Underwater Uncertainty: A Multi-Annotator Image Dataset for Benthic Habitat Classification

Galadrielle Humblot-Renaux[1,3], Anders Skaarup Johansen[1,3], Jonathan Eichild Schmidt[1], Amanda Frederikke Irlind[2], Niels Madsen[2], Thomas B. Moeslund[1,3], and Malte Pedersen[1,3]

[1] Visual Analysis and Perception Lab [2] Department of Chemistry and Bioscience
[1,2] Aalborg University, Denmark [3] Pioneer Center for AI, Denmark
{gegeh,asjo}@create.aau.dk jonei@dtu.dk {afir,nm}@bio.aau.dk
{tbm,mape}@create.aau.dk

Abstract. Continuous inspection and mapping of the seabed allows for monitoring the impact of anthropogenic activities on benthic ecosystems. Compared to traditional manual assessment methods which are impractical at scale, computer vision holds great potential for widespread and long-term monitoring. We deploy an underwater remotely operated vehicle (ROV) in Jammer Bay, a heavily fished area in the Greater North Sea, and capture videos of the seabed for habitat classification. The collected JAMBO dataset is inherently ambiguous: water in the bay is typically turbid which degrades visibility and makes habitats more difficult to identify. To capture the uncertainties involved in manual visual inspection, we employ multiple annotators to classify the same set of images and analyze time spent per annotation, the extent to which annotators agree, and more. We then evaluate the potential of vision foundation models (DINO, OpenCLIP, BioCLIP) for automating image-based benthic habitat classification. We find that despite ambiguity in the dataset, a well chosen pre-trained feature extractor with linear probing can match the performance of manual annotators when evaluated in known locations. However, generalization across time and place is an important challenge.

Keywords: Seabed monitoring · Benthic habitats · Label uncertainty

1 Introduction

Seabed integrity is challenged by anthropogenic activities causing habitat degradation such as fishing, mining and quarrying, and sustainable energy construction [19, 28, 40]. In the European Union, multiple directives are implemented to protect seabed integrity. The Marine Strategy Framework Directive requires member states to achieve Good Environmental Status (GES) in marine habitats [13] and the Habitat Directive and Nature 2000 focus on preserving species and habitats in Marine Protected Areas [10]. We specifically focus on Jammer Bay in the Greater North Sea (Skagerrak ICES area), a region heavily impacted

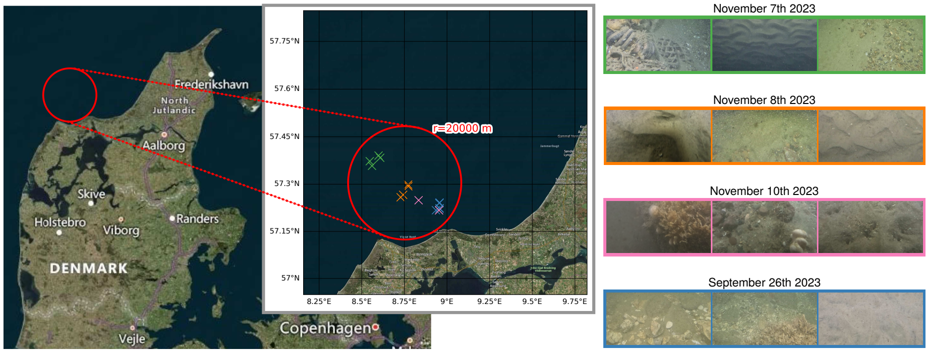


Fig. 1: Satellite view of the 16 data collection locations, color-coded by acquisition date and with image examples on the right. Map data from OpenStreetMap.

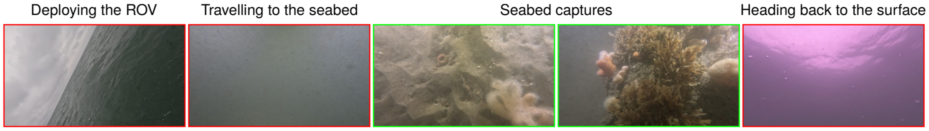


Fig. 2: Illustration of the recording procedure from deploying the ROV, diving to the seabed, capturing video data of the benthic habitats, and, lastly, returning to the surface. Only the data captured at the seabed is used in this project.

by human activities and thus crucial for management. Despite being included in the Marine Strategy Framework Directive and encompassing several Marine Protected Areas, monitoring efforts in this region are limited.

Monitoring the seabed requires comprehensive, systematic and reproducible data collection methods. Compared to sediment sampling which is invasive, costly and highly localized, recording and analysing video footage is a promising technique for marine biologists [1, 6, 7, 25, 34, 45, 47–49]. When combined with computer vision methods, imaging technology provides unprecedented possibilities to enhance and automate underwater monitoring [4, 5, 32].

Developing underwater image classifiers requires labelled data for training and evaluation [41]. While there is a growing number of underwater datasets for species identification [29, 35, 51], publicly available and annotated datasets of seabed habitats are much scarcer. The recently released BenthicNet [31] is a large-scale compilation of seabed imagery from different sources worldwide. However, the majority of labelled images are from Australia and Tasmania, with coral reefs and clear waters [15,26]; the only labelled images from European waters are in Spain. It is important to establish reference datasets in other environments targeted by anthropogenic activities. In Denmark, the main source of visual data for benthic habitats is the marine raw material database (MARTA) [21] - however the images are mostly provided uncurated, unlabelled, or in poor quality.

To address this gap, we present the JAMBO dataset - the first publicly available multi-annotator benthic habitat classification image dataset from temperate waters. We deploy an underwater remotely operated vehicle (ROV) equipped with lights and cameras to capture footage of the seabed in Jammer Bay. Each curated image is labelled by six annotators with different backgrounds and expertise. Their task is to distinguish between sandy substrates and stone reef, or to mark the image as "bad". Analysing their disagreement allows us to evaluate the difficulty and ambiguity of classifying benthic habitat types from ROV images in this area. We then evaluate several baseline models, leveraging pretrained vision transformers for feature extraction. To the best of our knowledge, this constitutes the first study of annotator disagreement and automatic benthic habitat classification in turbid environments resembling Jammer Bay.

2 The JAMBO Dataset

This section provides details on the data collection and annotation procedures. The JAMBO dataset can be found at `https://vap.aau.dk/jambo`.

2.1 Data Acquisition

The study took place in Jammer Bay, a part of the Greater North Sea off the Northwestern coast of Denmark, see Fig. 1. The water depths ranges from 14 to 34 meters at the study sites. The environmental conditions in the area is affected by the North Atlantic drift and currents are generally moving from South to North along the Danish West coast. These conditions, combined with weather and re-suspended particles, produce high degrees of variance in underwater turbidity and visibility. Relevant dive sites were determined by selecting three regions of management importance (commonly frequented by beam trawl, otter trawl, and Danish seine) and with limited habitat knowlegde, coupled with the utilization of a side-scan sonar to pinpoint areas featuring a combination of sandy substrates and stone reefs.

Data was collected with a Chasing M2 Pro Max underwater ROV featuring a 4K resolution front-facing camera solely used for navigation. A GoPro Hero Black 11 camera was mounted beneath the drone at a 45-degree angle to continuously record the seabed. Two 4000 lumen lights were placed in a downwards-facing position, spaced far apart, and angled in a way that ensured the seabed to be illuminated in the field of view of the GoPro camera while minimizing backscatter. Camera settings and a picture of the ROV can be found in Appendix A.1.

Recording was initiated prior to submerging the ROV into the water and continued until it resurfaced at the boat as illustrated in Fig. 2. Therefore, the videos included several stages, from deployment to recovery of the ROV, leading to a significant amount of data not suitable for benthic habitat classification. Each dive lasted 20-30 minutes and approximately seven hours of video data was recorded in total. We conducted manual curation to identify the relevant segments where the ROV was in proximity of the seabed. Subsequently, we carefully reviewed the remaining footage to locate and remove erroneous segments, e.g., if the seabed was not in the camera's field of view due to navigation. A frame was extracted for every two seconds from 12 video segments, resulting in a total of 3290 frames to be annotated. Frames were center-cropped at a resolution of 2048x1024 to exclude lens distortion effects from the image.

