

# Automated Detection of Antarctic Benthic Organisms in High-Resolution In Situ Imagery to Aid Biodiversity Monitoring

Cameron Trotter<sup>†\*</sup> Huw Griffiths<sup>†</sup> Tasnuva Ming Khan<sup>†‡</sup> Rowan Whittle<sup>†</sup>

<sup>†</sup>British Antarctic Survey

cater@bas.ac.uk\* h.j.g@bas.ac.uk

<sup>‡</sup>University of Cambridge

t.f.m.k2@cam.ac.uk roit@bas.ac.uk

## Abstract

*Monitoring benthic biodiversity in Antarctica is vital for understanding ecological change in response to climate-driven pressures. This work is typically performed using high-resolution imagery captured in situ, though manual annotation of such data remains laborious and specialised, impeding large-scale analysis. We present a tailored object detection framework for identifying and classifying Antarctic benthic organisms in high-resolution towed camera imagery, alongside the first public computer vision dataset for benthic biodiversity monitoring in the Weddell Sea. Our approach addresses key challenges associated with marine ecological imagery, including limited annotated data, variable object sizes, and complex seafloor structure. The proposed framework combines resolution-preserving patching, spatial data augmentation, fine-tuning, and postprocessing via Slicing Aided Hyper Inference. We benchmark multiple object detection architectures and demonstrate strong performance in detecting medium and large organisms across 25 fine-grained morphotypes, significantly more than other works in this area. Detection of small and rare taxa remains a challenge, reflecting limitations in current detection architectures. Our framework provides a scalable foundation for future machine-assisted in situ benthic biodiversity monitoring research.*

## 1. Introduction

Benthic communities, comprised of organisms that live in, on or around the seafloor, are highly biodiverse, play key roles within global nutrient cycling, and are a valuable food source [25]. Global anthropogenic change, e.g., ocean warming and acidification [44, 47], coupled with direct local and regional pressures such as harvesting and pollution, are negatively impacting the structure and function of benthic communities [11].

The Antarctic benthos is uniquely adapted to its isolated and frozen environment [2]. These cold-adapted species

face additional pressures through changes to the cryosphere that dominates their ocean, e.g., glacial melt and ice shelf collapse. These changes are most notable in the shallow benthic communities of the West Antarctic Peninsula, where changes to biodiversity, trophic structure, biomass, and distribution have been observed [19].

Historically, the exploration and monitoring of benthic environments has relied on invasive, non-quantitative methods such as dredging, or more quantitative yet slow-to-deploy instruments like corers and grabs [45]. In recent years, the adoption of imaging technologies, delivered via SCUBA, submersibles, towed or drop camera systems, remotely operated vehicles, and autonomous platforms, has significantly increased both the rate and scale of data acquisition. Photographic and video data enable rapid, in situ, and quantitative surveys of extensive seafloor areas.

Imaging techniques represent a non-destructive and repeatable survey method to monitor ecosystem change. To date, the usefulness of collected data has been restricted by the need for expert assessment of every image, which is time consuming [3, 52] and prone to fatigue and annotation bias [12, 15, 41]. This bottleneck is particularly evident for Antarctica, with highly diverse and endemic benthic species [2] and relatively few taxonomic experts capable of providing confident image-based identifications.

Additionally, the high logistical and financial costs associated with deep-sea data collection often results in comparatively small amounts of collected data. Antarctica's geographic isolation and extreme environmental conditions make fieldwork highly resource-intensive, limiting collection to infrequent, short-duration missions typically led by national research programmes.

Recent advances in deep learning and computer vision have enabled the development of machine-assisted in situ biodiversity monitoring tools, designed to automate parts of the data curation process and mitigate the annotation bottleneck faced by marine ecologists [49]. By leveraging manually curated data from previous surveys, researchers can now train models to detect benthic organisms commonly encountered in their study regions, supporting applications

---

\*Corresponding author

such as first-pass annotation workflows [26, 39, 40].

Given the high data collection and annotation costs associated with the Antarctic benthos however, there is a notable lack of publicly available datasets suitable for training automated biodiversity monitoring tools for these ecosystems. This limitation is further compounded by the region's high levels of endemism, which reduces the relevance and transferability of models trained on data from other regions.

In scenarios requiring the detection of small or densely aggregated organisms, high-resolution imagery is often utilised to enable finer-scale ecological observations. However, such data introduces additional complexity to the model development pipeline. High-resolution images place substantial computational demands on both training and inference processes, and the accurate detection of small or closely packed objects remains a persistent challenge for deep learning-based object detection systems [30].

In this study, we present an object detection framework designed to identify benthic organisms in high-resolution seafloor imagery from the Weddell Sea, Antarctica. Our approach accommodates large-scale inputs without downscaling through a patch-based processing methodology. The model is trained on a manually annotated dataset of only 100 images, which we release publicly as the first computer vision-ready benthic dataset from the Weddell Sea. The resulting model is capable of detecting a wide variety of benthic organisms, more than previous works in this area and to a higher level of granularity.

## 2. Related Work

Early studies into machine-assisted *in situ* benthic biodiversity monitoring used local features and hand-engineered pipelines to identify specific organisms of interest [13, 48]. Such methods require extensive adaptation for new organisms or environments, limiting their use in broad biodiversity monitoring surveys enabled by advances in underwater imaging and affordable data storage.

Data-driven deep learning models capable of generalised, automated feature extraction, e.g., Convolutional Neural Networks (CNNs), have accelerated the pace of machine-assisted *in situ* benthic biodiversity monitoring research [49]. Such models are often task-specific: image classifiers for taxonomic ID [41, 56], object detectors for abundance estimation [32, 55], semantic segmenters for habitat mapping and behaviour analysis [20, 35, 38], and instance segmenters for biomass estimation [31].

Few studies focus specifically on the Antarctic benthos, likely due to its remoteness and high fieldwork costs. [32] make use of a YOLOv5 [23] model pre-trained on the COCO dataset [28] to provide abundance estimates for two coarse-grained morphotypes, organism groupings classified based on shared morphological characteristics, in imagery from a stationary camera deployed in the Ross Sea. As im-

agery is downscaled, detection of small organisms may not be feasible using the presented methodology.

[31] apply a patching strategy to towed-camera imagery from the Weddell Sea to evaluate the effectiveness of synthetic data augmentation in training a CenterMask [27] model for instance segmentation of three coarse-grained morphotypes. However, their approach does not address potential detection failures at patch boundaries or support full-image ecological analysis. To address these limitations, we extend patching with overlap and postprocessing techniques that improve detection accuracy at patch edges. Additionally, we reproject detections back onto the original large-scale imagery, preserving spatial context for ecological analysis. Detailed methodology is provided in Sec. 4.

Further, while the aforementioned works demonstrate the feasibility of identifying Antarctic organisms using deep learning, they are limited to a small number of coarse-grained morphotypes. These studies also do not consider taxa with low abundance, potentially overlooking ecologically important but infrequently observed organisms. In contrast, we explore the development of object detection models that capture fine-grained taxonomic and morphological range, including rare organisms, examining the effect of abundance on model detection capability.

## 3. The Weddell Sea Benthic Dataset

Data used in this study were collected during expedition PS118 (cruises 69-1 and 6-9; see Fig. 1) of the RV *Polarstern* [43]. High-resolution benthic imagery (22 Megapixel, average filesize = 6.94 Megabyte (MB)) was captured in a top-down view using the Ocean Floor Observation and Bathymetry System (OFOBS) [42], a towed camera system operating just above the seafloor. The imagery captures a diverse range of environmental conditions, including variable turbidity, illumination levels, and substrate types (hard and soft). Some images exhibit mild distortion due to the motion of the OFOBS during capture.

A subset of the collected imagery, selected for their ecological rather than model training merit, was manually annotated to facilitate benthic community composition analysis [24]; this forms the ground truth dataset used in this study, which we name The Weddell Sea Benthic Dataset (WSBD). This dataset comprises 100 annotated images captured at a range of depths (421–2202 m) and seafloor inclinations (0–80°). Where images were not comprehensively annotated, e.g., due to distortion, the unlabelled regions were cropped, resulting in images of varying sizes (average = 3364×4545 px, 1.15 MB). Images are distinctly separated, with no overlap present between them. The original annotations were consolidated into 25 morphologically distinct classes, ranging from broad taxonomic groups to species level (see App. A).

The dataset presents substantial visual complexity, with

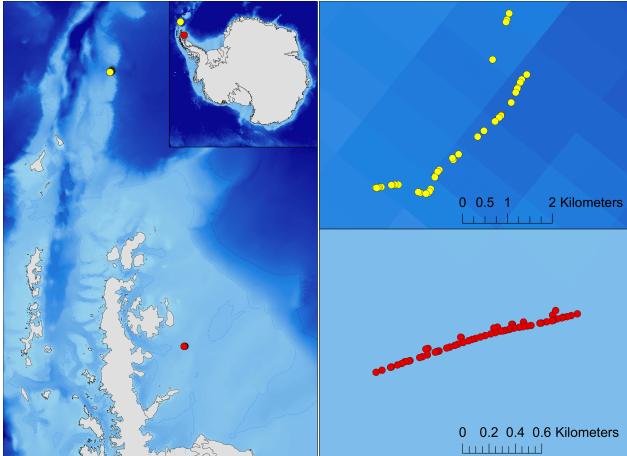


Figure 1. Map of PS118 image acquisition sites included in the Weddell Sea Benthic Dataset. Yellow dots show data collected during cruise 69-1, while red points correspond to cruise 6-9.

imagery characterised by high levels of background clutter, variable illumination, shadowing, and overlapping objects. These factors, plus the presence of fine-grained and morphologically similar taxa, make the WSBD a challenging and ecologically realistic benchmark for evaluating benthic object detection frameworks.

