



Fast and accurate mapping of fine scale abundance of a VME in the deep sea with computer vision

Nils Piechaud^{*}, Kerry L. Howell

School of Biological and Marine Sciences, University of Plymouth, Drake Circus, PL4 8AA, Plymouth, UK

ARTICLE INFO

Keywords:

Benthic ecology
Computer vision
Xenophyophores
Quantitative ecology
Mapping
Automated image analysis
Marine conservation

ABSTRACT

With growing anthropogenic pressure on deep-sea ecosystems, large quantities of data are needed to understand their ecology, monitor changes over time and inform conservation managers. Current methods of image analysis are too slow to meet these requirements. Recently, computer vision has become more accessible to biologists, and could help address this challenge. In this study we demonstrate a method by which non-specialists can train a YOLOV4 Convolutional Neural Network (CNN) able to count and measure a single class of objects. We apply CV to the extraction of quantitative data on the density and population size structure of the xenophyophore *Syringammina fragilissima*, from more than 58,000 images taken by an AUV 1200 m deep in the North-East Atlantic. The workflow developed used open-source tools, cloud-base hardware, and only required a level of experience with CV commonly found among ecologists. The CNN performed well, achieving a recall of 0.84 and precision of 0.91. Individual counts per image and size measurements resulting from model predictions were highly correlated (0.96 and 0.92, respectively) with manually collected data. The analysis could be completed in less than 10 days thus bringing novel insights into the population size structure and fine scale distribution of this Vulnerable Marine Ecosystem. It showed *S. fragilissima* distribution is patchy. The average density is 2.5 ind.m⁻² but can vary from up to 45 ind.m⁻² only a few tens of meter away from areas where it is almost absent. The average size is 5.5 cm and the largest individuals (>15 cm) tend to be in areas of low density. This study demonstrates how researchers could take advantage of CV to quickly and efficiently generate large quantitative datasets data on benthic ecosystems extent and distribution. This, coupled with the large sampling capacity of AUVs could bypass the bottleneck of image analysis and greatly facilitate future deep-ocean exploration and monitoring. It also illustrates the future potential of these new technologies to meet the goals set by the UN Ocean Decade.

1. Introduction

The deep sea, by convention, the portion of the oceans deeper than 200 m, covers most of the planet and provides many ecosystem services (Borja et al., 2016; Thurber et al., 2014) but is not understood as well as other ecosystems (Levin et al., 2019; Ramirez-Llodra et al., 2010, 2011). As anthropogenic pressure on these ecosystems is increasing (Halpern et al., 2007), the international scientific community is racing to acquire relevant biological data to inform the development of effective conservation strategies in line with the aims of the UN Decade of Ocean Science for Sustainable Development (Danovaro et al., 2017; Folkersen et al., 2018; Howell et al., 2020; Levin et al., 2019; Poore et al., 2015).

Recent research has highlighted the limitations of simple records of species presence in the sustainable management of the deep sea

(Danovaro et al., 2020; Howell et al., 2020; Levin et al., 2019; Miloslavich et al., 2018; Woodall et al., 2018). These authors argue that quantitative data are needed to understand complex phenomenon such response to disturbance, potential for dispersal, biomass fluctuations, or shift in standing stocks and spatial distribution, in the face of increasing human use and climate change. Quantitative data on deep-sea mega-faunal species is difficult to acquire. The low density of many deep-sea species necessitates survey over a large area to obtain sufficient observations to reliably measure key biological variables such as density and population size structure (McClain and Rex, 2015; Perkins et al., 2016; Rex and Etter, 2010). Robust sampling designs also necessitate replication in the measures made (Chapman and Underwood, 2008), which multiplies the size of datasets by at least a factor of 3, hence raising the challenge of collecting representative datasets in the deep sea.

^{*} Corresponding author.

E-mail address: Nils.Piechaud@plymouth.ac.uk (N. Piechaud).

<https://doi.org/10.1016/j.ecoinf.2022.101786>

Received 22 April 2022; Received in revised form 19 August 2022; Accepted 20 August 2022

Available online 24 August 2022

1574-9541/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Developments in Autonomous Underwater Vehicles (AUV) mean that scientists now have the capacity to survey large areas of seafloor in a short time using image-based methods (Huvenne et al., 2018; Jones et al., 2019; Morris et al., 2014, 2016; Wöfl et al., 2019; Wynn et al., 2014). With large datasets, questions that were previously difficult to address with statistical robustness can now be investigated and there is hope that new insights on benthic species distribution can be gained using these tools (Brandt et al., 2016; Danovaro et al., 2014; McClain and Rex, 2015). AUVs have the capacity to gather not only images, but other types of environmental data like hydrographic, oceanographic and topographic data from the exact same location (Wynn et al., 2014), thus, providing better access to the fine-scale environmental data sought by ecologists (Meyer et al., 2019; Milligan et al., 2016; Perkins et al., 2016).

Benthic megafauna studies that use images as samples require the translation of information contained in the pixels to semantic level (i.e. locate and identify object or events within the images) that can subsequently be turned into ecological metrics such as abundance or diversity (Gomes-Pereira et al., 2016). This step, referred to as image analysis or image annotation mainly relies on manual analysis that is both slow and biased (Culverhouse et al., 2003; Durden et al., 2016; Schoening et al., 2017). The resulting bottleneck is a well-known issue that restrict the pace of deep-sea exploration and, to an extent, prevents the exploitation of the large datasets collected by modern sampling tools, particularly AUVs. Solutions to this bottleneck have been developed including better image analysis software (Gomes-Pereira et al., 2016) or citizen science (Matabos et al., 2017) but automation of analysis through Computer Vision (CV) and Deep-Learning (DL) has also gained traction (Hoese and Kuenzer, 2020; Kattenborn et al., 2021; MacLeod et al., 2010; Weinstein, 2018). Indeed, automated image analysis can achieve accuracy comparable to manual analysis or slightly lower, but orders of magnitude faster (Favret and Sieracki, 2016; Kattenborn et al., 2021; Weinstein, 2018). The most promising algorithms currently used to analyse images rely on a specific type of neural network called a Convolutional Neural Network (CNN) (Krizhevsky et al., 2012). These models are capable of “learning” the features that differentiate one or more types of objects from the background of the image and thus are able to locate (or detect) and identify it within other images (LeCun et al., 2015). These innovations are influencing how remote sensing is used in ecology to circumvent the bottleneck of manual labour and CV is now commonly employed in vegetation monitoring (Kattenborn et al., 2021) and animals detection on camera trap footage (Norouzzadeh et al., 2018; Tabak et al., 2019; Whytock et al., 2021). In the marine environment, CV has been applied to the study of coral reefs (Beijbom et al., 2015; González-Rivero et al., 2020; Pavoni et al., 2020; Williams et al., 2019) and fish community diversity and abundance (Ditria et al., 2021; Marini et al., 2018). It has also successfully been used to study the diversity, abundance and distribution of benthic communities (Abad-Uribarren et al., 2022; Durden et al., 2021; Marburg and Bigham, 2016; Marini et al., 2022; Möller and Nattkemper, 2021; Piechaud et al., 2019; Schoening et al., 2012) and these authors report satisfying performances.

However, although these technologies have existed for more than a decade and their use regularly advocated (Gaston and O’Neill, 2004; MacLeod et al., 2010), their application to ‘practical’ case-studies (especially quantitative mapping of benthic megafauna with AUV) remain relatively exceptional and the methods used are often complicated to reproduce and require advanced knowledge of CV. With more accessible methods that can quickly yield usable results, CV can be deployed more regularly by the benthic ecology community to bypass the manual image analysis bottleneck and make full use of the large image datasets collected by AUVs. This will result in an increased amount of data on the extent and distribution of benthic megafauna available to inform ecological research and conservation strategies.

Within the deep-sea biome a number of taxa are recognised as Vulnerable Marine Ecosystems (VME) indicator taxa under the United Nations resolution 61/105, and states are encouraged to map their

extent and distribution (Ospar, 2008) in order to avoid significant adverse impacts occurring as a result of bottom trawl fishing. One such taxon is the xenophyophore *Syringammina fragilissima* (Brady, 1883). At up to 20 cm in diameter, *S. fragilissima* is possibly the largest single-celled organism on the planet. It lives on areas of soft sediment, deep in the Atlantic Ocean, and can form highly dense aggregations (Bett, 2001; Hughes and Gage, 2004); sometimes dominating the benthos as the main habitat-building organism (Howell et al., 2010). They are usually associated with areas of high surface productivity, which supply abundant organic carbon to the seabed (Levin and Gooday, 1992; Tendal, 1972). At very broad scales, they are thought to live near geological structures such as banks and margins (Tendal, 1972; Levin and Gooday, 1992b) but can also be found near canyons (Gooday et al., 2011) and seamounts (Davies et al., 2015). At finer scales, they are found near caldera, sediment mounds and walls (Levin and Gooday, 1992; Tendal, 1972). Their distribution is also known to vary over very short distances, in areas of complex topography and sedimentology, such as in the Darwin Mounds in the North Atlantic (Bett, 2001).

Little is known of their physiology, but they grow episodically, possibly in reaction environmental parameters changes (Gooday et al., 1993). Some species of xenophyophores are known to be very sensitive to changes in POC influx from the surface (Tsuchiya and Nomaki, 2021) and this may make them an important indicator of climate-related changes in POC flux. Their tests of agglomerated sediment can form a 3-dimensional frame, which is known to house other meiobenthic taxa. Additionally, a higher diversity of endofauna has been observed in the sediment directly surrounding them (Hughes and Gage, 2004; Levin, 1991; Levin and Gooday, 1992). The fragility of their structure makes them particularly vulnerable to physical damage and unsuitable for trawl-based studies (Roberts et al., 2000). Piechaud et al. (2019) showed that a CNN trained to identify various benthic taxa in AUV images performed unevenly but identified *S. fragilissima* with more than 90% accuracy with enough training data. This species, therefore, represents a potentially useful taxa for trialling the application of CV and DL to the generation of quantitative data to inform sustainable management of the deep sea.

The aims of this study are two-fold. 1) apply CV to extract quantitative data on the density and population size structure of a *S. fragilissima* aggregation in the north-east Atlantic from a large image dataset acquired by an AUV, and 2) Demonstrate that a simple open-source CV workflow applied to benthic images can replace manual analysis to generate accurate data and contribute to the future development of autonomous biological observations under the UN Ocean Decade.

