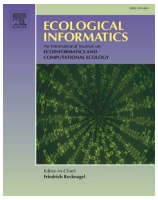
Ecological Informatics 95 (2026) 103749

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Testing the transferability of AI models for cold-water coral detection and classification

Kyran P. Graves [a,b,*], Louise Allcock [c], David K.A. Barnes [d], Amelia E.H. Bridges [a], Kerry L. Howell [a,b]

[a] School of Biological and Marine Science, Plymouth University, Plymouth, UK
[b] Plymouth Marine Laboratory, Plymouth, UK
[c] Ryan Institute and School of Natural Sciences, University of Galway, Galway, Ireland
[d] Department of Biodiversity, Evolution and Adaptation, The British Antarctic Survey (BAS), UKRI, Cambridge, UK

ARTICLE INFO

Keywords:
Cold-water coral
Object detection
Deep learning
Computer vision
YOLOv5
Model transfer

ABSTRACT

The proliferation of accessible deep-water imaging platforms has resulted in the acquisition of vast amounts of image data, resulting in an analysis bottleneck. Object detection is now being applied to assist the image annotation process, with the potential to reduce analysis time. However, for object detectors to effectively tackle the scale of the challenge, models need to be generalisable to allow the transfer between imaging platforms and in space. This study trains YOLOv5 object detection models to identify six coral morphology groups using annotated imagery collected by ROV ISIS in the UK (JC136). Model performance was tested with independent datasets to inspect different aspects of transferability. Imagery collected on Tropic Seamount near the Canary Islands (JC142) with the same ROV (ISIS) was used to test spatial transferability. Imagery collected with ROV Holland I (SeaRover Project) from the Irish deep sea was used to test the transferability of models between ROVs. Model performance was moderate, recalling 60% of human annotations when evaluated against the validation dataset with varying performance across morphological groups (Recall = 44–69%). However, when tested using the independent datasets, model performance falls, recalling only 23% to 34% of human annotations across transfer scenarios. The results suggest that the model performance when transferred was poor, arising because of high shape variability within some morphological groups and poor taxonomic representation across datasets. We discuss how a coordinated community effort could improve model transferability and potentially address the analysis bottleneck.

1. Introduction

Technological advances have seen the proliferation of accessible deep-water imaging platforms, such as towed cameras, Remotely Operated Vehicles (ROVs) and Autonomous Underwater Vehicles (AUVs). Such platforms are critical to surveying and monitoring ecologically important and sensitive areas, such as those supporting cold-water corals. Vast image datasets can be generated from a single research cruise, resulting in an analysis bottleneck where experienced analysts may take months to years to annotate imagery collected from a single survey. This challenge will ultimately be exacerbated as ocean-observing strategies become increasingly autonomous and data collection rates continue accelerating. This is set against the backdrop of the climate and biodiversity crises, expanding human uses of the oceans (e.g. deep-sea mining) and the ratification and implementation of the UN BBNJ Agreement. This places an urgency on the quantitative assessment of the status and change in marine ecosystems and the societal benefits they provide. To achieve this, it is critical to reduce and eliminate the data bottleneck problem. Now, within a deep-sea context, deep learning (DL) and computer vision (CV) are being applied to assist image annotation and tackle the analysis bottleneck (Bridges et al., 2025; Piechaud and Howell, 2022).

CV is a field of artificial intelligence (AI) that interprets and analyses visual data, such as images and videos. DL is a subset of machine learning that uses large neural networks to determine patterns and features in raw data and learn from them (Rubbens et al., 2023). Neural networks are inspired by the structure of the brain with an input layer (features), at least one 'hidden' layer and an output layer which

* Corresponding author at: School of Biological and Marine Science, Plymouth University, Plymouth, UK.
E-mail address: kyran.graves@plymouth.ac.uk (K.P. Graves).

https://doi.org/10.1016/j.ecoinf.2026.103749
Received 28 April 2025; Received in revised form 27 March 2026; Accepted 28 March 2026
Available online 30 March 2026

produces predictions. Algorithms designed to locate and classify objects within an image are referred to as "object detectors". Once trained, these algorithms can be applied to new data (e.g. an image) to detect and classify (predict) learnt objects. Typically, 'supervised' learning methods are used meaning that the data used to train such algorithms are labelled, i.e. objects within images have been manually annotated and classified.

Numerous studies have applied DL algorithms to seafloor imagery to assist or automate the process of image annotation for biological and geological data. Models have been trained to detect and classify a single group such as a particular coral species, *Dendrophyllia cornigera* (Abad-Uribarren et al., 2022; Lamarck, 1816), or xenophyophore species, *Syringammina fragilissima* Brady, 1883 (Piechaud and Howell, 2022). Multi-group models have also been trained to detect multiple objects (e.g. Beijbom et al., 2015; Cuvelier et al., 2024; Iyer et al., 2025; Mbani et al., 2023; Piechaud et al., 2019; Zhang et al., 2022). For example, Deo et al. (2024) trained a series of classification models to detect 33 groups; spanning a variety of taxa from feather stars to sea pens to tube worms; and included two non-living groups; coral rubble and woody debris. DL algorithms can also be trained to detect and classify benthic habitats (e.g. Game et al., 2024) and substrates (e.g. Jackett et al., 2023). As well as processing imagery to identify objects; DL algorithms can be used to generate count and density data (e.g. Marini et al., 2018; Piechaud and Howell, 2022), size data (e.g. Álvarez-Ellacuría et al., 2020; Piechaud and Howell, 2022), analyse timelapse image data (e.g. producing growth data; Osterloff et al., 2019); as well as be incorporated into larger data processing; analysis and modelling pipelines; such as species distribution modelling (Abad-Uribarren et al., 2022).

The scale of the image analysis bottleneck is vast. Piechaud and Howell (2022) demonstrate how DL can be utilised to reduce the time taken to annotate a single taxon across 60,000 images from hundreds of hours to less than 10 days – including model training. For DL to address the scale of the image analysis bottleneck, CV and DL models need to be generalised enough to transfer geographically and across imaging platforms, avoiding the need for new models to be trained for each newly collected dataset.

A time- and resource-effective way of tackling the bottleneck would be to develop DL models that perform consistently across most frequently used imaging platform types, e.g. different towed cameras, ROVs, or AUVs. However, this poses challenges regarding the comparability of imagery produced between platforms and whether model performance would be consistent. The comparability of seafloor imagery is affected by a platform's camera, lighting system and setup (e.g. camera resolution and angle, white balance, colour, brightness, distribution of light, etc.), the type of platform (e.g. fixed camera angle), and operational factors (e.g. zoom, distance from the seabed, etc.). In addition to comparability *between* platforms, the comparability of imagery acquired from the same platform may also change over time, for example, if camera or lighting systems are upgraded. These factors affect how an object will appear in an image, with the potential for an identical object to appear differently in imagery collected from different platforms, and therefore affect a model's transferability.

The spatial transferability of models also affects the ability of DL to address the scale of the bottleneck problem. Idealistically, the performance of a trained model would remain constant when transferred and applied to a new location, whether that be locally, regionally or at a basin scale. Factors affecting spatial transferability include the parameterisation of the model during training, e.g. models are not overfitted, and if the model training dataset is representative of the new transfer area, e.g. do the same model groups occur in the training (source domain) and transfer (target domain) datasets? Testing the dual aspects of spatial and cross-platform transferability of a single DL model is currently an unmet research gap in the context of deep-sea research.

1.1. Aims

This study investigates the application of DL models to the real-world challenge of cold-water coral surveying and monitoring. Specifically, we aim to test the transferability of DL models under two different scenarios to address the described research gap:

1. A transfer to a new location using the same imaging platform (spatial transfer).
2. A transfer to the same location but a different imaging platform.

2. Methods

2.1. Study area & data collection

Archived video transects were collated from five research cruises across three projects: Deep-Links (JC136; Howell et al., 2016); MarineE-Tech (JC142; Murton, 2016) and SeaRover (RH17001; RH18002; CE19015; O'Sullivan et al., 2017, 2018, 2019). Although all research cruises had individual objectives, all conducted high-definition video transects of the seabed using an ROV to investigate and characterise the benthos (Table 1, Fig. 1).

2.2. Image analysis

For each dataset, still frames were extracted every 20 s from each ROV transect video for annotation and deinterlaced using FFmpeg. An extraction interval of 20 s was used because when the ROV was moving, 20 s reduced the number of overlapping frames, whilst not losing a vast amount of image data. Black corals and octocorals (referred to herein as corals) were annotated with perpendicular boxes in BIIGLE 2.0 (Langenkämper et al., 2017). Images were not annotated where the image quality was too poor to identify corals, for example, where lighting was too low, or the camera was obscured by suspended material.

Corals were classified by expert annotators into one of the following morphological groups (classes): Arborescent, Bottlebrush, Branching-3D, Fan—2D, Mushroom and Unbranched (Fig. 2) based on the SMarTaR-ID Morphology Tree (Howell et al., 2019). These annotations were quality checked by a single annotator. This was carried out in BIIGLE using the Largo tool; which allows annotators to easily review annotations in a regular grid. The coral section of the SMarTaR-ID Morphology Tree is based on an unpublished morphological classification system developed by coral taxonomist Denis Opresko; and the Catami classification system (Althaus et al., 2015). Training was delivered by post-doctoral researchers within the research group led by Prof

Table 1
Details of research cruises where imagery was collected.

	Deep-Links (JC136)	*MarineE-Tech (JC142)*	*SeaRover (RH17001, RH18002,* CE19015*)*
Location	*UK*	*Tropic Seamount (Canary Islands)*	*Ireland*
ROV	*ISIS*	*ISIS*	*Holland I*
Camera	*Sony HDR-CX560V (1080i)*	*Sony HDR-CX560V (1080i)*	*Kongsberg OE14–502 (1080i)*
Lighting	*2 x DSPL Multi Sealite (LED), 3 x APHOS 16 LED*	*2 x DSPL Multi Sealite (LED), 3 x APHOS 16 LED*	*4 × 250-watt halogens, 2× Bowtech LED, 4 x APHOS LED lights*
Transects	*24*	*17*	*29 (12, 10, 7)*
Target Feature(s)	*Continental Slope, Seamounts, Ridges*	*Seamount*	*Continental Slope, Banks*

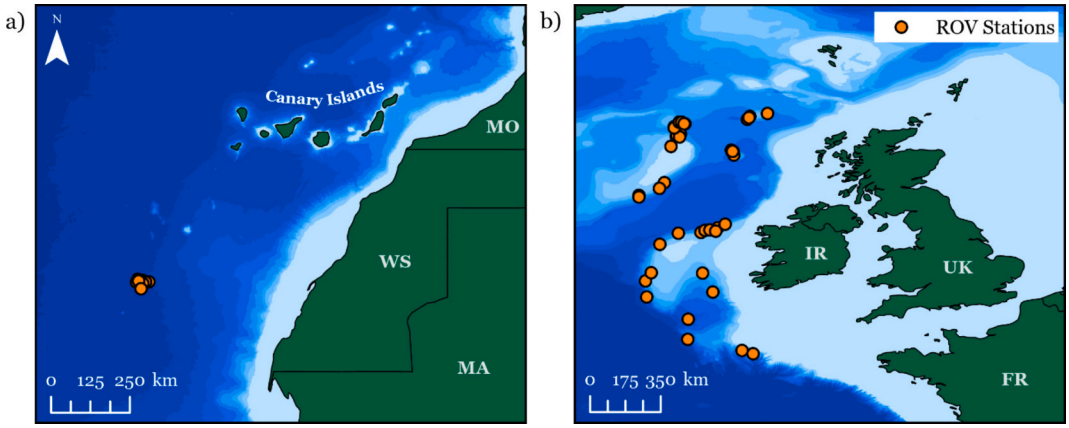


Fig. 1. a. Distribution of ROV video transects collected during the MarineE-Tech project (JC142) on Tropic Seamount (MO = Morocco, WS = Western Sahara, MA = Mauritania); Fig. 1b. The distribution of ROV video transects collected during the Deep-Links (JC136) and SeaRover (RH17001, RH18002, CE19015) projects (IR = Ireland, UK = The United Kingdom, FR = France).