Imagery is biased towards soft-substrate environments at shallower depths (420–500 m), comprising 61.00% of images but only 4.24% of annotations. The dataset also shows a bias towards low-inclination areas, with 52.00% of images taken on slopes  $\leq 10^\circ$ . Certain taxa are restricted to specific substrate.

Owing to the remoteness of the study site and limited anthropogenic disturbance under the Antarctic Treaty System, the WSBD contains high organism densities. The dataset contains 31,280 total bounding box annotations, with individual images containing between 5 and 1693 annotations (average = 312.8). This includes numerous overlapping bounding boxes, a known challenge for object detection systems [8, 21]. Further, the dataset exhibits significant class imbalance, following a long-tailed distribution consistent with ecological patterns. The number of annotated instances per class ranges from 13,295 for stylasterids to 10 for the ascidian *Cnemidocarpa verrucosa*. Addressing rare-class detection remains a critical issue in machine-assisted biodiversity monitoring [34, 50, 51].

Small object detection remains a challenging and largely unsolved problem in computer vision [30]. The WSBD dataset exemplifies these difficulties, exhibiting substantial inter-class size variation. Average bounding box areas range from 520 px<sup>2</sup> for cup corals to 68,092 px<sup>2</sup> for the ascidian *Distaplia*. Further, intra-class size variability is introduced by fluctuations in the OFOBS’ altitude during image capture, resulting in inconsistent scales which further

complicate detection tasks. We release the WSBD under an OGL-UK-3.0 license: <https://doi.org/10.5285/1BA97E4B-EFB7-460B-9F2D-90437E33CE09>.

## 4. Method

Our proposed methodology (see Fig. 2) enables us to exploit the high spatial fidelity of the WSBD whilst maintaining detection efficacy.

### 4.1. Dataset Preparation

To account for the imbalance in annotation volume between the two substrate types, the train, validation, and test sets were generated based on the proportion of total annotations rather than the number of images. These sets were then refined to ensure they remained representative of the geographic and environmental diversity present in the dataset, including variation in depth and seafloor inclination. This adjustment was made to enhance model generalisability and help prevent overfitting to specific environmental conditions, which is critical in biodiversity monitoring applications [37, 53]. The final annotation-level train-validation-test split was 68.71%, 18.93%, and 12.36%, respectively.

### 4.2. Image Patching

The WSBD provides high-resolution benthic imagery which, while crucial for classifying small and morphologically similar organisms, introduces substantial computational overhead. Conventional object detection architectures are typically optimised for lower-resolution inputs [29, 46] and thus struggle to process full-resolution WSBD images without exceeding memory constraints. Downscaling such imagery to meet these limitations results in the loss of visual features which is particularly detrimental to the detection of small organisms (see Sec. 5.7.1).

To retain visual features we implement a patch-based detection strategy, subdividing the original large-scale image into sub-images of uniform size via a sliding window with fixed horizontal and vertical strides. Image patching is a well-established technique within machine-assisted in situ benthic biodiversity monitoring research, though it has primarily been employed for coarse-grained tasks [18, 22, 31]. Patching also standardises input dimensions, mitigating image size variability from dataset generation and enabling more efficient training and inference.

To extend patch-based processing to object-level tasks, we adopt the Slicing Aided Hyper Inference (SAHI) methodology [1]. Designed to improve the performance of object detection models on high-resolution data, SAHI works by dividing large images into overlapping patches, applying patch-based detection, and subsequently merging results via postprocessing. This approach retains resolution

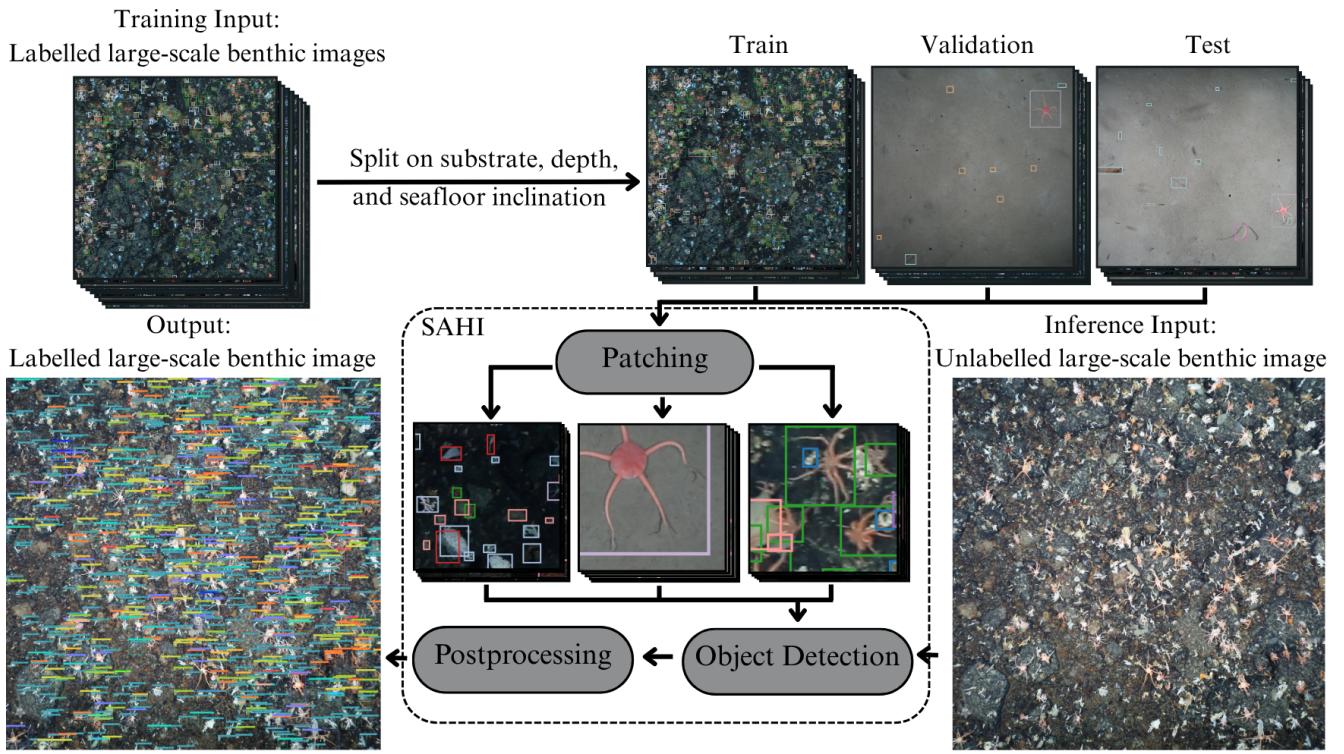


Figure 2. A high level overview of the proposed Antarctic benthic organism detection and classification framework. For large-scale visualisation of the output, see App. C.

and aids small object detection, a common problem in ecologically complex imagery. SAHI has demonstrated strong performance in object-level tasks involving high-resolution data across other domains [9, 17, 36]. We evaluate the optimal SAHI patching configuration, including patch size, overlap stride, and minimum bounding box visibility (the proportion of a ground truth bounding box required within a patch to be considered a valid object instance).

#### 4.3. Object Detection

An object detection model is trained using patches as input, allowing for the retention of fine-grained features necessary for accurately detecting small and densely clustered organisms. The model generates bounding boxes around proposed regions of interest per patch, accompanied by a predicted class label, corresponding to one of 25 defined organism morphotypes, and a confidence score. We evaluate a range of object detection architectures, including single-stage and two-stage detectors as well as CNN and transformer-based models, alongside various data augmentation strategies and model fine-tuning.

#### 4.4. Postprocessing

Following inference, patch-level detections are mapped back to their original coordinates on the large-scale in-

put image. To resolve redundant detections resulting from overlapping patches, we apply a Non-Maximum Merging (NMM) procedure, consolidating multiple detections of the same object into a single bounding box.

## 5. Experiments

We evaluate various methodological configurations to determine the optimal framework setup for the WSBD. Specifically, we examine the impact of different SAHI parameters, augmentation strategies, architectures, and the use of pre-trained weights for model fine-tuning. Throughout, we examine the effect of organism abundance on model performance. The final optimal setup uncovered represents a baseline benchmark for the WSBD.

### 5.1. Experimental Setup

Experiments were run on a single High Performance Computing node using one NVIDIA A2 GPU. Object detection models were implemented in Python using MMDetection [10], with data augmentation via Albumentations [6]. Training was performed for up to 200 epochs, with early stopping if no improvement was seen after 10 epochs. Our code is available at: <https://github.com/Trotts/antarctic-benthic-organism-detection/>.

Table 1. Test set Mean Average Precision (mAP) across key Intersection over Union (IoU) thresholds and object sizes for each dataset configuration. Bold indicates top performance per metric.

