out of the 30 from the regional pool) representing the 8 groups with 1–3 species, sorted randomly (Table 1).

Then, the simulation of abundance distribution through time of each species was led through 4 steps. First, the average duration of a presence slot was computed as the ratio between temporal occupancy (min) and the number of presence slots. Second, the duration of each presence slot was sorted from a normal distribution with the mean being the average presence time and the standard error as the square root of the mean time. Third, the time interval between two consecutive presence slots was iteratively sorted from a uniform distribution, ranging from the number of remaining time slots to the total remaining time. Lastly, the number of individuals within each presence slot was sorted from a binomial distribution with the duration of the time slot, incorporating the species' abundance.

These temporal abundances were simulated with a resolution of one second, which corresponds to the unit of measurement usually used during diver counts. They were reduced to 30 frames per second to simulate video recording at the site studied.

To demonstrate the robustness of our results, 10 video surveys were simulated to account for variability in species abundance through time. The number of replicates was limited to 10 to balance computational cost with reproducibility.

Simulation of video processing with detection and identification algorithms

For each video, we simulated an automated processing with a detection algorithm and then a classification algorithm with varying performances. For the detection model, we used recall to assess false negatives, ensuring that all present individuals are detected, and precision to evaluate false positives, minimizing incorrect detections. For the classification model, we relied on accuracy to measure the proportion of correctly classified images, providing an overall assessment of classification performance. This approach follows standard evaluation practices, where recall and precision are essential for detection tasks, while accuracy is commonly used for classification.

We first varied the processing rate of video processing with 7 levels from 0.25, 0.5, 1, 2, 5, 10, and 30 frames per second. The two first processing rates mean 1 image processed every 4 or 2 seconds, respectively. The last level is the processing of all the frames from a standard video recording. The subsampling consisted in keeping the first frame of each second, then evenly choosing among the next ones depending on the target processing rate.

The first step of the automated analysis was the detection of all individuals using a single-class detection algorithm. The recall measures the ability of a model to identify all relevant positive instances. The recall score of the algorithm varied from 0.60 (similar to the least efficient algorithm published) to 0.99 (i.e., hypothetically quasi-perfect algorithm) with 5 intermediate levels (0.70, 0.80, 0.85, 0.90, 0.95). To simulate this, each individual present in the image has a detection

probability calculated using a Bernoulli distribution with recall as a probability. False negatives, i.e., undetected individuals, are not treated in the rest of the simulations as will be the case in real-life pipelines.

The precision is the proportion of correctly labeled individuals among all those detected as present. The precision score of the algorithm also varied from 0.60 to 0.99 with 5 intermediate levels (0.70, 0.80, 0.85, 0.90, 0.95). We determined the number of false positives using the precision rate and the number of detections, then randomly added these false positives across all frames. We acknowledge that in real-world applications, false positives could be clustered, typically caused by recurring background features that resemble target objects.

Second, a species-level classification algorithm was simulated for each bounding box from the detection algorithm. The accuracy is the proportion of correctly labeled individuals among all individuals of this class. The algorithm had an accuracy score ranging from 0.60 to 0.99 with 5 intermediate levels (0.80, 0.70, 0.85, 0.90, 0.95). We thus simulated misidentification at random using 1-accuracy as a probability of misidentification and the identity of the erroneous species was randomly chosen among the remaining species from the regional pool (i.e., 29 species). We hence assume that the misidentifications were even across species (i.e., there was no species pair more likely to be confounded to each other). We acknowledge that in real-world applications misidentifications are often not random with on average lower accuracy for the classes with the fewer images in the train set and across classes with similar visible features.

Third, we implemented a post-processing step inspired by the confidence threshold approach of Villon (2020) that transfers some identifications (expected to be mostly misidentifications) to an unsure class. For each misidentification, there was a probability p (5 levels: 0.80, 0.85, 0.90, 0.95, 0.99) that it was eventually considered as “unsure”. Meanwhile, for each good identification, there was a probability $p/20$ that it was eventually considered as “unsure”. We also analyze results from detection and identification steps without this post-processing.

We thus simulated a total of 14,406 automated processing pipelines (7 processing rates * 7 detection recall * 7 detection precision * 7 classification accuracy * 6 post-processing thresholds) for each of the 10 simulated videos.

For each processing pipeline, we eventually computed the abundance of each species as the maximum number of individuals present in a single frame (i.e., MaxN metric commonly used in video-based surveys).

Assessing the performance of automated processing to estimate biodiversity metrics

As a preliminary analysis, we checked whether all individuals were detectable (i.e., present on at least one of the images) on the frames kept for analysis with each processing rate.

We assessed the influence of algorithms’ performance on three biodiversity metrics frequently used in the monitoring of ecosystems.

First, we computed total abundance as the sum of species abundances. Second, we computed species richness as the number of species seen along the video (i.e., at least one individual in at least one frame). Third, we computed species diversity using the Hill number version of the Shannon entropy [35]: $exp\left(-\sum_{i=1}^{S} p_i \times log\left(p_i\right)\right)$, where S is the total number of species present and p_i is the abundance of species i relative to the total number of individuals. Species diversity increases with increasing number of species and increasing evenness of their abundances.

These 3 indices were computed for data from ground truth (simulated communities) and for each output of each processing pipeline (species abundances provided by detection, classification, and post-processing algorithms).

To further investigate the outputs of the processing pipeline we computed Jaccard and Bray-Curtis similarity indices between species composition or abundance from each processing pipeline and species composition or abundances from ground truth.

The Jaccard index [36] quantifies the similarity between two sets of species by dividing the number of shared species by the total number of distinct species across both sets. It ranges from 0 (no species shared) to 1 (all species shared).

The Bray-Curtis similarity [37], quantifies differences between two sets of species abundances. it is defined as follows: $\frac{2\times\sum_{i=1}^{p} min(N_{ij},N_{ik})}{\sum_{i=1}^{n}(N_{ij}+N_{ik})}$, where N_{ij} is the number of individuals of species i at site j,N_{ik} is the number of individuals of species i at site k, and p is the total number of species in the samples. It ranges from 0 (sets have different species composition or abundance markedly differ for shared species) to 1 (all species have the same abundance in both sets).