Fig. 2. Example annotations of each coral morphology group. Where Row. A = Arborescent, B = Bottlebrush, C = Branching 3D, D = Fan 2D, E = Mushroom and F = Unbranched.

Howell, a co-creator of SMarTaR-ID. A morphological approach to classifying corals was taken for several reasons:

(1) Without physical specimens, identifying biota to high taxonomic resolution, e.g. species, from imagery is difficult – often impossible.

(2) Annotators often take a two-stage approach to annotating, as recommended by Howell et al. (2023). The first pass of the image data involves annotating to morphological groupings before refining such groups in a second pass, for example, into operational taxonomic units (OTUs).
(3) Object classification models have no taxonomic knowledge and learn only from features detectable from an image. It is theoretically, therefore, more relevant to DL that corals are labelled based on their morphological (physical) features and classified into morphologically similar groups.
(4) We assume that morphology is more generalisable than taxonomy. For example, two individual corals may be morphologically similar in imagery but be two separate species. Morphology would be more suitable in the context of spatial transferability because coral morphologies are geographically cosmopolitan, whereas coral species may not be.

2.3. Data preparation

For each dataset, annotation reports were downloaded from BIIGLE 2.0 for data preparation. In R (R Core Team, 2023), bounding box coordinates for each annotation were converted from BIIGLE (x_min, y_min, x_max and y_max in pixel coordinates) to YOLO format (x_centre and y_centre coordinates normalised by image height and width) – a subset of which were plotted onto images to ensure coordinates and scaling was correct. JC136 annotations were randomly subset by image into training (70%; Train-JC136) and validation (30%; Val-JC136) datasets. To test model performance and transferability, the JC142 (Test-JC142) and three SeaRover (Test-SR17/18/19) datasets were kept separate to act as independent test datasets. Libraries of corresponding images, labels, and YAML files were prepared and formatted for model training and evaluation. The number of annotations by dataset and morphological group are available in Table 2 which highlights the group (class) imbalance in the datasets used, representing typical deep-sea benthic imagery datasets.

2.4. Model training

A 'You Only Look Once' Version 5 (YOLOv5) Convolutional Neural Network (CNN) by Ultralytics (Jocher et al., 2022) was used to detect coral morphology groups (classes). When this study was initiated; YOLOv6 had just been released; but YOLOv5 was used because it is simple and fast to implement; and had been shown to perform well at detecting benthic megafauna in previous studies (Hou et al., 2023; Piechaud and Howell, 2022; Wang et al., 2024; Zhang et al., 2024). Additionally, YOLO only requires one pass of the new data through the CNN to make a prediction. This allows the algorithm to process imagery and make predictions in real-time, i.e. from live video. In the long term, working within the YOLO architecture leaves the potential for "on the fly" image annotation during data collection at sea, as demonstrated by Browne (2022).

Table 2
Number of annotations per dataset, per morphological group, and the number of images per dataset (bottom row).

Morphological Group	Train-JC136	Val-JC136	Test-JC142	Test-SR17	Test-SR18	Test-SR19
Arborescent	272	154	5	0	0	0
Bottlebrush	149	39	49	162	55	19
Branching-3D	1721	556	503	480	1745	165
Fan-2D	1425	409	245	193	153	40
Mushroom	138	39	8	14	11	26
Unbranched	965	331	219	3722	901	89
Total	4670	1528	1029	4571	2865	339
Number of Images	1539	513	186	965	788	175

Off-the-shelf, YOLOv5 is pre-trained using the COCO-2017 dataset (Lin et al., 2015), meaning re-training to update the CNN architecture to detect features of the custom image dataset (coral morphology) is required – a process referred to as transfer learning. A YOLOv5 model was re-trained using the training dataset (Train-JC136) within the PyTorch framework, using a Google Collaboratory Professional account. 'Google Colab' is a cloud-based service that allows remote access to a GPU and high RAM facilities that were unavailable locally.

Preliminary models were run to determine the largest model size (small, medium or large), batch size and number of epochs, and the highest image resolution that could be run using the computational resources available via Google Colab. No other hyperparameters, e.g. learning rate, were optimised. Three final YOLOv5-m (medium size) models were trained using image resolutions of 960 × 960 and 250 epochs, but with different batch sizes (8, 16 & 32; Mod-b8, Mod-b16, Mod-b32). Epoch refers to the number of complete passes of the training dataset through the CNN, and batch size refers to the number of images (samples) used in a single forward and backward pass of the CNN. Three models of equally spaced batch sizes were used within the computational limits (up to a batch of 32) to test the effect on model transferability and, by inference, model performance when transferred and tested using independent datasets.

2.5. Model evaluation & transfer

Performance metrics were obtained by running a modified YOLOv5 'val.py' script on the validation dataset (Val-JC136) and the independent datasets (Test-JC142 & Test-SR17/18/19) as follows. The 'best' weights of each trained model as determined by YOLO were used to make the detections (predictions) across each of the labelled image datasets (i.e., images with bounding boxes and classification labels), where 'best' weights are selected on the highest mean average precision (mAP) score. Detections for each image were saved as a text file containing the bounding box coordinates, predicted group and confidence score for each prediction. Using an Intersection over Union (IoU) of 0.5 and a variety of confidence thresholds (0.05, 0.1, 0.3, 0.5, 0.9), the number of true positives (TPs), false positives (FPs), false negatives (FNs), and Precision, Recall and F1 metrics were extracted and calculated (defined in Table 3) for each overall model and by group. Model detections and evaluation metrics were then saved locally.

In R, predicted bounding box coordinates were converted from YOLOv5 format back into BIIGLE format. The detections were then uploaded to BIIGLE on their corresponding images in a new project for visual inspection. Upon inspection, it was clear that the models had made correct detections that human annotators originally missed, therefore inaccurately penalising model performance. As a result, false

Table 3
Definitions of key evaluation metrics.

Intersection over Union (IoU)	The percentage of overlap between a ground truth and a predicted bounding box
True Positive (TP)	A predicted bounding box with an IoU equal to or greater than 0.5 and a correctly predicted group label
False Positive (FP)	An erroneous predicted bounding box with no corresponding ground-truthing box. This includes instances where the IoU < 0.5 hasn't been met and/or the predicted group label is incorrect
False Negative (FN)	An instance of a ground-truthing bounding box with no corresponding predictions, i.e. an annotated coral has not been detected
Precision	The proportion of predicted bounding boxes (positives) that are correct: $Pr = TP / (TP + FP)$
Recall	Proportion of annotations correctly classified: $Re = TP / (TP + FN)$
F1	Combines Precision and Recall scores to create a balanced metric: $F1 = 2 \times ((Pr \times Re) / (Pr + Re))$

positive predictions from a subset of images ($n = 50$), stratified by transect, made from each of the three models (Mod-b8, Mod-b16, Mod-b32) across the five datasets (Val-JC136, Test-JC142, Test-SR17/18/19) were manually reviewed and evaluated in BIIGLE. This allowed an estimate of the rate at which models have correctly detected corals missed by human annotators, i.e., false positives that are correct (true positives). This incidence rate (False Positive Correction Rate) was then used to make an estimated adjustment to the evaluation metrics for each validation and testing dataset, as per Bridges et al. (2025). Further details on this process are given in Appendix A.1.

Following Rainio et al. (2024); Friedman's tests were used in R to test for statistically significant differences ($p < 0.05$) in performance metrics (precision; recall and F1) between Mod-b8; Mod-b16 and Mod-b32 when transferred to the testing datasets (Test-JC142; Test-SR17/18/19). A Friedman's test was chosen because it is a non-parametric repeated measures test; so it does not assume the evaluation metrics are normally distributed; the test is suitable for a small number of testing datasets ($n = 4$) and is recommended by Rainio et al. (2024). Wilcoxon tests were run as a post-hoc test to determine which models were significantly different. These were run twice using a Holm (default) and Bonferroni p-value correction, with Bonferroni used to reduce the risk of false positives.

Firstly, the best performing model-confidence threshold combination was determined for the Val-JC136 dataset. This model-threshold combination was used to make predictions across the test datasets to create a strict transferability benchmark based on the validation-derived configuration.

Secondly, the best-performing model-threshold combinations were selected for each testing dataset. This second step was taken to optimise performance for each test dataset, considering the downstream application of the models: as a tool to assist ecologists annotate benthic imagery and reduce the time spent on the annotation process. Priority was given to models and thresholds that optimised recall metrics and reduced the number of False Negatives (FNs), retaining the highest confidence threshold possible. Should models perform well, they will be implemented as tools to make detections (predictions) across new, independent datasets that will ultimately require post-processing and cleaning. In the context of post-processing, correcting false negatives (i.e., drawing bounding boxes and labelling missed corals) is more time-intensive than correcting false positives (i.e., deleting incorrect coral detections or correcting labels of detected corals). As a result, detection was viewed as a more important task (drawing bounding boxes) than classification (labelling).

During the analysis of the results, it became clear that the difference in model performances across transfer scenarios, to some degree, was driven by the differences in taxonomic make-up of morphological groups across datasets. The identification of all coral taxa across the five different datasets was not possible within this study. However, to illustrate this important point (discussed in the results and discussion sections), corals within the Unbranched coral morphology groups were identified down to family or morphospecies across all the datasets where possible. The corals were identified by a single annotator, using the 'Largo' tool in BIIGLE. This was carried out in a two-stage process by a single annotator: Stage One, all unbranched coral annotations were sorted into three groups – Antipatharia (black corals), Octocorals or Pennatulacea (seapens). Stage Two, annotations in each group were refined further and re-labelled to either a taxonomic family or a morphospecies level.

3. Results

3.1. Model training & validation

Mod-b8 and Mod-b16 were automatically stopped short of running a full 250 epochs (Fig. 3) as YOLO determined that no increase in performance had been detected over the previous 50 epochs. Mod-b8 and Mod-b16 completed 173 and 182 epochs, respectively. Mod-b32 did complete 250 epochs, but precision and recall metrics had stabilised during the training process, reaching an asymptote (Fig. 3). Model performances based on metrics obtained on the Val-JC136 dataset indicate that no single model (batch size) outperformed all other models.

Performance metrics calculated on Val-JC136 indicate that the best-performing model was Mod-b16 with a threshold of 0.3 (Fig. 4),

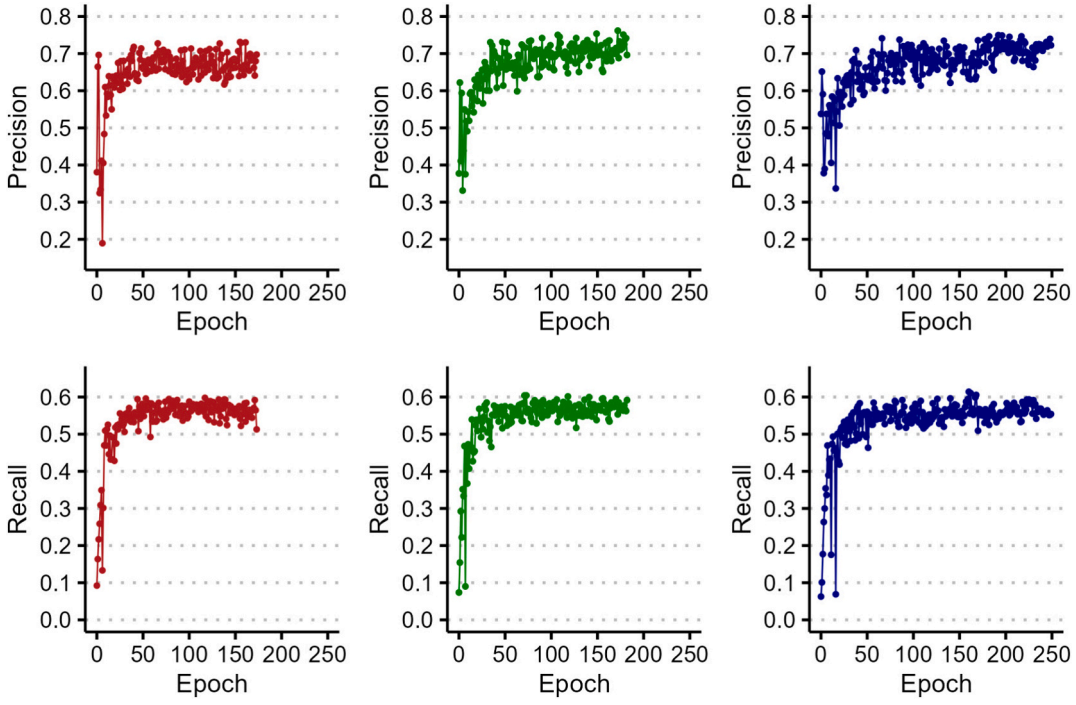


Fig. 3. Progression of recall and precision scores for each model during training. (Red = Mod-b8, Green = Mod-b16, Blue = Mod-b32).