Parameters	Num. Classes	Mean Average Precision (mAP)									
		@0.5:0.95		@0.5		All		Small		Medium	
		10	25	10	25	10	25	10	25	10	25
SAHI Patching	10	0.22	0.18	0.45	0.34	0.20	<b>0.24</b>	0.48	<b>0.35</b>	<b>0.54</b>	0.42
+ SAHI Postprocessing	10	0.21	<b>0.19</b>	0.45	<b>0.37</b>	0.20	0.23	0.50	0.33	0.52	<b>0.44</b>
+ Spatial Augmentation	10	0.21	0.18	0.45	0.33	0.22	0.19	<b>0.50</b>	0.32	0.49	<b>0.44</b>

Unless stated otherwise, all experiments used a Faster R-CNN architecture [46]. Evaluation used Mean Average Precision (mAP) across multiple Intersection over Union (IoU) thresholds and object sizes (Small, Medium, and Large), following the COCO format [28]. Models were evaluated on both the full 25-class set and a 10-class subset comprising the most abundant taxa. Metrics were computed after patch-level detections were reprojected back to their original large-scale image coordinates and postprocessed using NMM.

## 5.2. SAHI Patching Parameters

To implement SAHI effectively, several parameters must be defined to control how images and annotations are divided into patches. To determine the optimal configuration for the WSBD, we conducted a series of experiments training a model for each combination of three key parameters: patch size (250×250, 500×500, 750×750, and 1000×1000 px), overlap stride (0.0, 0.25, and 0.50), and minimum bounding box visibility (0.10, 0.25, and 0.50). The NMM IoU threshold was fixed at 0.5 across all configurations.

Evaluation revealed that a patch size of 500×500 px with a 0.50 stride and a minimum bounding box visibility of 0.25 achieved the highest overall performance (see Tab. 1). This configuration maintained robust performance across all object sizes relative to other tested permutations. While larger patch sizes yielded comparable results for detecting larger organisms, they were less effective for smaller taxa, suggesting a trade-off between patch size and sensitivity to fine-grained features. Additionally, larger patches and higher stride increased computational demands due to increased dataset size and model input parameters. A minimum bounding box visibility of 0.25 yielded the highest mAP. Lower thresholds introduced training noise by retaining extremely cropped objects, while higher thresholds excluded valid examples, disproportionately affecting small organisms that frequently occur near patch boundaries.

This setup generated 25,184 patches from the 100 WSBD images, with 17,819 used for training, 4310 for validation, and 3055 for testing.

## 5.3. SAHI Postprocessing Parameters

During SAHI postprocessing, predicted patch-level bounding boxes are reprojected to their original coordinates within the large-scale input image. Bounding boxes of the

same class that overlap by at least a specified IoU threshold are merged using NMM to reduce duplicate detections. To identify the optimal NMM IoU threshold, we applied SAHI postprocessing to the optimal patching model across a range of IoU values from 0.05 to 0.50 in 0.05 increments.

An IoU threshold of 0.20 was found to be sufficient, indicating a relatively low overlap threshold is optimal for merging duplicate detections after reprojection, particularly given the prevalence of small, densely clustered objects found within hard substrate imagery. Higher thresholds often failed to merge duplicate detections, resulting in inflated false positives, while lower thresholds erroneously merge distinct but nearby objects.

## 5.4. Data Augmentation Strategy

Given the limited volume of training data in the WSBD, we evaluated the effect of data augmentation, generating new samples by perturbing existing data, on model performance. Data scarcity is a persistent challenge for the development of machine-assisted in situ benthic biodiversity monitoring tools, though prior studies have shown that augmentation can enhance model performance [14, 16, 31, 38].

We evaluated three augmentation strategies: pixel-level, spatial-level, and a combined approach using both (see App. B). Each was assessed against a non-augmented baseline defined in Sec. 5.3. Spatial transformations yielded the best overall results. This is likely due to the WSBD’s existing artefacts, such as motion blur and shadow, reducing the effectiveness of additional pixel-level perturbations. In contrast, spatial transformations introduced beneficial variability without further obscuring key visual features. This approach achieved the highest mAP@0.5 for small objects and tied for the best performance on medium objects in the 10-class setup. These improvements are particularly valuable in the WSBD, where small, densely packed organisms are frequent and difficult to detect. Accordingly, spatial augmentation was adopted for all subsequent experiments.

## 5.5. Architecture Search

To explore a broader range of detection capabilities beyond the Faster R-CNN baseline, we evaluated a mix of single-stage and two-stage detectors, as well as CNN and transformer-based architectures, against WSBD performance.

The DINO [54] architecture achieved the highest mAP@0.5:0.95 for both the 10-class and 25-class configurations, sharing the top score with Cascade R-CNN [7] in the latter (See Tab. 2 Top). When evaluating at mAP@0.5, DINO also outperformed other models in the 10-class setup, while Deformable-DETR [57] achieved the highest performance in the 25-class setting. While no architecture surpassed the Faster R-CNN baseline in detecting small and medium objects under the 10-class configuration, Co-

Table 2. Test set Mean Average Precision (mAP) across key Intersection over Union (IoU) thresholds and object sizes for various model architectures (ordered by release), trained with optimal dataset settings, with and without COCO fine-tuning. Bold indicates top performance per metric; italics denote average performance.

Fine-tuned	Architecture	Mean Average Precision (mAP)									
		@0.5:0.95		@0.5		All		Small		Medium	
		10	25	10	25	10	25	10	25	10	25
$\times$	Faster R-CNN [46]	0.21	0.18	0.45	0.33	<b>0.22</b>	0.19	0.50	0.32	0.49	0.44
	Cascade R-CNN [7]	0.20	<b>0.19</b>	0.42	0.35	0.13	0.10	0.45	0.30	0.59	<b>0.55</b>
	RetinaNet [29]	0.18	0.12	0.38	0.24	0.09	0.11	0.42	0.24	0.61	0.35
	Deformable-DETR [57]	0.19	0.18	0.45	0.36	0.21	0.15	0.47	0.31	0.52	0.46
	DINO [54]	0.22	<b>0.19</b>	0.46	0.35	0.16	0.14	0.49	0.32	0.53	0.44
	CoDETR [33]	0.19	0.16	0.45	0.34	0.18	0.22	0.48	0.31	0.52	0.43
Average		0.20	0.17	0.44	0.33	0.17	0.15	0.47	0.30	0.54	0.45
$\checkmark$	Faster R-CNN [46]	0.21	0.18	0.48	0.36	<b>0.22</b>	0.21	0.49	0.31	0.56	0.44
	Cascade R-CNN [7]	0.22	0.17	0.46	0.34	0.18	0.18	0.48	0.29	0.53	0.45
	RetinaNet [29]	0.20	<b>0.19</b>	0.41	0.34	0.11	0.12	0.46	0.31	0.50	0.45
	Deformable-DETR [57]	0.22	<b>0.19</b>	<b>0.49</b>	<b>0.39</b>	0.21	<b>0.27</b>	<b>0.51</b>	<b>0.33</b>	0.55	0.47
	DINO [54]	<b>0.24</b>	<b>0.19</b>	0.47	0.33	0.19	0.13	0.49	0.29	0.57	0.49
	CoDETR [33]	0.22	<b>0.19</b>	0.46	0.33	0.17	0.12	0.49	0.30	<b>0.60</b>	0.44
Average		0.22	0.19	0.46	0.35	0.18	0.17	0.49	0.31	0.55	0.46

DETR [33] exhibited the best performance on small objects in the 25-class evaluation.

For medium-sized objects under the same configuration, the top-performing model was shared between Faster R-CNN and DINO. In the case of large object detection, RetinaNet [29] delivered the best results in the 10-class evaluation. However, its performance dropped substantially in the 25-class setting, where Cascade R-CNN significantly outperformed all other architectures.

## 5.6. Effect of Fine-tuning

Alongside data augmentation (see Sec. 5.4), fine-tuning is an effective strategy for improving model generalisability in object detection tasks where training data is limited [4]. Rather than initialising model weights randomly, requiring the network to learn fundamental visual representations from scratch, fine-tuned models leverage weights obtained from previous training on large-scale datasets. This facilitates transfer learning, wherein knowledge acquired in a source domain is applied to enhance performance in a target domain [4].

Given the challenge of limited labelled data in the development of machine-assisted *in situ* benthic biodiversity monitoring tools, fine-tuning has become a standard approach within the field [49]. Here, we evaluate the impact of fine-tuning on WSBD performance by comparing models initialised with random weights to those initialised on weights derived after training on the COCO dataset [28].