2. Material and method

2.1. Study area and data collection

The data were collected at station 26 of the DeepLinks Cruise (Howell et al., 2016) on the 29/05/2016, during mission M116 of the Natural Environment and Research Council’s (NERC) AUV Autosub6000. The study area lies northeast of Rockall bank (Fig. 1), is 1205 m deep on average (+ – 25 m) with relatively smooth terrain composed of fine sediment. Multibeam bathymetry and backscatter data at 2.5 m resolution was collected by Autosub6000 during dive M115 of the same cruise (Fig. 1). The area surveyed by multibeam formed a rectangle of 1.5 by 3.5 km with a data-gap in the in the southern edge where the topography was deemed too rough for AUV camera work. The total surface area for the study site is 8.6 km².

During mission M116 the AUV spent 22 h in the water and approximately 18 h near the seabed travelling a distance of 82 km. It performed 27 transects ranging from 1.7 to 3.2 km in length (Fig. 1), at an altitude of approximately 3 m above the seabed, and speed of 1.1 m.s⁻¹. Distance between transects ranged from 10 to 90 m.

Autosub 6000 was equipped with a Grasshopper 2 - GS2-GE-50S5C

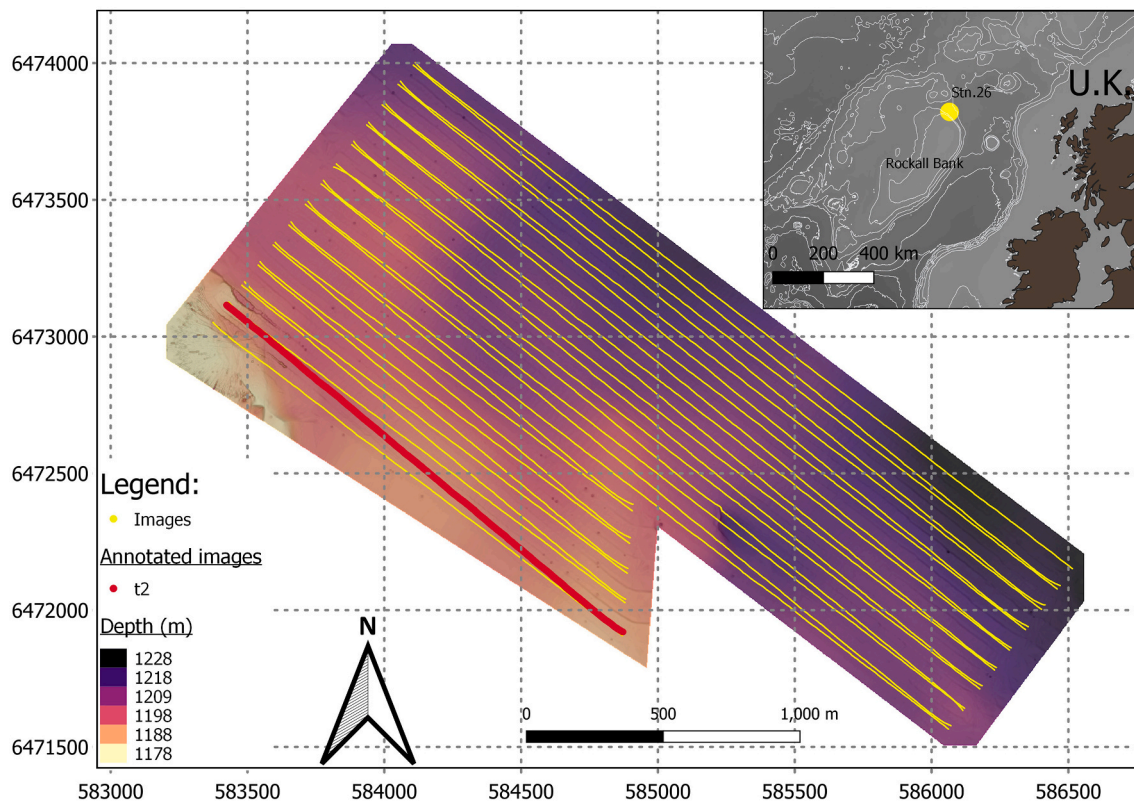


Fig. 1. Map of the study area (Station 26) in the general context of the Rockall Trough with EEZ boundaries (inset map, top-right corner) indicating the AUV transects, the annotated one (t2), as well as the local topography. Coordinates are UTM28 North.

camera looking vertically down at the seabed. Images of resolution 2448×2048 pixels, were taken every second giving near full coverage of the seabed along these tracks providing image data on $\sim 103,000 \text{ m}^2$ of seabed. Vehicle altitude, pitch and roll were recorded by on-board sensors, and a USBL system provided real time position data.

2.2. Image analysis

Upon recovery of the vehicle images were downloaded for analysis. Images were geolocated using AUV position, and colour-corrected using the default method implemented in the IrfanView software (Skiljan, 2012), before further processing.

Images taken at $< 2 \text{ m}$ and $> 3.5 \text{ m}$ from the seabed were excluded from the dataset to minimize challenges in animal identification and quantification resulting from inconsistent field of view and the potential for overlapping images. The resulting datasets consisted of 58,922 images of which 58,148 were suitable for analysis. The surface of each image in square meters (m^2) was calculated with the method described in Morris et al. (2014) using the vertical and horizontal acceptance angles of the cameras, the focal length of the camera and the altitude, pitch and roll of the vehicle. Without proper calibration, the exact error margin of this method as well as the effect of the distortion near the edge of the images are unknown but is estimated to be in the order of 10%.

A subset of 1718 images belonging to one of the 27 transects (transect 2 in the south west) were annotated by a single observer using the online software Biigle (Langenkämper et al., 2017). Individual xenophyophores were located in the images and marked with a circle annotation. *S. fragilissima* was very abundant at station 26 and almost all images selected for annotation contained at least one individual. Consequently, annotation was faster than it would have been if the target was rarer.

2.3. CNN training

We used a ‘You Only look Once’ Version 4 (YOLOV4) CNN to detect *S. fragilissima* (Bochkovskiy et al., 2020; Redmon et al., 2016). YOLOV4 is a highly flexible CNN framework capable of performing both object detection and classification. While not the only suitable architecture for the objectives of this study, nor the most accurate (Schneider et al., 2018), it is simple, fast to implement, and one of the most popular open-source architectures of CNN used for many different industrial and scientific tasks (Hoeser et al., 2020; Redmon et al., 2016).

We used transfer learning to re-train the last layer of an existing YOLOV4 CNN trained on the COCO dataset (Bochkovskiy et al., 2020). The CNN used to annotate the benthic images was trained within the Darknet (Redmon, 2013) framework through a Google Collaboratory Professional (Bisong, 2019) notebook with GPU acceleration and High Ram session enabled. This notebook acted as a virtual machine onto which a custom combination of command-line and python scripts prepared the environment, processed the data and trained the CNN. After purchasing ‘Pro’ membership (£8.10/month), this online interface required only a connection to the internet thus bypassing the need for advanced and expensive computing hardware and its setup on a local machine. Although the resource allocation in a Colab session is not directly controlled by the user, we most commonly had access to a Tesla P100 GPU (16 GB of VRAM memory). The Colab notebook was also set to ‘high RAM session’ to avoid memory limitations that could interrupt the training process.

Some model parameters were modified from the default configuration to balance use of resources and training speed. We retained a resolution of 704×704 , batch size 64, 32 subdivisions and trained up to 6000 iterations. Other parameters were kept to default or set according to guidance by the developers (<https://github.com/AlexeyAB/darknet>).

Manual annotations were used to generate training, validation and testing datasets. Biigle annotations (label name, centre x, center y,

radius – all expressed in raw pixels number with origin in top left corner) were downloaded and transformed to fit the format required by YOLO (class code name, center.x, center.y, width, height – all in fractions of images width and height scaled from 0 to 1) using a custom R script. Following recommendations in Piechaud et al. (2019) we aimed to use approximately 1000 annotations of the target species in training. Our training dataset contained exactly 997 individual *S. fragilissima* annotations from 262 images. A further 269 individuals annotated from 50 images formed the validation dataset used throughout the training process to monitor the performance changes by the Darknet implementation of YOLO.

2.4. Model evaluation

The testing dataset was composed of 500 manually annotated images from the same transect and was only used to calculate performances outside the Darknet framework by comparing CNN predictions with manual annotations.

Predictions come as a .JSON file containing the coordinate of the bounding box of each tentative annotation as well as a confidence score ranging from 0.05 at minimum (lower confidence scores introduced too much noise and therefore were not recorded) to 1 at maximum. Predictions made on the testing set were compared to the manual annotation of the same images.

In this exercise, a tentative annotation was considered a true positive if its centre was within a 30 pixel distance from a manual annotation centre in the same image. We used a custom R script to spatially match CNN predictions and manual annotations so that the number of True Positives (TP), False Positives (FP) and False Negatives (FN) could be counted in each image in the testing set. Hence, a true positive is an automatic annotation that falls within 30 pixels of a manual one. The 30 pixel distance allowed some leeway in matching annotations as the YOLO bounding boxes were rarely centred exactly where the manual annotations were centred. This distance is also small enough to avoid a single automated annotation to match multiple manual annotation since different individual *S. fragilissima* are never so closely packed. A false positive is an automated detection that does not fall within 30 pixels of a manual annotation. False negatives are manual annotations that were not matched to any automated annotations.

Production of values for TPs, FPs and FNs enabled us to calculate the recall (also referred to as sensitivity and is equal to $TP/(TP + FN)$) to know the true positive rate, and to calculate precision (also referred to as specificity and is equal to $TP/(TP + FP)$) to know the false positive rate. The F1 score or harmonic mean of recall and precision ($2 * (precision + recall) / (precision + recall)$) combines both values into a single value for clarity.

The Pearson correlation between the number of manually annotated *S. fragilissima* with the number of objects detected by the CNN within each image of the testing set was used to provide another measure of accuracy in CNN model predictions of the density of *S. fragilissima*. Note, however, that this metric does not inform on whether or not the same individuals were detected by the manual annotators and the CNN.

2.5. Model selection

CNNs are sensitive to overfitting (Domingos, 2012). Throughout the training process, the algorithm goes through cycles of training on the training set - evaluating performance on the validation set – adjusting CNN node weights – before repeating the entire cycle. These cycles are referred to as epochs or iterations when the training dataset is too large to be processed in one stroke and needs to be subdivided in several smaller batches as is the case in this study. There is an optimum number of iterations before the performance on the validation and training sets (independent data that the CNN has not seen) start to diverge as the CNN overfits and becomes more specialized at predicting the training set while becoming less able to predict the validation set (generalization).