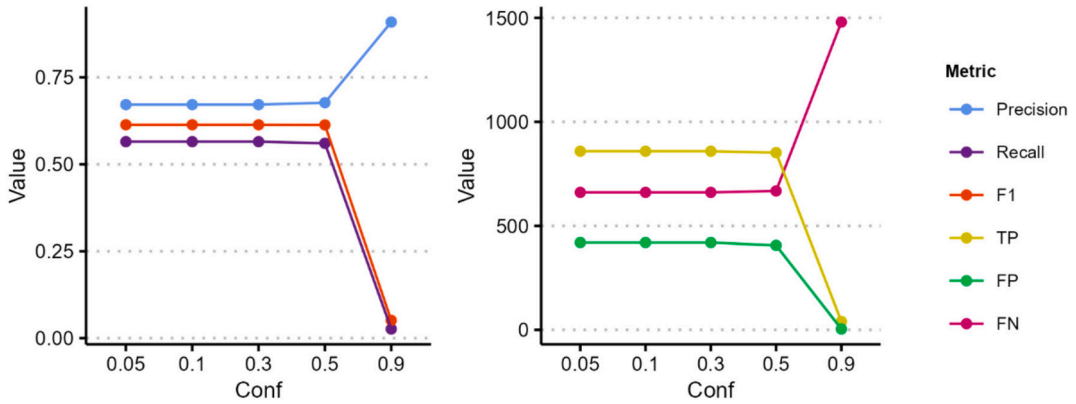


Fig. 4. Mod-b16 performance over various metrics and confidence thresholds. Note, there is a small decrease in model performance between confidence thresholds of 0.3 and 0.5 that is not discernible from this graph.

minimising the number of false negatives and maximising the recall score. A recall score of 0.57 means that 57% of the corals in the Val-JC136 dataset were detected and correctly classified, suggesting a low performance. However, a moderate precision score of 0.67 indicates that when the model makes a prediction, the prediction is correct 67% of the time.

3.2. Model testing

3.2.1. Model transfer benchmarking

The best performing model-confidence threshold combination for Val-JC136 (Mod-b16, Conf = 0.3) was used to make predictions across all four testing datasets, providing a benchmark test for transferability. This model performed poorly when tested in both transfer scenarios (Table 4a). Recall dropped from 0.57 to 0.10, and Precision dropped from 0.67 to 0.33 when tested with Test-JC142 (different region, same ROV). Model performance also dropped when tested using Test-SR17/18/19 (same region, different ROV). Recall varied from 0.12 to 0.30, and Precision varied from 0.33 to 0.49.

3.2.2. Model transfer optimisation

The performance of the three trained models (Mod-b8, Mod-b16, Mod-b32) were tested using the four independent datasets (Test-JC142, Test-SR17/18/19). The best model-threshold combination for each testing dataset was determined to optimise transfer performance for their intended application: an annotation assistance tool for ecologists.

Friedman's tests indicated that, across the four testing datasets, there were only statistically significant differences ($p < 0.05$) among models' performances for the precision metric ($p = 0.04$), but not for recall ($p = 0.17$) or F1 ($p = 1.00$). However, the subsequent post-hoc Wilcoxon tests indicated that there were no statistically significant differences between the model's precision scores for either Holm (all pairwise p-values = 0.380) or Bonferroni (Mod-8/Mod-16, $p = 0.38$; Mod-8/Mod-32, $p = 0.75$; Mod-16/32, $p = 0.38$) p-value correction methods. However, while not statistically significant, the effect sizes were large for both Holm and Bonferroni methods ($r = 0.73$–0.91). The difference in results between the Friedman's and post-hoc Wilcoxon tests is likely the result of only testing three models (Rainio et al., 2024).

The statistical tests suggest that no single model (Mod-b8, Mod-b16 or Mod-b32) outperformed other models when transferred and independently tested. As a result, the model/threshold combinations with the best performance metrics (with an emphasis on recall score) were identified for each testing dataset. These were Mod-b8 using a threshold of 0.1 for Test-JC142, and Mod-b16 and thresholds of 0.1/0.1/0.05 for datasets Test-SR17/18/19, respectively. *Re*-thresholding and model selection for each test dataset were carried out to maximise the recall metric for each dataset (Table 4b).

Compared with performance on the Val-JC136 dataset, the performance of models on all test datasets was greatly reduced across all metrics despite optimisation (Fig. 5). Additionally, the performance of models on the test datasets is less stable over increasing thresholds, suggesting reduced confidence in predictions on the testing datasets. No predictions were made with a confidence of 0.9+ for the Test-JC142 dataset.

The trained models performed best when tested on the Test-SR18 dataset (Fig. 5, green line) across both precision and recall metrics. However, it is unclear from these results whether there is a difference in transferability between ROVs or locations. There was a large variation in performance between the Test-SR17/18/19 image dataset sets collected

Table 4
Comparison of performance metrics when (a) the same model and threshold is applied to benchmark model transferability, and (b) the best-performing model and threshold combination for each testing dataset is applied to optimise model performance. Conf = Confidence threshold, TP = True Positive, FP = False Positive, FN = False Negative.

	Dataset	Batch	Conf	Precision	Recall	F1	Instances	TP	FP	FN
a)	Val-JC136	16	0.30	0.67	0.57	0.61	1520	859	420	661
	Test-JC142			0.33	0.10	0.15	1029	100	199	929
	Test-SR17			0.39	0.12	0.19	4571	560	892	4011
	Test-SR18			0.49	0.30	0.37	2865	846	896	2019
	Test-SR19			0.33	0.20	0.25	339	67	139	272
	Total						10,324	2432	2546	7892
b)	Val-JC136	16	0.30	0.67	0.57	0.61	1520	859	420	661
	Test-JC142	8	0.10	0.32	0.17	0.22	1029	174	374	855
	Test-SR17	16	0.10	0.33	0.16	0.21	4571	731	1511	3840
	Test-SR18	16	0.05	0.44	0.33	0.38	2865	938	1191	1927
	Test-SR19	16	0.10	0.31	0.21	0.25	339	71	159	268
	Total						10,324	2773	3655	7551

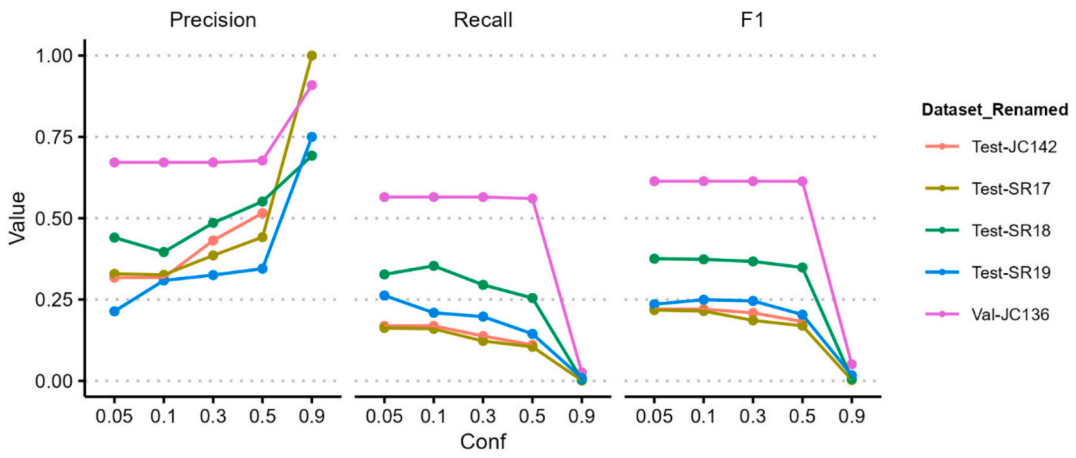


Fig. 5. Precision, recall and F1 metrics for the best performing model for each dataset (Val-JC136 = Mod-b16; Test-JC142 = Mod-b8; Test-SR17/18/19 = Mod-b16) across various tested confidence thresholds (Conf).

using ROV Holland I. When thresholded to maximise performance (0.1/0.1/0.05), Test-SR17/18/19 dataset precision scores varied between 0.31 and 0.44, and recall scores between 0.16 and 0.33; a variation in which the precision and recall scores for the Test-JC142 (ROV ISIS), 0.32 and 0.17 respectively, sit. As the performances of the models under the two transfer scenarios are not distinct, with neither scenario outperforming the other, the differences in performance are likely driven by dataset-specific factors, rather than a specific transfer scenario.

However, despite having minimal effects on the Val-JC136 performance, results suggest that batch size can impact the performance and transferability of a model to new independent datasets. For comparison, when all three models (Mod-b8, Mod-b16, Mod-b32) were applied to Test-SR18, and predictions thresholded at a confidence level of 0.1, precision and recall scores varied from 0.40 to 0.52 and 0.27–0.35, respectively – differences of 12% & 8%. Full testing results for each model on each dataset are in Appendix A.2. Additionally, taking the additional optimisation step did improve model performance, albeit not statistically different. By optimising models and thresholds for each dataset, FN decreased by 341 across all the datasets compared to the Val-JC136 transfer benchmark results.

3.3. Manual evaluation & adjustment

From each dataset, 50 images were randomly sampled (stratified across transects) for manual evaluation to determine the rate at which each model correctly identified corals missed by human annotators, i.e. FPs that are TPs. The FP Correction Rate – the proportion of FPs that were considered TPs after manual evaluation – varies across datasets, from 5.2% to 21.3% (Table 5). When metrics are recalculated using the correction rate, the performance across all metrics is improved. As calculating precision considers TPs and FPs, precision is more greatly affected by the corrections than recall, which considers TP and FN. The largest FP Correction Rates were observed in the Val-JC136 and Test-SR17 datasets, with 21.3% and 12.6%, respectively. These results indicate that although the model performance may be poor overall, when tested independently, the models are still capable of detecting and correctly classifying corals missed by human annotators (e.g. Test-SR17).

3.4. Performance by morphological group

Model performance by morphological group was highly variable (Table 6). Results from the Val-JC136 dataset indicated variability in recall (0.54 to 0.71) and precision (0.63 to 0.81) scores between groups. The best-performing group by recall was 'Mushroom' (0.71), and the worst, 'Bottlebrush', 'Branching 3D' and 'Fan 2D' (0.54). By precision, the best-performing group was 'Bottlebrush' (0.81), and the worst-performing group was 'Fan 2D' (0.63). Similar to the overall model performances, performance across groups decreased when tested with independent datasets. However, the reduced model performance was not uniform across test datasets. Except for the 'Mushroom' group, recall for the Test-JC142 across all other groups was lower than the Test-SR17/18/19 datasets.

Confusion matrices depict the degree of confusion by the model, representing the proportion of ground truthing annotations assigned across the predicted coral groups. Outside correlations between the same groups, the only correlations between two groups greater than or equal to 10% occurred between Mushroom-Branching 3D and, more consistently, Branching 3D-Fan 2D (Table 7). However, confusion was highest between groups and 'background', where 'background' is deemed any image pixels outside a ground truthing bounding box. Confusion between true background and predicted groups was higher in the testing datasets than in the validation dataset, reflecting the relative increase in false presences and reduced precision scores. Complete confusion matrices for each dataset are available in Appendix A.3.

The taxonomic breakdown of the Unbranched coral morphological group is detailed in Fig. 6. The most dominant taxa for each dataset were as follows: *Stichopathes gravieri* (Molodtsova, 2006) represents 71% of the JC136 unbranched corals; *Stichopathes* msp. 1 represents 73% of the Test-JC142 unbranched corals; and *Stichopathes* msp. 2 represents 70%, 67% and 39% of the Test-SR17, SR18 and SR19 unbranched corals, respectively. Example annotations of the three dominant Stichopathes

Table 5
Performance metrics (precision, recall, F1) for each dataset and the adjusted metrics according to the dataset's FP Correction Rate.