Overall, COCO fine-tuning resulted in a slight improvement in average model performance (see Tab. 2 Bottom). With the exception of large object detection in the 25-class evaluation, the highest metrics across evaluation categories were achieved by fine-tuned models. Notably, in contrast to the non-fine-tuned models where optimal performance varied depending on the specific evaluation scenario (e.g., object size or number of classes), the use of fine-tuning consistently elevated Deformable-DETR to either the top-

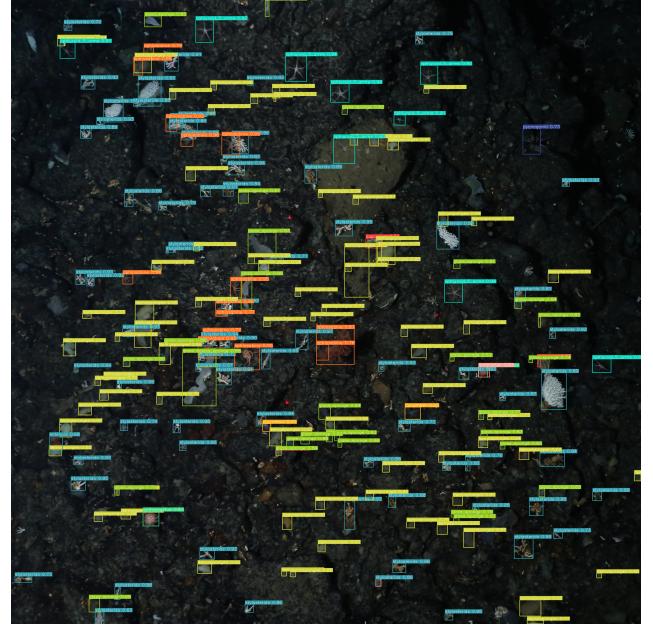


Figure 3. Example WSBD test set image output. Predicted organism bounding boxes, class labels, and confidence scores shown after reprojection and postprocessing. Confidence threshold = 0.60. For large-scale visualisations, see App. C.

performing model or among the top three performers across nearly all categories. This suggests that while the absolute gains from fine-tuning may be limited, the approach contributes to increased robustness and consistency in model performance. Crucially, it enables the identification of a single architecture, Deformable-DETR, as the most effective model overall, providing a clear candidate for subsequent deployment. An example output from this model can be seen in Fig. 3.

## 5.7. Ablation Study

To further evaluate model performance and verify that each component of the proposed framework contributes positively, we conducted a series of additional ablation studies.

### 5.7.1. Image-level Downscaling

Although the use of patching is intended to aid the detection of small objects, the results presented in Tab. 2 indicate that all evaluated model architectures continue to exhibit limited performance when detecting those present in the WSBD. To verify patching contributed positively, an additional Deformable-DETR model was trained using non-patched imagery, spatially augmented and fine-tuned on the COCO dataset, for comparison. To ensure consistent image size we downscale the data to 1635×1635 px, the smallest WSBD image.

Substantial declines in detection performance were observed across all object categories (see Tab. 3), indicating a critical loss of discriminative features resulting from image

Table 3. Test set Mean Average Precision (mAP) across key Intersection over Union (IoU) thresholds and object sizes for each ablation study. The baseline model corresponds to the optimal Deformable-DETR configuration. Bold indicates top performance per metric.

Experiment	Num. Classes	Mean Average Precision (mAP)											
		@0.5:0.95		@0.5		All		Small		Medium		Large	
		10	25	10	25	10	25	10	25	10	25	10	25
Baseline		<b>0.22</b>	<b>0.19</b>	<b>0.49</b>	<b>0.39</b>	<b>0.21</b>	<b>0.27</b>	<b>0.51</b>	<b>0.33</b>	<b>0.56</b>	<b>0.47</b>		
Image-level Downscaling		0.06	0.07	0.13	0.12	0.03	0.02	0.12	0.06	0.19	0.19		
Non-Maximum Suppression		0.21	<b>0.19</b>	<b>0.48</b>	0.34	0.20	0.20	<b>0.51</b>	0.33	0.53	0.39		
No Postprocessing		0.13	0.13	0.26	0.22	0.13	0.16	0.27	0.19	0.32	0.29		

downscaling. These findings reinforce the necessity of employing a patching strategy to preserve resolution and maintain object-level detail.

### 5.7.2. SAHI Postprocessing Algorithm

Following patch-level detection, bounding boxes are reprojected to their corresponding large-scale image coordinates. To address duplicate detections resulting from patch overlap, boxes with identical class labels, overlapping with an IoU  $\geq 0.20$ , are postprocessed using NMM. However, SAHI also allows for the use of Non-Maximum Suppression (NMS), where only the overlapping box with the highest confidence level is retained. Additionally, detections may be reprojected without any postprocessing applied. To verify merging was the correct approach, we evaluated the use of NMS and no postprocessing after reprojection.

Substituting NMM with NMS resulted in either a slight decrease or no measurable improvement in detection performance, particularly under the more challenging 25-class evaluation (see Tab. 3). The decline was most evident for both small and large object categories, where the baseline model employing merging demonstrated superior performance. Further, the use of no postprocessing significantly reduces performance across all evaluated metrics. These findings suggest that merging methods more effectively preserve localisation quality in cases where object instances are fragmented across overlapping patches.

## 6. Discussion

Based on the results presented in Sec. 5, we propose an optimal framework configuration for the fine-grained detection of benthic organisms in high-resolution towed camera imagery from the Weddell Sea, Antarctica. The recommended approach involves subdividing large-scale images into 500x500 px patches with a 0.50 horizontal and vertical overlap stride, alongside a minimum bounding box visibility threshold of 0.25. Dataset splitting is stratified by substrate type, depth, and seafloor inclination to ensure geographic and environmental diversity. The resulting patches are spatially augmented and used to train a Deformable-DETR object detection model, with initial weights derived from the COCO dataset. Following inference, detec-

tions are reprojected to their original locations on the full-resolution image. Overlapping same-class bounding boxes are then postprocessed using NMM with an IoU threshold of 0.20.

### 6.1. Small Object Detection

The resulting model is trained to detect 25 distinct morphotypes found in the Weddell Sea. However, notable performance limitations for small organisms are present, even when employing the SAHI methodology. While these limitations may partly stem from the restricted size of the training dataset, a common constraint in machine-assisted in situ benthic biodiversity monitoring, they are likely exacerbated by the logistical and environmental challenges of data collection in Antarctica.

The observed underperformance for small object detection, despite the use of high-resolution imagery, advanced patching strategies, data augmentation, and fine-tuning, suggests current object detection architectures are limited in their ability to extract meaningful features from small instances in visually complex benthic environments. This restricts accurate learning and detection of ecologically important taxa, and highlights the need for new architectural approaches tailored to small object representation.

### 6.2. Effect of Abundance and Morphology

Additionally, organism abundance was found to have a notable influence on model performance. This is evident when comparing the results of the 10-class and 25-class evaluations. The average number of annotations per class in the 10-class evaluation is 2068.5. In contrast, overall average abundance for the 25-class configuration is 859.4, dropping to just 53.4 for organisms present in the 25-class set only. Examining the optimal model’s class confusion reveals that although the overall number of missed detections is high, especially for rare organisms, the rate of misclassification among detected abundant instances is low (see App. D). This suggests that when the model places a bounding box, it is likely to contain a valid organism and to assign it the correct label.

Where misclassification does occur it is typically between morphologically similar organisms, e.g., demosponges and glass sponges, which share structural features and can be difficult to distinguish visually, even for trained experts. In contrast, misclassifications between taxonomically related but visually distinct organisms, e.g., *Ophiosabine* and other ophiuroids, are relatively rare. This indicates that the model relies primarily on visual cues rather than taxonomic proximity when assigning class labels. Interestingly, we observe strong detection performance for pycnogonids, despite this class being the fourth least abundant. However this may be due to a lack of morphological variation between the dataset splits for this class.

### 6.3. SAHI Postprocessing Limitations

Unlike domains where SAHI is commonly applied, e.g., satellite imagery [9, 17, 36], data in the WSBD is captured from varying altitudes above the seafloor due to changes in OFOBS platform depth and seafloor topography. This introduces significant intra-class size variation. For large organisms, a single instance may span many patches. Experimental observations show SAHI occasionally struggles to accurately postprocess duplicate detections into a single coherent bounding box when an organism is divided across a large number of patches (see App. E). This negatively affects overall model performance and may bias abundance estimates if not addressed during manual post-hoc review.

### 6.4. Expert Labelling Agreement

It is important to note that evaluation metrics reported in this study, as in other automated biodiversity monitoring research, e.g., [5], reflect the degree of agreement between the model and the human annotator rather than an absolute measure of detection accuracy. Given the high densities of organisms, the prevalence of small-bodied taxa, and the well-documented issues of fatigue and subjectivity in manual annotation processes for benthic imagery [12, 15, 41], it is likely some valid organisms were omitted from the ground truth, leading to correct model detections being penalised and artificially lowering performance metrics.

Due to the high level of taxonomic expertise required to accurately annotate Antarctic benthic fauna and the significant time investment needed (averaging approximately eight hours per image), it was not feasible to obtain multiple independent expert annotations to reduce potential labelling bias. With the time savings afforded by our framework, future labelling can incorporate consensus agreement.

### 6.5. Potential Framework Application

Despite these challenges, the resulting model remains highly valuable to benthic ecologists. The proposed framework offers the potential for substantial time and cost savings, particularly when applied to the processing of extensively backlogged survey data, totalling in the tens of thousands of images. As a result, the framework is well-suited for use in first-pass, human-in-the-loop analyses, allowing ecologists to focus on completing remaining annotations instead of reviewing full images manually.

A promising direction for future work involves integrating the proposed framework into an active learning pipeline, wherein unlabelled imagery is prioritised for annotation based on predefined selection criteria, automatically annotated using the framework, then refined by expert ecologists. Selection strategies may incorporate both ecological relevance and expected contribution to framework performance. As more archival data is processed through this iterative approach, the resulting enlarged dataset could serve

as a valuable fine-tuning resource, especially for currently rare organisms where model performance may benefit from increased abundance, enabling iterative improvements in model performance as annotation efforts progress.