Results

All 15 species simulated on each video were detectable (i.e., present on at least 1 of the images) for analyses with a processing rate of 1 image per second or higher. For the 2 lowest processing rates (0.25 and 0.5 images per second), 14.3 (sd = 0.67) and 14.8 (sd = 0.42) species on average were detectable, respectively (S1 Table).

For total abundance, the ground truth was 65 individuals present in each simulated community. Estimated abundance increased with the frame processing rate for all combinations of deep learning model performance, reaching a maximum of 187 individuals for a detection model with a recall of 0.99 and a precision of 0.6 (fifth row on Fig 1), and an identification model with an accuracy of 0.6 (red dot on Fig 1).

For all processing pipelines for which the processing frame rate was 2 and above, the estimated number of species was 30, which corresponds to the number of species learned by the classification model. For lower processing rates (0.25, 0.5 or 1), estimated species richness was lower than 30 and even reached 16 for a recall of 0.6, a precision of 0.99 and an accuracy of 0.99.

Species diversity was 8.71 for the ground truth. For all combinations of model parameters, the estimates of this metric were all above the ground truth and increased with the processing frame rate. It was at a minimum of 10.0 for a detector with a recall of 0.6 and a precision of 0.99 (second row on Fig 1) and a classification model with an accuracy of 0.99 (pink dot on Fig 1) for a processing frame rate of 0.25. Estimated diversity even reached a maximum of 27.8 for a detector with a recall of 0.6 and a precision of 0.6 (top row on Fig 1) and a classification model with an accuracy of 0.6 (red dot on Fig 1) for a processing frame rate of 30.

The Jaccard similarity index between actual species composition and those estimated after automated processing was constant at 0.5 except for 36 models with a processing rate of 0.25 or 0.5 frames per second, for which the Jaccard similarity was between 0.32 and 0.49 (S2 Fig).

The Bray Curtis similarity between actual abundances and those estimated after automated processing decreased with increasing processing rate from 0.72 for 1 frame processed by second to 0.64 for 30 frames processed by second (Fig 2).

Given these results, we focus hereafter on the effects of algorithm performance for a processing rate of 1 image per second to avoid missing out on rare species.

For total abundance, using a 0.99 confidence threshold post-processing, it was possible to get closer to the ground truth for all the detection models. A detection model with a recall of 0.6 and a precision of 0.6 (top row on Fig 3) estimated 7% more abundance than the ground truth, regardless of the accuracy of the classification model (colors of dots on Fig 3), when the post-processing threshold was between 0.8 and 0.95.

For a detection model with a precision of 0.6 (first and third rows on Fig 3) the post-processing threshold had no impact on the richness for a threshold under 0.95. Setting threshold above 0.99 led the species richness being closer to the ground truth [15], averaging 23.9 for a threshold of 0.99. For detection models with a precision of 0.99, the impact of threshold post-processing was greater when the accuracy of the classification model was close to 1 (Fig 3).

The Jaccard similarity between actual species composition and those estimated after automated post-processing decreased regardless of the classification model performance (S3 Fig). With a post-processing threshold of 0.99, the maximum increase was 0.45, going from 0.51 to 0.96 for a model with a recall of 0.99, a precision of 0.99, and a classification accuracy of 0.99.

Using a post-processing threshold increased the Bray-Curtis similarity (Fig 4) between actual abundances and those estimated after automated processing for all performances of detection or classification models. With a post-processing