Dataset	Precision	Recall	F1	Precision Adjusted	Recall Adjusted	F1 Adjusted	FP Correction Rate
Val-JC136	0.67	0.57	0.61	0.76	0.60	0.67	21.3%
Test-JC142	0.32	0.17	0.22	0.33	0.18	0.23	5.2%
Test-SR17	0.33	0.16	0.21	0.37	0.18	0.24	12.6%
Test-SR18	0.44	0.33	0.38	0.47	0.34	0.40	6.8%
Test-SR19	0.31	0.21	0.25	0.33	0.22	0.26	5.2%

Table 6
Performance metrics (P = precision, R = recall) for individual morphological groups for validation and test datasets.

Morphological Group	Validation		Test							
	JC136		JC142		SR17		SR18		SR19	
	P	R	P	R	P	R	P	R	P	R
Arborescent	0.69	0.62	0.00	0.00	NA	NA	NA	NA	NA	NA
Bottlebrush	0.81	0.54	0.40	0.16	0.45	0.20	0.39	0.38	0.00	0.00
Branching 3D	0.71	0.54	0.48	0.21	0.39	0.40	0.65	0.43	0.44	0.34
Fan 2D	0.63	0.54	0.20	0.15	0.16	0.53	0.15	0.50	0.08	0.38
Mushroom	0.70	0.71	0.67	0.50	0.40	0.14	0.59	0.27	0.67	0.23
Unbranched	0.64	0.57	0.16	0.09	0.39	0.11	0.23	0.10	0.14	0.13

Table 7
Instances of confusion between morphological groups equal to or greater than 10% across validation and test datasets.

Dataset	*True Group*	*Predicted Group*	*Confusion (%)*
Val-JC136	*Branching 3D*	*Fan 2D*	*11*
	Fan 2D	*Branching 3D*	*14*
Test-JC142	*Mushroom*	*Branching 3D*	*12*
Test-SR17	*Branching 3D*	*Fan 2D*	*21*
	Fan 2D	*Branching 3D*	*19*
Test-SR18	*Fan 2D*	*Branching 3D*	*21*
Test-SR19	*Branching 3D*	*Fan 2D*	*12*
	Fan 2D	*Branching 3D*	*10*

taxa are available in Fig. 7.

4. Discussion

4.1. Model training & validation

This study, using transfer learning, trained YOLOv5 algorithms to detect six morphological groups of cold-water coral using imagery collected using ROV ISIS on the research cruise JC136 and validated those models. These models were then tested using four independent image datasets from four different research cruises: JC142 (collected with ROV ISIS) to test spatial transferability, and SeaRover 2017/2018/2019 (collected with ROV Holland I) to test transferability between

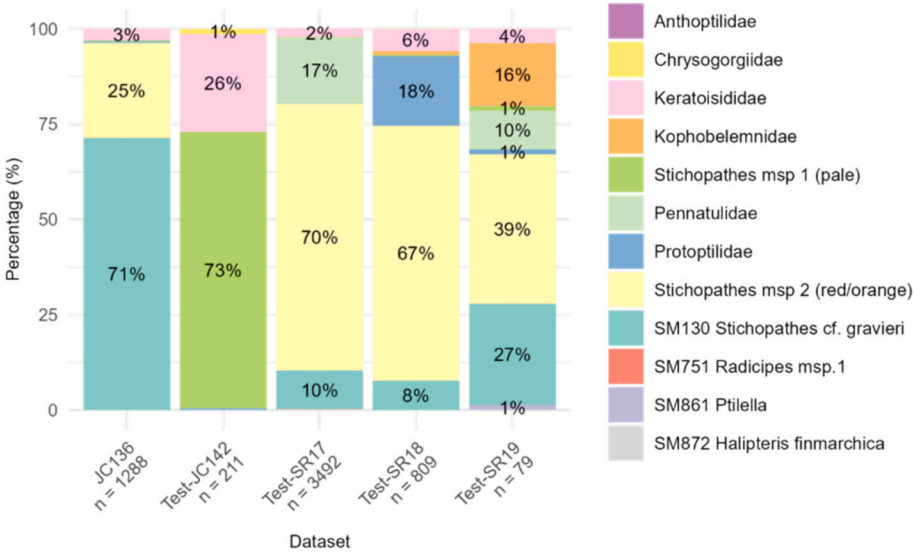


Fig. 6. Breakdown of the taxa comprising the Unbranched morphology group for each dataset, where n = number of annotations.

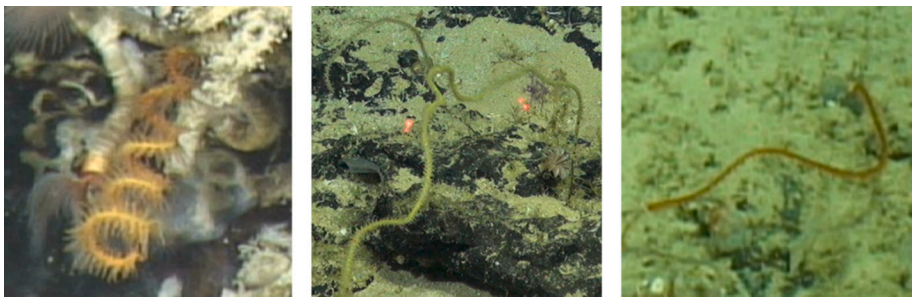


Fig. 7. Example annotations of the dominant unbranched taxa across the annotation datasets: Stichopathes gravieri (left - orange, thickly coiled), Stichopathes msp.1 (middle – thin, not coiled, pale, dull colour), and Stichopathes msp. 2 (right – thin, not coiled, orange or red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

imaging platforms (ROVs).

The overall performance of models trained and validated using the JC136 dataset was only moderate for precision (Pre = 0.67, Adj-Pre = 0.76) and poorer for recall (Re = 0.57, Adj-Re = 0.60). When comparing the performance of the three individual models – Mod-b8, Mod-b16 and Mod-b32 – validated on the Val-JC136 dataset, there was no obvious divergence in performance (Fig. 3), suggesting that batch size had little effect on the performance of this dataset and only remained moderate. For comparison, the DeepSee Model (Iyer et al., 2025), which utilises a YOLOv8 architecture, is trained from ROV imagery to detect 15 morphospecies, including "Gersemia" – a genus of soft coral, has precision and recall scores for the overall model of 0.85 and 0.78, respectively, and 0.87 and 0.82 for the "Gersemia" group.

4.2. Model testing

Results indicate poor model transferability when benchmarked with independent datasets, with a substantial drop in performance compared to the validation data, which was only initially moderate. After model optimisation – selecting the best model-confidence threshold – for each testing dataset, model transferability and performance remained poor.

No model – Mod-b8, Mod-b16 or Mod-b32 – consistently outperformed other models when tested independently; Mod-16 performed best on Val-JC136, Mod-16 performed best across the Test-SR datasets, and Mod-8 performed best on Test-JC142. These results suggest that, at least within those tested, a one-size-fits-all model parameterisation approach does not yield the best results. Across the three models and batch sizes used in this study, recall performance varied by up to 8% when independently tested. When considering that image datasets from a single research cruise can be composed of 10,000 s of images and 100,000 s of annotations, an improvement of 8% recall performance could potentially reduce annotation time significantly. It is therefore recommended that, where computational resources and time allow, a variety of model parameter configurations should be optimised before being applied as an annotation assistance tool and predictions (detections) are made across new datasets.

The direct comparison between published models is difficult because of the different image datasets, classification systems, model architectures and number of groups used, and the different reporting of metrics. However, where some degree of comparison is possible, the models presented in this study perform seemingly poorly. For example, Cuvelier et al. (2024) present MAIA models each individually trained to detect five groups; including "Alcyonacea" and "Actiniaria/Corallimorpharia"; with recall scores of 0.95 and 0.81 and precision scores of 0.12 and 0.04; respectively. Abad-Uribarren et al. (2022) trained two models (Faster RCNN and RetinaNet architectures) to detect a single coral species; *Dendrophyllia cornigera* (Lamarck, 1816); and utilised augmentation approaches. Evaluation metrics for the FasterRCNN model were Re = 0.84; Pre = 0.86 and F1 = 0.85; whilst metrics for the Retinanet model were Re = 0.98; Pre = 0.40 and F1 = 0.57. The FaunD-Fast model (Mbani et al., 2023) was trained to detect 10 groups of abyssal morphotypes; including "coral"; from images collected from a towed camera platform. Model performance was presented for small; medium and large objects (by pixel size) with average recall and precision scores of 0.16/0.50/0.73 and 0.12/0.42/0.69; respectively. Deo et al. (2024) trained ResNet, DenseNet, Inception, and Inception-ResNet models to detect 33 groups, including "Anthomastus-like coral", "Arborescent coral", "free-living coral", "sea pen", "sea whip coral", and "simple branching coral". The Inception-ResNet model provided a mean precision across groups of 0.65, but no precision or recall scores are provided for each group.

However, these published models were not tested with independent datasets. Instead, these were tested using a subset of the dataset retained during model training or using data from a separate transect, but from the same dataset. Few published studies, e.g., Katija et al. (2022); have tested benthic fauna object detection models using independent datasets. As an applied case study for the FathomNet dataset; Katija et al. (2022) fine-tuned a RetinaNet object detection model to detect 20 "high-level benthic supercategories" (Woodward, 2021); including "sea fan"; "sea pen"; "soft coral"; and "stony coral". This model was tested with an independent dataset. Although evaluation metrics are not reported; Katija et al. (2022) reported the model made many FPs and FNs predictions on the target (test) dataset, "sometimes overwhelming the number of correct annotations of a given class (group)". They conclude this resulted from subtle differences between the training (source) and testing (target) datasets, including ROV, camera angle relative to the seafloor, groups not represented in the training data and high densities of some fauna, particularly the "sea fan" group. Such distribution shifts - changes in data distributions between training (source) and testing (target) datasets - cause poor model performance in independent tests or on transfer because models are not sufficiently generalised.

4.3. Manual evaluation

Manual evaluation of predictions on a subset of each dataset indicated that across all datasets, models detected corals that were previously missed by human annotators, i.e. False Positives that are True Positives. The False Positive Correction Rate (Bridges et al., 2025) was variable across the datasets, from 5.2% to 21.3%, resulting in the undermarking of model performance before correction. For example, in the instance of Val-JC136, after adjusting for the FP Correction Rate (21.3%), recall increased from 0.57 to 0.60, and precision increased from 0.67 to 0.76. However, it is noteworthy that a higher FP Correction Rate is not necessarily an indicator of good model performance. Test-SR17 has an FP Correction Rate of 12.6%, resulting in adjusted recall and precision scores of 0.18 and 0.37, respectively – an example of a relatively high FP Correction Rate, but poor overall model performance. Therefore, given the results of this exercise, where it is appropriate, we recommend that a sample of predictions be manually evaluated to adjust and correct model performance. As a secondary benefit, additional analysis could help identify datasets or taxonomic groups with elevated rates of error and help guide and shape training for image annotators.

4.4. Understanding model performance

Although not specifically tested in this study, the poor performance of models when transferred – irrespective of ROV or geography – will, to some degree, be driven by differences in image-specific factors between research cruises. As previously set out, these factors include lighting and camera setups (e.g. camera resolution and angle, white balance, colour, brightness, distribution of light, etc.), the type of platform (e.g. fixed camera angle), and operational factors (e.g. zoom, distance from the seabed, etc.) – even subtle differences in training and testing datasets can result in a drop in model performance (Katija et al., 2022). However; image augmentation and enhancement have been shown to combat the effects of these domain shifts on model performance (Folkman et al., 2025; Walker et al., 2024), and should be considered in future studies. Additional factors will also affect the model's performance between datasets, for example, in high coral density areas, where annotation becomes more difficult to tell overlapping corals apart.

The decrease in performance across the testing datasets is not uniform, with Test-SR18 performing better than all other test datasets (Table 5). However, the results do not follow any distinct patterns across either geographical location or gear type. Interrogation of results and the underlying datasets suggests that the decrease in performance is likely not just attributed to a change in location or ROV – the morphological classification approach used may not be as generalisable as assumed, and poor testing performance is inherent of this.