## 7. Conclusion

We address the challenge of detecting and classifying Antarctic benthic organisms in high-resolution, top-down imagery captured using a towed camera system in the Weddell Sea. Through the creation of the first publicly available computer vision-ready dataset of Antarctic seafloor ecology, we develop and assess a comprehensive object detection framework specifically designed for the complexities of benthic imagery. The proposed pipeline integrates the SAHI methodology, patching to retain spatial resolution and reduce computational expense, alongside spatial data augmentation and model fine-tuning to support generalisability under data-scarce conditions. Postprocessing using NMM enhances detection coherence after bounding box reprojection from patch-level back to the original large-scale image. Our framework demonstrates strong performance in detecting medium and large benthic morphotypes.

Persistent underperformance on small and rare taxa, even when enhancement strategies were applied, highlights fundamental limitations in current object detection architectures when applied to ecologically complex imagery. These findings underscore the need for targeted research into small object representation, as well as the potential value of active learning approaches to enable faster processing of backlogged, unprocessed field imagery, refining rare organism performance while uncovering new ecological insights. By providing our data and models open-source, we hope to encourage community efforts to improve the detection of such organisms in complex marine imagery. Nevertheless, our proposed framework offers a scalable, generalised solution for automated analysis of high-resolution benthic imagery, with significant potential to accelerate biodiversity monitoring and enable better protection of the unique benthic ecosystems found in Antarctica and beyond.

## Acknowledgements

We thank Miao Fan, Alfred Wegener Institutue (AWI), for providing the bathymetry maps used to subset the WSBD by seafloor inclination. We also thank Autun Purser, AWI, and all crew of the RV *Polarstern* PS118 cruise for their data collection efforts. CT, HJG and RJW are funded by the UKRI Future Leaders Fellowship MR/W01002X/1 ‘The past, present and future of unique cold-water benthic (sea floor) ecosystems in the Southern Ocean’ awarded to RJW. For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) license to any Accepted Manuscript version arising.

## References

- [1] Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. Slicing Aided Hyper Inference and Fine-Tuning for Small Object Detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 966–970, Bordeaux, France, 2022. IEEE. 3
- [2] M Alcaraz. Biogeographic Atlas of the Southern Ocean. *Scientific Committee on Antarctic Research, Cambridge, XII*, page 498, 2014. 1
- [3] Romero-Ramirez Alicia, Morales Luna Hadrys Laura, Kuklinski Piotr, Chelchowski Maciej, and Balazy Piotr. Image analysis and benthic ecology: Proceedings to analyze in situ long-term image series. *Limnology and Oceanography: Methods*, 21(4):169–177, 2023. 1
- [4] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. 6
- [5] Justine Boulent, Bertrand Charry, Malcolm McHugh Kennedy, Emily Tissier, Raina Fan, Marianne Marcoux, Cortney A. Watt, and Antoine Gagné-Turcotte. Scaling whale monitoring using deep learning: A human-in-the-loop solution for analyzing aerial datasets. *Frontiers in Marine Science*, 10:1099479, 2023. 8
- [6] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and Flexible Image Augmentations. *Information*, 11(2):125, 2020. 4, 2
- [7] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018. 5, 6
- [8] Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ram-prasaath R. Selvaraju, Dhruv Batra, and Devi Parikh. Counting Everyday Objects in Everyday Scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4428–4437, Honolulu, HI, 2017. IEEE. 3
- [9] Divyansh Chaurasia and B.D.K. Patro. Real-time Detection of Birds for Farm Surveillance Using YOLOv7 and SAHI. In *2023 3rd International Conference on Computing and Information Technology (ICCIT)*, pages 442–450, Tabuk, Saudi Arabia, 2023. IEEE. 4, 8
- [10] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tian-heng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahu Lin. MMDetection: Open MMLab Detection Toolbox and Benchmark, 2019. 4
- [11] Daniel Crespo and Miguel Ângelo Pardal. Ecological and Economic Importance of Benthic Communities. In *Life Below Water*, pages 313–323. Springer International Publishing, Cham, 2022. 1
- [12] Phil F. Culverhouse. Human and machine factors in algae monitoring performance. *Ecological Informatics*, 2(4):361–366, 2007. 1, 8
- [13] Matthew Dawkins, Charles Stewart, Scott Gallager, and Amber York. Automatic scallop detection in benthic environments. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 160–167, Clearwater Beach, FL, USA, 2013. IEEE. 2
- [14] Heather Doig, Oscar Pizarro, Jacquomo Monk, and Stefan Williams. Detecting Endangered Marine Species in Autonomous Underwater Vehicle Imagery Using Point Annotations and Few-Shot Learning, 2024. 5
- [15] Jm Durden, Bj Bett, T Schoening, Kj Morris, Tw Nattkemper, and Ha Ruhl. Comparison of image annotation data generated by multiple investigators for benthic ecology. *Marine Ecology Progress Series*, 552:61–70, 2016. 1, 8
- [16] Jennifer M. Durden, Brett Hosking, Brian J. Bett, Danelle Cline, and Henry A. Ruhl. Automated classification of fauna in seabed photographs: The impact of training and validation dataset size, with considerations for the class imbalance. *Progress in Oceanography*, 196:102612, 2021. 5
- [17] Bao Tran Gia, Tuong Bui Cong Khanh, Hien Ho Trong, Thuyen Tran Doan, Tien Do, Duy-Dinh Le, and Thanh Duc Ngo. Enhancing Road Object Detection in Fisheye Cameras: An Effective Framework Integrating SAHI and Hybrid Inference. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 7227–7235, Seattle, WA, USA, 2024. IEEE. 4, 8
- [18] Manuel González-Rivero, Oscar Beijbom, Alberto Rodriguez-Ramirez, Dominic E. P. Bryant, Anjani Ganase, Yeray Gonzalez-Marrero, Ana Herrera-Revels, Emma V. Kennedy, Catherine J. S. Kim, Sebastian Lopez-Marcano, Kathryn Markey, Benjamin P. Neal, Kate Osborne, Catalina Reyes-Nivia, Eugenia M. Sampayo, Kristin Stolberg, Abbie Taylor, Julie Vercelloni, Mathew Wyatt, and Ove Hoegh-Guldberg. Monitoring of Coral Reefs Using Artificial Intelligence: A Feasible and Cost-Effective Approach. *Remote Sensing*, 12(3):489, 2020. 3
- [19] Huw J. Griffiths, Vonda J. Cummings, Anton Van de Putte, Rowan J. Whittle, and Catherine L. Waller. Antarctic benthic ecological change. *Nature Reviews Earth & Environment*, 5(9):645–664, 2024. 1
- [20] Dominica Harrison, Fabio Cabrera De Leo, Warren J. Gallin, Farin Mir, Simone Marini, and Sally P. Leys. Machine Learning Applications of Convolutional Neural Networks and Unet Architecture to Predict and Classify Demosponge Behavior. *Water*, 13(18):2512, 2021. 2
- [21] Jeroen P. A. Hoekendijk, Benjamin Kellenberger, Geert Aarts, Sophie Brasseur, Suzanne S. H. Poiesz, and Devis Tuia. Counting using deep learning regression gives value to ecological surveys. *Scientific Reports*, 11(1):23209, 2021. 3
- [22] Chris Jackett, Franziska Althaus, Kylie Maguire, Moshiur Farazi, Ben Scoulding, Candice Untiedt, Tim Ryan, Peter Shanks, Pamela Brodie, and Alan Williams. A benthic substrate classification method for seabed images using deep learning: Application to management of deep-sea coral reefs. *Journal of Applied Ecology*, 60(7):1254–1273, 2023. 3
- [23] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, Imyhxy, Lorna, Zeng Yifu, Colin Wong, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UglyKitDe, Victor Sonck, Tkianai,