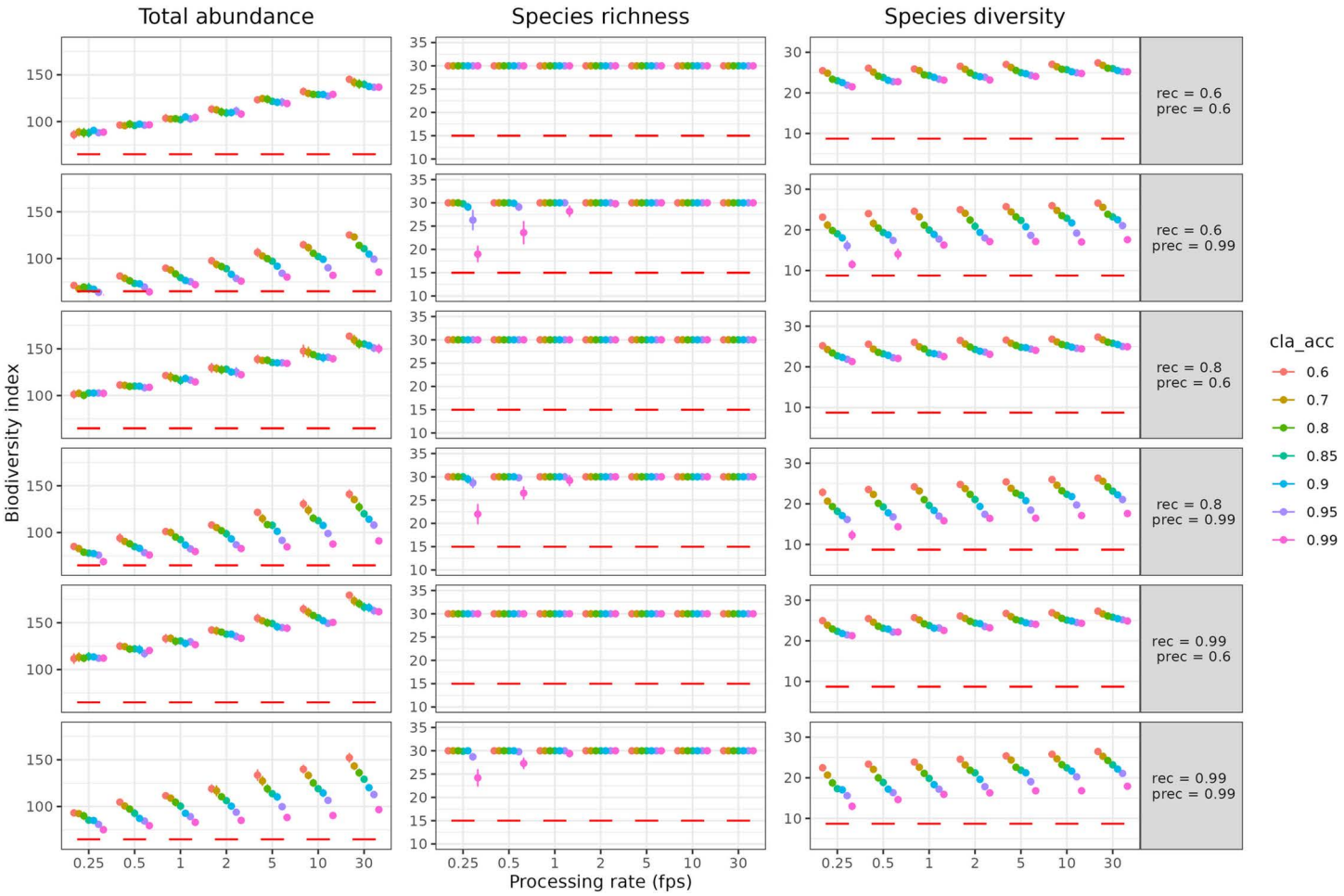

Fig 1. Effects of processing rate and algorithms performance on estimates of biodiversity. *Total abundance, species richness and species diversity estimates as a function of video processing rate (fps, frame per second) for 42 automated analysis models resulting from 6 detection models, with their respective recall ("rec") and precision ("prec") performances (rows) and 7 classification models (accuracy ("cla_acc") as colors). Each dot represents the average over the 10 simulated videos with corresponding standard error as vertical bars. The red horizontal bar represents the ground truth (i.e., diversity facet present on each video). Total abundance is the sum of the maxN (maximum number of individuals of a species seen in one image) of each species, richness is the number of species seen, and species diversity is the exponential of the Shannon entropy index.*

https://doi.org/10.1371/journal.pone.0327577.g001

threshold of 0.8, the minimum augmentation was 0.003, going from 0.791 to 0.794 for a model with a recall of 0.6, a precision of 0.95 and a classification accuracy of 0.99. With a post-processing threshold of 0.99, the maximum augmentation was 0.24, going from 0.61 to 0.85 for a model with a recall of 0.99, a precision of 0.6 and a classification accuracy of 0.6.

Although only 10 simulations were performed, the results were highly consistent across replicates, as indicated by the very low standard deviations around the mean values shown in the figures.

Discussion

In our simulations, the scarcest species were not always detectable by the models processing an image every 2 seconds (processing frame rate = 0.5) or an image every 4 seconds (processing frame rate = 0.25). Such species moving briefly across the camera field-of-view, such as fast-swimming predators (tuna, trevally) in the marine environment contribute