In this study, black and octocorals were classified morphologically. Although coral groups occur across training and testing datasets at a morphological level, some corals are not represented (or not as abundant) in both training and testing datasets at a taxonomic level and,

therefore, represent a distribution shift (Katija et al., 2022). For example; the "Unbranched" morphology group is dominated (71%) by the black coral *Stichopathes gravieri* (Molodtsova, 2006; Fig. 7) – a coiled, thick, orange coral – in the JC136 dataset. Comparatively, Test-JC142 has no occurrences of *S. gravieri* but includes other morphospecies of *Stichopathes msp. 1* (73%; inconspicuous, thin, pale, and uncoiled; Fig. 7) and bamboo corals – taxa that either do not occur or are not well represented in the training dataset. The same also occurs across the Test-SR17/18/19 datasets, which is dominated by *Stichopathes msp. 2* (39–70%), which is thin, uncoiled and red or orange (Fig. 7). This suggests that, at least in this study, the "Unbranched" morphological group is not inherently generalisable and that datasets also need to maintain some taxonomic parity. This may also apply to other morphological groups where many taxa occur within it, such as "Branching 3D", but this was not formally tested in this study.

Variability in taxonomic occurrences between datasets will be driven by location as a function of species range and suitable habitat, but also by project or survey objectives and aims. Typically, research cruises and surveys are time and resource limited. This results in strict criteria when developing survey plans, often leading to the targeting of habitats and geomorphological features. As a result, the datasets collected are biased towards the target feature(s) and the biological communities that characterise or inhabit it, leading to group (class) imbalances once annotated as well as variable image 'background'. This will place limitations on how generalisable object detection models can be when trained using typical datasets collected in deep-sea ecological research. For example, the Train-JC136 dataset is dominated by coral taxa found on hard or coral rubble substrates such as antipatharians (black corals), compared to Test-SR19, which is mostly characterised by coral taxa found on soft substrates such as sea pens.

4.4.1. Performance by morphological groups

Group performance by recall and precision in the Val-JC136 dataset variable (0.54 to 0.71 and 0.63 to 0.81, respectively). Performance across all morphological groups is reduced when models are applied to testing datasets and are considerably more variable, with recall varying from 0.00 to 0.53 and precision varying from 0.00 to 0.67. Across Test-SR17/18/19 datasets, Fan-2D consistently ranked the best group for recall (0.38 to 0.53) but ranked poorly for precision (0.08 to 0.20).

Group (class) imbalance, the disproportionate ratio of instances across groups, exists in the Train-JC136 dataset; the smallest training group is Mushroom (n = 138), and the largest training group is Branching-3D (n = 1721). This is representative of real-world ecological datasets where natural variation in abundance occurs. This is due to the inherent rarity of some taxa compared with others, but also because of the often limited and targeted nature of deep-sea research cruises, as previously discussed. Group (class) imbalance can affect model performance "detrimentally" (Buda et al., 2018) as models can become biased towards the larger; majority groups; reducing model performance and generalisation. Different methods can address group (class) imbalance and be split into two groups: data-level and classifier-level. Data-level methods involve the re-balancing of the groups in the dataset by over-sampling or under-sampling. Classifier-level methods require adjustments to model parameters; such as altering the weights of misclassifications for different groups or applying different thresholding techniques (Buda et al., 2018; Johnson and Khoshgoftaar, 2019).

A clear limitation of this study is that there have been no attempts to address the group (class) imbalance that occurs within the training dataset and future studies should consider this. However, the purpose of this study is to determine what is achievable with datasets typical of those acquired within a research cruise or survey. As a result, it would be expected that model performance would be biased to larger groups, such as Branching-3D. However, Val-JC136 results show the Mushroom group (n = 138) has the highest recall score (0.71) compared to other morphological groups, outperforming Branching-3D (n = 1721) in recall score (0.54). This relationship is the same when tested using the Test-JC142 dataset, however, Branching-3D outperforms Mushroom when tested using the Test-SR datasets. The training group size of Bottlebrush (n = 149) is similar to Mushroom (n = 138), but has a lower validation recall score (0.54) of any group and a mixed performance across all the test datasets.

These varying performances imply that the results are more complex than just the effect of group (class) imbalance. In the instance of Mushroom and Branching-3D performance, disparities are likely attributed to the morphological variation found within the annotations of each of those groups. In this study, the Mushroom group comprises individual corals of similar structural morphology and colour, with inter-group morphological variance coming mostly from whether individual Mushroom corals have their polyps extended or retracted. Comparatively, there is significant morphological variability within the Branching-3D group and it contains many different taxa (Fig. 2, Row C). Therefore, although it is a much larger group size, there is much more morphological variance within the group to be learned during model training compared to Mushroom, which is a more homogenous group.

This morphological variety also occurs in the Fan-2D group, with the morphological group characterised by numerous different taxa. However, morphology between Branching-3D and Fan-2D are less distinct compared to other morphological coral groups, with instances of the same taxa, e.g. *Leiopathes* sp., occurring across both groups because of relatively subtle differences in their morphology. This lesser distinction between the two groups is apparent when observing confusion matrices, where four of the five datasets have 10–21% confusion between the two.

4.5. Next steps: model applications and improving performance

Despite the relatively poor performance of the models, they may still be useful tools in some applications. For example, as a foundation model to act as a first pass of newly acquired image data. In this instance, the models could be used to make a first pass to quickly obtain a small subset of annotations from the new data. The new annotations could then be used as a transfer learning dataset to update and optimise the model, potentially improve performance and then reapply across the new dataset. A second potential use is group-specific. For example, in some transfer scenarios, the models in this study still performed moderately for some classes, such as "Fan 2D" in the Test-SR17/18 and "Mushroom" in the Test-JC142 dataset. In these instances, Recall scores of 50% + were obtained. If applied to a newly acquired dataset, even moderate Recall performances of 50% results in halving the number of corals that require manual annotation. However, improving model performance is still needed.

Increasing the initial performance of the trained model when validated requires improving the quality of the model training data, particularly the degrees of heterogeneity within coral groups. Where group sizes are large and comprise morphological variability and confusion, i.e. Branching-3D and Fan—2D, two potential solutions for improving model performance concerning the datasets include,

1. *Re*-classify annotations to additional, nested morphological groups to create more homogenous groups. However, this would also increase the number of overall groups, which can reduce model performance (Piechaud et al., 2019) and the number of instances per group to train a model on; or,
2. Aggregate groups together to create a larger group, e.g., 'Coral' or 'Animal', to create a single object detection model. This has been shown in other studies to reduce annotation complexity without substantially compromising model performance (Bridges et al., 2025).

Overcoming these group performance issues is dataset-specific, but also dependent on the application of the model (Bridges et al., 2025). For assisting the annotation process, the decision depends on whether the human effort should be front-loaded, taking more time to create

more homogenous and numerous groups, or whether a training set of fewer groups should be quickly created (or consolidated to one, e.g. 'Coral'), with more human effort afterwards to clean predictions and further resolve annotations to lower morphological or taxonomic groups.

The more considerable room for improvement in this study is the overall performance of the models when transferred and tested by datasets collected with a different ROV (Test-SR) or in a different location (Test-JC142). As discussed, a fundamental challenge in tackling this is the limitations, biases and differences between the training and testing datasets. In addition to trialling whether training performance improves with different levels of classification, i.e. single-group or additional resolved groups, the generalisation of the models with varying levels of classification should also be tested against new locations and imaging platforms. The datasets used in this study could also be combined to create a larger training dataset that is more representative, but this could increase morphological variability within groups if they are not resolved further. Additionally, a future study could investigate how further transfer learning could improve model performance. For example, an interesting question would be, how much additional data is required from the testing datasets in transfer learning to see an improvement in model performance? The fast-paced nature of this area of research means that since this study's inception, many newer DL algorithms, e.g. YOLOv11, have come online. Future studies should also test if these newer algorithms, by nature of their improved architectures, inherently perform transfers and handle domain shifts better.

However, taking a 'go it alone' approach to tackling this challenge will likely only take developments so far, given the image-bottleneck issue is universal, and the urgent need to process image-based data comes from pressures at a local to an ocean basin scale. For object detection models to be transferable enough to address the scale of this challenge, they require training datasets that are truly representative at an ocean basin or global scale, not just locally or regionally. Developing representative models will require the collating of imagery datasets and the standardisation of image annotations between datasets. Existing projects such as SMarTaR-ID (Howell et al., 2019); CATAMI (Althaus et al., 2015) and FathomNet (Katija et al., 2022) represent community-led efforts to address these challenges, such as establishing standardised marine image databases, unified species catalogues and classification schemes, and open-sourced image databases for training and testing object detection models.

Furthermore, international groups such as the Challenger 150 Megafaunal Image-based Technical Working Group (www.challenger150.world) are taking coordinated steps to make accepted standards and quality control methods of image annotation, including supporting the use of automation using AI. Actions from this working group include training in the use of annotation software and identification, development of training materials, the sharing and further development of reference identification guides and libraries, standardisation workshops, establishing networks of annotators and taxonomists, and coordinating funding applications.

5. Conclusion

The object detection models trained in this study to detect and classify black and octocoral morphologies performed moderately. However, when models were transferred geographically and between ROVs, performance was poor. There was no discernible difference in model performance based on spatial or image platform transfer. Interrogation of the results suggests that, in addition to the effects of model transfer, model performance was poor because of high morphological and taxonomic variability within the "Unbranched" coral group and that the training dataset (source domain) was not representative of the testing datasets (target domain). While this was not tested across all morphological groups, this should be considered when developing training-testing datasets and transferring models.

To tackle the image analysis bottleneck, a community effort will be required to develop representative training datasets at an ocean basin scale. With representative training datasets, the community will then be able to test which coral classification approach (e.g. morphological or taxonomic) and level will produce the most transferable models when applied to unanalysed existing or newly collected image-based data. This will reduce the human and time resources needed to annotate image datasets, allowing scientists to extract biodiversity information from images as effectively as possible and address the conservation and management needs of the day.

Acknowledgements & Funding

We would like to thank Charlie Keeney, an undergraduate student at the University of Plymouth, for contributing to the image analysis. We would also like to thank the scientists, officers, and crews of all the research cruises that contributed to the collection of imagery data used in this study. The research cruise JC136 was funded by the UK Natural Environment Research Council, grant number NE/K011855/1 – DeepLinks project. The research cruise JC142 was funded by the UK Natural Environment Research Council (NERC) MarineE-Tech project, grant number NE/MO1151/1, awarded to Bramley Murton, National Oceanography Centre (NOC). Imagery data from the SeaRover programme acquired offshore Ireland during 2017, 2018 and 2019 were kindly made available by the Government of Ireland in support of this research. The Sensitive Ecosystem Assessment and ROV Exploration of Reef (SeaRover) was commissioned by the Marine Institute in partnership with the National Parks and Wildlife Service (NPWS) and funded by the European Maritime and Fisheries Fund (EMFF), Department of Agriculture, Food and the Marine (DAFM) and NPWS. The project was coordinated by the Department of Environment, Climate & Communications, funded INFOMAR programme team, with research support from the University of Galway, Plymouth University, and the Institute of Marine Research, Norway. INFOMAR is jointly managed by the Marine Institute & Geological Survey Ireland. This work was supported and funded by the Natural Environment Research Council, the ARIES Doctoral Training Partnership and the University of Plymouth; grant number NE/S007334/1.

CRediT authorship contribution statement

Kyran P. Graves: Writing – original draft, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Louise Allcock:** Writing – review & editing, Supervision, Funding acquisition. **David K. A. Barnes:** Writing – review & editing, Supervision, Funding acquisition. **Amelia E.H. Bridges:** Writing – review & editing, Data curation. **Kerry L. Howell:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

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.

Appendix A. Appendices

A.1. False positive correction rate

For each dataset, 50 images were randomly selected for manual evaluation to calculate the False Positive Correction Rate. For each subset image, the accompanying YOLO predictions (from the best-performing model) and human annotations were compiled.

Images, human annotation and YOLO predictions were uploaded to BIIGLE. A new volume was created for each dataset, e.g. Test-JC142. YOLO predictions and human annotations were uploaded onto two separate but duplicated SMarTaR-ID Morphology label trees. This allowed for the model and human annotations to have distinct coloured annotations, making it easier to tell the two sets of annotations apart.

