Benthic Benchmark: AI-Powered Optical Survey for Benthic Ecosystem Modeling

Heng Lian[1], Daniel M. Runfola[1], Roger Mann[2], Deborah Hart[3], Yi He[1]

[1]Department of Data Science, William & Mary
[2]Virginia Institute of Marine Science, William & Mary
[3]Northeast Fisheries Science Center, NOAA Fisheries
hlian01@wm.edu, dsmillerrunfol@wm.edu, rmann@vims.edu, deborah.hart@noaa.gov, yihe@wm.edu

Abstract

Long-term monitoring of benthic ecosystems is essential for understanding marine biodiversity. Traditional methods such as trawling and physical sampling are invasive, costly, and inefficient. Although deep learning methods offer a promising non-invasive alternative for automated monitoring, most existing approaches rely on datasets unrelated to benthic species and lack systematic evaluation in realistic underwater environments. To address this gap, we present **Benthic Benchmark**, a large-scale, high-resolution underwater image dataset covering seven representative benthic species, under three task settings: image classification, object detection, and image recognition using large multimodal models. We systematically evaluate 14 representative model architectures, such as ResNet, ViT, YOLO-v11, DETR, SAM, LLaVA, and identify three common challenges in benthic imagery: poor visibility, incomplete object integrity, and lack of size standardization. This benchmark serves as a standardized platform for marine visual understanding and facilitates future research on multimodal approaches to ecological monitoring.

1 Introduction

The benthos play a vital role in maintaining biodiversity, contributing to nutrient cycling, stabilizing sediments, preventing erosion, and maintaining seabed structure (Kristensen et al. 2012). Estimating the spatial distribution of benthic populations is essential for assessing and sustaining the health of marine ecological-economic infrastructure. To wit, in fiscal year 2022, commercial landings of Atlantic sea scallop yielded 31.6 million pounds of meats and generated $478 million in revenue, according to the National Oceanic and Atmospheric Administration Fisheries (NOAA 2022). Traditional benthic survey methods, such as dredging (Hart and Rago 2006), will physically disturb the seabed and hence be destructive to benthic habitats (Chang, Shank, and Hart 2017). Their utility is further stretched by limited spatial coverage and high labor costs, as they rely on point sampling rather than continuous observation. In response, underwater optical survey (UOS) systems have emerged as non-invasive alternatives for benthic monitoring (Davis, Gallager, and Solow 1992; Gallager et al. 2005; Taylor et al. 2008; Singh, Örnólfsdóttir, and Stefansson 2013).

A standard UOS pipeline involves acquiring high-resolution seafloor imagery and subsequently identifying and enumerating benthic organisms from these images.

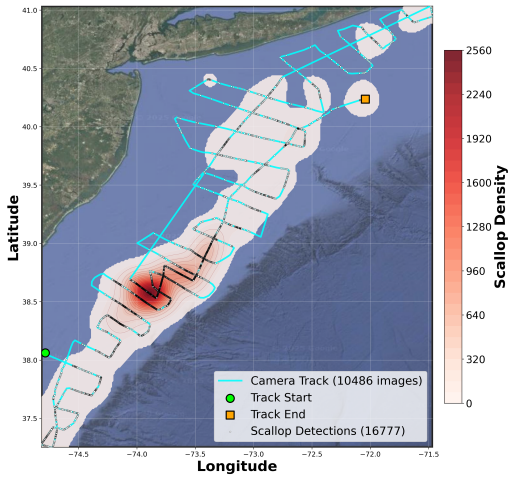


Figure 1: Comparison of scallop distribution across the Mid-Atlantic Bight using human annotations (blue stripes) and AI-powered UOS (heatmap). A rough consistency is observed between the two, motivating this study: *while AI systems offer scalability for processing vast volumes of seafloor imagery, how closely can their outputs align with those of trained human annotators?*

Alas, this process is currently performed by trained human annotators, making it difficult to scale across expansive marine regions that routinely yield millions of images (Chang et al. 2016; Roman, Lin, and Rudders 2024). Recent advances in artificial intelligence (AI) have led to an abundance of computer vision (CV) algorithms, spanning image classification, object recognition, and semantic segmentation, which are seemingly plausible solutions to automate the UOS pipeline. For example, as shown in Figure 1, human annotation typically relies on stratified sampling along predefined transect stripes, where each stripe serves as a sampling stratum. Abundance estimates are then extrapolated using autocorrelation-based techniques (Chang et al. 2016; Chang, Shank, and Hart 2017). In contrast, we post-train YOLOv11 (Jocher and Qiu 2024) using only 2,000 labeled scallop images, evaluate its performance on over 15,000 images spanning a much broader spatial extent than the original sampled strata. We note a rough spatial consistency between the two distributions.