- YxNONG, Piotr Skalski, Adam Hogan, Dhruv Nair, Max Strobel, and Mrinal Jain. Ultralytics/yolov5: V7.0 - YOLOv5 SOTA Realtime Instance Segmentation. Zenodo, 2022. [2](#)
- [24] Tasnuva Ming Khan, Huw J. Griffiths, Rowan J. Whittle, Nile P. Stephenson, Katie M. Delahooke, Autun Purser, Andrea Manica, and Emily G. Mitchell. Network analyses on photographic surveys reveal that invertebrate predators do not structure epibenthos in the deep (~2000m) rocky Powell Basin, Weddell Sea, Antarctica. *Frontiers in Marine Science*, 11:1408828, 2024. [2](#)
- [25] Orlando Lam-Gordillo, Ryan Baring, and Sabine Dittmann. Ecosystem functioning and functional approaches on marine macrobenthic fauna: A research synthesis towards a global consensus. *Ecological Indicators*, 115:106379, 2020. [1](#)
- [26] Daniel Langenkämper, Martin Zurowietz, Timm Schoening, and Tim W. Nattkemper. BIIGLE 2.0 - Browsing and Annotating Large Marine Image Collections. *Frontiers in Marine Science*, 4:83, 2017. [2](#)
- [27] Youngwan Lee and Jongyoul Park. CenterMask: Real-Time Anchor-Free Instance Segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13903–13912, Seattle, WA, USA, 2020. IEEE. [2](#)
- [28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing, Cham, 2014. [2, 5, 6](#)
- [29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. [3, 6](#)
- [30] Yang Liu, Peng Sun, Nickolas Wergeles, and Yi Shang. A survey and performance evaluation of deep learning methods for small object detection. *Expert Systems with Applications*, 172:114602, 2021. [2, 3](#)
- [31] Mona Lütjens and Harald Sternberg. Deep Learning based Detection, Segmentation and Counting of Benthic Megafauna in Unconstrained Underwater Environments. *IFAC-PapersOnLine*, 54(16):76–82, 2021. [2, 3, 5](#)
- [32] Simone Marini, Federico Bonofoglio, Lorenzo P. Cognati, Andrea Bordone, Stefano Schiaparelli, and Andrea Peirano. Long-term automated visual monitoring of Antarctic benthic fauna. *Methods in Ecology and Evolution*, 13(8):1746–1764, 2022. [2](#)
- [33] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional DETR for Fast Training Convergence, 2023. [6](#)
- [34] Zhongqi Miao, Ziwei Liu, Kaitlyn M. Gaynor, Meredith S. Palmer, Stella X. Yu, and Wayne M. Getz. Iterative human and automated identification of wildlife images. *Nature Machine Intelligence*, 3(10):885–895, 2021. [3](#)
- [35] Katsunori Mizuno, Kei Terayama, Seiichiro Hagino, Shigeru Tabeta, Shingo Sakamoto, Toshihiro Ogawa, Kenichi Sugimoto, and Hironobu Fukami. An efficient coral survey method based on a large-scale 3-D structure model obtained by Speedy Sea Scanner and U-Net segmentation. *Scientific Reports*, 10(1):12416, 2020. [2](#)
- [36] M. Muzammul, Abdulmohsen Algarni, Yazeed Yasin Ghadi, and Muhammad Assam. Enhancing UAV Aerial Image Analysis: Integrating Advanced SAHI Techniques With Real-Time Detection Models on the VisDrone Dataset. *IEEE Access*, 12:21621–21633, 2024. [4, 8](#)
- [37] Ennio Ottaviani, Marco Francescangeli, Nikolla Gjeci, Joaquin Del Rio Fernandez, Jacopo Aguzzi, and Simone Marini. Assessing the Image Concept Drift at the OBSEA Coastal Underwater Cabled Observatory. *Frontiers in Marine Science*, 9:840088, 2022. [3](#)
- [38] G. Pavoni, M. Corsini, N. Pedersen, V. Petrovic, and P. Cignoni. Challenges in the deep learning-based semantic segmentation of benthic communities from Ortho-images. *Applied Geomatics*, 13(1):131–146, 2021. [2, 5](#)
- [39] Gaia Pavoni, Massimiliano Corsini, Federico Ponchio, Alessandro Muntoni, Clinton Edwards, Nicole Pedersen, Stuart Sandin, and Paolo Cignoni. TagLab: AI-assisted annotation for the fast and accurate semantic segmentation of coral reef orthoimages. *Journal of Field Robotics*, 39(3):246–262, 2022. [2](#)
- [40] Malte Pedersen, Joakim Bruslund Haurum, Rikke Gade, and Thomas B Moeslund. Detection of Marine Animals in a New Underwater Dataset with Varying Visibility. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–26, Long Beach, USA, 2019. [2](#)
- [41] N Piechaud, C Hunt, Pf Culverhouse, NI Foster, and KI Howell. Automated identification of benthic epifauna with computer vision. *Marine Ecology Progress Series*, 615:15–30, 2019. [1, 2, 8](#)
- [42] Autun Purser, Yann Marcon, Simon Dreutter, Ulrich Hoge, Burkhard Sablotny, Laura Hehemann, Johannes Lemburg, Boris Dorschel, Harald Biebow, and Antje Boetius. Ocean Floor Observation and Bathymetry System (OFOBS): A New Towed Camera/Sonar System for Deep-Sea Habitat Surveys. *IEEE Journal of Oceanic Engineering*, 44(1):87–99, 2019. [2](#)
- [43] Autun Purser, Simon Dreutter, Huw Griffiths, Laura Hehemann, Kerstin Jerosch, Axel Nordhausen, Dieter Piepenburg, Claudio Richter, Henning Schröder, and Boris Dorschel. Seabed video and still images from the northern Weddell Sea and the western flanks of the Powell Basin. *Earth System Science Data*, 13(2):609–615, 2021. [2](#)
- [44] Carl J. Reddin, Martin Aberhan, Nussaibah B. Raja, and Ádám T. Kocsis. Global warming generates predictable extinctions of warm- and cold-water marine benthic invertebrates via thermal habitat loss. *Global Change Biology*, 28(19):5793–5807, 2022. [1](#)
- [45] H. L. Rees, editor. *Guidelines for the Study of the Epibenthos of Subtidal Environments*. Number 42 in ICES Techniques in Marine Environmental Sciences. International Council for the Exploration of the Sea, Copenhagen, 2009. [1](#)
- [46] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. [3, 5, 6](#)

- [47] Yuntian Shi and Yaowu Li. Impacts of ocean acidification on physiology and ecology of marine invertebrates: A comprehensive review. *Aquatic Ecology*, 58(2):207–226, 2024. 1
- [48] Daniel Smith and Matthew Dunbabin. Automated Counting of the Northern Pacific Sea Star in the Derwent Using Shape Recognition. In *9th Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications (DICTA 2007)*, pages 500–507, 2007. 2
- [49] Cameron Trotter, Huw J. Griffiths, and Rowan J. Whittle. Surveying the deep: A review of computer vision in the benthos. *Ecological Informatics*, 86:102989, 2025. 1, 2, 6
- [50] Grant Van Horn and Pietro Perona. The Devil is in the Tails: Fine-grained Classification in the Wild, 2017. 3
- [51] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist Species Classification and Detection Dataset. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, Salt Lake City, UT, 2018. IEEE. 3
- [52] Ivor D. Williams, Courtney S. Couch, Oscar Beijbom, Thomas A. Oliver, Bernardo Vargas-Angel, Brett D. Schumacher, and Russell E. Brainard. Leveraging Automated Image Analysis Tools to Transform Our Capacity to Assess Status and Trends of Coral Reefs. *Frontiers in Marine Science*, 6:222, 2019. 1
- [53] Mathew Wyatt, Ben Radford, Nikolaus Callow, Mohammed Bennamoun, and Sharyn Hickey. Using ensemble methods to improve the robustness of deep learning for image classification in marine environments. *Methods in Ecology and Evolution*, 13(6):1317–1328, 2022. 3
- [54] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection, 2022. 5, 6
- [55] Lijun Zhang, Jiawen Fan, Yi Qiu, Zhe Jiang, Qingsong Hu, Bowen Xing, and Jingxiang Xu. Marine zoobenthos recognition algorithm based on improved lightweight YOLOv5. *Ecological Informatics*, 80:102467, 2024. 2
- [56] Zhinuo Zhou, Ge-Yi Fu, Yi Fang, Ye Yuan, Hong-Bin Shen, Chun-Sheng Wang, Xue-Wei Xu, Peng Zhou, and Xiaoyong Pan. EchoAI: A deep-learning based model for classification of echinoderms in global oceans. *Frontiers in Marine Science*, 10:1147690, 2023. 2
- [57] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection, 2021. 5, 6

# Automated Detection of Antarctic Benthic Organisms in High-Resolution In Situ Imagery to Aid Biodiversity Monitoring

## Supplementary Material

### A. Dataset Composition

Table S1. Counts and areas for the Weddell Sea Benthic Dataset classes, for both the original large-scale and patched images. Bolded counts denote the most abundant classes used for 10-class evaluation.

Class Label	Whole Image Dataset				Area (px <sup>2</sup> )			Patched Dataset				Area (px <sup>2</sup> )		
	All	Train	Validation	Test	Min	Max	Avg	All	Train	Validation	Test	Min	Max	Avg
actiniarian	165	118	22	25	111	51328	6821	777	553	105	119	111	51328	5512
alcyonium	280	266	6	8	266	104355	6773	1269	1202	24	43	60	98203	5508
anthomastus	89	60	22	7	198	30478	4666	371	260	80	31	106	30478	3810
ascidian.cnemidocarpa.verrucosa	10	2	4	4	4689	91344	21113	55	15	20	20	1517	91344	15311
ascidian.distaplia	32	24	1	7	1303	1432444	68092	186	146	4	36	539	250000	34927
ascidian.pyura.bouvetensis	66	41	21	4	981	188218	16652	351	203	126	22	275	153083	12022
asteroidia	156	111	23	22	144	569242	9572	643	482	74	87	26	250000	8279
astrochlamys	<b>720</b>	528	127	65	475	136010	13402	<b>3366</b>	2439	556	371	110	126777	10608
benthic.fish	71	52	14	5	309	179118	40603	474	349	92	33	309	139500	23828
bryozoan	15	8	5	2	386	45004	14134	88	45	35	8	386	45004	9918
crinoid	26	19	3	4	491	20989	6422	119	78	14	27	232	20989	5377
crustaceans	<b>461</b>	338	79	44	554	399481	12799	<b>2469</b>	1817	424	228	253	173560	9494
cucumber	355	287	41	27	62	9041	1522	<b>1447</b>	1166	164	117	56	9041	1415
cup.coral	<b>4757</b>	2807	1611	339	31	13756	520	<b>18552</b>	11262	6039	1251	12	13756	482
demospinges	<b>2211</b>	1517	340	354	7	258424	2283	<b>8960</b>	6216	1312	1432	7	194988	2003
echinoid	11	6	2	3	1975	12170	4098	50	24	12	14	817	12170	3607
glass.sponge	<b>2308</b>	1612	477	219	29	92647	1603	<b>9144</b>	6295	1930	919	13	92647	1508
gorgonian	<b>1144</b>	903	113	128	62	303363	9045	<b>5396</b>	4248	546	602	62	219445	7274
hydroid.solitary	25	17	5	3	2519	97165	14968	133	94	25	14	378	97165	11492
ophiosabine	<b>3075</b>	1853	694	528	219	50859	3125	<b>12783</b>	7630	2897	2256	68	50859	2768
ophiuroid_5_arms	<b>1885</b>	1293	280	312	93	748890	17419	<b>8819</b>	5961	1349	1509	72	250000	13686
pencilurchin	78	51	20	7	700	36549	6419	338	219	86	33	130	36549	5296
pycnogonid	11	7	2	2	2381	25210	13854	59	40	9	10	747	25210	9264
stylasterids	<b>13295</b>	9547	2011	1736	4	74999	1720	<b>53523</b>	38529	7543	7451	4	69524	1560
worm.tubes	35	19	2	14	168	25355	4591	157	82	8	67	168	25355	3958
Total	31280	21486	5925	3869				129529	89355	23474	16700			

## B. Data Augmentation Strategy Details

Tab. S2 presents the Albumentations-based [6] data augmentation techniques used in this study, categorised by the augmentation strategy in which they were employed. Probabilities for all augmentations were set to 0.50. For *Random Sized BBox Safe Crop*, the height and width parameters were set to the patch size. All other parameters were set to the Albumentations default.