2.2 Annotation Scheme

The material that makes up the sediment and substrate of a given habitat plays a vital role in the classification procedure [11] (referred to as level 2 in the European nature information system (EUNIS) hierarchy). The noise and lighting from the ROV may affect the behavior of organisms, potentially leading to escape responses, which complicates classification procedures that consider flora and fauna (level 3 and above in the EUNIS hierarchy). Thus, the aim is to build a pipeline for automatically classifying the primary material of the seabed. However, determining the true habitat type from visual inspection of images from Jammer Bay is not trivial due to the turbid and low light conditions in the area. Therefore, we have defined two general habitat types covering Jammer Bay, based on the EUNIS key navigation [18] and GEUS habitats definitions [30], resulting in the following class labels:

- **Sand** habitats are characterized as primarily sand or muddy sand with less than 5% clay and less than 30% cover of stones/boulders, vegetation, and mussel beds. The habitat is stable but under the influence of tidal streams, waves, and turbidity. This habitat classification is modified from the EUNIS classification *MC52 Atlantic circalittoral sand* [17] and corresponds to the 'sand dunes' habitat as defined by GEUS [30].
- **Stone** reef habitats are characterized by having more than 30% seabed cover of stones and boulders and are to some extent affected by tidal streams, waves, and turbidity. This classification is modified from the EUNIS classification *MC12 Atlantic circalittoral rock* [16] and corresponds to GEUS' 'reef' classification [30].
- **Bad** is a class used to label images that cannot be confidently annotated as containing one of the aforementioned habitat types by the annotator due to poor image quality, turbidity, or similar.

To investigate the uncertainty involved in labeling images of benthic habitats in turbid waters, we asked six annotators to annotate the same set of images. The group of annotators consisted of three marine biologists (*Bio1, Bio2, Bio3*) and three computer vision researchers (*CV1, CV2, CV3*) who were all familiar with the framework for classifying habitats, while the level of background knowledge regarding habitat classification varied among the annotators. One participant, *Bio1*, was a field expert with knowledge about the habitats in the Jammer Bay area and was also acquainted with the videos. The other two marine biologists, *Bio2* and *Bio3*, were not acquainted with the videos but had prior knowledge about marine habitats in general. The three computer vision researchers, *CV1, CV2*, and *CV3*, had viewed and processed the videos during the initial sorting phase but had no experience with marine habitat classification.

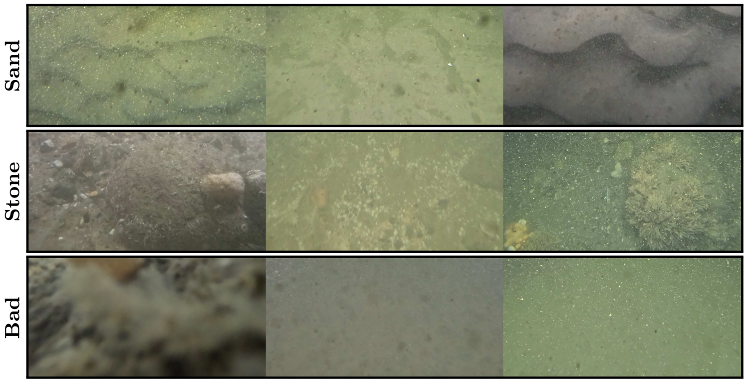


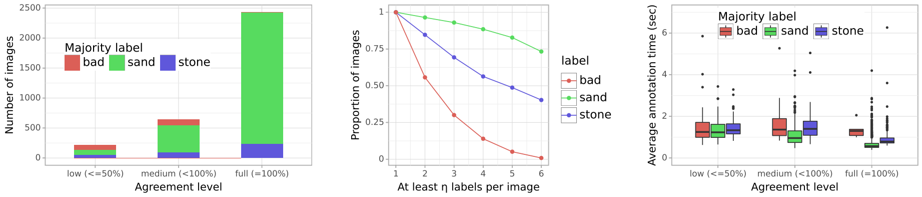
Fig. 3: Dataset samples illustrating variations in turbidity, lighting, and texture of the seabed. These images were unanimously labelled as sand (top), stone (middle) or bad (bottom) by all annotators.

Images were displayed to the annotators at a fixed size of 1920x960, with different monitors used across annotators. For each image, we stored the class label assigned by each of the annotators, as well as the time spent assigning the label. Image examples from the three categories are presented in Fig. 3 showcasing some of the variations encountered within and between the classes.

2.3 Dataset analysis

To gain insights into the characteristics and quality of the collected data, we first perform an analysis of the image content and manual annotations. We investigate the level of agreement between annotators as well as discrepancies between them, shedding light on the ambiguity associated with image-based habitat classification. We quantify the prevalence of sand vs. stone reefs, the proportion of *bad* images, and the time spent on manual annotation.

Annotator Agreement - Each image in our dataset is associated with six labels (one from each annotator). Its majority label is taken as the most frequent label; in cases where 2 or 3 classes appear with the same frequency across annotators, the majority label is chosen at random. We quantify the level of annotator agreement for every image by taking the ratio of annotators who chose the majority label to the total number of annotators involved (six). For example, if three annotators labeled an image as *sand* while the remaining three selected different labels, the agreement level would be 50%. As illustrated in Fig. 4a, complete label agreement was observed for 2431 images (73.89% of the dataset), while high disagreement was noted for 209 images (6.35%). We see a notable prevalence of *sand* habitats, accounting for 82.67% (2720 images) of the data. Conversely, *stone* habitats are less frequently encountered, comprising only 11.88% of the dataset (391 images), resulting in a stone-to-sand ratio of 14.38% (almost 7 sand images for every stone image). A subset of 179 images (5.44%)



(a) Number of images and label distribution for three agreement levels.

(b) Proportion of images with at least η label(s) for a given class.

(c) Annotation time for different agreement levels. Outliers beyond 7 seconds are not shown for readability.

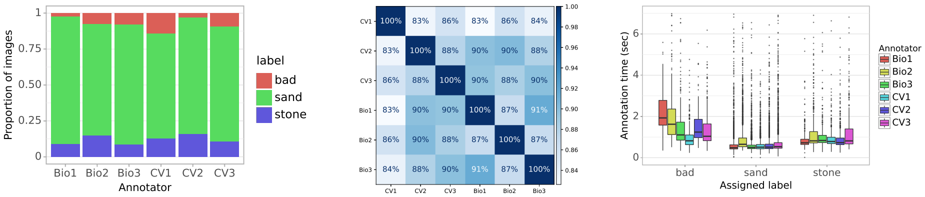
Fig. 4: Quantitative analysis of inter-annotator agreement.

has been categorized as *bad* by the majority of annotators, but only a small number (6) of images were unanimously labeled as *bad*. The graph presented in Fig. 4b elaborates this point by providing an indication of the uncertainty level for each of the classes. The vertical axis represents the proportion of images with respect to the number of images having at least one label of the given class, while the horizontal axis denotes the minimum number of labels present. Notably, approximately 75% of the *sand* images received unanimous labeling by all annotators, whereas for the *stone* images, this proportion is less than 50%. Additionally, the *bad* images show minimal unanimity among annotators.

On average, an image takes less than a second to annotate and *sand* images tend to be the most quickly identified. More time is spent labeling *stone* images, while ambiguous or low-quality images subject to the *bad* class, take the longest time to annotate. Looking at the time spent per label across three agreement levels, we see that images unanimously labeled by all annotators require the least processing time, as depicted in Fig. 4c. This trend holds true for all three classes, with noticeably less variation in the time spent annotating images when everyone agrees compared to when there are discrepancies between annotators.