No.	Dataset	# Images	# Classes	Overlapped
1	Pascal VOC	10K+	20	None
2	CIFAR-100	60K	100	Crab, Flatfish
3	COCO	330K	80	Crab
4	Open Images	9M+	600	Crab
5	ImageNet	14M+	1000	Crab
6	Marine Animal	806	9	None
7	SUIM	1,635	8	None
8	URPC2021	8,200	4	Scallop
9	WildFish	54,459	1000	None
10	Sea Animals	13K+	23	Eel, Crab

Table 1: Summary of common public vision datasets and marine-focused datasets regarding marine-related categories and overlapping species with our datasets.

Motivated by this, we propose *Benthic Benchmark*, aiming to field-test the effectiveness and standardize the use of representative vision algorithms in automating benthic organism detection and counting with high accuracy. We collaborate with NOAA Fisheries to curate a large-scale collection of over 720 thousand benthic images, spanning over 10 years and covering two ecologically and commercially significant regions, namely, Mid-Atlantic Bight and Georges Bank. These images are acquired using Habitat Mapping Camera (HabCam), a towed dual-camera system that captures high-resolution (2.7K) seafloor images in real time (30 frames per second). We identify seven economically vital species, including scallop, crab, whelk, eel, skate fish, flatfish, and roundfish. These species pose varying levels of difficulty for current AI algorithms in terms of detection, classification, and segmentation, presenting a diverse and realistic challenge for automated benthic analysis.

2 Existing Testbeds and Gap

Recent advances in automated image analysis have led to the widespread adoption of both traditional computer vision models, like image classification and object detection models (Ren et al. 2015; He et al. 2016; Redmon et al. 2016), and large multi-modal models (LMMs) (Touvron et al. 2023; Liu et al. 2023), with impressive performance demonstrated across various domains. However, their effectiveness in benthic imagery analysis remains largely unexplored and uncertain. A comprehensive evaluation of existing models on benthic imagery is essential, serving two purposes: 1) Select the optimal CV model architectures for the benthic modeling tasks; and 2) Verify if LMMs are more effective modeling alternatives before we scale them up.

Standard CV models are now open-source and ready to use, their performance largely depends on the data they were trained on. However, as summarized in Table 1, existing public datasets have very little overlap with benthic imagery, and most of our seven target species are not included. Given this gap, even if a model can extract useful features from benthic images, its classifier is likely to make random or incorrect predictions due to the lack of relevant training data. To address this, it is necessary to build a dedicated benthic dataset, retrain representative CV models on it, and evaluate their performance in order to identify the most suitable model architectures. However, no prior work has conducted a comprehensive evaluation of whether these models can accurately capture the features of benthic imagery, how well they perform in this domain, and finally which architectural designs are most effective for this benthic task.

Large multi-modal models trained on vast and heterogeneous internet-scale datasets (Radford et al. 2021; Kirillov et al. 2023), offer a compelling alternative to conventional computer vision approaches. LMMs are therefore considered capable of showing strong generalization across various downstream tasks, such as directly recognizing benthic organisms without additional fine-tuning. Recent studies (Cheng et al. 2024) also suggest that LMMs possess open-vocabulary capabilities, allowing them to identify previously unseen objects through descriptive prompts; for example, a scallop may be recognized as "an orange, fan-shaped marine organism." Despite these advances, the deployment cost of LMMs is substantially higher than traditional CV models; for example, deploying LLaMA2-70B requires around 42,000 GPU hours, costing approximately $218K (Agrawal et al. 2024). Therefore, it is necessary to first evaluate their suitability for benthic image recognition and determine which model family offers better performance in this domain before deployment. However, no existing work has systematically evaluated the recognition performance of LMMs on benthic imagery, nor examined whether their open-vocabulary capability remains effective for benthic species as well.