A new evaluation label tree was created for evaluating the YOLO predictions to denote false negatives (FN), false positives (FP) and true positives (TP).

- FN_record = Corals missed by the model
- FP_but_detected = Coral detected but incorrectly classified
- FP_duplicate = Coral detected and classified more than once
- FP_nothing_not_coral = Detection made, but it is not a coral
- TP_legit = Correctly detected and classified coral
- TP_missed = Correctly detected and classified coral that a human annotator missed

To evaluate, each image was cycled through, and every YOLO annotation was assigned an additional label from the evaluation label tree in accordance with its performance, e.g. TP_legit. Where the model missed corals identified by human annotators, they were marked with an "FN_record" annotation. See Fig. A.1. for an example.

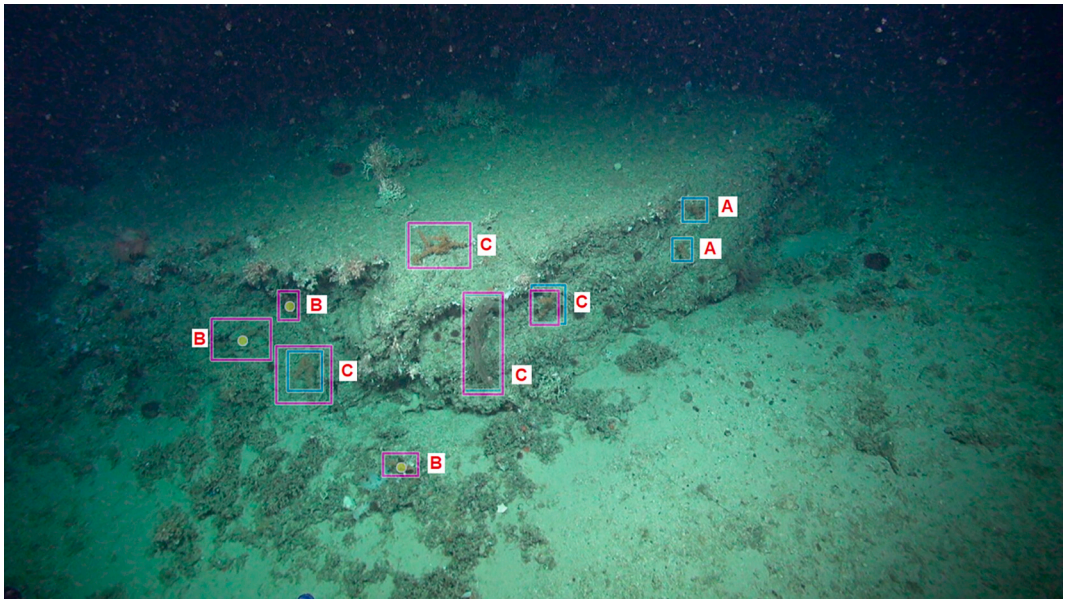


Fig. A.1. Example image from Val-JC136 after manual evaluation. Annotations marked [A] are an example of a corrected FP, i.e. the model correctly detected a coral that was missed by human annotators; [B] an example of a FN, i.e. the model did not predict a coral annotated by human annotators; and [C] an example of a TP, i.e. the model correctly detected and labelled a coral.

Once all subset images were evaluated, BIIGLE reports were generated for the evaluation label tree, giving a breakdown of the number of FNs, FPs and TPs per image. From these annotations, evaluation metrics (Recall, Precision, F1) were calculated.

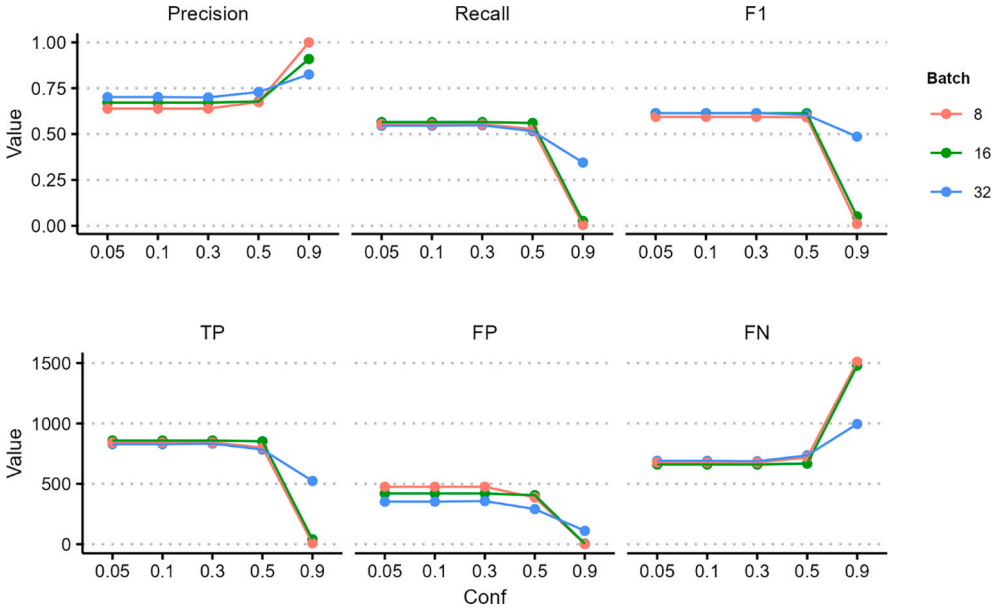
For each dataset, the number of corals that were correctly detected and labelled by the models were calculated, i.e. TP_missed. In the initial evaluations, these corals would have been determined FPs because they were not accompanied by a ground-truthed, human annotation. The proportion of the original FPs that had been re-evaluated as TPs, or the FP Correction Rate, was calculated for each of the image datasets by.

$$\text{FP Correction Rate} = (\text{TP\_missed}/(\text{TP\_missed} + \text{All FPs})).$$

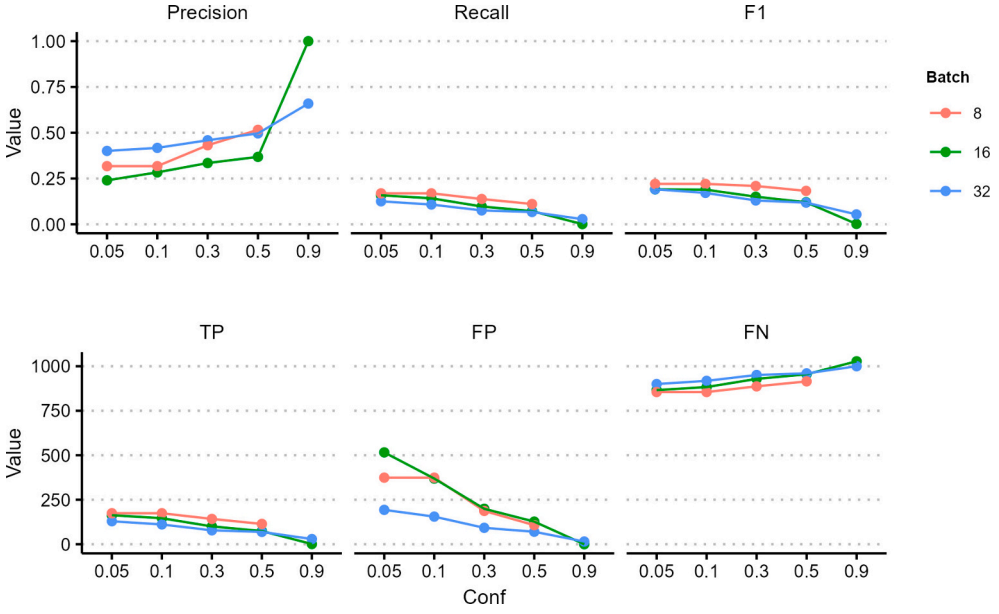
The correction rate calculated from each subset of data was then extrapolated across the complete datasets to generate estimated evaluation metrics. Two main limitations of this are that (1) it is assumed that the correction rate in the subset of images is the same across the whole dataset, and (2) the subsampling of images was not stratified by morphology, meaning that a correction rate was not calculated for each morphological group.

A.2. All Model Results.

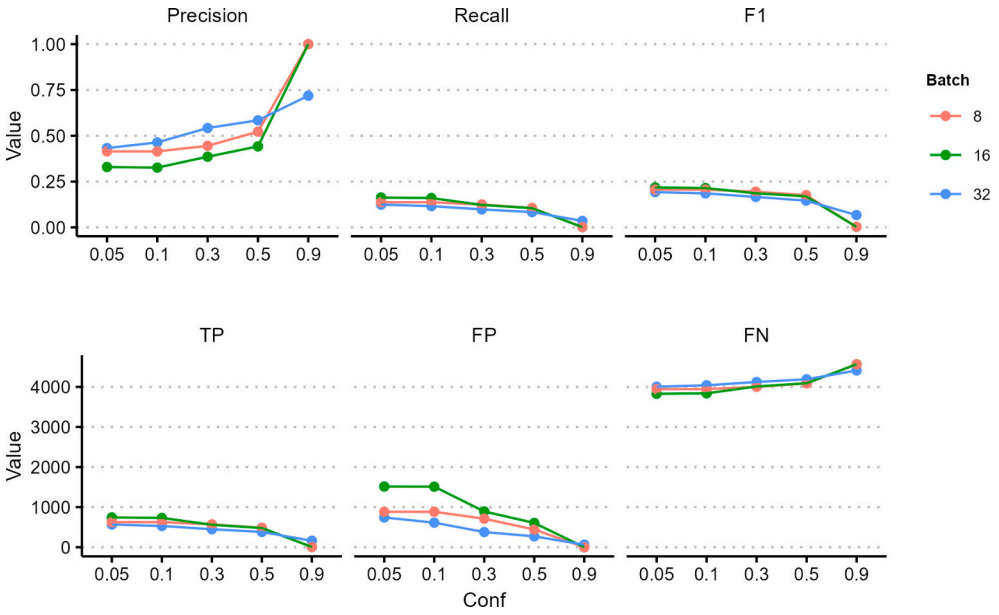
A.2.1. Val-JC136



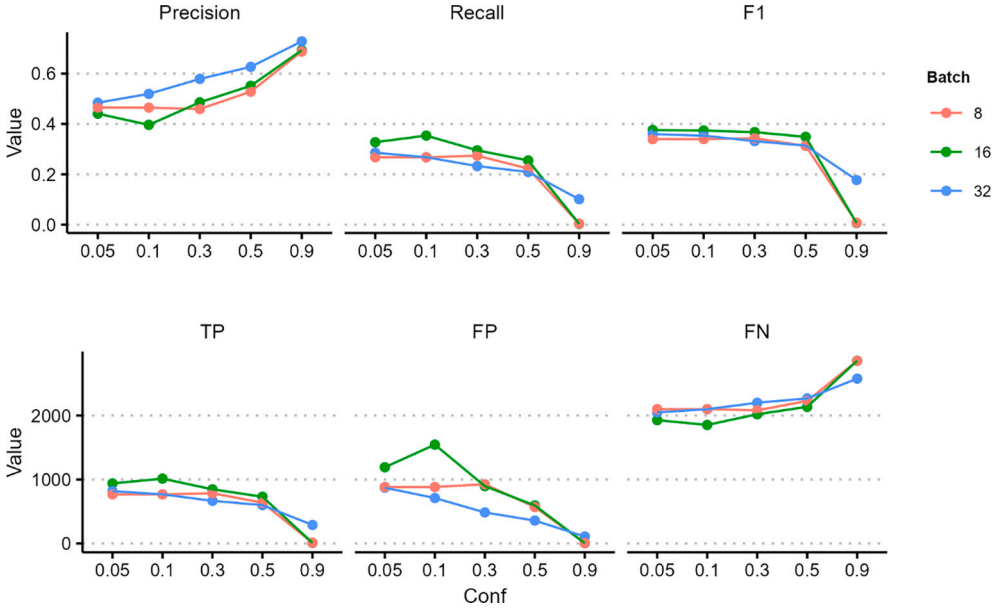
A.2.2. Test-JC142.