In summary, this suggests that annotating *stone* images poses a greater challenge than *sand* images. This complexity may arise from uncertainties in assessing variables such as the extent of stone coverage on the seafloor or correctly identifying objects within the frame (is it actually a stone?). Conversely, *sand* images seem to present fewer interpretative difficulties, likely attributable to their relatively clean nature.

Comparing Individual Annotators - We examine the class distribution at the annotator level to identify potential disparities attributable to educational background. Initially, we look at Fig. 5a, which shows the proportion of images that have been assigned the respective classes by each of the six annotators. We see that *Bio1* and *CV2* were the least likely to label an image as *bad*, but had a significant difference in the number of images labeled as *stone*. The same pattern is seen between *Bio2, Bio3* and *CV3* who agree on the number of *bad* images, while the sand-to-stone ratio varies. Generally, the sand-to-stone ratio varies across annotators from 10.17% (*Bio1*) to 19.31% (*Bio2*) and there is no clear link between the class ratios and the background of the annotator.



(a) Proportion of images assigned to each of the three labels across the six annotators.

(b) Percentage of labels matching between pairs of annotators.

(c) Annotation time for different classes. Outliers >7 seconds not shown for readability.

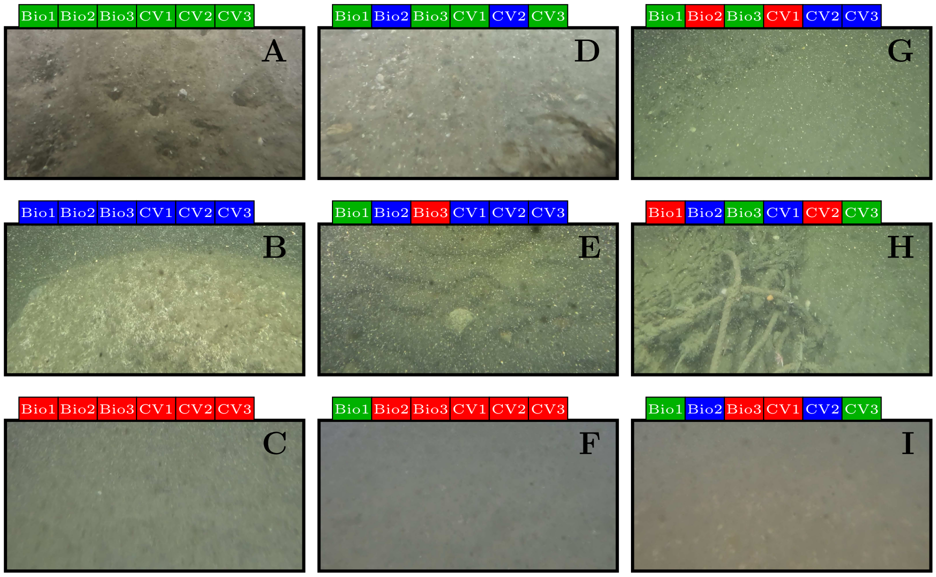
Fig. 5: Quantitative analysis of individual annotation patterns.

Taking a closer look at the pair-wise agreement between the individual annotators, we see the same pattern, which is quantified in Fig. 5b. The general tendency is an agreement level close to 90%, except for *CV1* who is more likely to disagree with the rest, with an average agreement level of 84.26%. However, *CV1* is also the annotator with the highest number of *bad* labels, which could partially explain this pattern. Interestingly, there appears to be no discernible correlation between the background of the annotators and their agreement regarding classifications; annotators with a computer vision (CV) background are not necessarily more likely to agree with each other than with annotators specialized in marine biology (Bio), and vice-versa. Bio1 and Bio3 are the most likely pair of annotators to agree, but no pair exceeds 91% of matching labels.

Besides looking at the distribution of classes, the time spent annotating the images provides further insights regarding the labeling patterns of the individual annotators. In Fig. 5c we can see that the time spent per image varies both depending on the class and the individual annotator. Annotators who are the most conservative with assigning the *bad* label (*Bio1, Bio2* and *CV2*) take the longest time to do so with more than a second on average. On the contrary, *CV1*, who labeled more than 14% of the images as *bad*, typically completed this task in less than one second, suggesting less tolerance for ambiguity. Across all classes and annotators, we see outliers that take significantly longer time to annotate, which suggests a higher label uncertainty for a subset of the images.

Nine concrete examples with varying levels of agreement between the annotators are presented in Fig. 6. The first column contains three images with full agreement. Images (A) and (B) contain little motion blur and include salient features of the seabed, leading to a unanimous labeling of *sand* and *stone*, respectively. In image (A), the seabed is predominantly composed of sand with few small stones, while in image (B), a prominent overgrown stone is visible despite the backscatter caused by resuspended particles. The challenges posed by motion blur, low light conditions, and a scarcity of distinct features render the interpretation of image (C) challenging, resulting in a unanimous *bad* label.

The second column contains images where the majority of the annotators agree on one of the classes. Image (D) contains some motion blur but is otherwise of decent quality with a reasonable amount of light and low turbidity. The image contains a blurred dark object in the lower right corner, smaller stones



(a) Full agreement (100%). **(b)** Medium agreement (≥ 50%). **(c)** Low agreement (< 50%).

Fig. 6: Image examples for different agreement levels across the six annotators. The slots in the colored bars represent the label assigned by *Bio1, Bio2, Bio3, CV1, CV2,* and *CV3*, respectively. The green, blue, and red colors represent sand, stone, and bad.

scattered across the seabed, and no distinct sand ridges. These are all factors that are assumed to contribute to the contradictory labels with four votes on *sand* and two on *stone*. In the second row, image (E) has been labeled as *stone* by the majority, but has also been assigned a *bad* and *sand* label. The label discrepancy is likely due to the image being affected by heavy backscatter, while simultaneously featuring a distinct stone in the center of the image and relatively clear sand ridges. Image (F) is uniformly dark, making it ambiguous whether the background is the seabed or turbid water. Additionally, there are no distinct habitat markers, except for subtle wave-like patterns in the bottom half (possibly sand ridges) and some salt- and pepper-like spots, which can be interpreted either as particles in the water, or small stones or shells on the seabed.

The last column features images with low annotator agreement. The first image, (G), is labeled as *sand* (2) or *bad* (1) by the marine biologists, whereas the annotators with a background in computer vision label it as *stone* (2) or *bad* (1). Most of the image is uniformly sandy colored and there is a decent amount of backscattering particles. However, in the upper left corner, there is a structure that may be interpreted as a formation of rocks, which is assumed to be the cause of the label discrepancy. We defined no label for abandoned fishing nets and this is likely causing some confusion when labeling image (H). The image is of decent quality with good lighting and only little backscatter, however some observers may view part of the structure as being stones or expect the net to be entangled in stones, while others classify the image based on the sandy right part of the image and ignore the net. Image (I) is blurry and contains few features, except for bright spots, which may be interpreted as either stones or shells.

In summary, after examining the dataset's images, and as exemplified by the images in Fig. 6, we identify three primary motivations for assigning the *bad* label, all contributing to the complexity of image-based habitat classification: (1) issues related to image quality such as darkness, blur, or heavy backscatter (2) objects obscuring the seabed (such as ghosts nets) and (3) a lack of salient features. Disagreement between *sand* and *stone* appears to be influenced by different interpretations of what 30% stone coverage looks like. Distinguishing between small stones, suspended particles, or organisms on the seabed can be difficult, especially when images are blurred and lack distinctive features.

3 Benthic Habitat Classification

3.1 Experimental set-up

We describe the models used for classifying JAMBO images along with the evaluation procedure.