3 Problem Statement and Challenges

To effectively evaluate CV models on benthic organism imagery, two major issues must be addressed: the lack of a large, well-annotated dataset tailored to benthic species, and the unique visual characteristics of seafloor imagery that complicate model assessment.

First, constructing a comprehensive dataset of representative benthic species is a fundamental challenge due to the large proportion of seafloor images containing no visible organisms and the significant manual effort required for accurate annotation. Most existing CV models ranging from classification to object detection networks and LMMs have yet to be thoroughly evaluated in the context of benthic organism research. Such a dataset is essential both for training conventional models and for benchmarking the zero-shot capabilities of LMMs on benthic imagery. The initial step of filtering out empty frames is labor-intensive, and precise annotations demand marine science expertise since they must include species-level labels and accurate localization for object detection tasks, greatly increasing dataset curation costs.

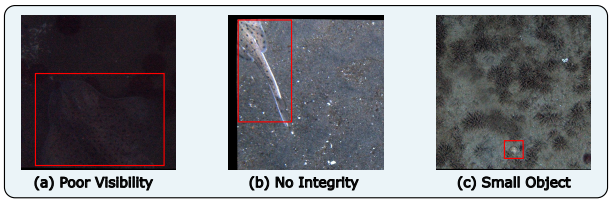


Figure 2: These examples illustrate three key challenges. (a) A large skate whose body is barely visible due to the dark background. The underexposed image severely reduces target visibility. (b) A large skate, whose head and partial body are invisible due to its size exceeding the image frame. (c) A scallop that is difficult to detect due to its small size.

Even when evaluating LMMs, a well-annotated dataset is indispensable, providing the ground truth necessary to verify whether these models generate correct and meaningful interpretations of benthic images.

Second, several characteristics inherent to benthic imagery may further degrade the performance of both CV models and LMMs. Although recent advances in underwater imaging have reduced issues such as blue-green distortion caused by light absorption and scattering (Anwar and Li 2020), other quality-related challenges persist. Visibility is frequently compromised by turbid water conditions from sediment disturbance and by species-specific behaviors such as partial or full burial in the substrate - common among scallops and skates. Additional degradations stem from underexposure, overexposure, and low visual contrast between organisms and their environments (e.g., scallops camouflaging among shell hash), all of which obscure object boundaries. Image integrity is also a recurring issue: because benthic imagery is captured using towed camera systems, organisms often appear only partially in frame - such as a skate's tail or a scallop's shell edge - limiting the availability of key visual features like shape, size, and texture. Compounding these challenges is the lack of scale standardization; wide-angle imaging used to capture larger organisms often renders smaller individuals (e.g., juvenile flatfish) too small to resolve with adequate detail. Collectively, these factors create a highly complex visual domain that diverges substantially from the clean, well-structured datasets on which most models are trained, as illustrated in Figure 2.