This optimum point for best performances was sought by interrupting the CNN training before it loses capacity to generalize, or early stopping (Ying, 2019). Performances were calculated for CNNs trained with 1000, 2000, 3000, 4000, 5000 and 6000 iterations in order to detect the point where performance was best and thus avoid overfitting. During testing, every object detected with a confidence value >0.05 were recorded so that higher thresholds could later be applied. Performances were calculated with confidence thresholds of 0.05, 0.3 (the default), 0.5 and 0.9 and compared to select the threshold that maximized the TP while keeping FP to a reasonable level.

The weights that offered the best compromise between precision and recall were used to make predictions on the whole dataset of 58,148 images. This was also performed in the same Colab notebook. To limit space taken on Cloud storage without compromising performance, they were reduced in resolution to a 1224×1024 .

2.6. Size and surface area measurement

In the CNNs predictions, the size of individual *S. fragilissima* was measured as the smallest side of the bounding box. After examination of the training set, it became apparent that the sizing of these annotations was too inaccurate to compare to the CNN's bounding boxes which tended to fit much more closely to the animal. Indeed, manually fitting the circle to the exact size of the individual *S. fragilissima* size is difficult to achieve quickly and consistently during manual annotation. Consequently, 100 images containing 467 individuals from the testing set were annotated again with particular care given to fitting the circle around the individual *S. fragilissima* and accurately measuring their size by manual mean. Evaluation on the accuracy of the sizing were conducted by comparing the CNNs predictions and precise manual measures in this small subset exclusively. The size in pixels was converted to m based on calculated image size. The area covered by each individual was approximated by calculating the surface of circle ($\pi * radius^2$) of a radius equal to half the measured size of the individuals. The Pearson correlation between manually and automatically measured sizes in pixels was used to express the accuracy of the size measurement.

2.7. Data analysis and mapping

The selected CNN was used to analyse all AUV images providing a georeferenced spatially explicit dataset from which metrics describing the local *S. fragilissima* population could be calculated. Density of *S. fragilissima* was calculated for each image (nb. individuals/ image surface area) and expressed in individuals per meters square ($ind.m^{-2}$). The percentage of seabed covered by the combined *S. fragilissima* individuals in each image was calculated as (sum of *S. fragilissima* surface area / image surface) x 100). The relative abundance of different sizes-classes of *S. fragilissima* were explored using histograms. The relationship between density and relative seabed cover was explored with Pearson, Spearman and Kendall Correlations. All data analysis was performed in R (Team, 2021) with the “tidyverse” package (Wickham, 2017).

Continuous maps of the density, percentage surface cover and average sizes of *S. fragilissima* at 10 m resolution (and down-sized to 2.5 m to be overlaid with the resolution of the existing bathymetry available for the area) were produced with ordinary Kriging. These continuous surfaces were then overlaid with topographic information (i. e. bathymetry) to visualize pattern in *S. fragilissima* distribution, density, and population size structure over the local terrain. This was performed in R with the “automap v. 1.0-14” package (Hiemstra and Hiemstra, 2013) to automatically fit the variogram and the “gstat v. 2.0-8” (Pebesma, 2004) package to interpolate the values between discrete points of measurement and produce a complete surface. QGIS (QGIS Development Team, 2020) was then used to visualize the resulting spatial layers and export maps.

3. Results

3.1. Model evaluation and selection

The number of iterations for which the CNN was trained has a strong influence on its performances as well as on the confidence it gives to its predictions (Fig. 2). In general, longer model training past 1000 iterations tended to give higher recall but lower precision.

We used the CNN constructed using 3000 iterations, and a confidence threshold of 0.3 as the best compromise in terms of maximising recall and precision. The selected model had comparable performance to one constructed using 4000 iterations but trains faster (Fig. 2). Although a higher confidence threshold gave a slightly higher overall performance (f1-score) the 0.3 threshold was selected to favour recall (true positives) over precision as the latter was above 0.9 regardless of the threshold.

The recall and precision of our selected model was 0.836 and 0.912 respectively. This means that 84% of the manually detected individuals were also detected by the CNN (84% True positive rate) and 91% of automatically detected individuals were also manually detected (9% false positive rate).

The correlation between manually counted and automatically counted xenophyophores in each image for which manual measures existed reached 0.96. Although this suggests this model performs well it also shows manual and automated annotations will both miss individuals that the other method does detect. With the chosen configuration of number of iterations and confidence thresholds, the CNN detected 8.5% less *S. fragilissima* individuals than the manual annotation. The manually annotated individuals that the CNN failed to detect were predominantly small ones.

The Pearson correlation between the manually and automatically measured size of *S. fragilissima* was 0.92. An example of predictions performed on one of the most densely occupied images in the dataset is shown in Fig. 3 to illustrate the predictions made by YOLOV4.

3.2. Density and population size structure of an *S. fragilissima* aggregation in the north East Atlantic

261,049 individuals were observed in the complete image dataset. Their density is on average 2.5 ind.m⁻² and in most images, the density was close to this value and higher abundances much rarer (Fig. 4). Some patches of seafloor supported a very high density with up to 45 ind.m⁻² in some instances.

3.3. Size-abundance relationship

The correlation between density and relative seabed cover (Table 1) is high. The average diameter of the individual *S. fragilissima* encountered at station 26 was near 5 cm. The majority of the individuals (50.7%) measured 3.5 to 6 cm and only 2151 (less than 1%) were larger than 10 cm. The largest observed individual confirmed to be a *S. fragilissima* was measured at 16.2 cm in diameter (some of largest detections were false positives). Although both manual and automated methods cannot consistently detect individuals smaller than 1 cm, the size-classes distribution shows the average-sized individuals are also the most abundant (Fig. 5). The largest individuals were observed in images where density was low (Fig. 6).

3.4. Mapping community distribution

A continuous map of the distribution of *S. fragilissima* Figure 9 shows important spatial variability. The percentage of the seabed covered by *S. fragilissima* (Fig. 7) largely follows the density pattern (can be seen in Fig. 9 in appendix). Together, all the *S. fragilissima* observations represent 562 m² of seabed. On average, they cover less than 1% of the surface of the image but this value can go up to 5% in some areas. *S. fragilissima* occupy the whole survey area in various patch sizes but four high density patches, 100 to 300 m across, stand out. The transition between these areas and their surroundings, where the density is much lower, can be very abrupt and go from close to 0 to more than 20 individuals in few tens of meters. The patches of high density are also the areas where the average size of *S. fragilissima* individuals is lowest (Fig. 10 in appendix). Overlaying measured density over the local topography shows that high density patches are located in small depressions slightly deeper than the surrounding seabed (Fig. 8 or view in interactive map in Appendix S1). However, density can be very low in some depressions and average on some of the raised features.

3.5. Application of deep-learning computer vision approaches to benthic image analysis

S. fragilissima is very abundant at station 26 and almost all images (~85%) selected for annotation contained at least one individual. Consequently, the annotation of the training, testing and validation sets was relatively fast. Note that, for a simple task like counting xenophyophores, manual analysis is quick and hundreds of annotations of

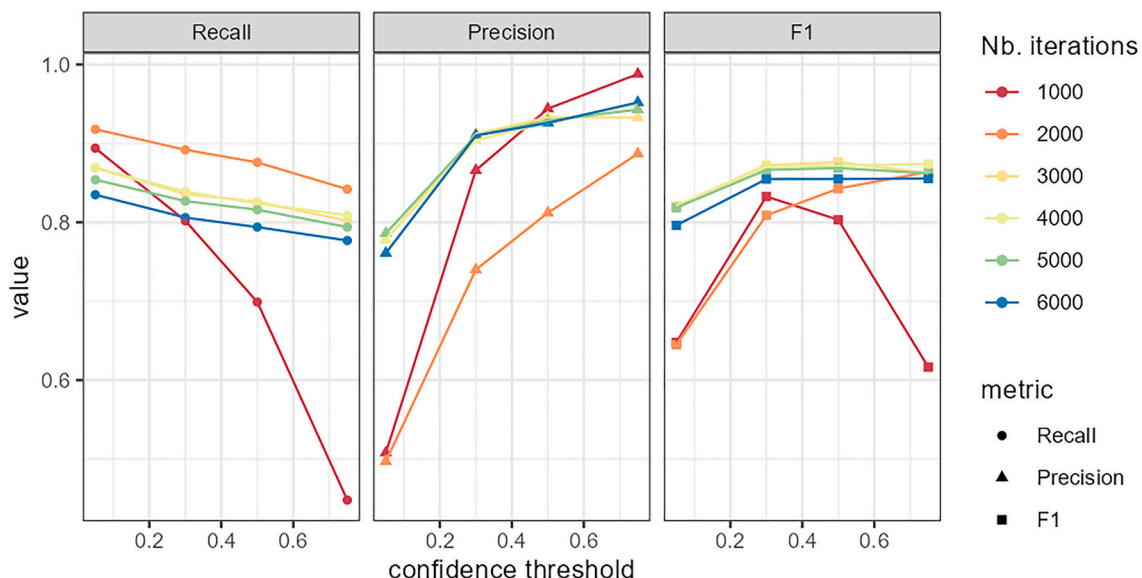


Fig. 2. Performances of the CNNs with varying numbers of iteration and different confidence thresholds.