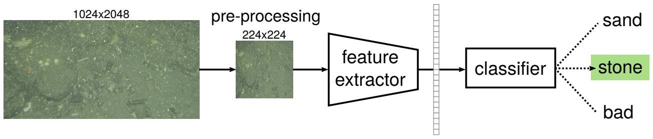


Fig. 7: A pre-trained encoder converts the high-dimensional input image into a 1-dimensional vector capturing its essential characteristics. This feature vector is fed to a linear classifier trained on the JAMBO dataset, which outputs the predicted class.

Models - We apply a three-step pipeline to automatically classify the seabed images, as illustrated in Fig. 7. First, images are resized to 224x224, following standard practice. Second, considering the limited size of our dataset, we leverage pre-trained vision encoders for feature extraction, which we apply out-of-the-box (frozen weights). Following [46], we compare several models for feature extraction, all using a ViT-B/16 vision transformer [14] as image encoder but differing in their training data and objective (implementation details in Appendix A.2):

- **Sup-IN21K** - a model trained with a standard supervised classification objective on ImageNet-21K [12].
- **DINO** [8] - a general vision-only foundation model, trained in a self-supervised manner on ImageNet-1K [43] using self-distillation.
- **OpenCLIP** [24] - open-source version of CLIP [38] trained on LAION-400M [44] (crawled from random web pages). CLIP is a general multimodal foundation model which learns joint representations of image-text pairs using contrastive learning.

- **BioCLIP** [46] - a multimodal foundational model for the biology domain, trained with a species-classification CLIP objective on TreeOfLife-10M.

Lastly, we train a L2-regularized logistic regression classifier (linear probing) to map the frozen features to class logits. We chose linear probing as it gave consistently better results than a k-Nearest Neighbour (k-NN) classifier on this dataset (see Appendix A.4 for a comparison).

Ground truth - The classifier requires well-defined ground truth labels for training and evaluation. However, the ground truth in our dataset is often ambiguous due to disagreement amongst the six annotators leading to no clear-cut and objective ground truth. Therefore, we compare several supervision schemes:

- **Consensus**: only training and evaluating on images with full agreement between annotators, discarding the rest.
- **Majority**: using the entire dataset, with the majority label as ground truth.
- **Expert**: using the entire dataset, taking the annotations of *Bio1*, who has the highest level of expertise, as ground truth.
- **Noisy**: using all the annotators' labels during training instead of a single label per image. That is, every training image is observed six times by the classifier, each time with a potentially different label. Since there is no single ground truth in this scheme, it is excluded from the quantitative evaluation.

Evaluation & Dataset splits - Considering the class imbalance, we evaluate classification performance on unseen images in terms of class-wise and macro F1 score. We first evaluate stratified cross-validation performance with 20 random folds (90% train, 10% test). We then consider 3 more challenging but realistic dataset splits, where images from a specific date (Sep.26, Nov.7 and Nov.10) are held out as a test set while the other dates are used as training. See Appendix A.3.

3.2 Quantitative Evaluation

We benchmark the classification models against individual human annotators on the JAMBO dataset. The results are summarized in Fig. 8 at the class level and Tab. 1 at the test set level. Note that in the consensus scenario, the annotators have perfect performance by definition, as the subset of images has been unanimously labelled by all the annotators. We also note that there is a large variation in classification performance for the *bad* class due to the small number of samples (only six images in the dataset with unanimous *bad* labels).

Looking at cross-validation performance, the classifiers excel at identifying sand and stone images, even surpassing the performance of individual annotators in the majority and expert settings. That is, learning to automate classification from a single annotator can be more accurate than asking a second annotator to manually perform the classification. Models perform the best when following the consensus scheme, aside from the *bad* class (likely due to few *bad* samples). The strong performance was expected, as ambiguous images are excluded from both training and evaluation. Looking at Fig. 8, we see that class-wise performance echoes the annotator agreement and class distribution analysis: *sand* images

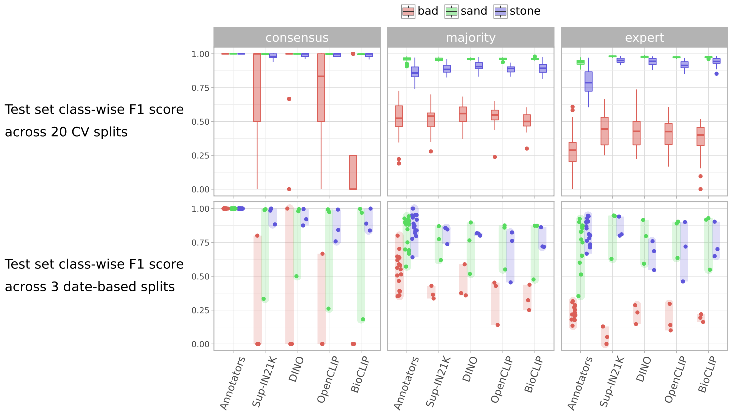


Fig. 8: Class-wise classification performance in the cross-validation setting (top) and on the 3 date-based test sets (bottom). Due to the small sample size, performance on date-based test sets is shown as a scatter plot (one point per test set and model/annotator).

Table 1: 2-class macro F1-score (excluding the *bad* class) in % on different test sets. For the cross-validation (CV) setting, we record the mean score across the 20 test sets. Best score per test set is in bold, and scores greater than the annotators' are underlined.

label scheme	consensus				majority				expert			
test sets	CV	Sep.26	Nov.07	Nov.10	CV	Sep.26	Nov.07	Nov.10	CV	Sep.26	Nov.07	Nov.10
Annotators	*100*	*100*	*100*	*100*	*91.16*	*87.7*	*87.17*	*80.74*	*86.17*	*82.25*	*82.43*	*71.12*
Sup-IN21K	99.03	98.74	**99.75**	60.82	92.25	85.79	75.54	**73.81**	**96.69**	**94.49**	**87.25**	**71.90**
DINO	99.46	99.61	92.75	**71.05**	**93.51**	85.75	78.24	66.63	95.90	74.09	73.08	67.56
OpenCLIP	**99.64**	99.31	90.81	50.92	92.31	81.83	65.70	68.68	94.64	90.18	67.57	67.71
BioCLIP	99.33	**99.90**	92.91	50.98	92.83	**86.75**	**79.66**	59.69	95.73	91.06	81.45	59.85

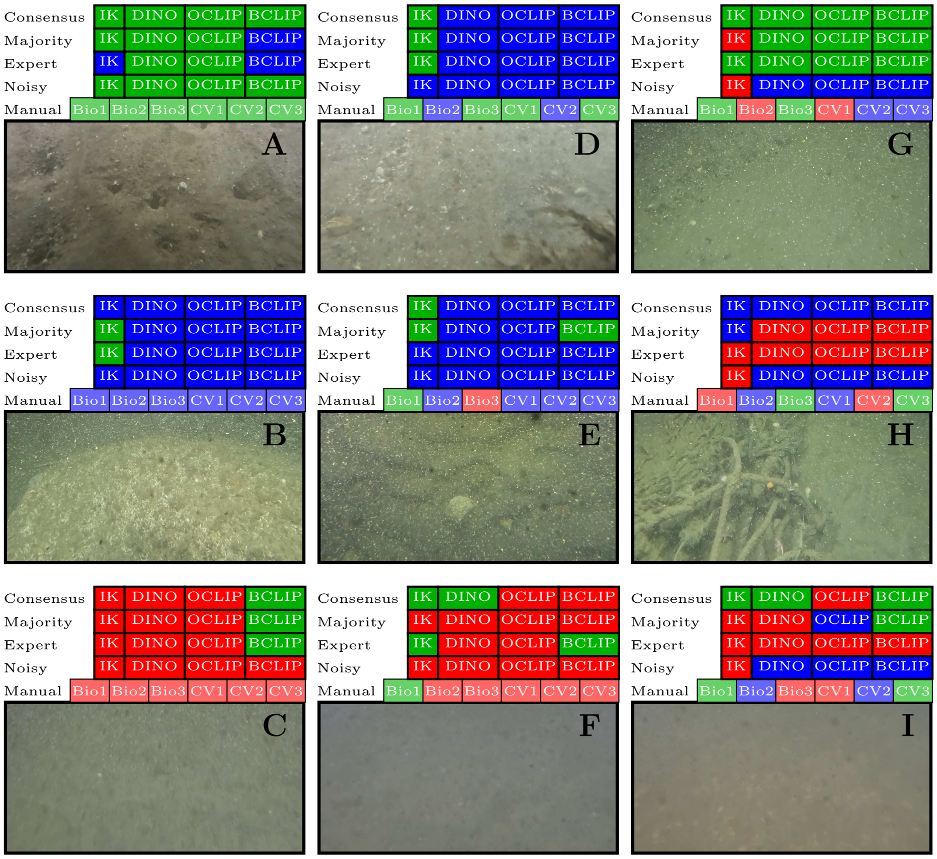
have the highest level of annotator agreement and the highest prevalence in the dataset, and are also the easiest to classify. The performance gap between feature extractors widens for the *stone* and *bad* class, which are rarer and more prone to disagreement. When trained and evaluated on images from different dates, classification performance significantly drops compared to the cross-validation setting. The Nov.10 test set is especially challenging due to the dark, cloudy and crowded images, introducing a strong domain shift (cf. examples in Appendix A.3).