4 Field-Tests and Evaluation

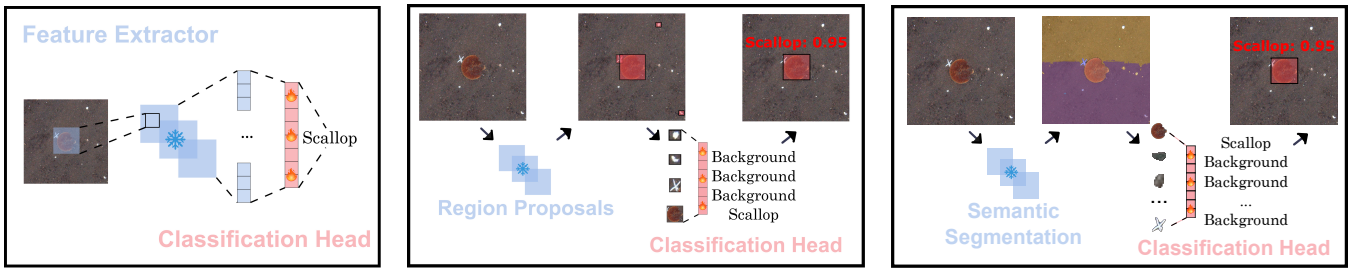
4.1 Datasets

The dataset used in this study was collected by the HabCam, a specialized towed imaging system capturing high-resolution seafloor images at six frames per second, producing continuous visual records of the benthic environment. Since most raw images do not contain benthic organisms, marine science experts manually filtered out empty frames and annotated the remaining images with species labels and precise organism locations. Data are organized by year, with 2015, 2016, and 2022 selected for experiments. Images from 2015 and 2022, prioritized for their higher quality and richer annotations, include over 90,000 cropped 640×640 sub-images used to evaluate overall model performance and challenges related to image integrity and size standardization. The 2016 dataset, affected by increased sedimentation and lower image quality, contains over 20,000 sub-images divided into high- and low-visibility subsets for testing model robustness under visibility challenges. To ensure consistent evaluation, each annotated organism was cropped into a 640×640 sub-image centered on its annotation. Objects were categorized as 'Large' or 'Small' based on bounding box area, and sub-images with crop regions extending beyond the original image boundaries were labeled as 'Non-Integrity'. The dataset includes seven benthic species: scallop, crab, whelk, skate fish, flatfish, roundfish, and eel. Please refer to Section 1 in the appendix for details on dataset collection, processing, and statistics.

4.2 Models and Implementation

Models We select a set of widely used open-source models across classification, object detection, and multi-modal learning, with an emphasis on architectural diversity and practical relevance. For **classification models**, we include convolutional architectures such as **(1)** ResNet (He et al. 2016), **(2)** MobileNet (Howard 2017), and **(3)** EfficientNet (Tan and Le 2019), which cover both standard and lightweight designs. We also include **(4)** Vision Transformer (ViT) (Dosovitskiy et al. 2020) and **(5)** Swin Transformer (Swin) (Liu et al. 2021) to represent attention-based models. For **object detection models**, we consider both two-stage detectors such as **(1)** Faster R-CNN (Ren et al. 2015) and **(2)** Cascade R-CNN (Cai and Vasconcelos 2018), and a one-stage detector, **(3)** YOLOv11 (Jocher and Qiu 2024; Redmon et al. 2016), to cover different detection paradigms. We further include **(4)** DETR (Carion et al. 2020) as a representative transformer-based detector. In addition, we evaluate a semantic segmentation model, **(5)** SAM (Kirillov et al. 2023), which outputs instance-level masks without classification. It is treated as a special case of object detection and is paired with a classification model to assign species labels. For **multi-modal models**, we start with **(1)** CLIP (Radford et al. 2021), a lightweight contrastive model for image-text alignment. We further include three more advanced vision-language models with stronger multi-modal reasoning capabilities: **(2)** LLaMA (Touvron et al. 2023), **(3)** LLaVA (Liu et al. 2023), and **(4)** Janus (Wu et al. 2024).

Training Protocol Our study evaluates four distinct model types. The classification model processes each input sub-image holistically, applying a feature extractor followed by a classification head to predict a single label (Figure 3a). In contrast, the object detection model identifies and classifies multiple regions within a sub-image, outputting bounding boxes, class labels, and confidence scores for each detected object (Figure 3b). Regions without valid targets are assigned a 'background' class. The segmentation model outputs object-level masks, which we evaluate by measuring spatial overlap (e.g., Intersection over Union) with ground-truth annotations and verifying label correctness. This combined assessment of localization and classification enables us to treat the segmentation model as an object detector in



(a) Classification Model (b) Object Detection Model (c) Semantic Segmentation Model

Figure 3: Classification, object detection, and segmentation models in this study. All models share a common setting where the pretrained backbones are frozen, and only the task-specific classification heads are trained for their respective recognition tasks.

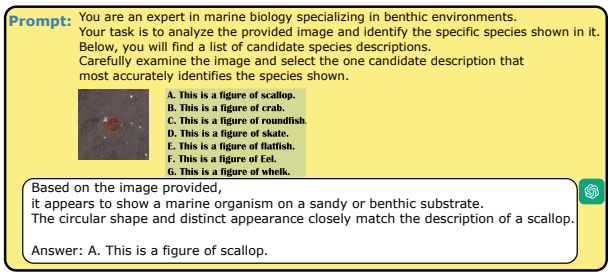


Figure 4: Example of large multi-modal model performing visual identification via text prompts.

the context of this work. Blue modules in Figure 3 denote pretrained components (e.g., feature extractors) that are kept frozen during training, while red modules (e.g., classification heads) represent trainable components fine-tuned for downstream tasks. Figure 4 shows how we evaluate LMMs. We input each image to the model along with a fixed question asking it to identify the species shown. The LMM processes the image and directly returns an answer, selecting the most appropriate label from a predefined list.