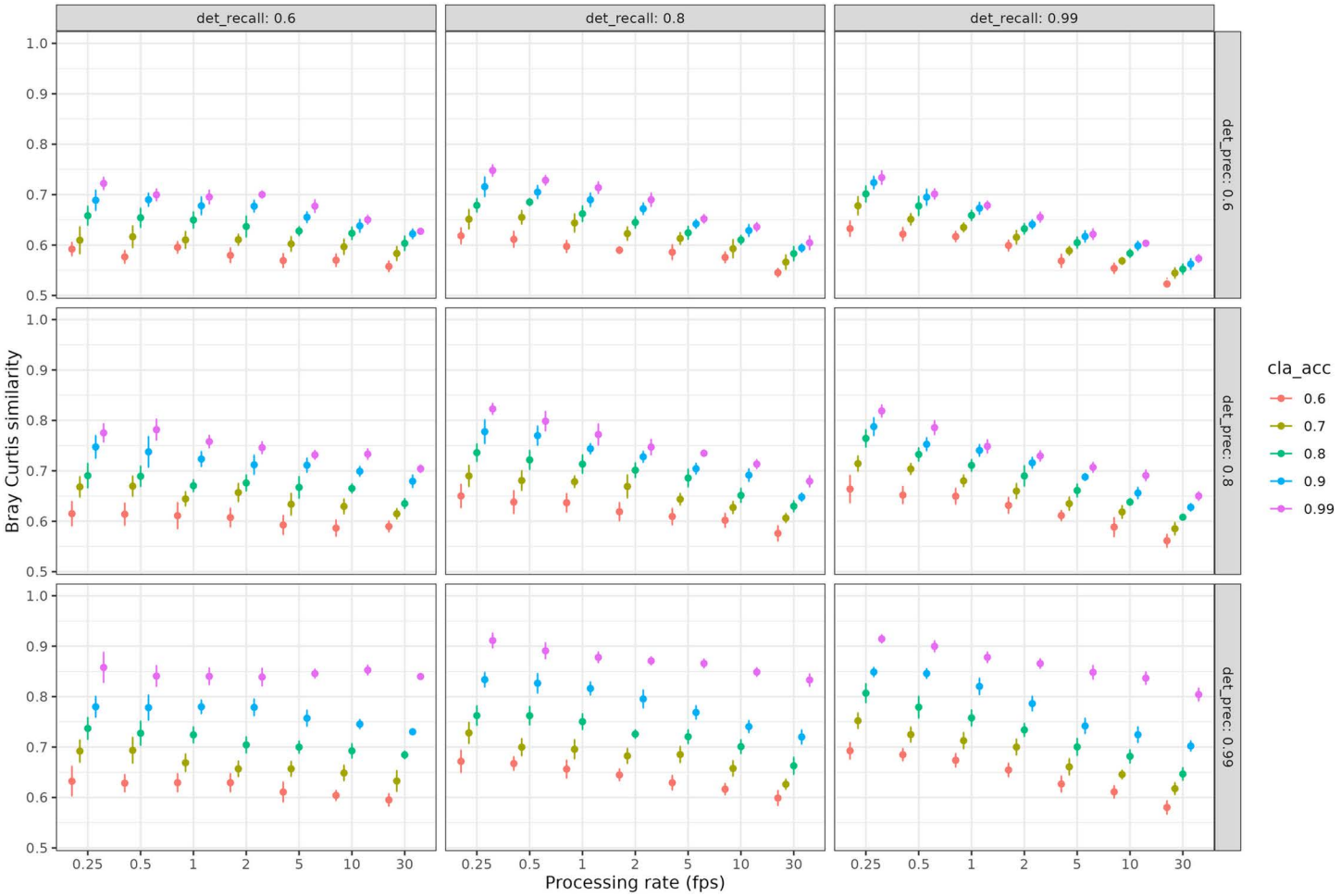

Fig 2. Effect of video processing rate and algorithms performance on the estimation of the abundance-structure of simulated communities. *Estimation of abundance-structure as a function of video processing rate (fps, frame per second) for 45 automated analysis models resulting from 9 detection models, with their respective recall ("det_recall" – columns) and precision ("det_prec" – rows) performances and 5 classification models (accuracy ("cla_acc") as colors). Each dot represents the average over the 10 simulations with corresponding standard error as vertical bars. The Bray-curtis similarity measures the difference in species abundances estimated after automated processing using algorithms and species abundances actually visible on videos.*

https://doi.org/10.1371/journal.pone.0327577.g002

more than 10% to the species richness [18] and play important functions for ecosystems [38]. Hence, in most studies, processing rates should at least be 1 image per second to detect scarce species.

Eventually, with all processing rates above 1 frame per second, all species were detected by detection models. Regardless of the performance of the detection models, for processing rates of 0.25 and 0.5 image per second, only the rare species present < 1% of the time, were not detected. Hence, overall, the false negatives from the detection model (i.e., individual missed on images) did not markedly affect the estimated composition of the assemblage. It is important to note that although errors in detecting rare species at the level of individual frames may result in underestimating their presence, combining observations from multiple frames typically diminishes this bias when assessing broader ecological indicators. Consequently, despite the occurrence of detection errors at the frame level, their overall influence on biodiversity metrics can be reduced through repeated sampling. This partly accounts for our simulations indicating a general overestimation rather than underestimation of species richness.