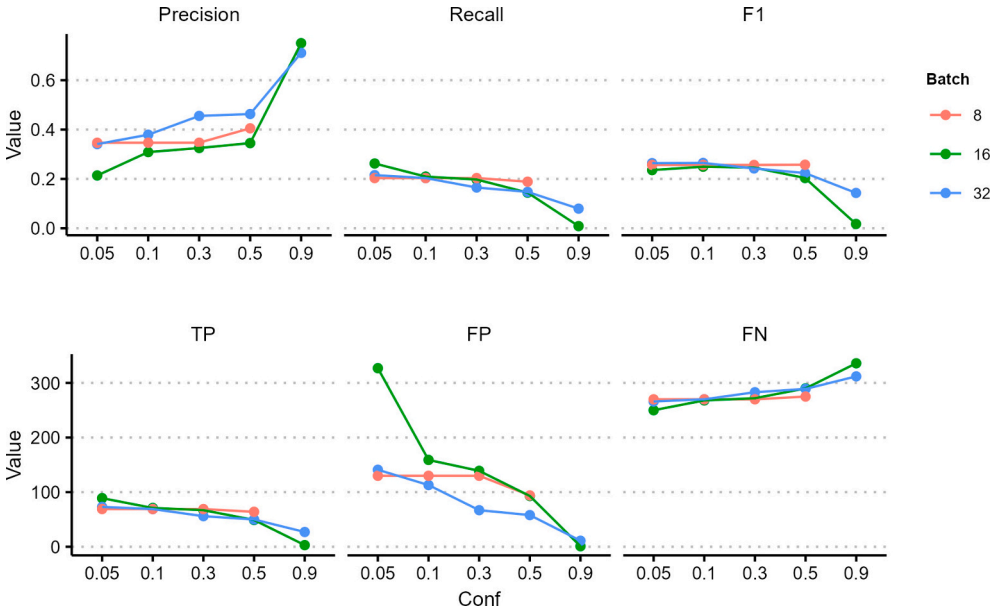
A.2.3. Test-SR17.



A.2.4. Test-SR18.

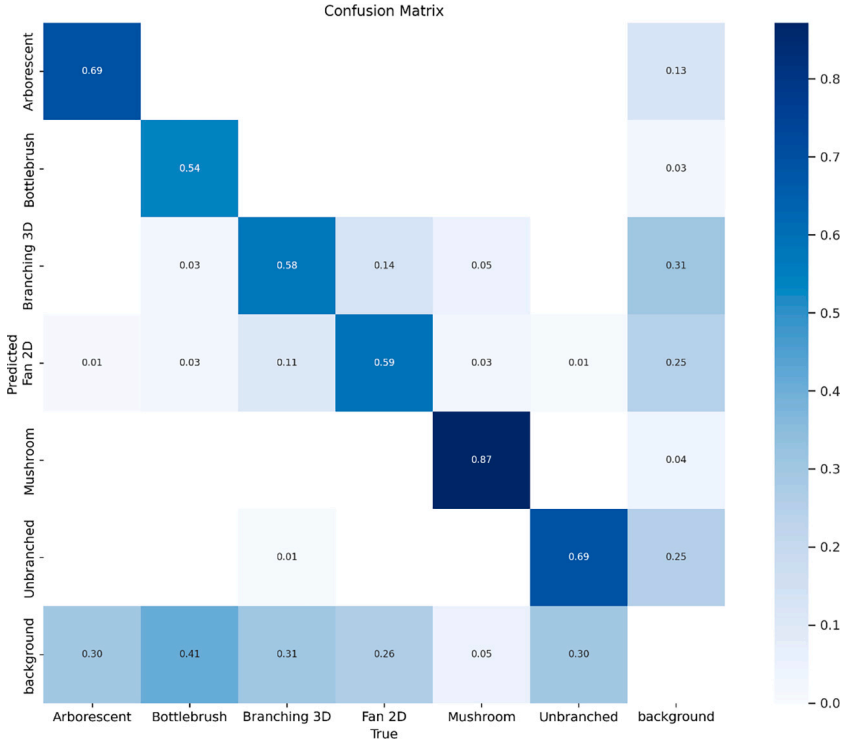


A.2.5. Test-SR19.

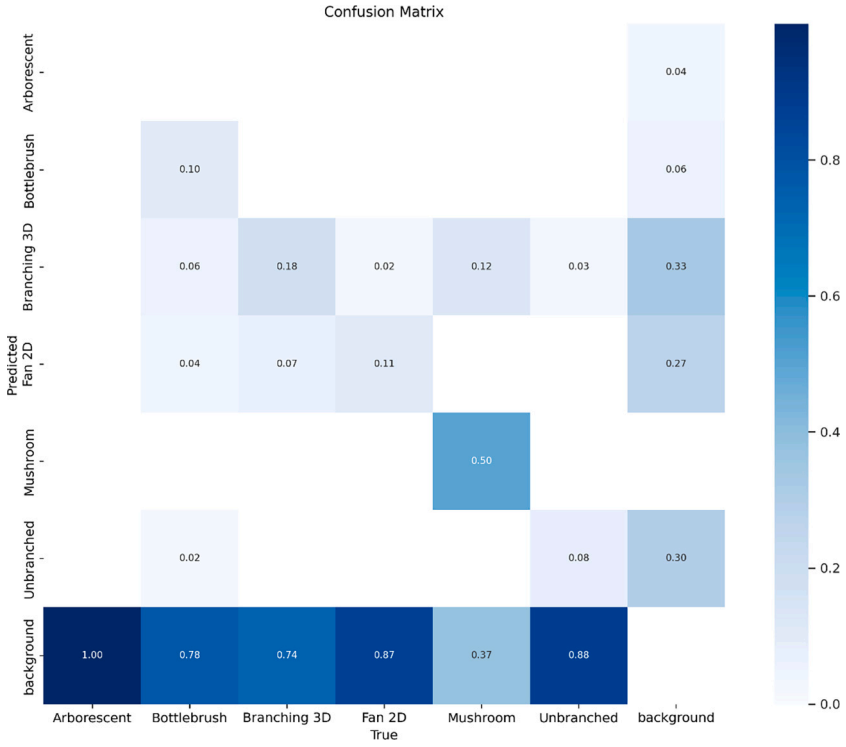


A.3. Confusion Matrices.

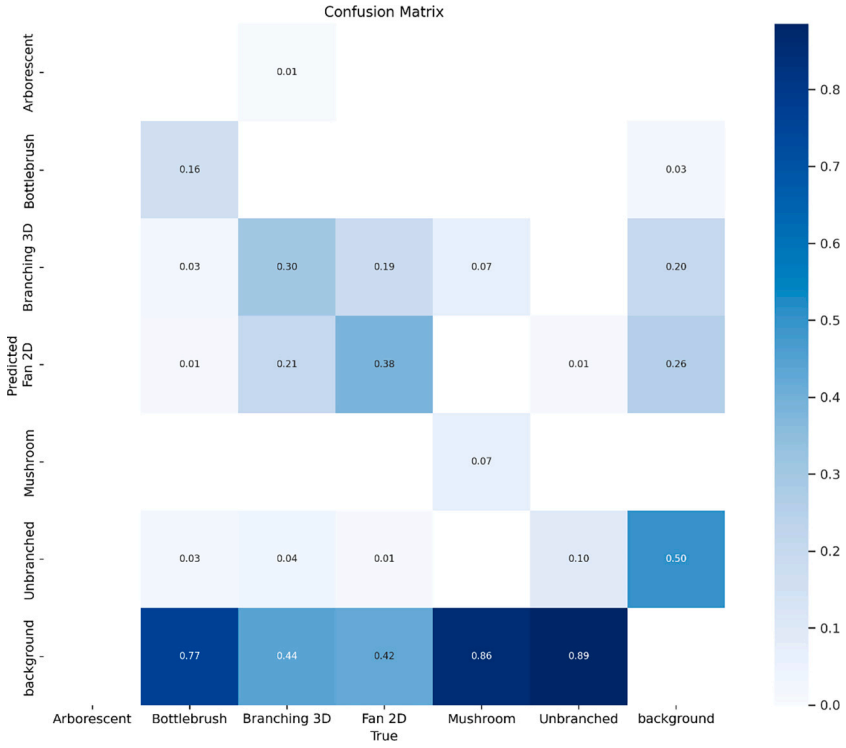
A.3.1. Val-JC136 Confusion Matrix.



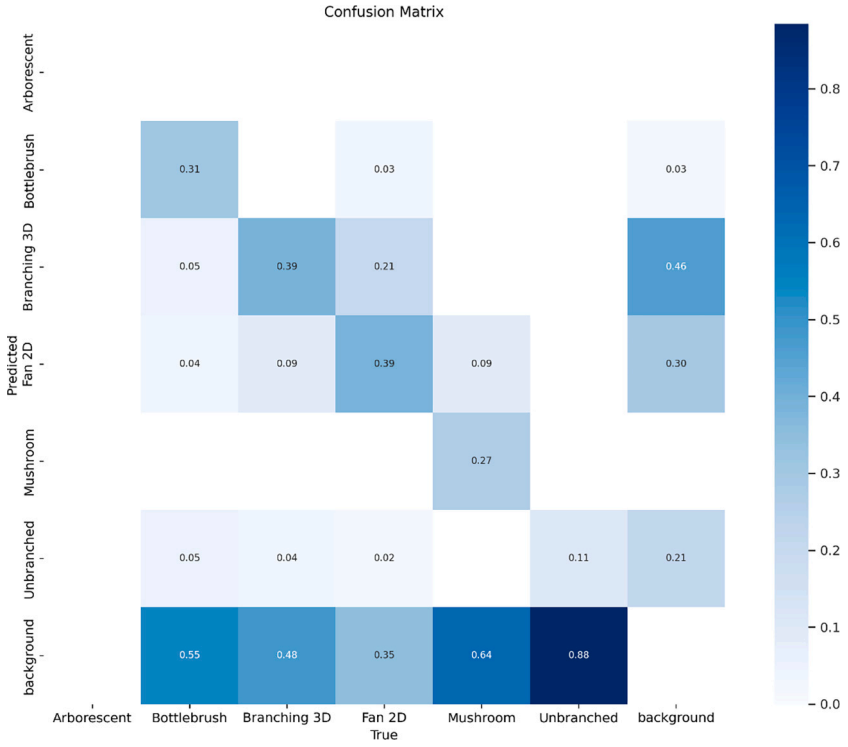
A.3.2. Test-JC142.



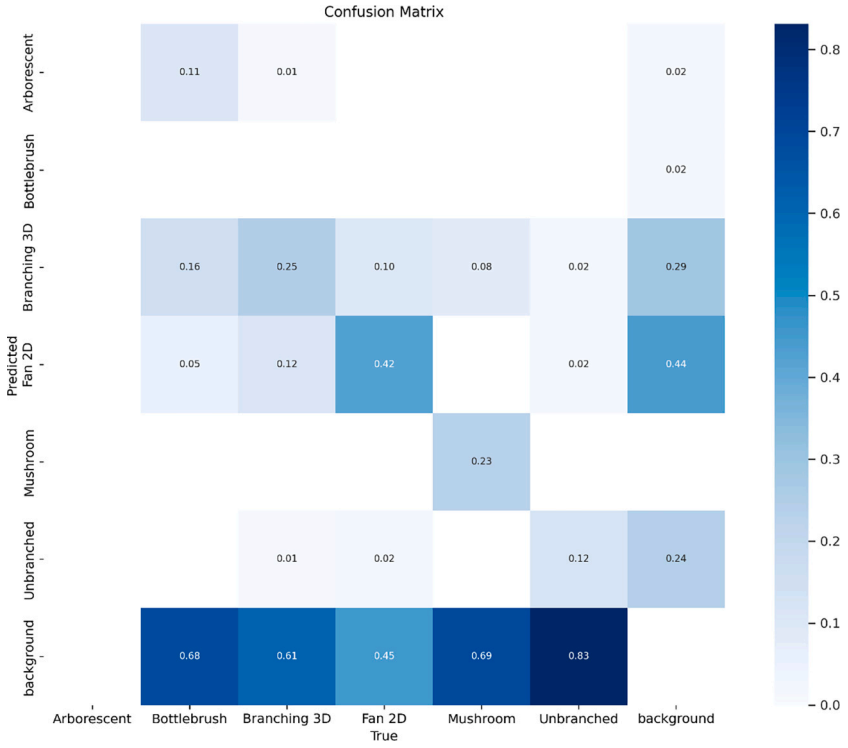
A.3.3. Test-SR17



A.3.4. Test-SR18.