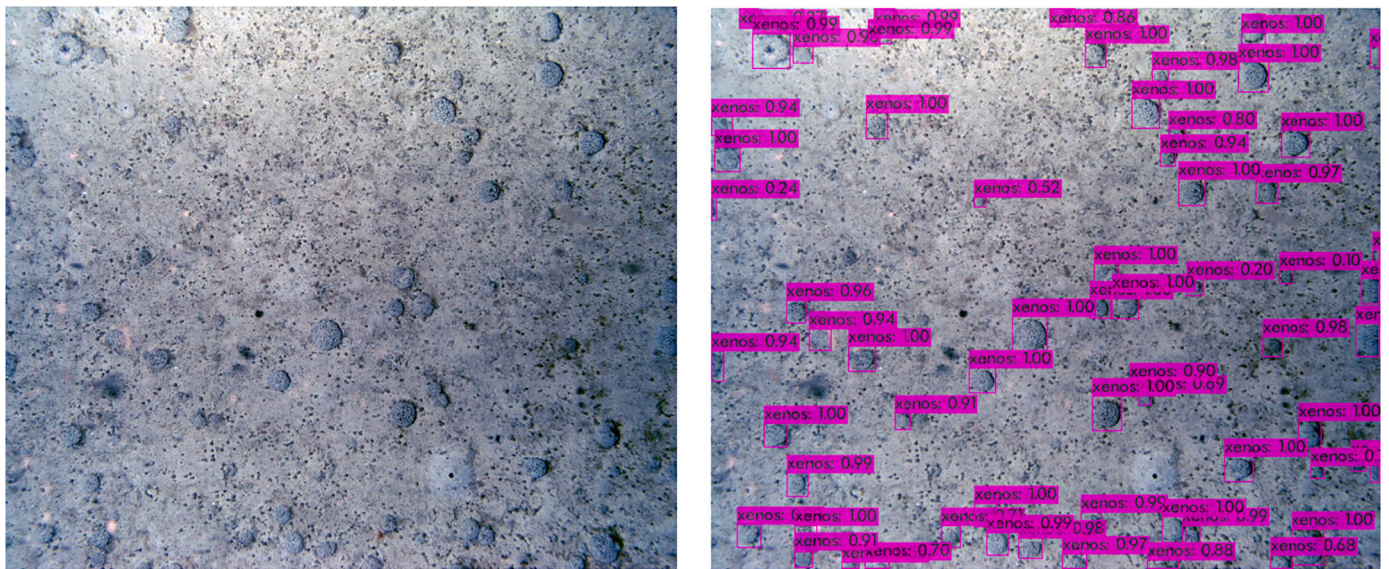


Fig. 3. Images of the highest recorded density without (left) and with (right) the bounding boxes of the CNN’s predictions overlaid over individual *S. fragilissima*.

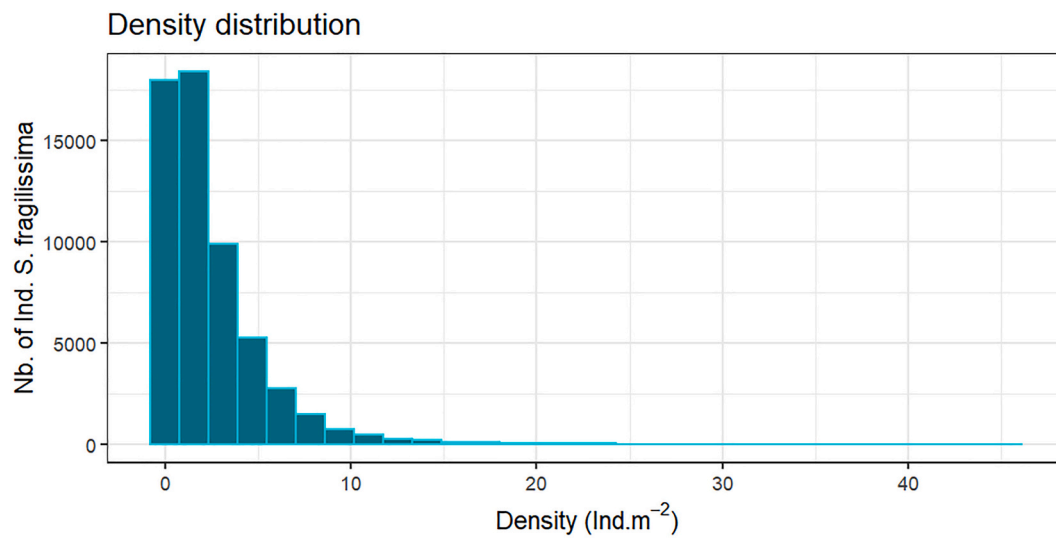


Fig. 4. Density classes relative abundance of *S. fragilissima* at the study site. The bars indicate the number of individuals (count) in each abundance classes.

Table 1
Correlations between density and relative seabed cover by *S. fragilissima* at station 26.

Method	Value
Pearson	0.86
spearman	0.92
Kendall	0.76

individuals can be gathered in single hour in areas of high density. However, this repetitive task is straining for the annotator who may lose focus and consistency, particularly regarding the detection of the smallest individuals. Besides, this speed depends largely on the density of the target species and it would take longer to form a training set for a less abundant target. Thus, the time taken to annotate the training and validation sets can vary widely, but it is estimated that the entire process from taking the images out of the AUV to having them ready to train the CNN could be accomplished in less than 10 days or even less for experienced analysts used to performing this task routinely.

Training the CNN took approximately 18 h to complete the 6000 iterations. Measuring the accuracy of the different CNNs and the effect of different confidence thresholds took several hours but could be further automated. Predictions on the 58,148 images took approximately 10 h however the training and predictions phases do not require constant supervision and can be performed overnight or while the analyst is free to attend to other tasks. Finally, the calculation-intensive nature of the training, and to a lesser extent, prediction phases makes their duration largely dependent on the hardware used and thus may vary between users. Overall, the whole pipeline could be completed within one or two weeks (10 days) once the analysts are familiar with it.

4. Discussion

Automated annotation of nearly 60,000 images by a CNN trained on a relatively small training set produced highly accurate counts and measurements of the target species in a very short time that only marginally deviated from the manual annotations it was compared to. The dataset collected reveals interesting insights into the ecology of *S. fragilissima* and demonstrates the feasibility of applying deep-learning

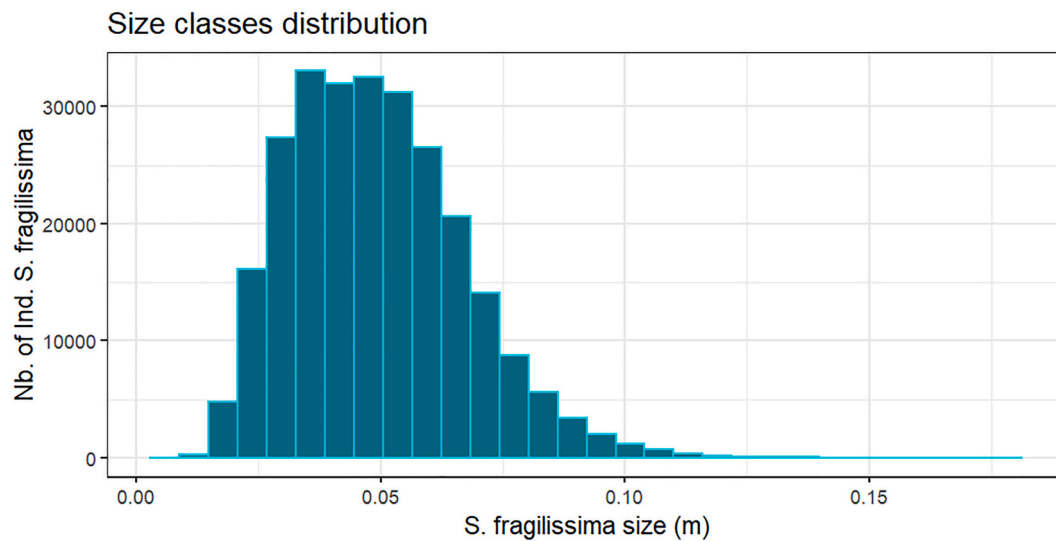


Fig. 5. Histogram of *S. fragilissima* population size (in meters) structure. The bars indicate the number of individuals (count) in each size classes.

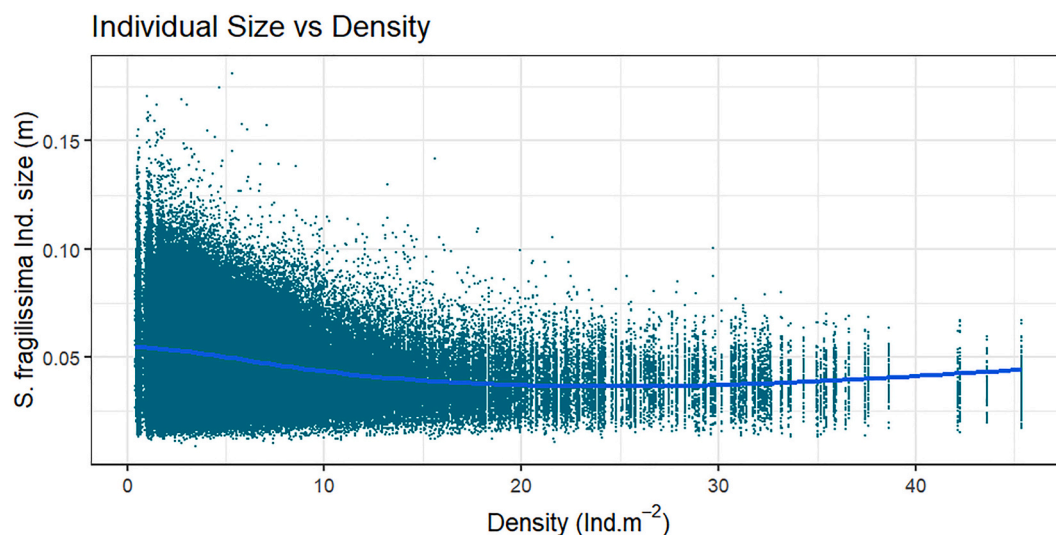


Fig. 6. Relationship between local (in-image) density of *S. fragilissima* and their individual sizes.

computer vision approaches to benthic image analysis.

4.1. Distribution of *S. fragilissima*, density, sizes and seabed cover in the survey area

Our results suggest the distribution of *S. fragilissima* is patchy within the study area. Dense aggregations tend to be formed by individuals of small to average sizes and the large individuals were only found in areas of low density. Abundances above 45 ind.m⁻² were observed, which are higher than the previously published values for this species in this area. Roberts et al. (2000) reported densities of 10 ind.m⁻² and Bett (2001a) reported densities of up to 32.8 ind. 100 m⁻² (0.3 ind.m⁻²) with clear local variations. The maximum densities observed within the study area are higher although the use of different imaging platforms should be noted when comparing these values. It is possible that the environmental conditions within the study area could be closer to optimum for *S. fragilissima* than previously studied locations. The patchiness of their distribution and sharp variation of density (over tens of meters) have been reported previously and have been hypothesized to be a response to environmental drivers at very fine scale (Bett, 2001; Levin and Gooday, 1992). Large concentrations in small areas are likely linked to

local environmental drivers forming favourable niches at a microscale.

Although the drivers of xenophyophore distributions are not known with certainty. At broad scales, (~1 km × 1 km) Ashford et al. (2014) identified depth, oxygen availability, nitrate concentration, organic carbon, and temperature to be the most important parameters driving the distribution of xenophyophores. Ross et al. (2015) determined that, topographic factors influencing the distribution of *S. fragilissima* were the same at broad (750x750m) and fine (200x200m) scale and concluded that the main drivers (here, depth and bathymetric position index) of *S. fragilissima* distribution operated at either of those scales.