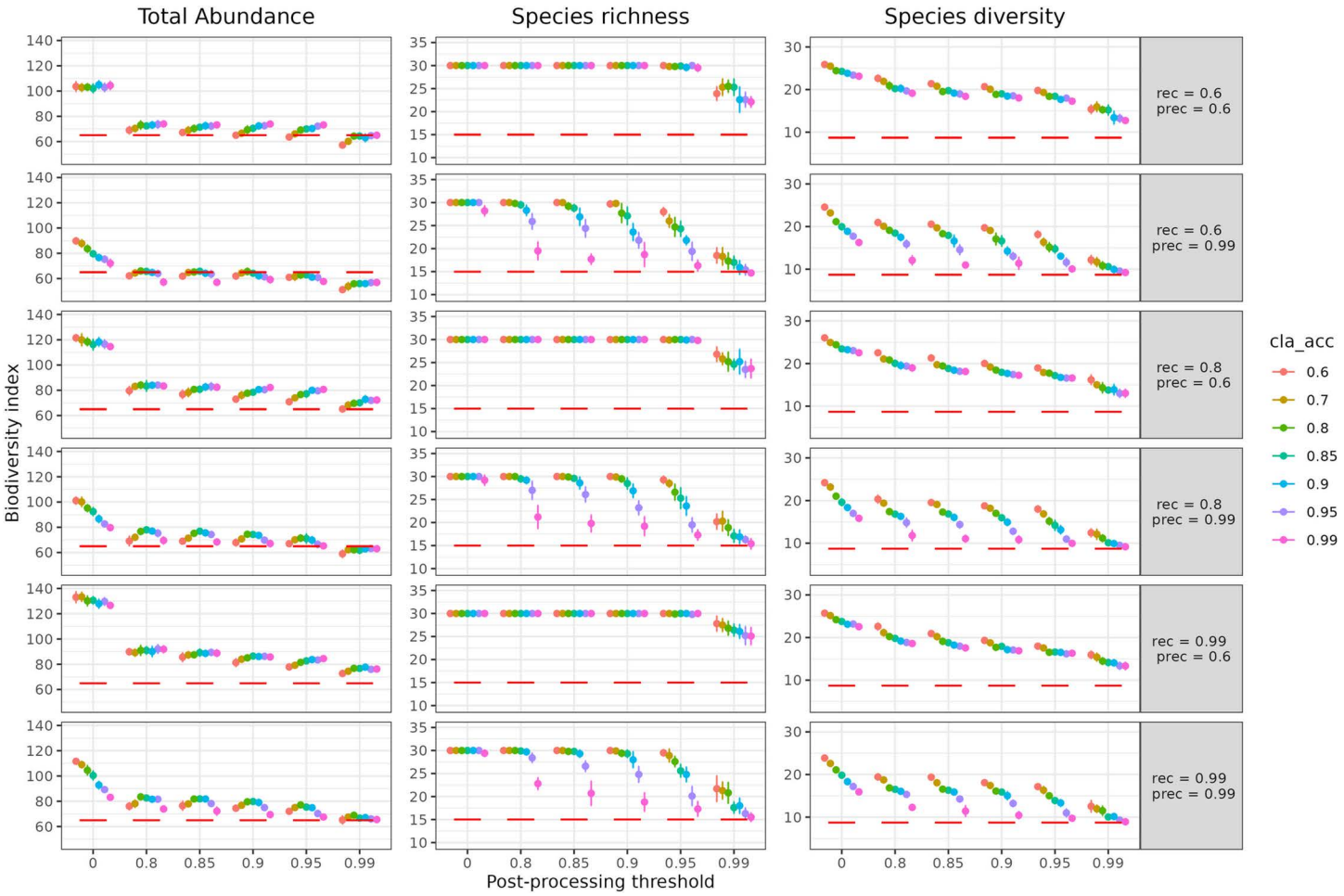

Fig 3. Effect of post-processing model outputs on biodiversity estimates. *Total abundance, species richness and species diversity estimates as a function of post-processing threshold for 42 automated analysis models resulting from 6 detection models, with their respective recall ("rec") and precision ("prec") performances (rows) and 7 classification models (accuracy ("cla_acc") as colors). Each dot represents the average over the 10 simulated videos with corresponding standard error as vertical bars. The red horizontal bar represents the ground truth. Total abundance is the sum of the maxN (maximum number of individuals of a species seen in one image) of each species, richness is the number of species seen, and species diversity is the exponential of the Shannon entropy index. Post-processing applies a confidence threshold to outputs of the identification models, discarding those with the lowest confidence scores to minimize misidentifications.*

https://doi.org/10.1371/journal.pone.0327577.g003

The near-perfect detection model (both recall and precision of 0.99) yielded on average 53 false positives (i.e., objects being actually part of the background) for 5,295 detectable fish when the 10-minutes videos were processed at 1 frame per second. There were up to 527 false positives for 52,927 detectable fish when videos were processed at 10 frames per second and 1,586 false positives for 158,789 detectable fish when processed at 30 frames per second. According to our simulation design, the number of false positives is proportional to the number of frames analyzed, which is realistic as false positives occur mostly because of textured sessile organisms (e.g., algae, corals, sponges) [39]. Therefore, we recommend using a processing frame rate of 1 frame per second for video analysis, in order to ensure all species are detected while minimizing the number of false positives. However, if fast-moving species are not present or are not in the scope of monitoring, it is possible to analyze fewer images per second. In such contexts, using a lower processing rate (e.g., 0.5 fps) does not compromise species detection, while significantly reducing the number of analyzed frames and the number of false positives, thereby reducing the proportion of error in biodiversity estimates.

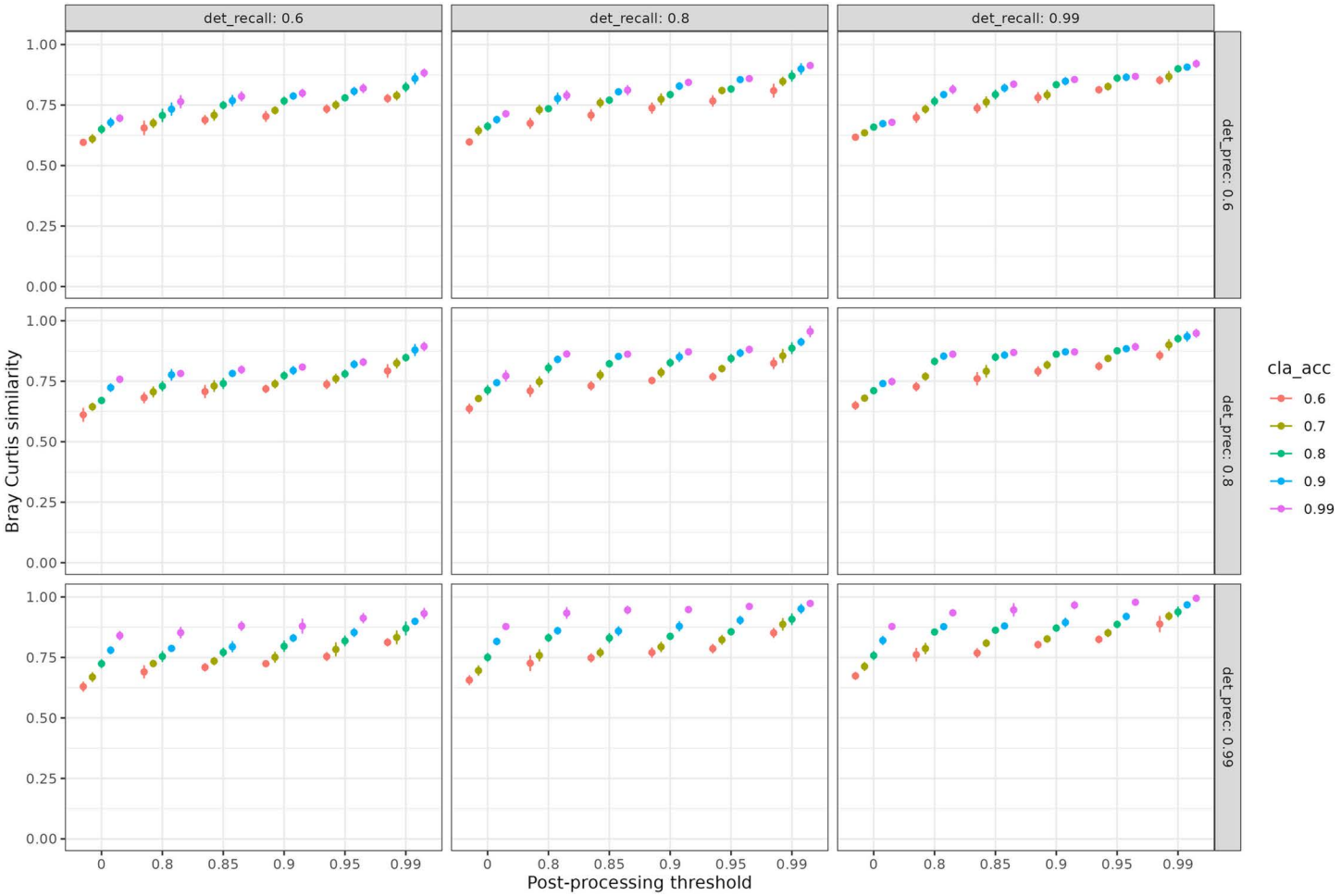

Fig 4. Effect of post-processing threshold and algorithms performance on the communities abundance-structure estimation. *Estimation of abundance-structure as a function of post-processing threshold for 45 automated analysis models resulting from 9 detection models, with their respective recall ("det_recall" – columns) and precision ("det_prec" – rows) performances and 5 classification models (accuracy ("cla_acc") as colors) for 1 frame per second processing rate. Each dot represents the average over the 10 simulations and the error bars are shown. The Bray-curtis similarity measures the difference in species abundances estimated after automated processing using algorithms and species abundances actually visible on videos. Each panel gathers results for a detection algorithm (recall in column and precision in row) and a classification algorithm (color corresponding to its accuracy). Post-processing applies a confidence threshold to outputs of the identification models, discarding those with the lowest confidence scores to minimize misidentifications.*