Despite its name, *Pixel Dropout* is classified as a spatial transformation. This augmentation operates non-uniformly across the image by randomly selecting specific spatial coordinates at which to drop pixels. Consequently, it alters the spatial structure of the image rather than applying a uniform change across all pixels.

Table S2. A list of data augmentation techniques provided by the Albumentations library, along with the augmentation strategy in which each technique was applied.

Augmentation \ Strategy	Pixel	Spatial	Both
Horizontal Flip	✗	✓	✓
Motion Blur	✓	✗	✓
Pixel Dropout	✗	✓	✓
Random Brightness and Contrast	✓	✗	✓
Random Shadow	✓	✗	✓
Random Sized BBox Safe Crop	✗	✓	✓
Vertical Flip	✗	✓	✓

## C. Weddell Sea Benthic Dataset High-Resolution Examples

Example Weddell Sea Benthic Dataset test set image outputs. Predicted organism bounding boxes, class labels, and confidence scores shown after reprojection and postprocessing. Confidence threshold = 0.60.

Figure S1. HOTKEY\_2019\_03\_31\_at\_13\_30\_13\_IMG\_0853. Original size: 3799x3798 px.

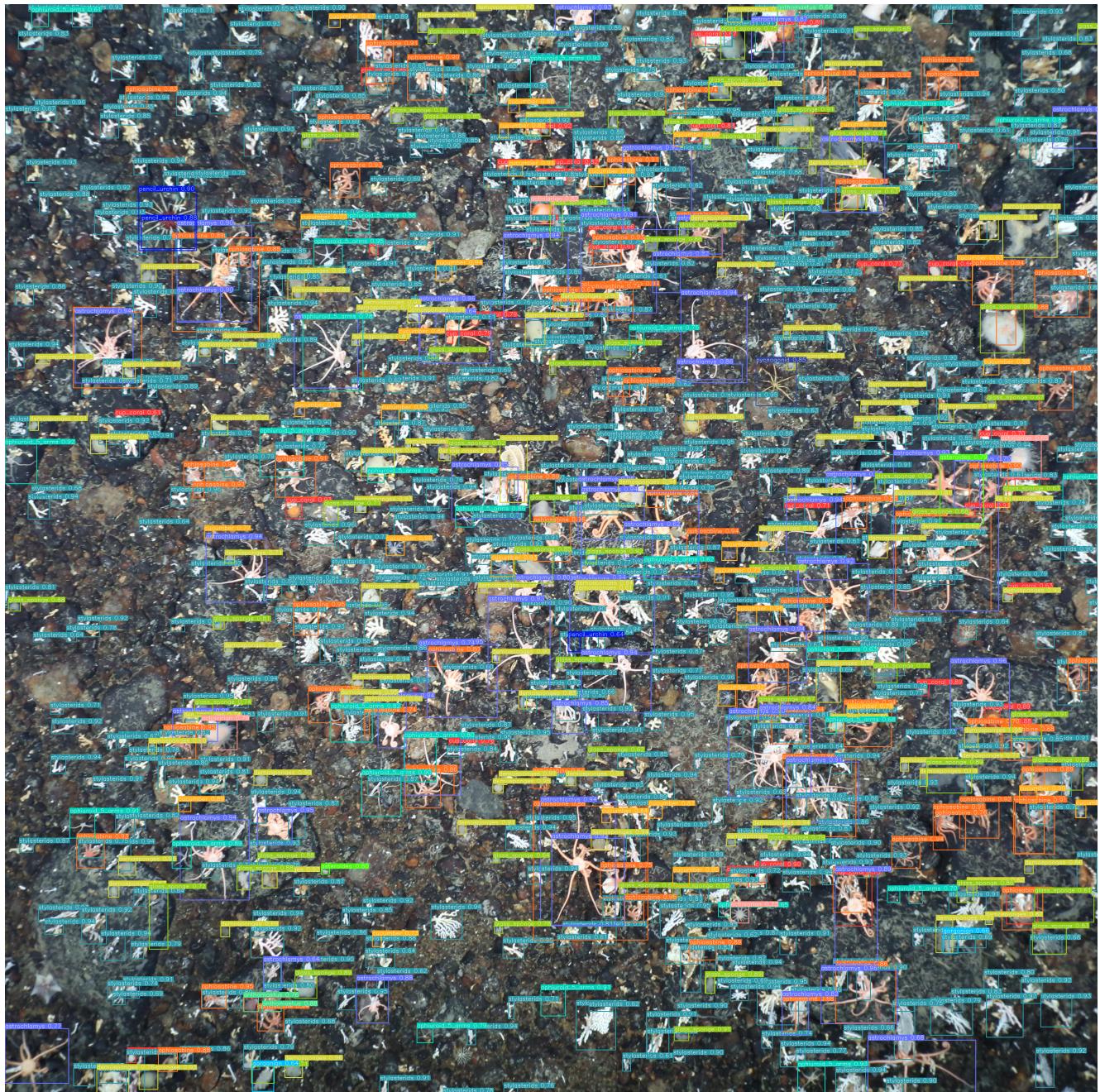


Figure S2. HOTKEY\_2019\_03\_31\_at\_13\_21\_23.IMG\_0816. Original size: 2975x2964 px.