Other in situ observations suggest they inhabit flat muddy terrain, such as near the Darwin Mounds (Bett, 2001; Huvenne et al., 2016), as well as steep slopes in Canyons (Gooday et al., 2011). Terrain could be an important driver of their distribution, as observed in other areas where their close proximity with raised features has been reported at fine and broad scale (Ross and Howell, 2013; Davies et al., 2015; Huvenne et al., 2016), although not systematically. It has long been hypothesized that the dense yet patchily distributed aggregations of xenophyophores observed in specific locations, like the Western Darwin Mounds, were associated with the scoured tails created by currents behind raised structures (Bett, 2001). A counter example exists where

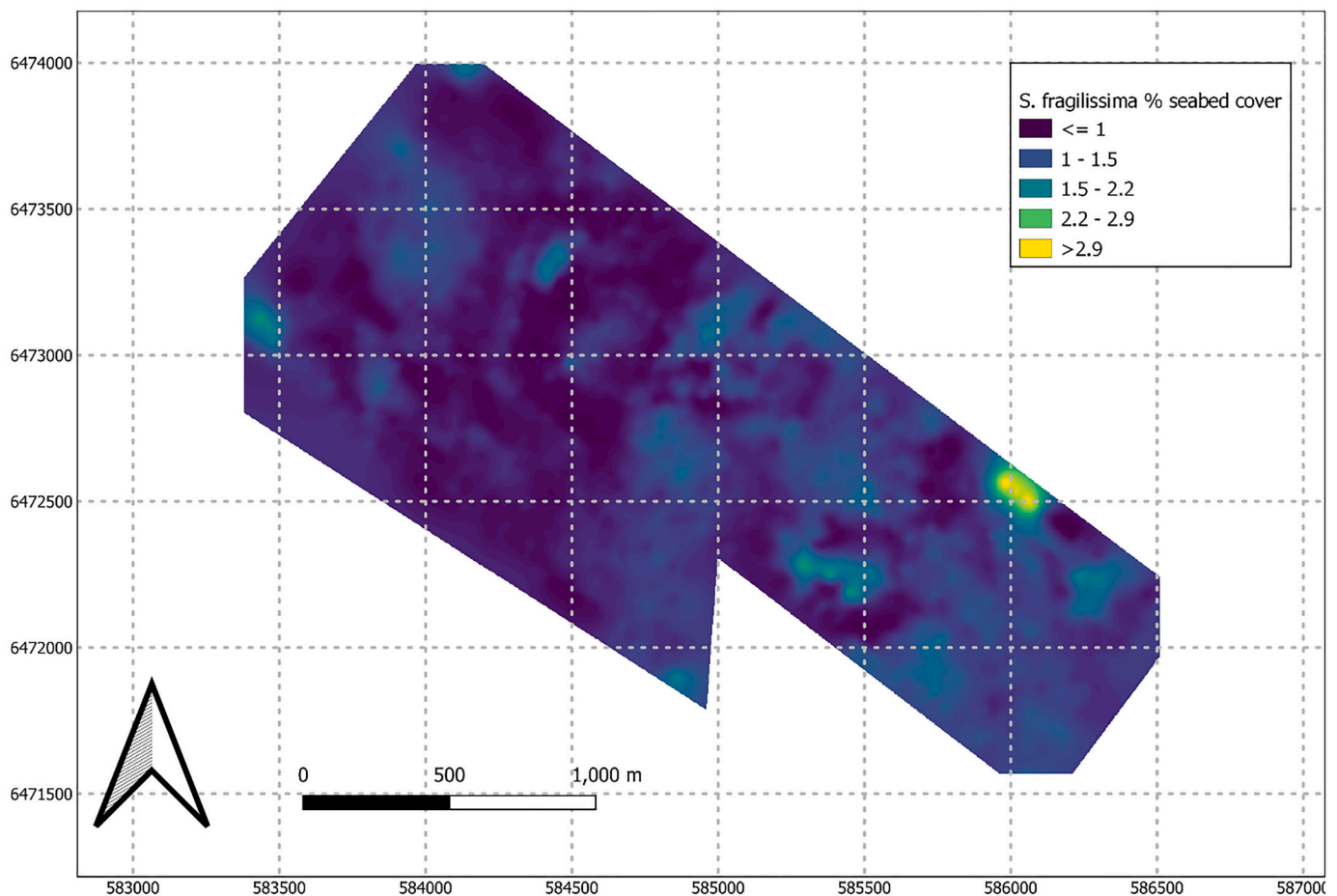


Fig. 7. Percentage (%) of the seabed cover distribution. Combination of size and density to evaluate the abundance of *S. fragilissima* and how much of the seabed it occupies. Coordinates are UTM28 North.

this association was not observed and the xenophyophores were very widespread (Howell et al., 2014).

Within the study site, patches of medium or high density were found both on local peaks and troughs (with elevations differences of no more than 25 m), while the relatively small size of the area makes a significant effect of depth, temperature, and chemical parameters unlikely. It therefore remains to be precisely determined why some areas within the surveyed perimeter are almost devoid of xenophyophores while other are occupied by higher densities. There are no previous studies that have focused on variations at such fine scale and thus we cannot resolve the data presented here. Whether the niches that favour these high concentrations of *S. fragilissima* are stable in time or are transitional is also unknown. Furthermore, the distribution of *S. fragilissima* may not be entirely environmentally driven and could be shaped by biotic interactions that can only be observed if the other organisms of the community were considered in the analysis. Replicating the analysis of datasets of benthic images and topography would better inform fine scale spatial variability in the distribution of *S. fragilissima* and its drivers.

The proportion of seabed cover is correlated to density more than the size of the individuals and, interestingly, the largest individuals tended to live in the less dense patches rather than the presumably more suitable areas where the density is high. On average, this cover in the surveyed area is 0.5%. This could represent several thousand square meters of seabed in this small area alone if the pattern seen in the images is repeated outside. If this station is representative of the whole Rockall Basin within the depth band occupied by these xenophyophores (which is not clearly defined either), their contribution to the local ecosystem

could be considerable and warrants a closer examination of their functional role, population dynamics, sensitivity to disturbance, particularly climatic changes, and resilience. More knowledge on their ecology is necessary to interpret these results and adequately manage their conservation. For example, studies such as those of Tsuchiya and Nomaki (2021) on metabolism of pacific xenophyophores *S. fragilissima* could be replicated in the Atlantic enabling metabolic calculations for individuals and subsequently populations, thus linking density and seabed cover to metabolic variables. *S. fragilissima* and other xenophyophore populations have been found to be sensitive to organic carbon influx from the surface, reproducing and growing rapidly following increased flux of organic carbon to the seabed (Gooday et al., 1993; Tsuchiya and Nomaki, 2021).

Large scale surveys of density and sizes measurement between two points in time (repeated monitoring) would give statistical and spatially explicit visualisation of this phenomenon and complement the in-situ studies at individual scales. In this context, repeating surveys with the presented method could detect subtle variations over time of the biometric variables that the CNNs can measure, and offer a proxy for ecosystem metabolism and carbon uptake as well as reveal important information on their ecology and life history. Thus, these quick and cost-effective measurements could, in future, be applied to monitoring of ecosystem health and impacts of climate change.

There are many taxa considered VME indicator taxa that form high density monospecific aggregations, which are considered VME when a threshold density is reached. These thresholds are, however, not clearly defined. Although *S. fragilissima* aggregations are a widely recognised as a VME (Ospar, 2008), there is no precise criterion of density or relative

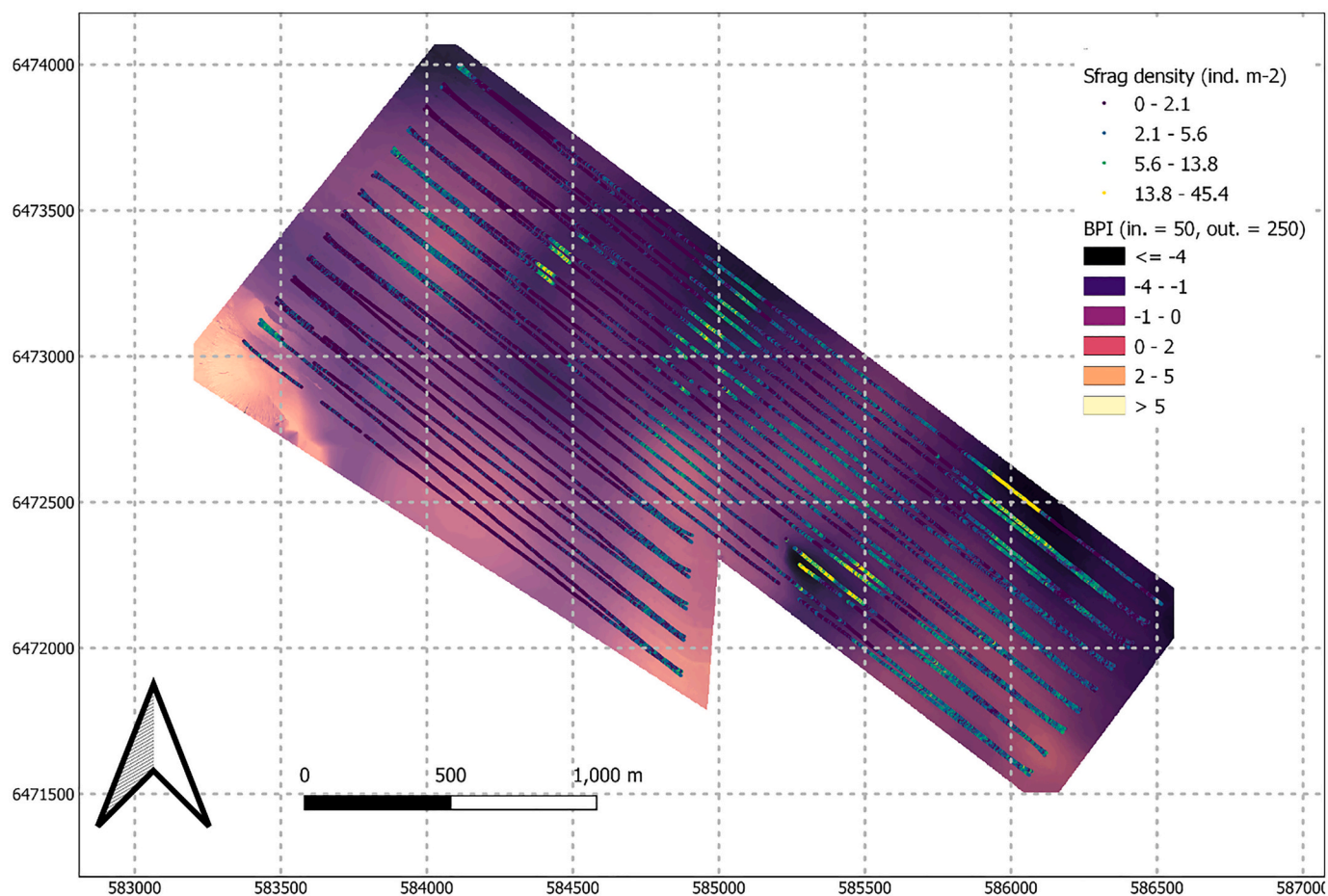


Fig. 8. Image location coloured by density overlaid onto the local topography. The Bpi index reflects the shape of the terrain and 0 indicates flat terrain at those scale (here, over 250 m). Negative values are dips, troughs and gullies and positive values are peaks and crests. Coordinates are UTM28 North.

seabed cover that is used to make a clear decision on where this VME is and isn't present. This complicates the assessment of its extent and distribution and is detrimental to its conservation. If applied to other known examples of *S. fragilissima* aggregations this method could quickly quantify densities of *S. fragilissima* and provide numerical data to help precisely define the density at which the VME indicator taxa is considered a VME.