Comparing the four encoders, there is no clear "winner". Although BioCLIP is specifically for the biology domain and includes underwater images, it does not bring consistent performance gains compared to the general-purpose encoders - in fact, it performs the poorest under strong distribution shift (Nov.10 test set), and on the *bad* class (Fig. 8). Sup-IN21K and DINO, both trained on ImageNet, exhibit stronger overall performance on the date-based test sets than CLIP-based models. Sup-IN21K gives the best performance under the expert scheme, while DINO obtains the highest score when averaging across classes, test splits and supervision schemes.

3.3 Qualitative Evaluation

We take a closer look at specific predictions by classifiers trained on the JAMBO dataset. Revisiting the examples from Fig. 6, we hold out these nine images as a test set, and train linear classifiers on extracted features from the rest of the images in the dataset. We obtain sixteen different trained models (four different feature extractors, four supervision schemes) whose predictions on the test set are visualized in Fig. 9. Annotator labels are also shown for reference.

It is interesting to note that although predictions vary across models and supervision schemes, they are rarely unreasonable or unacceptable when compared to the annotators. BioCLIP is the only variant whose predictions does not always match with at least one of the manual annotators: in (C), BioCLIP



(a) Full agreement (100%). **(b)** Medium agreement (≥ 50%). **(c)** Low agreement (< 50%).

Fig. 9: Predictions on images from Fig. 6 by the four models trained on the rest of the dataset with 2 different supervision schemes (noisy labels vs. majority labels). Model names are abbreviated due to space constraints. The green, blue, and red colors represent sand, stone, and bad, respectively.

models predict *sand*, while all annotators agree that the image is *bad* due to poor visibility of the seafloor. As the agreement between annotators decreases, so does the agreement between the sixteen models.

Looking at specific supervision schemes, the noisy label scheme seems to smooth out differences between feature extractors, giving identical predictions across DINO and CLIP-based models. We also observe that under expert supervision, the models' predictions are not necessarily consistent with the manual labels of *Bio1*, despite being trained with the labels of this annotator. This warrants investigating the extent to which features from different pre-trained encoders are able to capture not only the presence of certain objects/cues (e.g. a stone or sand ridge) but also their percentage cover in the image.

4 Discussion

Through our analysis of multi-annotator labels, we identified two key factors contributing to disagreement between annotators: ambiguity in the images and ambiguity in the interpretation of the image content related to the habitat definitions. On one hand, the seabed is not always clearly or fully visible. Currents or movement of the camera may cause motion blur, and turbidity caused by re-suspended particles and organic matter in the water degrades image clarity by scattering the light. Attenuation of the natural light or uneven illumination from artificial lights can obscure details and distort colors. These issues are all amplified as the distance between the camera and seabed increases. On the other hand, even when images are clear, we saw that annotators interpreted the content differently with respect to the *sand*, *stone*, and *bad* categories, leading to conflicting annotations. This was especially true for images featuring both sand and stones (is it *stone* enough?), or images with poor visibility containing salient features indicating the habitat type (is it *bad* enough?). Special cases such as abandoned fishing gear or muddy habitats were also not explicitly included in our label definition, which may have introduced additional doubt.

While the quality of seabed images captured in turbid waters is difficult to improve, ways to reduce some of the disagreement between annotators could be explored. First, labelling instructions could be defined more comprehensively manner and covers unforeseen scenarios, e.g., by including an *other* class to handle special cases. Second, the labelling interface could provide additional visual cues to annotators, e.g., by displaying the image in multiple resolutions/scales to make high or low frequency features more apparent, or including adjacent frames in the form of a short video clip to provide contextual information. Other and more informative annotation schemes could also be considered. For example, regression-based annotation where the *stone* or *sand* coverage is recorded from 0% to 100%, or a more detailed procedure where objects of interest, such as stones, flora, and fauna, are individually segmented at the instance level. Alternatively, an unstructured labelling scheme could be followed, where annotators provide a free-form text description rather than a single pre-defined category per image. Improved annotation schemes may facilitate more precise evaluations of stone and sand reef coverage and size, which is relevant for assessing the condition of the reefs. For instance, the structural complexity of stone reefs offers insights into their role as nurseries, shelters, and foraging sites, aspects that may be overlooked when conducting simple image classification. However, the manual labelling process becomes costlier with the inclusion of additional information, and the automated analysis pipeline grows more complex.

The models we have evaluated for automating benthic habitat image classification are simple, since they only require training a linear classifier, and achieve promising baseline results. However, the JAMBO dataset opens interesting challenges for future work due to heavy class imbalance, distribution shift, and lack of a single objective ground truth. Class imbalance occurs naturally due to the prevalence of sand habitats in the region and the rarity of bad images in curated video clips. Limited generalization to images from unfamiliar locations is also a well-known problem when deploying computer vision systems in the wild [3, 27], which is amplified in the underwater domain due to the cost and scarcity of training data. Observer bias and disagreement, while well-studied in biology [9,22,34,37], is often overlooked in computer vision datasets. An interesting direction is to consider the uncertainty inherent in this dataset as something that is valuable to model [23, 33] and evaluate [2]. This could, for example, be achieved by encoding label uncertainty with soft labels [20] or formulating habitat classification as a multiple hypothesis prediction task [42], where the model can produce multiple plausible interpretations for a single image.

Lastly, while this work focuses on habitat classification from seabed images, an equally relevant task to automate is the curation of raw underwater videos, as only a small fraction of recorded frames actually capture the seabed. Manually sifting through this largely uneventful footage to identify relevant segments is laborious and time-consuming, hindering the adoption of computer vision-based classification. Future research could explore techniques such as keyframe extraction or novelty detection to identify relevant images automatically.

5 Conclusion

We have presented JAMBO, the first publicly available and multi-annotator benthic habitat classification image dataset from the Greater North Sea. This is a valuable resource for researchers and practitioners in the cross-disciplinary fields of seabed mapping and underwater vision. Through detailed analysis, we have not only provided insights into the appearance of benthic habitats in this important area, but also into the uncertainties arising when annotating images of the seabed. Lastly, we have explored the potential of pre-trained vision models for classifying seabed images. These provide a solid baseline without needing to train or fine-tune a feature extractor. However, automatic benthic habitat classification under distribution shift and label ambiguity remains an open problem.

Acknowledgements

We would like to express our gratitude to the volunteer annotators, who made invaluable contributions to this research project. This research was funded by the European Union, the European Maritime and Fisheries Fund (EMFF), and the Ministry of Food, Agriculture, and Fisheries in Denmark under the case number: 33113-B-23-190. The work was also supported by the Pioneer Centre for AI in Denmark. Galadrielle Humblot-Renaux was supported by the Danish Data Science Academy (DDSA), which is funded by the Novo Nordisk Foundation (NNF21SA0069429) and VILLUM FONDEN (40516).