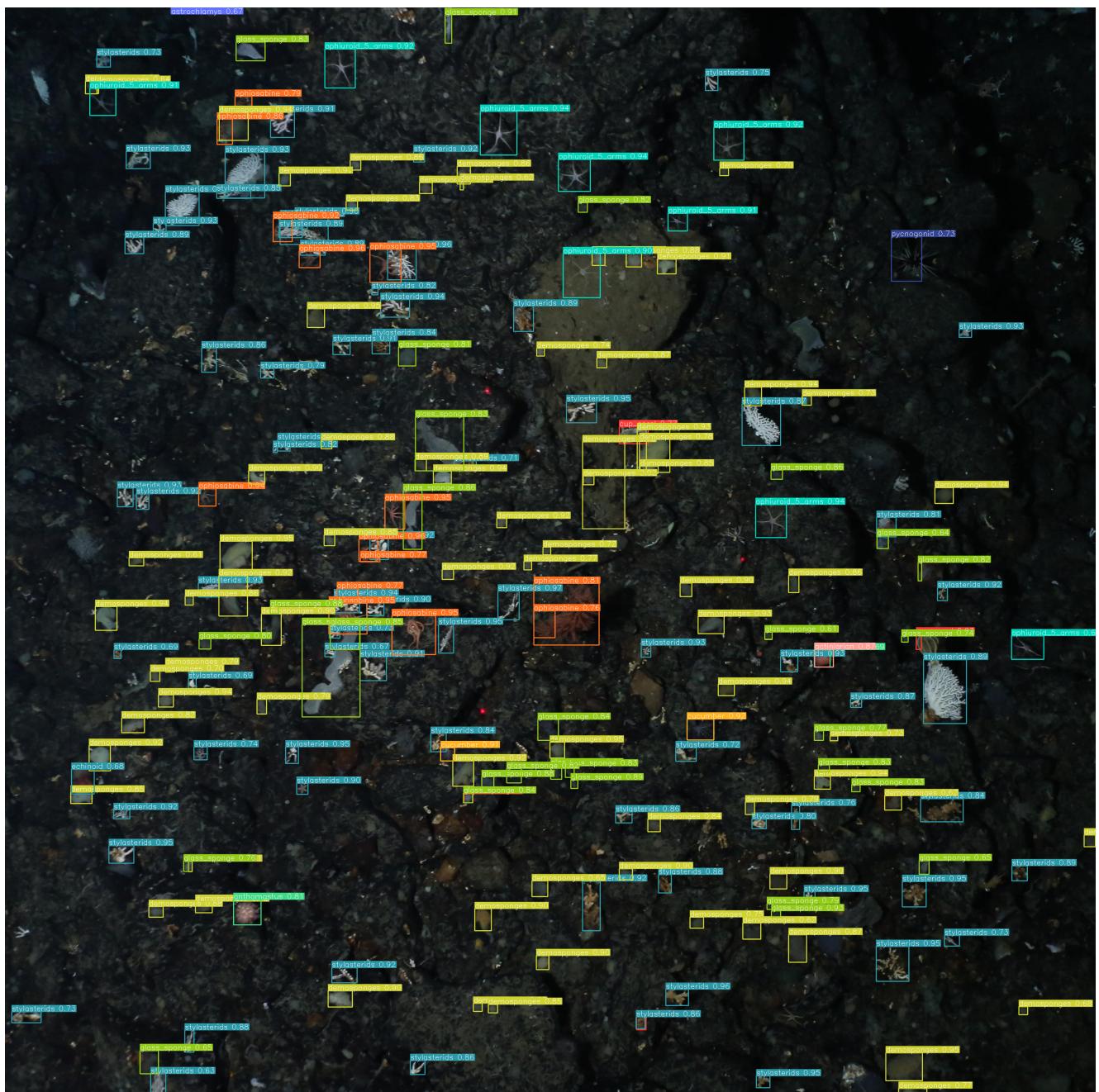
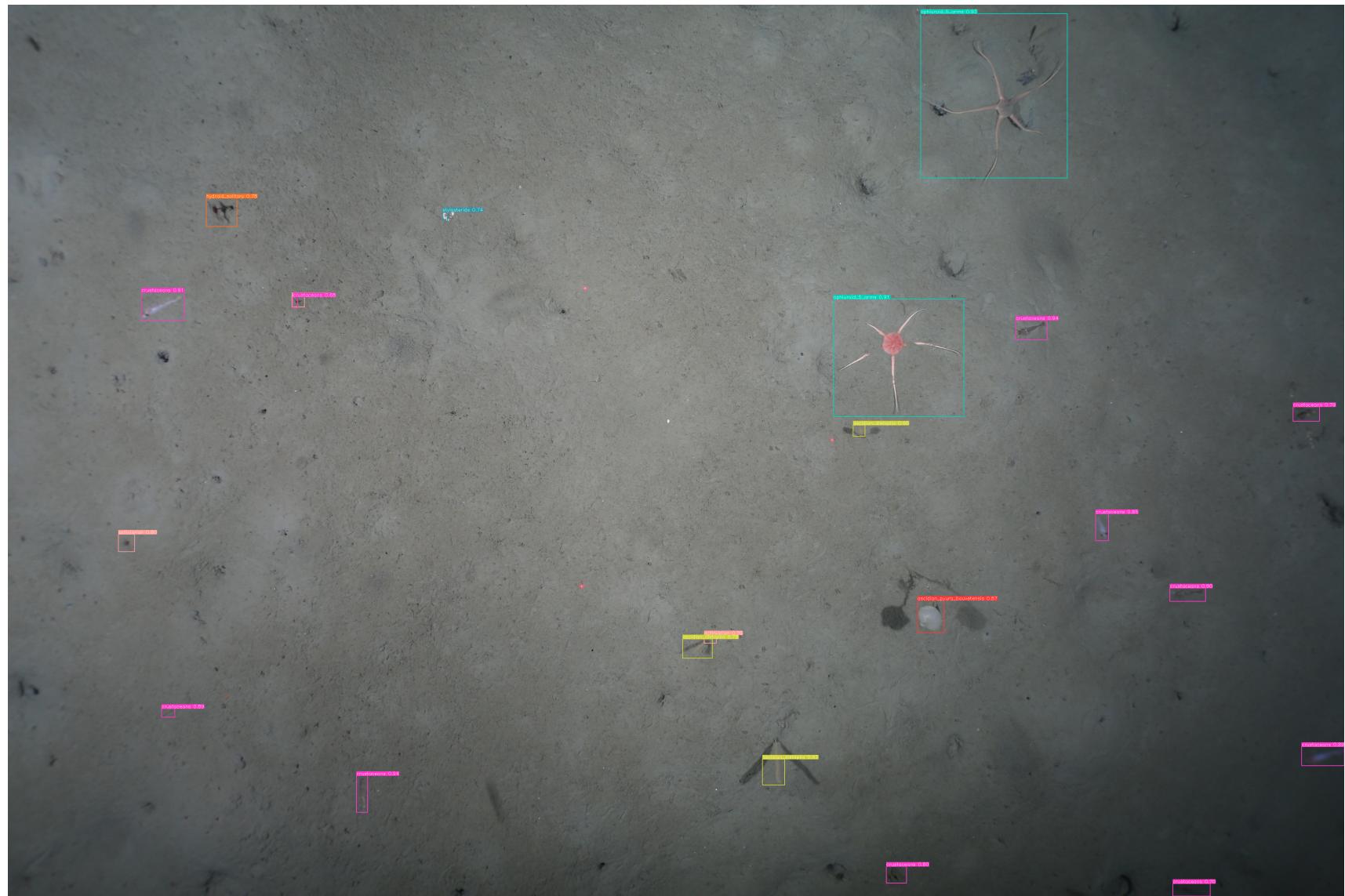


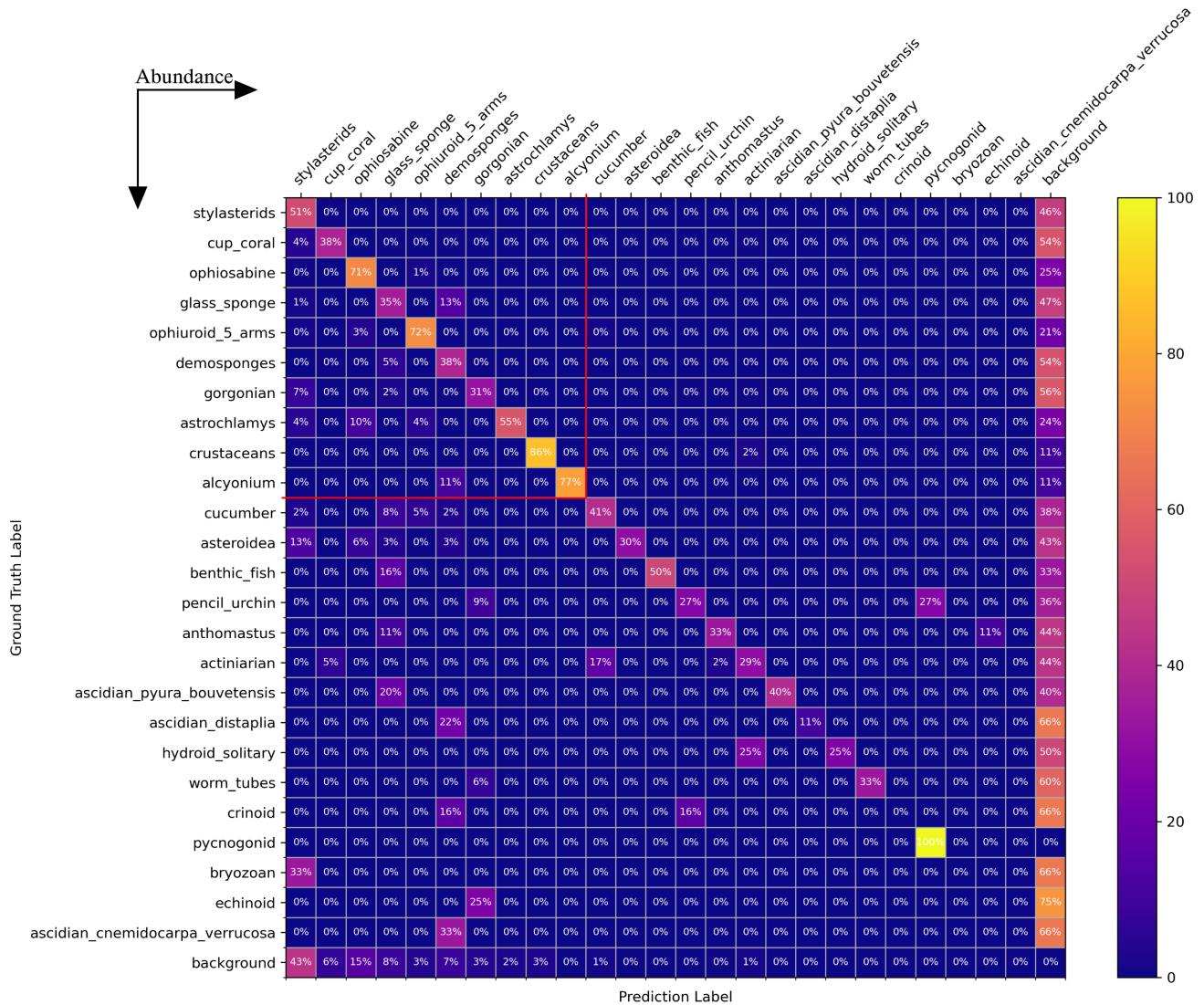
Figure S3. TIMER\_2019\_03\_06\_at\_05\_40\_47\_IMG\_0253. Original size: 5760x3840 px.



## D. Effect of Abundance on Model Performance

Organism abundance was shown to have a large effect on framework performance. See 6.2 for further discussion.

Figure S4. Confusion matrix for the optimal framework configuration, ordered by abundance. Red lines indicate the top-10 most abundant classes. Confidence threshold = 0.60.



## E. SAHI Postprocessing Limitations

Large objects, split over a high number of patches, may fail to merge into a single coherent bounding box after Non-Maximum Merging via SAHI. See [6.3](#) for further discussion. Confidence threshold = 0.60.

Figure S5. A single ophiuroid\_5\_arms, represented by two bounding boxes after postprocessing. Cropped and enlarged for clarity.