4.2. The application of deep-learning computer vision approaches to benthic image analysis

CV based techniques (involving DL or not) for image analysis have been available as a tool for a long time (Gaston and O'Neill, 2004; MacLeod et al., 2010) but its complexity and difficulty to access trained staff for long term projects may have limited its application to the field of deep-sea ecology. This workflow uses open-source tools and a remote cloud computing platform which makes its implementation easy and cost-efficient while achieving high accuracy. A research team using a combination of YOLO and Google Colab as we did here can thus have access to a CV workflow without hiring a trained computer scientist and acquiring the costly hardware needed to run the algorithm. Consequently, more ecologists can now use CV and workflow such as this one to quickly gather large datasets of ecological data with an accuracy comparable to that of manual annotation.

Implementing our CV workflow onto this dataset yielded excellent results in terms of raw enumeration with 96% correlation between manual and automated counting but the measures of recall and precision indicate that the real accuracy is slightly lower. This is interpreted as the

sign that the CNN and human annotators, in agreement for the most part, do detect or miss different individuals (or possibly make different mistakes). It is worth noting that two equally experienced human annotators are unlikely to produce identical datasets and may even disagree by a comparable margin (Culverhouse et al., 2003, 2014; Durden et al., 2016). Further exploration of the biases of human vs machine are necessary to understand the sources of these differences and what impact it could have on comparison between manually and automatically annotated data. This also raises the question of what the gold standard to which annotations are compared to should be (Schoening et al., 2016) and a dataset combining the input of multiple human annotators would give a less biased evaluation of performances. Investigating the performances of CNNs trained with increasing numbers of iterations also showed default parameters (6000 iterations) set by Bochkovskiy et al. (2020) may not apply to all cases. Here, testing with less iteration showed that slightly higher recall and precision could be achieved with half the training time before overfitting starts affecting accuracy. This is a known phenomenon that affects CNNs (Domingos, 2012). Closely monitoring the performances throughout training and applying early stopping (Ying, 2019) can yield a significant gain when it saves 8 or 9 h and maximizes recall and precision. Picking an arbitrarily low number of iterations could save time in the analysis phase if speed of analysis is an objective in itself and the best performance are not necessarily required. An example situation would be during an exploration phase of a sampling exercise when the area of highest concentration of a target for physical sampling is sought. Quickly mapping the concentration of the target with a CNN such as the one used here would quickly provide usable information in the same way the automated

processing of acoustic data from multibeam echo sounders (Mayer, 2006) can quickly produce a visual representation of local topography. However, it seems good practice to investigate how overfitting affects the performance to avoid relying on unnecessarily inaccurate CNNs.

Depending on the objectives of the study, one could favour a combination of number of iterations and confidence threshold that promote a high recall (more true positives) or high precision (less false positives) or a compromise between the two (highest F1 score). For example, a study aiming at generating reliable presence records, would likely minimize the number of false positives so that the detections made are almost certainly the target species even though individuals are missed. Such studies would rather apply a high confidence threshold. On the other hand, studies like this one aiming at quantifying the target would prefer a balance between the two so the final abundance is as close as possible to the actual count.

The AUV survey took 22 h from deployment to recovery. The whole pipeline of analysis including manual annotation and quality assessment of the training set, and training and testing of the model itself could take 5 to 10 days depending on the density of the target species in the images, the hardware available and the experience of the operator. The most time-consuming tasks are the training sets annotations and the training of the CNN. Making predictions on 60,000 images was performed in 10 h. This means that the presented map of species abundance and relative seabed cover could be available in less than a day after the AUV came back from its mission provided a working CNN is available prior to the launch. In that case, these results would require very little time investment and be available almost immediately.

Without the use of automated annotation, manually processing the near 60,000 images collected by Autosub6000 at station 26 could have taken hundreds of hours of manual work. Although the resulting dataset would have been potentially more accurate, it could also be less consistent as annotators focus varies throughout the process. This manual analysis would not have been the best use of time and resources available to the researchers and would, most likely, not have been carried out. Indeed, the known issue of the image analysis bottleneck partially cancels the advantages of AUVs (particularly of the cruising class (Huvenne et al., 2018) that offer the best coverage) for benthic habitat surveys. This somewhat limits the usefulness of these AUVs, particularly if one is reminded of their notorious drawbacks such as lack of manoeuvrability, incapacity to safely survey rough terrain, within-mission adaptability and their occasional unreliability (Przeslawski et al., 2018).

Although the potential contribution of easily implemented CV is enticing and could have large benefits for benthic ecology, there are a number of drawbacks that must be kept in mind by potential future users. Assessment of CV capacities and limitations to offer a clear picture of what it can achieve are necessary to avoid introducing unknown biases in the datasets it generates and make inappropriate use of the results. As the dataset generated also increase in size, subsequent analysis may also need proportionally more computing power which could also become a limitation. Therefore, future studies using CV will need to consider using methods and platforms that can accommodate their large datasets or risk simply shifting the data analysis bottleneck down the analysis pipeline. Subsampling (taking a representative subset of the data that replicates the trend found in the whole dataset) would be an obvious solution (Hurlbert, 1984; Morrissey et al., 1992). This requires a larger than necessary dataset from which the right sample size to get statically robust results can be extracted. This can be impractical or even wasteful of the annotators time if not all the images they analyse are use. On the other hand, once a CNN is trained, analysing 5000 or 50,000 images makes little difference for the annotators as, although the computer will need more time (a matter of hours), the inference requires no supervision. Gathering a larger than necessary dataset can thus be achieved at no extra cost if the images have been collected in sufficient number, which is often the case with AUV sampling (Morris et al., 2014; Pizarro et al., 2013).

In this study we trained a CNN to identify a single taxa that previous research had shown was well identified by another CNN trained on the same dataset (Piechaud et al., 2019). However, this previous work showed that performance of CV was taxon-specific and always plateaued some way below 100% and sometimes too low for the results to be considered usable. Hence, ecologists wishing to focus on a single taxon must consider their target carefully and always keep in mind that CV may not be accurate enough and, consequently, resort to using manual or semi-automated annotations. Moving forward, training of CNNs to accurately classify multiple classes of benthic megafauna (rather than a single taxon) would be more efficient and ecologically more valuable although other challenges must be addressed with this approach. For example, the uneven performances observed in multi-species classifiers across classes is likely to make interpretation of the resulting ecological metrics complicated (Durden et al., 2021; González-Rivero et al., 2020; Piechaud et al., 2019). The amount of training data needed for such classifiers is also far greater in order to sufficiently represent the rarer taxa (Durden et al., 2021) and they are sensitive to class imbalance (Menardi and Torelli, 2014; Schneider et al., 2020).

In this study, we also used CV to measure the size of individuals, made possible because this species always grows in a circular shape, and the bounding box of the annotation gave a reasonable estimate of the size of the individual. However, using the bounding box to approximate sizes would be inaccurate for taxa with irregular shapes like corals or sponges or for taxa whose perceived size can be altered by movements and changes of orientation like crustaceans or fishes. Other CV methods like semantic segmentation have a more complex perception of the shape of the object they detect and thus be more flexible (Pavoni et al., 2019) but would still not be appropriate to measure the size of every marine taxa on two dimensional images.

Finally, although we have trained a CNN to reliably identify our target taxa in images obtained by the Autosub6000 AUV at one site visited on one research cruise, transferring this model to a new dataset (generalizing or adapting its domain) is unlikely to be successful (Li et al., 2017; Schneider et al., 2020). Use of CNNs trained on a specific dataset should only be used on other datasets with great care. The consequence of this is that the long procedure of training the CNN will have to be implemented on a dataset-by-dataset basis. This highlights the need to retain expert knowledge for annotation but also raises a number of questions regarding the best strategy to adopt in the future as to the most sensible way to gather, combine and expand on existing datasets of annotated benthic imagery and this topic will require more research (González-Rivero et al., 2020; Kattenborn et al., 2021; Li et al., 2017). In the future, a specific CNN trained from the scratch from a large dataset of expert annotations of marine taxa or functional classes on underwater images would be the best solution (Beyan and Browman, 2020; Katija et al., 2021) but requires a concerted effort to standardize annotation data and international cooperation through projects like 'Challenger 150' for the UN Decade of Ocean Science for Sustainable Development (Howell et al., 2019, 2021).

5. Conclusion

This study addresses the need for more data to study deep sea ecosystems and especially VMEs. It uses an open source, easily accessible yet powerful CV pipeline to quickly and accurately count and measure individuals of the VME indicator species *S. fragilissima* in images collected by an AUV. This relatively large dataset brings new knowledge on the fine scale distribution and population size structure of this species in the north-east Atlantic and was collected and analysed in a very short time. It illustrates how application of CV to AUV images can potentially yield more data than analysis, and could, therefore, replace this method in specific scenarios. This increased speed of analysis comes at no extra staff cost, little hardware investment and is accessible to scientists without advanced programming skills. Hence, it becomes more logistically feasible to investigate aspects of deep-sea ecology that require

these large datasets. We are now in the second year of the UN Decade of Ocean Science for Sustainable Development, and lack of data on deep-sea ecosystems has been highlighted as a key challenge to be addressed (Howell et al., 2020) as well as the role of new technology in doing so. This study provides a first step, hinting at what might be possible by 2030.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 862428 (MISSION ATLANTIC). This output responsible for any use that may be made of the information contained therein.