References

1. Aronson, R.B., Edmunds, P.J., Precht, W.F., Swanson, D.W., Levitan, D.R.: Large-scale, long-term monitoring of caribbean coral reefs: Simple, quick, inexpensive techniques. Atoll Research Bulletin **421**, 1–19 (1994). `https://doi.org/10.5479/si.00775630.421.1`
2. Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., Uma, A.: We need to consider disagreement in evaluation. In: Church, K., Liberman, M., Kordoni, V. (eds.) Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future. pp. 15–21. Association for Computational Linguistics, Online (Aug 2021). `https://doi.org/10.18653/v1/2021.bppf-1.3`
3. Beery, S., Cole, E., Gjoka, A.: The iwildcam 2020 competition dataset. arXiv preprint arXiv:2004.10340 (2020)
4. 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.: Towards automated annotation of benthic survey images: Variability of human experts and operational modes of automation. PLOS ONE **10**(7), e0130312 (July 2015). `https://doi.org/10.1371/journal.pone.0130312`
5. Beijbom, O., Treibitz, T., Kline, D.I., Eyal, G., Khen, A., Neal, B., Loya, Y., Mitchell, B.G., Kriegman, D.: Improving automated annotation of benthic survey images using wide-band fluorescence. Scientific Reports **6**(1) (Mar 2016). `https://doi.org/10.1038/srep23166`
6. Buhl-Mortensen, L., Buhl-Mortensen, P., Dolan, M., Gonzalez-Mirelis, G.: Habitat mapping as a tool for conservation and sustainable use of marine resources: Some perspectives from the mareano programme, norway. Journal of Sea Research **100**, 46–61 (2015). `https://doi.org/10.1016/j.seares.2014.10.014`
7. Buhl-Mortensen, L., Ellingsen, K.E., Buhl-Mortensen, P., Skaar, K.L., Gonzalez-Mirelis, G.: Trawling disturbance on megabenthos and sediment in the barents sea: chronic effects on density, diversity, and composition. ICES Journal of Marine Science **73**(suppl_1), i98–i114 (2016). `https://doi.org/10.1093/icesjms/fsv200`
8. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9650–9660 (October 2021)
9. Cherrill, A., McClean, C.: Between-observer variation in the application of a standard method of habitat mapping by environmental consultants in the uk. Journal

of Applied Ecology **36**(6), 989–1008 (1999). https://doi.org/10.1046/j.1365-2664.1999.00458.x
10. Council Directive: 92/43/eec of 21 may 1992 on the conservation of natural habitats and of wild fauna and flora. Official Journal of the European Communities (2013), http://data.europa.eu/eli/dir/1992/43/2013-07-01
11. Davies, C.E., Moss, D., Hill, M.O.: Eunis habitat classification revised 2004. European environment agency-European topic centre on nature protection and biodiversity pp. 127–143 (2004), https://www.eea.europa.eu/data-and-maps/data/eunis-habitat-classification-1/documentation/eunis-2004-report.pdf
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
13. Directive, S.F.: Directive 2008/56/ec of the european parliament and of the council. Journal). Council Decision of (2008)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021). https://doi.org/10.48550/arXiv.2010.11929
15. Dunlap, M.J.: Coral reef photoquads from the main hawaiian islands (2013), https://www.st.nmfs.noaa.gov/aiasi/PIFSC_benthic.html
16. (EEA), E.E.A.: Atlantic circalittoral rock, https://eunis.eea.europa.eu/habitats/30972, accessed on 12-03-2024
17. (EEA), E.E.A.: Atlantic circalittoral sand, https://eunis.eea.europa.eu/habitats/31202p, accessed on 12-03-2024
18. (EEA), E.E.A.: Habitat types key navigation, https://eunis.eea.europa.eu/habitats-key.jsp2, accessed on 12-03-2024
19. Frigstad, H., et al.: Three decades of change in the skagerrak coastal ecosystem, shaped by eutrophication and coastal darkening. Estuarine, Coastal and Shelf Science **283**, 108193 (2023). https://doi.org/10.1016/j.ecss.2022.108193
20. Gao, B.B., Xing, C., Xie, C.W., Wu, J., Geng, X.: Deep label distribution learning with label ambiguity. IEEE Transactions on Image Processing **26**(6), 2825–2838 (2017). https://doi.org/10.1109/TIP.2017.2689998
21. Geological Survey of Denmark and Greenland (GEUS): Seabed sediment map, https://eng.geus.dk/mineral-resources/danish-raw-materials/seabed-sediment-map, accessed on 16-02-2024
22. Hearn, S., Healey, J., McDonald, M., Turner, A., Wong, J., Stewart, G.: The repeatability of vegetation classification and mapping. Journal of environmental management **92**(4), 1174–1184 (2011). https://doi.org/10.1016/j.jenvman.2010.11.021
23. Herde, M., Huseljic, D., Sick, B.: Multi-annotator deep learning: A probabilistic framework for classification. Transactions on Machine Learning Research (2023). https://doi.org/10.48550/arXiv.2304.02539
24. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). https://doi.org/10.5281/zenodo.5143773, https://doi.org/10.5281/zenodo.5143773, if you use this software, please cite it as below.
25. Jac, C., Desroy, N., Duchêne, J.C., Foveau, A., Labrune, C., Lescure, L., Vaz, S.: Assessing the impact of trawling on benthic megafauna: comparative study of video

surveys vs. scientific trawling. ICES Journal of Marine Science **78**(5), 1636–1649 (April 2021). https://doi.org/10.1093/icesjms/fsab033

26. Jackett, C., Althaus, F., Maguire, K., Farazi, M., Scoulding, B., Untiedt, C., Ryan, T., Shanks, P., Brodie, P., Williams, A.: 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 (Apr 2023). https://doi.org/10.1111/1365-2664.14408
27. Katija, K., Orenstein, E., Schlining, B., Lundsten, L., Barnard, K., Sainz, G., Boulais, O., Cromwell, M., Butler, E., Woodward, B., et al.: Fathomnet: A global image database for enabling artificial intelligence in the ocean. Scientific reports **12**(1), 15914 (2022)
28. Kenny, A.J., Jenkins, C., Wood, D., Bolam, S.G., Mitchell, P., Scougal, C., Judd, A.: Assessing cumulative human activities, pressures, and impacts on north sea benthic habitats using a biological traits approach. ICES Journal of Marine Science **75**(3), 1080–1092 (2018). https://doi.org/10.1093/icesjms/fsx205
29. Khan, F.F., Li, X., Temple, A.J., Elhoseiny, M.: Fishnet: A large-scale dataset and benchmark for fish recognition, detection, and functional trait prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20496–20506 (October 2023)
30. Leth, J., Jensen, D., Lomholt, S., Al-Hamdani, Z., Nørgaard-Pedersen, N., Jensen, J., Andresen, C., Tougaard, L., von Platen-Hallermund, F., Andersen, M., Larsen, B.: Nyt digitalt havbundssedimentkort. Geoviden - Geologi og Geografi **2014**(2), 9–11 (2014)
31. Lowe, S.C., Misiuk, B., Xu, I., Abdulazizov, S., Baroi, A.R., Bastos, A.C., Best, M., Ferrini, V., Friedman, A., Hart, D., Hoegh-Guldberg, O., Ierodiaconou, D., Mackin-McLaughlin, J., Markey, K., Menandro, P.S., Monk, J., Nemani, S., O'Brien, J., Oh, E., Reshitnyk, L.Y., Robert, K., Roelfsema, C.M., Sameoto, J.A., Schimel, A.C.G., Thomson, J.A., Wilson, B.R., Wong, M.C., Brown, C.J., Trappenberg, T.: Benthicnet: A global compilation of seafloor images for deep learning applications (2024), https://arxiv.org/abs/2405.05241
32. Misiuk, B., Brown, C.J.: Benthic habitat mapping: A review of three decades of mapping biological patterns on the seafloor. Estuarine, Coastal and Shelf Science **296**, 108599 (January 2024). https://doi.org/10.1016/j.ecss.2023.108599
33. Mostafazadeh Davani, A., Díaz, M., Prabhakaran, V.: Dealing with disagreements: Looking beyond the majority vote in subjective annotations. Transactions of the Association for Computational Linguistics **10**, 92–110 (2022). https://doi.org/10.1162/tacl_a_00449
34. Ninio, R., Delean, S., Osborne, K., Sweatman, H.: Estimating cover of benthic organisms from underwater video images: variability associated with multiple observers. Marine Ecology Progress Series **265**, 107–116 (2003). https://doi.org/10.3354/meps265107
35. Pedersen, M., Haurum, J.B., Gade, R., Moeslund, T.B., Madsen, N.: Detection of marine animals in a new underwater dataset with varying visibility. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
36. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011). https://doi.org/10.48550/ARXIV.1201.0490