https://doi.org/10.1371/journal.pone.0327577.g004

All 30 species present in the environment were always detected and identified when analyzing 1 frame per second, regardless of the recall and precision scores of the detection model, and the accuracy of the classification model. There were hence 15 false positive species, arising from classification errors, including those from false positives from the detection model. However, even with a near-perfect detection model (recall and precision = 0.99) and a near-perfect classification model (accuracy = 0.99), there were on average 29.4 (sd = 0.7) species identified at the end of the processing pipeline. With a near-perfect classification model (accuracy = 0.99), the errors were mostly due to false positives in the detection models. Indeed, background objects erroneously detected as objects of interest only lead to a misclassification, likely evenly picked among the list of putative species (i.e., present in the training set of the classification algorithm). In our simulations, we considered that errors in the confusion matrix were distributed randomly. In real cases, there are some

species pairs that are confused more than others (e.g., species with similar shapes and colors, or species with the most images in the training set). In such cases the final error in species richness assessment would depend on the number and co-occurrence of these confounded species: if they are often co-occurring (i.e., present in the same habitats), confusion will not lead to bias in richness spatial patterns, but if they have different habitats, the species richness will be inflated in all habitats (i.e., the 2 species will be erroneously detected everywhere). Therefore, as there were a large number of false positives all species were eventually erroneously estimated to be present. If classification of false positives from the detection tend to be towards a small subset of species (e.g., to cryptic species because of their similarity with benthos), such overestimation of species occurrence could be lower, but will still be an issue for biodiversity monitoring. One way to limit errors for these species is to apply a filter based on the size of the detected objects. While our study may not fully represent all the real-world complexities, such as non-random error patterns or unique community structures, the restriction in computational resources limited the scope of scenarios we could explore. However, the simulation code is openly available and can be adapted to test additional or more complex cases.

In addition to biases in species occurrence, and hence richness of communities, detection and classification models can lead to biases in species abundance. Here we found that total abundance was overestimated by 160% when video was processed at 1 image per second by detection and classification models with low performance (recall and precision of detection both <0.8 and classification accuracy <0.8). Even with a near perfect detection model (recall and precision both 0.99) and a near perfect classification model (classification accuracy = 0.99), abundance was still overestimated by at least 128% due to false positives.

Similarly, species diversity, which accounts for the evenness of species abundances, was overestimated by all models although less than species richness.

This lower sensitivity of abundance-based metrics to model errors resulted from two potentially cumulative patterns. First, a species' abundance is overestimated only if a false positive of this species is added to one of the video frames where this species was the most abundant. Second, false negatives due to detection or classification models could compensate for false positives.

Importantly, our simulations revealed that post-processing outputs from the identification model markedly improved estimation of the three biodiversity metrics (Fig 3, S3 Fig).

A detection model with moderate performance (precision and recall at 0.6) and a moderate classification model (accuracy at 0.6) with a post-processing threshold at 0.99, which overestimated species richness by 160%, outperformed a near-perfect detection model (precision and recall at 0.99) and a near-perfect classification model without post-processing, which overestimated species richness by 200%. Furthermore, the pipeline with models having moderate performance and a strict post-processing provided more accurate estimates of total abundance and species diversity, with error of 38% and 82% respectively, than the pipeline with high-performance models without post-processing (error of 95% and 74%) (Fig 3). This comparison highlights the importance of implementing post-processing of model outputs based on a confidence threshold.

For all detection models with a given precision, the gain with post-processing was the same (Fig 4). Indeed, for a model with a recall of 0.8, Bray Curtis similarity increased from 0.83 to 0.93 with post-processing, whereas for a recall of 0.9, Bray Curstis similarity increased from 0.83 to 0.97 with post-processing. We therefore recommend improving precision rather than recall for a detection model when using a post-processing on the outputs of the classification model. This aligns with previous works (e.g., [40]), which showed that conventional metrics like classification error may not fully capture a model's ability to reliably estimate key ecological indicators.

The post-processing method used in this study is similar to the one proposed by Villon et al. [26] which requires only splitting the image dataset into 2 subsets, one for training the model and one for setting the confidence threshold. Given our findings, we recommend setting the threshold to 0.99 to minimize the number of false positives and misclassifications. This aligns with previous works (e.g., [41]), which showed that applying threshold on classification outputs improve

accuracy, particularly for rare classes. However, this comes at the cost of reduced recall, which may lead to nondetection of some rare species, a trade-off that should be carefully considered depending on the monitoring goal.

Findings from our simulations involving a succession of two models — detection and classification — are also valid with a detection model alone or a classification model alone. A detection model alone serves as a processing pipeline where an expert performs classification with near-perfect accuracy (0.99), while a classification model alone represents detection by a human expert with near-perfect precision and recall (both 0.99). Multiclass detection models [23] provide an alternative to a single-class detection followed by a classification approach [22], sharing an overall recall for the detection and the classification and species-specific accuracy, thus aligning with our approach and conclusions. However, post-processing these models is challenging due to difficulties in independently training them twice to establish confidence thresholds for all species, especially rare ones with limited images [29]. Moreover, classification accuracy in multiclass detectors is generally lower compared to separate-step pipelines, which reinforces current recommendations to separate detection and classification tasks [26].