The authors acknowledge the officers and crew of RV 'James Cook' as well as the Autosub support team who assisted in the collection of the data. The DeepLinks project is funded by NERC NE/K011855/1. NP was funded by a Plymouth University studentship and by MISSION ATLANTIC. Data analysis was conducted under the MISSION ATLANTIC project funded by the European Union's Horizon 2020 Research and Innovation Program (No. 862428). We thank Kirsty Morris, Erik Simon Lledo, Brian Bett and Marcus Shirley for their advice on processing the images prior to analysis. Help from the developers of Biigle, Daniel Langenkämper and Martin Zurowietz throughout the development of the workflow was equally instrumental and appreciated. We also thank Jamie Cowle and Rhodri Irranca-Davies whose undergraduate and masters projects respectively helped with annotation and testing of early stage methodology for this study. Finally, we thank Chunxu Li, David Hutchinson, Emily Hutchinson and Erin Browne, for fruitful discussions and advice on implementing CV.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ecoinf.2022.101786>.

References

- Abad-Uribarren, A., et al., 2022. Deep learning-assisted high resolution mapping of vulnerable habitats within the Capbreton canyon system, Bay of Biscay. *Estuar. Coast. Shelf Sci.* 275, 107957 <https://doi.org/10.1016/j.ecss.2022.107957>.
- Ashford, O.S., Davies, D.J., Jones, D.O., 2014. Deep-sea benthic megafaunal habitat suitability modelling: A global-scale maximum entropy model for xenophyophores. *Deep Sea Res. Part I Oceanogr. Res. Pap.* 94, 31–44.
- Beijbom, O., et al., 2015. Towards automated annotation of benthic survey images: variability of human experts and operational modes of automation. *PLoS One* 10 (7), e0130312.
- Bett, B.J., 2001. UK Atlantic margin environmental survey: introduction and overview of bathyal benthic ecology. *Cont. Shelf Res.* 21 (8–10), 917–956.
- Beyan, C., Browman, H.I., 2020. Setting the stage for the machine intelligence era in marine science. *ICES J. Mar. Sci.* 77 (4), 1267–1273. <https://doi.org/10.1093/icesjms/fsaa084>.
- Bisong, E., 2019. Google Colaboratory. In: Bisong, E. (Ed.), *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*. Apress, Berkeley, CA, pp. 59–64. https://doi.org/10.1007/978-1-4842-4470-8_7.
- Bochkovskiy, A., Wang, C.-Y., Liao, H., 2020. YOLOv4: optimal speed and accuracy of object detection. *ArXiv, arXiv:2004.10934* [Preprint].
- Borja, A., et al., 2016. Bridging the gap between policy and science in assessing the health status of marine ecosystems. *Front. Mar. Sci.* 3 (175) <https://doi.org/10.3389/fmars.2016.00175>.
- Brady, H.B., 1883. IV. Note on siringammina, a new type of arenaceous rhizopoda. *Proceedings of the Royal Society of London* 35 (224–226), 155–161.
- Brandt, A., et al., 2016. Cutting the umbilical: new technological perspectives in benthic Deep-Sea research. *J. Marine Sci. Eng.* 4 (2), 36.
- Chapman, M.G., Underwood, A.J., 2008. Scales of variation of gastropod densities over multiple spatial scales: comparison of common and rare species. *Mar. Ecol. Prog. Ser.* 354, 147–160.
- Culverhouse, P.F., et al., 2003. Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. *Mar. Ecol. Prog. Ser.* 247, 17–25.
- Culverhouse, P.F., et al., 2014. An empirical assessment of the consistency of taxonomic identifications. *Mar. Biol. Res.* 10 (1), 73–84.
- Danovaro, R., Snelgrove, P.V.R., Tyler, P., 2014. Challenging the paradigms of deep-sea ecology. *Trends Ecol. Evol.* 29 (8), 465–475.
- Danovaro, R., et al., 2017. The deep-sea under global change. *Curr. Biol.* 27 (11), R461–R465.
- Danovaro, R., et al., 2020. Ecological variables for developing a global deep-ocean monitoring and conservation strategy. *Nat. Ecol. Evol.* 4 (2), 181–192.
- Davies, J.S., et al., 2015. Benthic assemblages of the Anton Dohrn seamount (NE Atlantic): defining deep-sea biotopes to support habitat mapping and management efforts with a focus on vulnerable marine ecosystems. *PLoS One* 10 (5), e0124815.
- Ditria, E.M., et al., 2021. Annotated video footage for automated identification and counting of fish in unconstrained seagrass habitats. *Front. Mar. Sci.* 8, 160.
- Domingos, P., 2012. A few useful things to know about machine learning. *Commun. ACM* 55 (10), 78–87.
- Durden, J.M., et al., 2016. Comparison of image annotation data generated by multiple investigators for benthic ecology. *Mar. Ecol. Prog. Ser.* 552, 61–70.
- Durden, J.M., et al., 2021. Automated classification of fauna in seabed photographs: the impact of training and validation dataset size, with considerations for the class imbalance. *Prog. Oceanogr.* 102612.
- Favret, C., Sieracki, J.M., 2016. Machine vision automated species identification scaled towards production levels. *Syst. Entomol.* 41 (1), 133–143. <https://doi.org/10.1111/syen.12146>.
- Folkersen, M.V., Fleming, C.M., Hasan, S., 2018. The economic value of the deep sea: a systematic review and meta-analysis. *Mar. Policy* 94, 71–80. <https://doi.org/10.1016/j.marpol.2018.05.003>.
- Gaston, K.J., O'Neill, M.A., 2004. Automated species identification: why not? *Philosop. Trans. Roy. Soc. Lond. B: Biol. Sci.* 359 (1444), 655–667.
- Gomes-Pereira, J.N., et al., 2016. Current and future trends in marine image annotation software. *Prog. Oceanogr.* 149, 106–120.
- González-Rivero, M., et al., 2020. Monitoring of coral reefs using artificial intelligence: a feasible and cost-effective approach. *Remote Sens.* 12 (3) <https://doi.org/10.3390/rs12030489>.
- Goody, A., Bett, B., Pratt, D., 1993. Direct observation of episodic growth in an abyssal xenophyophore (Protista). *Deep-Sea Res. I Oceanogr. Res. Pap.* 40 (11–12), 2131–2143.
- Goody, A.J., Aranda da Silva, A., Pawlowski, J., 2011. Xenophyophores (Rhizaria, foraminifera) from the Nazaré canyon (Portuguese margin, NE Atlantic). *Deep-Sea Res. II Top. Stud. Oceanogr.* 58 (23), 2401–2419. <https://doi.org/10.1016/j.dsr2.2011.04.005>.
- Halpern, B.S., et al., 2007. Evaluating and ranking the vulnerability of global marine ecosystems to anthropogenic threats. *Conserv. Biol.* 21 (5), 1301–1315.
- Hiemstra, P., Hiemstra, M.P., 2013. Package "automap". *compare* 105, 10.
- Hoerer, T., Kuenzer, C., 2020. Object detection and image segmentation with deep learning on earth observation data: a review—part I: evolution and recent trends. *Remote Sens.* 12 (10) <https://doi.org/10.3390/rs12101667>.
- Hoerer, T., Bachofer, F., Kuenzer, C., 2020. Object detection and image segmentation with deep learning on earth observation data: a review—part II: applications. *Remote Sens.* 12 (18) <https://doi.org/10.3390/rs12183053>.
- Howell, K.L., Davies, J.S., Narayanaswamy, B.E., 2010. Identifying deep-sea megafaunal epibenthic assemblages for use in habitat mapping and marine protected area network design. *J. Mar. Biol. Assoc. U. K.* 90 (1), 33–68.
- Howell, K.L., Bullimore, R.D., Foster, N.L., 2014. Quality assurance in the identification of deep-sea taxa from video and image analysis: response to Henry and Roberts. *ICES J. Marine Sci.: J. Conseil* 71 (4), 899–906.
- Howell, K.L., et al., 2016. RRS James Cook, Cruise No. JC136, 14th May – 23rd June, DEEPLINKS: Influence of Population Connectivity on Depth-dependent Diversity of Deep-Sea Marine Benthic Biota. https://www.bodc.ac.uk/resources/inventories/cruise_inventory/report/16050/.
- Howell, K.L., et al., 2019. A framework for the development of a global standardised marine taxon reference image database (SMarTaR-ID) to support image-based analyses. *PLoS One* 14 (12), e0218904.
- Howell, K.L., et al., 2020. A blueprint for an inclusive, global deep-sea ocean decade field program. *Front. Mar. Sci.* 999.
- Howell, K.L., et al., 2021. A decade to study deep-sea life. *Nat. Ecol. Evol.* 5 (3), 265–267.
- Hughes, D.J., Gage, J.D., 2004. Benthic metazoan biomass, community structure and bioturbation at three contrasting deep-water sites on the northwest European continental margin. *Prog. Oceanogr.* 63 (1–2), 29–55. <https://doi.org/10.1016/j.pocean.2004.09.002>.
- Hurlbert, S.H., 1984. Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* 54 (2), 187–211.
- Huvenne, V.A., et al., 2018. ROVs and AUVs. In: *Submarine Geomorphology*. Springer, pp. 93–108.
- Huvenne, V.A.I., et al., 2016. Effectiveness of a deep-sea cold-water coral Marine protected area, following eight years of fisheries closure. *Biol. Conserv.* 200, 60–69.
- Jones, D.O., et al., 2019. Autonomous marine environmental monitoring: application in decommissioned oil fields. *Sci. Total Environ.* 668, 835–853.