37. Poole, G.C., Frissell, C.A., Ralph, S.C.: In-stream habitat unit classification: inadequacies for monitoring and some consequences for management. Journal of the American Water Resources Association **33**(4), 879–896 (1997). https://doi.org/10.1111/j.1752-1688.1997.tb04112.x
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021)
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html
40. Reker, J., Murray, C., Gelabert, E.R., Abhold, K., Korpinen, S., Peterlin, M., Vaughan, D., Andersen, J.: Marine messages ii: Navigating the course towards clean, healthy and productive seas through implementation of an ecosystem-based approach. EEA Topic Report (2019). https://doi.org/ReportNo17/2019,at34âĂŞ44
41. Rubbens, P., et al.: Machine learning in marine ecology: an overview of techniques and applications. ICES Journal of Marine Science **80**(7), 1829–1853 (08 2023). https://doi.org/10.1093/icesjms/fsad100, https://doi.org/10.1093/icesjms/fsad100
42. Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., Hager, G.D.: Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 3611–3620 (2017). https://doi.org/10.1109/ICCV.2017.388
43. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
44. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Jitsev, J., Komatsuzaki, A.: LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. In: Proceedings of NeurIPS Data-Centric AI Workshop (2021)
45. Sheehan, E.V., Vaz, S., Pettifer, E., Foster, N.L., Nancollas, S.J., Cousens, S., Holmes, L., Facq, J.V., Germain, G., Attrill, M.J.: An experimental comparison of three towed underwater video systems using species metrics, benthic impact and performance. Methods in Ecology and Evolution **7**(7), 843–852 (2016). https://doi.org/10.1111/2041-210x.12540
46. Stevens, S., Wu, J., Thompson, M.J., Campolongo, E.G., Song, C.H., Carlyn, D.E., Dong, L., Dahdul, W.M., Stewart, C., Berger-Wolf, T., Chao, W.L., Su, Y.: Bioclip: A vision foundation model for the tree of life. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19412–19424 (June 2024)
47. Taormina, B., Marzloff, M.P., Desroy, N., Caisey, X., Dugornay, O., Metral Thiesse, E., Tancray, A., Carlier, A.: Optimizing image-based protocol to monitor macroepibenthic communities colonizing artificial structures. ICES Journal of Marine Science **77**(2), 835–845 (2020). https://doi.org/10.1093/icesjms/fsz249

48. Tiano, J.C., van der Reijden, K.J., O'Flynn, S., Beauchard, O., van der Ree, S., van der Wees, J., Ysebaert, T., Soetaert, K.: Experimental bottom trawling finds resilience in large-bodied infauna but vulnerability for epifauna and juveniles in the frisian front. Marine Environmental Research **159**, 104964 (2020). https://doi.org/10.1016/j.marenvres.2020.104964
49. Williams, A., Althaus, F., Schlacher, T.A.: Towed camera imagery and benthic sled catches provide different views of seamount benthic diversity. Limnology and Oceanography: Methods **13**(2), 62–73 (2015). https://doi.org/10.1002/lom3.10007
50. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020), https://www.aclweb.org/anthology/2020.emnlp-demos.6
51. Šiaulys, A., Vaičiukynas, E., Medelytė, S., Olenin, S., Šaškov, A., Buškus, K., Verikas, A.: A fully-annotated imagery dataset of sublittoral benthic species in svalbard, arctic. Data in Brief **35**, 106823 (Apr 2021). https://doi.org/10.1016/j.dib.2021.106823

A Appendix

This appendix contains supporting figures and implementation details for reproducibility.

A.1 Data collection set-up

Fig. 10 shows the ROV used to collect the JAMBO dataset, and Fig. 11 shows a screenshot of the labelling interface used by all annotators. The ROV's GoPro was configured to record with a frame rate of 60 FPS, a resolution of 3840x2160 pixels, and using the *linear* digital lens setting with a field of view of 92°, 61°, and 100° in the horizontal, vertical, and diagonal direction, respectively.

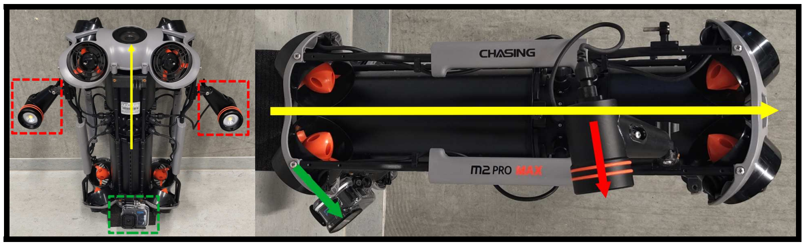
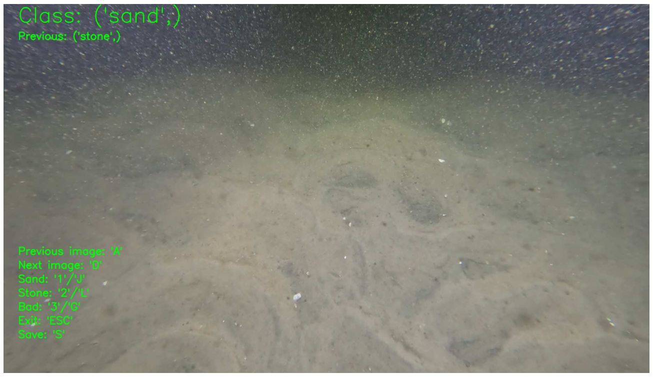


Fig. 10: The underwater ROV captured from two different angles. The position and orientation of the camera and light sources are highlighted by green and red markings, respectively. The yellow arrows indicate the forward orientation of the drone.



Class: ('sand',)
Previous: ('stone',)

Previous image: 'A'
Next image: 'D'
Sand: '1'/'J'
Stone: '2'/'L'
Bad: '3'/'G'
Exit: 'ESC'
Save: 'S'

Fig. 11: Screenshot of the labelling interface.

A.2 Classification pipeline

Before being fed to a pre-trained feature extractor, images are resized to a resolution of 224 × 224 using bilinear interpolation. The baseline models in our experiments were implemented as follows:

- The Sup-IN21K and DINO models are taken from the HuggingFace library [50] (model cards `google/vit-base-patch16-224-in21k` and `facebook/dino-vitb16` respectively).
- The OpenCLIP and BioCLIP models are taken from the OpenCLIP library (`open_clip.create_model_and_transforms('ViT-B-16', pretrained='laion400m_e32')` and `open_clip.create_model_and_transforms('hf-hub:imageomics/bioclip')` respectively).
- For all four pre-trained backbones, a 1-dimensional feature vector of size 768 is extracted by taking the pooling output (we do not apply the projection layer in the CLIP models).
- The logistic regression classifier is trained using sklearn's L-BFGS implementation [36] with cross-entropy loss as training objective, and a maximum of 1000 iterations (similarly to [39]). Samples are inversely weighted based on class frequency to account for the heavy class imbalance in our dataset. The inverse regularization parameter C is kept to its default value of 1.0 (cf. Appendix A.4).
- As an alternative to logistic regression, the k-Nearest Neighbour classifier is trained using sklearn's implementation [36] based on Euclidean distance. Features are zero-centered based on training data statistics and normalized before being fed to the classifier. The parameter K is kept to its default value of 5 (cf. Appendix A.4).