A.3.5. Test-SR19.



Data availability

Code and links to image data and annotations are available and stored in a GitHub repository (https://github.com/kyrangraves/Graves_et_al_2026_AI_Coral_Morphology).

References

Abad-Uribarren, A., Prado, E., Sierra, S., Cobo, A., Rodríguez-Basalo, A., Gómez-Ballesteros, M., Sánchez, F., 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.

Althaus, F., Hill, N., Ferrari, R., Edwards, L., Przeslawski, R., Schönberg, C.H.L., Stuart-Smith, R., Barrett, N., Edgar, G., Colquhoun, J., Tran, M., Jordan, A., Rees, T., Gowlett-Holmes, K., 2015. A standardised vocabulary for identifying benthic biota and substrata from underwater imagery: the CATAMI classification scheme. PLoS One 10, e0141039. https://doi.org/10.1371/journal.pone.0141039.

Álvarez-Ellacuría, A., Palmer, M., Catalán, I.A., Lisani, J.-L., 2020. Image-based, unsupervised estimation of fish size from commercial landings using deep learning. ICES J. Mar. Sci. 77, 1330–1339. https://doi.org/10.1093/icesjms/fsz216.

Beijbom, O., Edmunds, P.J., Roelfsema, C., Smith, J., Kline, D.I., Neal, B.P., Dunlap, M.J., Moriarty, V., Fan, T.-Y., Tan, C.-J., Chan, S., Treibitz, T., Gamst, A., Mitchell, B.G., Kriegman, D., 2015. Towards automated annotation of benthic survey images: variability of human experts and operational modes of automation. PLoS One 10, e0130312. https://doi.org/10.1371/journal.pone.0130312.

Brady, H.B., 1883. IV. Note on syringammina, a new type of arenaceous rhizopoda. Proc. R. Soc. Lond. 35, 155–161. https://doi.org/10.1098/rspl.1883.0031.

Bridges, A.E.H., Cross, E., Graves, K.P., Piechaud, N., Raymont, A., Howell, K.L., 2025. Practical application of artificial intelligence for ecological image analysis: Trialling different levels of taxonomic classification to promote convolutional neural network performance. Eco. Inform. 88, 103146. https://doi.org/10.1016/j.ecoinf.2025.103146.

Browne, E., 2022. A step towards automating real-time collection of ecological data on observational platforms: a pilot study on deep-sea benthic species syringammina fragilissima. Thesis. University of Plymouth. https://doi.org/10.24382/818.

Buda, M., Maki, A., Mazurowski, M.A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011.

Cuvelier, D., Zurowietz, M., Nattkemper, T.W., 2024. Deep learning–assisted biodiversity assessment in deep-sea benthic megafauna communities: a case study in the context of polymetallic nodule mining. Front. Mar. Sci. 11. https://doi.org/10.3389/fmars.2024.1366078.

Deo, R., John, C.M., Zhang, C., Whitton, K., Salles, T., Webster, J.M., Chandra, R., 2024. Deepdive: leveraging pre-trained deep learning for Deep-Sea ROV biota identification in the great barrier reef. Sci. Data 11, 957. https://doi.org/10.1038/s41597-024-03766-3.

Folkman, L., Pitt, K.A., Stantic, B., 2025. A data-centric framework for combating domain shift in underwater object detection with image enhancement. Appl. Intell. 55, 272. https://doi.org/10.1007/s10489-024-06224-0.

Game, C.A., Thompson, M.B., Finlayson, G.D., 2024. Machine learning for non-experts: a more accessible and simpler approach to automatic benthic habitat classification. Eco. Inform. 81, 102619. https://doi.org/10.1016/j.ecoinf.2024.102619.

Hou, C., Guan, Z., Guo, Z., Zhou, S., Lin, M., 2023. An improved YOLOv5s-based scheme for target detection in a complex underwater environment. J. Mar. Sci. Eng. 11, 1041. https://doi.org/10.3390/jmse11051041.

Howell, K.L., Taylor, M., Crombie, K., Faithfull, S., Golding, N., Nimmo-Smith, W.A., Perrett, J., Piechaud, N., Ross, R.E., Stashchuk, N., Turner, D., Vlasenko, V., Foster, N.L., 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. Plymouth University Marine Institute.

Howell, K.L., Davies, J.S., Allcock, A.L., Braga-Henriques, A., Buhl-Mortensen, P., Carreiro-Silva, M., Dominguez-Carrió, C., Durden, J.M., Foster, N.L., Game, C.A., Hitchin, B., Horton, T., Hosking, B., Jones, D.O.B., Mah, C., Marchais, C.L., Menot, L., Morato, T., Pearman, T.R.R., Piechaud, N., Ross, R.E., Ruhl, H.A., Saeedi, H., Stefanoudis, P.V., Taranto, G.H., Thompson, M.B., Taylor, J.R., Tyler, P., Vad, J., Victorero, L., Vieira, R.P., Woodall, L.C., Xavier, J.R., Wagner, D., 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, e0218904. https://doi.org/10.1371/journal.pone.0218904.

Howell, K.L., Bridges, A.E.H., Davies, J., Parimbelli, A., Piechaud, N., 2023. An ecologist's guide to BIIGLE. https://doi.org/10.5281/zenodo.7728927.

Iyer, K.H., Marnor, C.M., Schmid, D.W., Hartz, E.H., 2025. Detecting and quantifying deep sea benthic life using advanced object detection. Front. Mar. Sci. 11. https://doi.org/10.3389/fmars.2024.1470424.

Jackett, C., Althaus, F., Maguire, K., Farazi, M., Scoulding, B., Untiedt, C., Ryan, T., Shanks, P., Brodie, P., Williams, A., 2023. A benthic substrate classification method for seabed images using deep learning: application to management of deep-sea coral reefs. J. Appl. Ecol. 60, 1254–1273. https://doi.org/10.1111/1365-2664.14408.

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., Michael, K., TaoXie, Fang, J., imyhxy, Lorna, Yifu), 曾逸夫(Zeng, Wong, C., V, A., Montes, D., Wang, Z., Fati, C., Nadar, J., Laughing, UnglvKitDe, Sonck, V., tkianai, yxNONG, Skalski, P., Hogan, A., Nair, D., Strobel, M., Jain, M, 2022. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation. https://doi.org/10.5281/zenodo.7347926.

Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J. Big Data 6, 27. https://doi.org/10.1186/s40537-019-0192-5.

Katija, K., Orenstein, E., Schlining, B., Lundsten, L., Barnard, K., Sainz, G., Boulais, O., Cromwell, M., Butler, E., Woodward, B., Bell, K.L.C., 2022. FathomNet: a global image database for enabling artificial intelligence in the ocean. Sci. Rep. 12, 15914. https://doi.org/10.1038/s41598-022-19939-2.

Lamarck, J.-B., 1816. Histoire naturelle des animaux sans vertèbres. Verdière, Paris. https://doi.org/10.5962/bhl.title.12712.

Langenkämper, D., Zurowietz, M., Schoening, T., Nattkemper, T.W., 2017. Biigle 2.0 - browsing and annotating large marine image collections. Front. Mar. Sci. 4.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P., 2015. Microsoft COCO: common objects in context. https://doi.org/10.48550/arXiv.1405.0312.

Marini, S., Fanelli, E., Sbragaglia, V., Azzurro, E., Del Rio Fernandez, J., Aguzzi, J., 2018. Tracking fish abundance by underwater image recognition. Sci. Rep. 8, 13748. https://doi.org/10.1038/s41598-018-32089-8.

Mbani, B., Buck, V., Greinert, J., 2023. An automated image-based workflow for detecting megabenthic fauna in optical images with examples from the Clarion–Clipperton Zone. Sci. Rep. 13, 8350. https://doi.org/10.1038/s41598-023-35518-5.

Molodtsova, T.N., 2006. Black Corals (Antipatharia: Anthozoa: Cnidaria) of North-East Atlantic, in: Biogeography of the North Atlantic Seamounts.

Murton, B.J., 2016. JC142 Cruise Report. MarineE-Tech Project: To Map the Cobalt-Rich Ferromanganese Crusts of Tropic Seamount, NE Atlantic Ocean. National Oceanography Centre.

Osterloff, J., Nilssen, I., Järnegren, J., Van Engeland, T., Buhl-Mortensen, P., Nattkemper, T.W., 2019. Computer vision enables short- and long-term analysis of lophelia Pertusa polyp behaviour and colour from an underwater observatory. Sci. Rep. 9, 6578. https://doi.org/10.1038/s41598-019-41275-1.

O'Sullivan, D., Leahy, Y., Guinan, J., Party, S.S, 2017. Sensitive Ecosystem Assessment and ROV Exploration of Reef Survey Report 2017.

O'Sullivan, D., Leahy, Y., Healy, L., Party, S.S, 2018. EMFF Offshore Reef Survey 'SeaRover' Cruise Report 2018.

O'Sullivan, D., Healy, L., Leahy, Y., Party, S.S, 2019. EMFF Offshore Reef Survey 'SeaRover' Cruise Report 2019.

Piechaud, N., Howell, K.L., 2022. Fast and accurate mapping of fine scale abundance of a VME in the deep sea with computer vision. Eco. Inform. 71, 101786. https://doi.org/10.1016/j.ecoinf.2022.101786.

Piechaud, N., Hunt, C., Culverhouse, P.F., Foster, N.L., Howell, K.L., 2019. Automated identification of benthic epifauna with computer vision. Mar. Ecol. Prog. Ser. 615, 15–30. https://doi.org/10.3354/meps12925.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rainio, O., Teuho, J., Klén, R., 2024. Evaluation metrics and statistical tests for machine learning. Sci. Rep. 14, 6086. https://doi.org/10.1038/s41598-024-56706-x.

Rubbens, P., Brodie, S., Cordier, T., Destro Barcellos, D., Devos, P., Fernandes-Salvador, J.A., Fincham, J.I., Gomes, A., Handegard, N.O., Howell, K., Jamet, C., Kartveit, K.H., Moustahfid, H., Parcerisas, C., Politikos, D., Sauzède, R., Sokolova, M., Uusitalo, L., Van den Bulcke, L., van Helmond, A.T.M., Watson, J.T., Welch, H., Beltran-Perez, O., Chaffron, S., Greenberg, D.S., Kühn, B., Kiko, R., Lo, M., Lopes, R.M., Möller, K.O., Michaels, W., Pala, A., Romagnan, J.-B., Schuchert, P., Seydi, V., Villasante, S., Malde, K., Irisson, J.-O., 2023. Machine learning in marine ecology: an overview of techniques and applications. ICES J. Mar. Sci. 80, 1829–1853. https://doi.org/10.1093/icesjms/fsad100.

Walker, J.L., Zeng, Z., Wu, C.L., Jaffe, J.S., Frasier, K.E., Sandin, S.S., 2024. Underwater object detection under domain shift. IEEE J. Ocean. Eng. 49, 1209–1219. https://doi.org/10.1109/JOE.2024.3425453.

Wang, W., Sun, Y.F., Gao, W., Xu, W., Zhang, Y., Huang, D., 2024. Quantitative detection algorithm for deep-sea megabenthic organisms based on improved YOLOv5. Front. Mar. Sci. 11.

Woodward, Lundsten Orenstein, 2021. MBARI Benthic Supercategory Object Detector. https://doi.org/10.5281/zenodo.5571043.

Zhang, J., Yongpan, W., Xianchong, X., Yong, L., Lyu, L., Wu, Q., 2022. YoloXT: a object detection algorithm for marine benthos. Eco. Inform. 72, 101923. https://doi.org/10.1016/j.ecoinf.2022.101923.

Zhang, L., Fan, J., Qiu, Y., Jiang, Z., Hu, Q., Xing, B., Xu, J., 2024. Marine zoobenthos recognition algorithm based on improved lightweight YOLOv5. Eco. Inform. 80, 102467. https://doi.org/10.1016/j.ecoinf.2024.102467.