- Katija, K., et al., 2021. FathomNet: an open, underwater image repository for automated detection and classification of midwater and benthic objects. *Mar. Technol. Soc. J.* 55 (3), 136–137.
- Kattenborn, T., et al., 2021. Review on convolutional neural networks (CNN) in vegetation remote sensing. *ISPRS J. Photogramm. Remote Sens.* 173, 24–49. <https://doi.org/10.1016/j.isprsjprs.2020.12.010>.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. 'Imagenet classification with deep convolutional neural networks', in: *Adv. Neural Inf. Proces. Syst.* 1097–1105.
- Langenkämper, D., et al., 2017. BIGLE 2.0 - browsing and annotating large marine image collections. *Front. Mar. Sci.* 4 (83) <https://doi.org/10.3389/fmars.2017.00083>.
- LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *nature* 521 (7553), 436.
- Levin, L.A., 1991. Interaction between metazoans and large, agglutinating protozoans: implications for the community structure of deep-sea benthos. *Am. Zool.* 31, 886–900.
- Levin, L.A., Gooday, A.J., 1992. Possible roles for Xenophyophores in deep-sea carbon cycling. In: Rowe, G.T., Pariente, V. (Eds.), *Deep-Sea Food Chains and the Global Carbon Cycle*. Kluwer, Netherlands, pp. 93–104.
- Levin, L.A., et al., 2019. Global observing needs in the Deep Ocean. *Front. Mar. Sci.* 6 (241) <https://doi.org/10.3389/fmars.2019.00241>.
- Li, D., et al., 2017. Deeper, broader and artier domain generalization. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5542–5550.
- MacLeod, N., Benfield, M., Culverhouse, P., 2010. Time to automate identification. *Nature* 467 (7312), 154–155.
- Marburg, A., Bigham, K., 2016. Deep learning for benthic fauna identification. In: *OCEANS 2016 MTS/IEEE Monterey*. IEEE, pp. 1–5.
- Marini, S., et al., 2018. Automated estimate of fish abundance through the autonomous imaging device GUARD1. *Measurement* 126, 72–75. <https://doi.org/10.1016/j.measurement.2018.05.035>.
- Marini, S., et al., 2022. Long-term automated visual monitoring of Antarctic benthic fauna. *Methods Ecol. Evol.* 13 (8), 1746–1764. <https://doi.org/10.1111/2041-210X.13898>.
- Matabos, M., et al., 2017. Expert, crowd, students or algorithm: who holds the key to deep-sea imagery 'big data' processing? *Methods Ecol. Evol.* 8 (8), 996–1004.
- Mayer, L.A., 2006. Frontiers in seafloor mapping and visualization. *Mar. Geophys. Res.* 27 (1), 7–17.
- McClain, C.R., Rex, M.A., 2015. Toward a conceptual understanding of β -diversity in the Deep-Sea benthos. *Annu. Rev. Ecol. Syst.* 46 (1), 623–642. <https://doi.org/10.1146/annurev-ecolsys-120213-091640>.
- Menardi, G., Torelli, N., 2014. Training and assessing classification rules with imbalanced data. *Data Min. Knowl. Disc.* 28 (1), 92–122.
- Meyer, H.K., et al., 2019. Spatial patterns of arctic sponge ground fauna and demersal fish are detectable in autonomous underwater vehicle (AUV) imagery. *Deep-Sea Res. I Oceanogr. Res. Pap.* 153, 103137.
- Milligan, R.J., et al., 2016. High resolution study of the spatial distributions of abyssal fishes by autonomous underwater vehicle. *Sci. Rep.* 6, 26095. <https://doi.org/10.1038/srep26095>.
- Miloslavich, P., et al., 2018. Essential Ocean variables for global sustained observations of biodiversity and ecosystem changes. *Glob. Chang. Biol.* 24 (6), 2416–2433.
- Möller, T., Nattkemper, T.W., 2021. ALMI—A generic active learning system for computational object classification in marine observation images. *Sensors* 21 (4). <https://doi.org/10.3390/s21041134>.
- Morris, K.J., et al., 2014. A new method for ecological surveying of the abyss using autonomous underwater vehicle photography. *Limnol. Oceanogr. Methods* 12 (11), 795–809.
- Morris, K.J., et al., 2016. Landscape-scale spatial heterogeneity in phytodetrital cover and megafauna biomass in the abyss links to modest topographic variation. *Sci. Rep.* 6, 34080.
- Morrissey, D.J., et al., 1992. Spatial variation in soft-sediment benthos. *Marine Ecol. Progr. Ser.* Oldendorf 81 (2), 197–204.
- Norouzzadeh, M.S., et al., 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. In: *Proceedings of the National Academy of Sciences*, p. 201719367.
- Ospar, 2008. OSPAR list of threatened and/or declining species and habitats. *OSPAR Agreement* 2008-06 1–4.
- Pavoni, G., et al., 2019. Semantic segmentation of benthic communities from orthomosaic maps. In: *2019 Underwater 3D Recording and Modelling A Tool for Modern Applications and CH Recording*. Copernicus GmbH, pp. 151–158.
- Pavoni, G., et al., 2020. On improving the training of models for the semantic segmentation of benthic communities from orthographic imagery. *Remote Sens.* 12 (18), 3106.
- Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. *Comput. Geosci.* 30 (7), 683–691.
- Perkins, N.R., et al., 2016. Image subsampling and point scoring approaches for large-scale marine benthic monitoring programs. *Estuar. Coast. Shelf Sci.* 176, 36–46. <https://doi.org/10.1016/j.ecss.2016.04.005>.
- Piechoud, N., et al., 2019. Automated identification of benthic epifauna with computer vision. *Mar. Ecol. Prog. Ser.* 615, 15–30.
- Pizarro, O., et al., 2013. Benthic monitoring with robotic platforms—The experience of Australia. In: *2013 IEEE International Underwater Technology Symposium (UT)*. IEEE, pp. 1–10.
- Poore, G.C., et al., 2015. Invertebrate diversity of the unexplored marine western margin of Australia: taxonomy and implications for global biodiversity. *Mar. Biodivers.* 45 (2), 271–286.
- Przeslawski, R., et al., 2018. Comparative assessment of seafloor sampling platforms. Report to the National Environmental Science Programme. Mumbai: Nat. Envi. Sci. Prog. 57.
- QGIS Development Team, 2020. QGIS'. Available at: <https://qgis.org/en/site/>.
- Ramirez-Llodra, E., et al., 2010. Deep, diverse and definitely different: unique attributes of the world's largest ecosystem. *Biogeosciences* 7 (9), 2851–2899.
- Ramirez-Llodra, E., et al., 2011. Man and the last great wilderness: human impact on the deep sea. *PLoS One* 6 (8), e22588.
- Redmon, J., 2013. Darknet: Open Source Neural Networks in c.
- Redmon, J., et al., 2016. 'You only look once: unified, real-time object detection', in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 779–788.
- Rex, M.A., Etter, R.J., 2010. *Deep-Sea Biodiversity: Pattern and Scale*. Harvard University Press.
- Roberts, J.M., et al., 2000. Seabed photography, environmental assessment and evidence for deep-water trawling on the continental margin west of the Hebrides. *Hydrobiologia* 441 (1–3), 173–183. <https://doi.org/10.1023/A:1017550612340>.
- Ross, L.K., et al., 2015. The influence of data resolution on predicted distribution and estimates of extent of current protection of three 'Listed' Deep-Sea habitats. *PLoS One* 10 (10), e0140061.
- Ross, R.E., Howell, K.L., 2013. Use of predictive habitat modelling to assess the distribution and extent of the current protection of "listed" deep-sea habitats. *Divers. Distrib.* 19 (4), 433–445. <https://doi.org/10.1111/ddi.12010>.
- Schneider, S., Taylor, G.W., Kremer, S.C., 2018. Deep learning object detection methods for ecological camera trap data. *15th Conference on computer and robot vision (CRV)* 321–328.
- Schneider, S., et al., 2020. Three critical factors affecting automated image species recognition performance for camera traps. *Ecol. Evol.* 10 (7), 3503–3517.
- Schoening, T., et al., 2012. Semi-automated image analysis for the assessment of megafaunal densities at the Arctic deep-sea observatory HAUSGARTEN. *PLoS One* 7 (6), e38179.
- Schoening, T., Osterloff, J., Nattkemper, T.W., 2016. RecoMIA—recommendations for marine image annotation: lessons learned and future directions. *Front. Mar. Sci.* 3, 59.
- Schoening, T., et al., 2017. Report on the marine imaging workshop 2017. *Res. Ideas Outcomes* 3, e13820.
- Skiljan, I., 2012. IrfanView. URL: <http://irfanview.tuwien.ac.at>. Retrieved June [Preprint].
- Tabak, M.A., et al., 2019. Machine learning to classify animal species in camera trap images: applications in ecology. *Methods Ecol. Evol.* 10 (4), 585–590. <https://doi.org/10.1111/2041-210X.13120>.
- Team, R.C. (2021) 'R: A language and environment for statistical computing (R Version 4.0.3), R Foundation for Statistical Computing, Vienna, Austria, 2020'.
- Tendal, O., 1972. Xenophyophore. *Natures Verden* 1972 (8–9), 302–310.
- Thurber, A.R., et al., 2014. Ecosystem function and services provided by the deep sea. *Biogeosciences* 11 (14), 3941–3963.
- Tsuchiya, M., Nomaki, H., 2021. Rapid response of the giant protist xenophyophores (foraminifera, Rhizaria) to organic matter supply at abyssal depths revealed by an in situ dual stable isotope labeling experiment. *Deep-Sea Res. I Oceanogr. Res. Pap.* 176, 103608.
- Weinstein, B.G., 2018. A computer vision for animal ecology. *J. Anim. Ecol.* 87 (3), 533–545.
- Whytock, R.C., et al., 2021. Robust ecological analysis of camera trap data labelled by a machine learning model. *Methods Ecol. Evol.* 12 (6), 1080–1092. <https://doi.org/10.1111/2041-210X.13576>.
- Wickham, H. (2017) 'Tidyverse: easily install and load tidyverse packages', R Package Version, 1(1).
- Williams, I.D., et al., 2019. Leveraging automated image analysis tools to transform our capacity to assess status and trends of coral reefs. *Front. Mar. Sci.* 6 (222) <https://doi.org/10.3389/fmars.2019.00222>.
- Wölfel, A.-C., et al., 2019. Seafloor mapping – the challenge of a truly Global Ocean bathymetry. *Front. Mar. Sci.* 6 (283) <https://doi.org/10.3389/fmars.2019.00283>.
- Woodall, L.C., et al., 2018. A multidisciplinary approach for generating globally consistent data on mesophotic, deep-pelagic, and bathyal biological communities. *Oceanography* 31 (3), 76–89.
- Wynn, R.B., et al., 2014. Autonomous underwater vehicles (AUVs): their past, present and future contributions to the advancement of marine geoscience. *Mar. Geol.* 352, 451–468.
- Ying, X., 2019. An overview of overfitting and its solutions. *J. Phys. Conf. Ser.* 1168, 022022. <https://doi.org/10.1088/1742-6596/1168/2/022022>.