In conclusion, selecting a processing rate of 1 image per second optimizes the trade-off of minimizing false positives and subsequent misidentifications while ensuring the detection of rare species. Using a detection model with a high recall then applying a post-processing on outputs of the model classification using a strict confidence threshold eventually allows faithful estimates of key biodiversity metrics. Overall, these results demonstrate that automated processing of video with deep learning models having high, although not perfect performances, will ease the assessment of disturbance gradients and the benefits of protection on biodiversity.

Supporting information

S1 Fig. Example of the distribution of species abundances within a video. The number of individuals of each species is illustrated with shades of green (white slots indicate absence of species).
(TIF)

S1 Table. Number of detectable species according to the processing rate of videos. Values are the number of simulations (out of 10) yielding 12, 13, 14, or 15 detectable species (i.e., at least 1 individual present on at least one frame). For each of the 7 processing rates (frames per second).
(DOCX)

S2 Fig. Effect of video processing rate and algorithms performance on the estimation of species composition of simulated communities. Estimation of the species composition as a function of video processing rate (fps, frame per second) for 45 automated analysis models resulting from 9 detection models, with their respective recall ("det_recall" – columns) and precision ("det_prec" – rows) performances and 5 classification models (accuracy ("cla_acc") as colors). Each dot represents the average over the 10 simulations with corresponding standard error as vertical bars. The Jaccard similarity index was used to measure the difference in composition of species present on the videos and the composition estimated after automated processing using algorithms.
(TIF)

S3 Fig. Effect of post-processing threshold and algorithms performance on the species composition estimation. Estimation of the species composition as a function of post-processing threshold for 45 automated analysis models resulting from 9 detection models, with their respective recall ("det_recall" – columns) and precision ("det_prec" – rows) performances and 5 classification models (accuracy ("cla_acc") as colors) for 1 frame per second processing rate. Each dot represents the average over the 10 simulations and the error bars are shown. Each panel gathers results for a detection algorithm (recall in column and precision in row) and a classification algorithm (color corresponding to its accuracy). Post-processing applies a confidence threshold to outputs of the identification models, discarding those with the lowest

confidence scores to minimize misidentifications. The Jaccard similarity index was used to measure the difference in composition of species present on the videos and the composition estimated after automated processing using algorithms.
(TIF)

Acknowledgments

The authors thank the two anonymous reviewers who helped improve and clarify this manuscript.

Author contributions

Conceptualization: Valentine Fleuré, Sébastien Villéger.

Methodology: Valentine Fleuré.

Visualization: Valentine Fleuré.

Writing – original draft: Valentine Fleuré.

Writing – review & editing: Valentine Fleuré, Kévin Planolles, Thomas Claverie, Baptiste Mulot, Sébastien Villéger.

References

1. Bush A, Sollmann R, Wilting A, Bohmann K, Cole B, Balzter H, et al. Connecting Earth observation to high-throughput biodiversity data. Nat Ecol Evol. 2017;1(7):1–9.
2. Sparrow BD, Edwards W, Munroe SEM, Wardle GM, Guerin GR, Bastin J-F, et al. Effective ecosystem monitoring requires a multi-scaled approach. Biol Rev Camb Philos Soc. 2020;95(6):1706–19. https://doi.org/10.1111/brv.12636 PMID: 32648358
3. Spalding M. World atlas of coral reefs. 2001.
4. Trenkel V, Vaz S, Albouy C, Brind'Amour A, Erwan D, Laffargue P, et al. We can reduce the impact of scientific trawling on marine ecosystems. Marine Ecol Progress Series. 2018;609.
5. Cristescu ME, Hebert PDN. Uses and misuses of environmental DNA in biodiversity science and conservation. Ann Rev Ecol Evol Syst. 2018;49:209–30.
6. Rostami B, Nansen C. Application of active acoustic transducers in monitoring and assessment of terrestrial ecosystem health—A review. Methods Ecol Evol. 2022;13(12):2682–91.
7. Wetz JJ, Ajemian MJ, Shipley B, Stunz GW. An assessment of two visual survey methods for documenting fish community structure on artificial platform reefs in the Gulf of Mexico. Fisheries Res. 2020;225:105492.
8. Besson M, Alison J, Bjerge K, Gorochowski TE, Høye TT, Jucker T, et al. Towards the fully automated monitoring of ecological communities. Ecol Lett. 2022;25(12):2753–75. https://doi.org/10.1111/ele.14123 PMID: 36264848
9. Bilodeau SM, Schwartz AWH, Xu B, Paúl Pauca V, Silman MR. A low-cost, long-term underwater camera trap network coupled with deep residual learning image analysis. PLoS One. 2022;17(2):e0263377. https://doi.org/10.1371/journal.pone.0263377 PMID: 35108340
10. Shortis M, Harvey ES, Abdo DA. A review of underwater stereo-image measurement for marine biology and ecology applications. Oceanogr Marine Biol. 2009;47:257–92.
11. Griffin RA, Robinson GJ, West A, Gloyne-Phillips IT, Unsworth RKF. Assessing fish and motile fauna around offshore windfarms using stereo baited video. PLoS One. 2016;11(3):e0149701. https://doi.org/10.1371/journal.pone.0149701 PMID: 26934587
12. Magneville C, Leréec Le Bricquir ML, Dailianis T, Skouradakis G, Claverie T, Villéger S. Long-duration remote underwater videos reveal that grazing by fishes is highly variable through time and dominated by non-indigenous species. Remote Sens Ecol Conserv. 2023;9(3):311–22.
13. Dunkley K, Dunkley A, Drewnicki J, Keith I, Herbert-Read JE. A low-cost, long-running, open-source stereo camera for tracking aquatic species and their behaviours. Methods Ecol Evol. 2023;14(10):2549–56.
14. Maslin M, Louis S, Godary Dejean K, Lapierre L, Villéger S, Claverie T. Underwater robots provide similar fish biodiversity assessments as divers on coral reefs. Remote Sens Ecol Conserv. 2021;7(4):567–78. https://doi.org/10.1002/rse2.209
15. Desgarnier L, Mouillot D, Vigliola L, Chaumont M, Mannocci L. Putting eagle rays on the map by coupling aerial video-surveys and deep learning. Biol Conserv. 2022;267:109494.
16. Schiele M, Rowcliffe JM, Clark B, Lepper P, Letessier TB. Using water-landing, fixed-wing UAVs and computer vision to assess seabird nutrient subsidy effects on sharks and rays. Remote Sens Ecol Conserv. 2023;10(3):416–30. https://doi.org/10.1002/rse2.378
17. Gilbert NA, Clare JDJ, Stenglein JL, Zuckerberg B. Abundance estimation of unmarked animals based on camera-trap data. Conserv Biol. 2021;35(1):88–100. https://doi.org/10.1111/cobi.13517 PMID: 32297655