A.3 Dataset splits

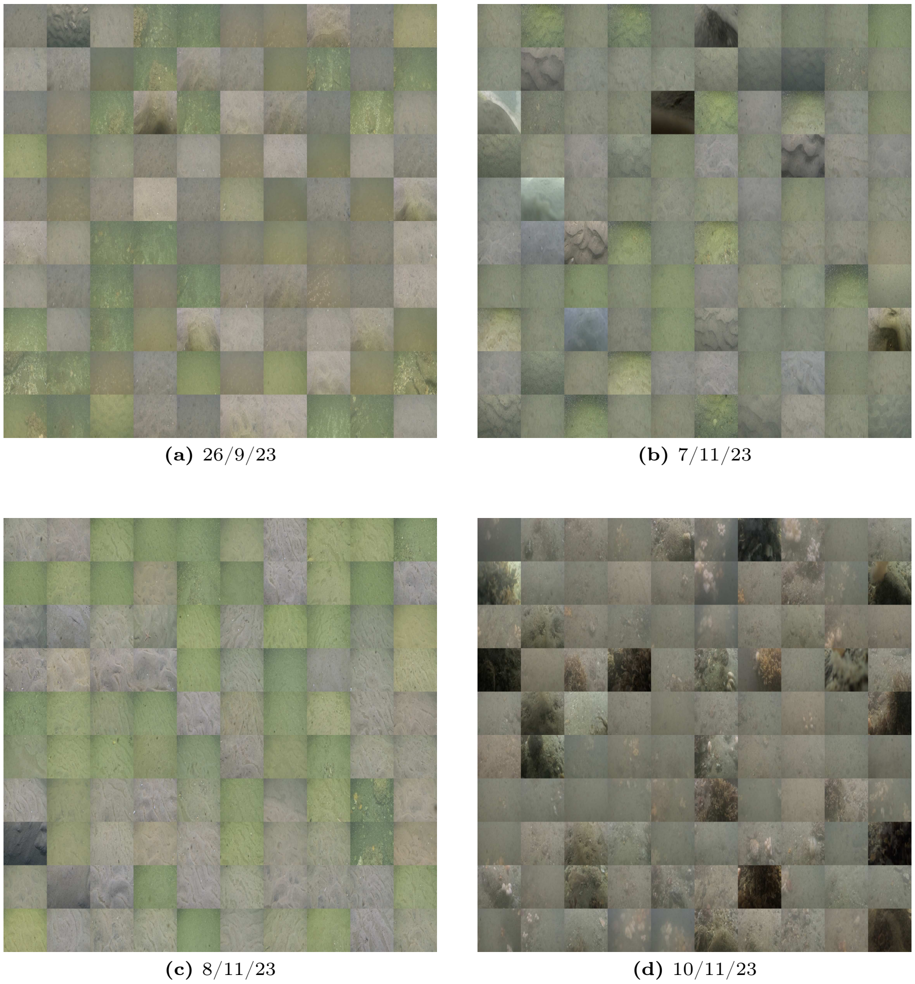
Here we describe the train/test splits used in the 2 benthic habitat classification experiments. Cross-validation is less challenging, since the test set can contain images taken at a very similar location to images in the training set. The date-based splits are designed to evaluate generalization to new locations.

Cross-validation - We apply random stratified cross-validation as implemented by sklearn [36] via the `StratifiedShuffleSplit` method, with 20 splits and a test size of 10%. That is, 20 different train/test splits are created randomly with 90% of images in the train set and 10% in the test set, while also preserving the original class distribution in each set. Using randomized cross-validation rather than K-fold cross-validation ensures that the minority *bad* class has at least one example per train and test test, despite there only being six images unanimously labelled as *bad* in the whole dataset. This means that some test examples are repeated across different splits.

Date-based test splits - As illustrated in Fig. 1, videos were collected on 4 different days, covering a different area each day. To evaluate model generalization, we create three train/test splits by holding out data from a specific day (cf. Tab. 2). Images from the 8/11 is kept as training data in all three splits due to the large number of images collected on that day. Examples from each day are shown in Fig. 12.

Table 2: Number of dataset images for each data collection day, and the three dataset splits used to evaluate generalization.

date	26/9/23	7/11/23	8/11/23	10/11/23	**total**
num. images	710	289	2189	102	**3290**
split Sep.26	**test**	train	train	train	
split Nov.07	train	**test**	train	train	
split Nov.10	train	train	train	**test**	



(a) 26/9/23

(b) 7/11/23

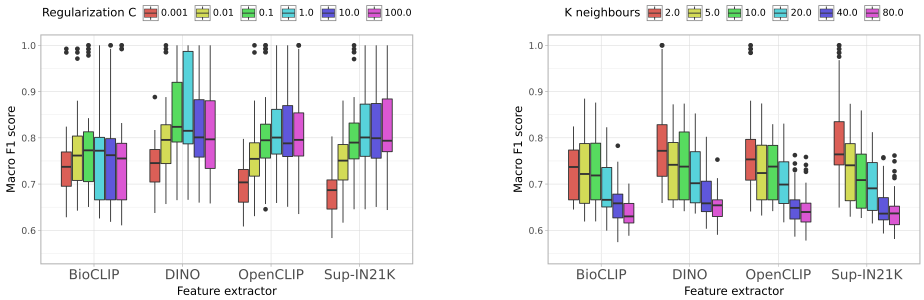
(c) 8/11/23

(d) 10/11/23

Fig. 12: 100 randomly selected dataset images from each data collection day, after being resized to 224x224.

A.4 Classifier hyper-parameters

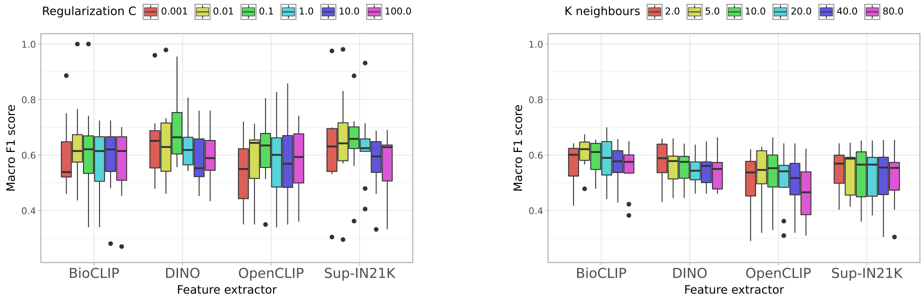
Here we zoom in to the choice of classifier hyperparameters: C in the logistic regression classifier determines the inverse of the regularization strength (stronger regularization as C decreases), and K in the KNN classifier determines the number of neighbours used for majority voting (larger K leads to smoother decision boundary). We sweep 6 values of C and K and record classification performance on the cross-validation splits in Fig. 13 and the date-based splits in Fig. 14.



(a) Logistic regression classifier

(b) KNN classifier

Fig. 13: Effect of the classifier's hyperparameter values on test set classification performance across the **20 cross-validation splits** and across the 3 supervision schemes ($20 \times 3 = 60$ points per boxplot).



(a) Logistic regression classifier

(b) KNN classifier

Fig. 14: Effect of the classifier's hyperparameter values on test set classification performance across the **3 date-based test splits** and across the 3 supervision schemes ($3 \times 3 = 9$ points per boxplot).