18. Magneville C, Brissaud C, Fleuré V, Loiseau N, Claverie T, Villéger S. A new framework for estimating abundance of animals using a network of cameras. Limnol Oceanogr. 2024;22(4):268–80.
19. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44. https://doi.org/10.1038/nature14539 PMID: 26017442
20. Kaur R, Singh S. A comprehensive review of object detection with deep learning. Digital Signal Process. 2023;132:103812.
21. Sharma S, Guleria K. Deep learning models for image classification: comparison and applications. In: 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). 2022: 1733–8. https://ieeexplore.ieee.org/document/9823516
22. Jenrette J, Liu ZYC, Chimote P, Hastie T, Fox E, Ferretti F. Shark detection and classification with machine learning. Ecol Inform. 2022;69:101673.
23. Roy AM, Bhaduri J, Kumar T, Raj K. WilDect-YOLO: An efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection. Ecol Inform. 2023;75:101919.
24. Catalán IA, Álvarez-Ellacuría A, Lisani J-L, Sánchez J, Vizoso G, Heinrichs-Maquilón AE, et al. Automatic detection and classification of coastal Mediterranean fish from underwater images: good practices for robust training. Front Mar Sci. 2023;10. https://doi.org/10.3389/fmars.2023.1151758
25. Marrable D, Barker K, Tippaya S, Wyatt M, Bainbridge S, Stowar M, et al. Accelerating species recognition and labelling of fish from underwater video with machine-assisted deep learning. Front Mar Sci. 2022;9. https://doi.org/10.3389/fmars.2022.944582
26. Gadot T, Istrate Ș., Kim H, Morris D, Beery S, Birch T, et al. To crop or not to crop: Comparing whole-image and cropped classification on a large dataset of camera trap images. IET Comput Vision. 2024;18(8):1193–208.
27. Ditria EM, Lopez-Marcano S, Sievers M, Jinks EL, Brown CJ, Connolly RM. Automating the analysis of fish abundance using object detection: optimizing animal ecology with deep learning. Front Mar Sci. 2020;7:429.
28. Cowans A, Lambin X, Hare D, Sutherland C. Improving the integration of artificial intelligence into existing ecological inference workflows. Methods Ecol Evol. 2024. https://doi.org/10.1111/2041-210x.14485
29. Villon S, Mouillot D, Chaumont M, Subsol G, Claverie T, Villéger S. A new method to control error rates in automated species identification with deep learning algorithms. Sci Rep. 2020;10(1):10972. https://doi.org/10.1038/s41598-020-67573-7 PMID: 32620873
30. Fleuré V, Magneville C, Mouillot D, Villéger S. Automated identification of invasive rabbitfishes in underwater images from the Mediterranean Sea. Aquatic Conserv. 2024;34(1):e4073.
31. Gill R, Logeshwaran J, Aggarwal P, Srivastava D, Verma A. Potential of deep learning for image processing. In: 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). 2024: 1–6. https://ieeexplore.ieee.org/document/10522154
32. Priede IG, Bagley PM, Smith A, Creasey S, Merrett NR. Scavenging deep demersal fishes of the Porcupine Seabight, north-east Atlantic: observations by baited camera, trap and trawl. J Marine Biol Assoc United Kingdom. 1994;74(3):481–98.
33. Bürgi K, Bouveyron C, Lingrand D, Derijard B, Precioso F, Sabourault C. Towards a fully automated underwater census for fish assemblages in the Mediterranean Sea. Ecol Inform. 2025;85:102959.
34. Rigoudy N, Dussert G, Benyoub A, Besnard A, Birck C, Boyer J. The DeepFaune initiative: a collaborative effort towards the automatic identification of European fauna in camera trap images. Eur J Wildl Res. 2023;69(6):113.
35. Jost L. Entropy and diversity. Oikos. 2006;113(2):363–75.
36. Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et du Jura. Bulletin de la Société Vaudoise des Sciences Naturelles. 1901;37(142):547.
37. Bray JR, Curtis JT. An ordination of the upland forest communities of southern Wisconsin. Ecol Monographs. 1957;27(4):325–49.
38. Mouillot D, Bellwood DR, Baraloto C, Chave J, Galzin R, Harmelin-Vivien M, et al. Rare species support vulnerable functions in high-diversity ecosystems. PLoS Biol. 2013;11(5):e1001569. https://doi.org/10.1371/journal.pbio.1001569 PMID: 23723735
39. Salman A, Siddiqui SA, Shafait F, Mian A, Shortis MR, Khurshid K. Automatic fish detection in underwater videos by a deep neural network-based hybrid motion learning system. ICES J Marine Sci. 2020;77(4):1295–307.
40. Bevan PA, Pantazis O, Pringle H, Ferreira GB, Ingram DJ, Madsen E. Deep learning-based ecological analysis of camera trap images is impacted by training data quality and quantity. arXiv. 2025. http://arxiv.org/abs/2408.14348
41. Willi M, Pitman RT, Cardoso AW, Locke C, Swanson A, Boyer A. Identifying animal species in camera trap images using deep learning and citizen science. Methods Ecol Evol. 2019;10(1):80–91.