Metrics We adopt three evaluation metrics in our experiments: average classification accuracy (ACC), detection rate (DR), and average object detection accuracy (AODA).

Average Accuracy (ACC) measures the proportion of correctly classified samples across all categories. It is computed as: Avg Acc $= \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i)$, $\text{Acc}_{\text{class}} = \frac{N_{\text{correct}}^{\text{class}}}{N_{\text{total}}^{\text{class}}}$ where N is the total number of samples, and $\mathbb{I}(\cdot)$ is the indicator function, which returns 1 if $\hat{y}_i = y_i$ and 0 otherwise. $N_{\text{correct}}^{\text{class}}$ is the number of correctly predicted samples in a class, and $N_{\text{total}}^{\text{class}}$ is the total number of samples.

Detection Rate (DR) is used for object detection models to measure how many images contain at least one valid detection: $\text{DR} = \frac{N_{\text{detected}}}{N_{\text{total}}}$ where N_{total} is the total number of test images and N_{detected} is the number of images in which the model produces at least one high-confidence detection that matches a ground-truth object.

Average Object Detection Accuracy (AODA) evaluates the classification correctness over detected objects with sufficient confidence: AODA $= \frac{1}{N_{\text{detected}}}\sum_{i=1}^{N_{\text{detected}}}\mathbb{I}\left(C_i > 0.5 \wedge \hat{y}_i = y_i\right)$ where C_i is the classification confidence of the selected detection in the i-th detected image. The indicator function $\mathbb{I}(\cdot)$ returns 1 if the confidence exceeds 0.5 and the predicted label $\hat{y}_i$ matches the ground truth y_i.

4.3 Results and Analysis

Table 2 summarizes the performance of all three model types on the benthic datasets. Overall, CV models consistently outperform the LMMs across all evaluation conditions. Among classification models, Swin Transformer achieves the highest overall accuracy at 67.2%, followed by ViT at 58.2% and MobileNet at 55.1%. ResNet and EfficientNet yield slightly lower accuracies, reaching 56.1% and 54.2%, respectively. This suggests that transformer-based architectures, particularly those employing hierarchical designs, are able to extract more informative features than convolutional models, leading to improved performance in benthic species classification tasks. All models demonstrate strong performance on Eels, with Swin achieving 93.9% accuracy and ResNet achieving 90.4%. In contrast, Whelk becomes more challenging. For example, ResNet reports only 41.0% accuracy on Whelk, suggesting insufficient discriminative features in these categories. Under high-visibility conditions, all models exhibit a significant performance boost, with an average accuracy improvement of 12.7% compared to low-visibility scenarios. Object integrity also emerges as a critical factor: when benthic organisms are fully intact and unobstructed, classification accuracy improves across all models, resulting in an average increase of 15.9% relative to cases involving partially occluded or truncated instances. For standardization, larger objects do not lead to improved classification performance. One possible explanation is that the small object group is dominated by scallops, whereas the large group contains a higher proportion of three fish species. Because of better performance for scallops compared to fish, the overall accuracy in the small-object setting is higher, despite the small size.

For object detection models, the performance trends are generally consistent with those observed in classification models. All detectors perform well on scallop instances, with an average accuracy of 85.8%, while performance drops significantly for categories such as whelk and flatfish,

Table 2: Performance comparison of vision and multi-modal models on the benthic benchmark dataset

Model	Category							Avg(/DR)	Visibility		Integrity		Standardization	
	Scallop	Crab	Whelk	SF	FF	RF	Eel		Low	High	No	Yes	Small	Large
ResNet	60.6	40.5	41.0	69.5	25.7	35.7	90.4	56.1	57.9	66.8	46.8	57.8	50.3	50.9
EfficientNet	52.7	53.5	33.0	57.5	34.9	40.4	76.8	51.2	47.3	50.7	45.1	51.7	50.8	48.9
MobileNet	58.5	60.9	31.4	61.3	41.1	29.6	84.2	55.1	55.2	60.8	47.0	55.8	50.9	49.9
ViT	59.3	62.4	42.3	79.0	68.0	47.9	87.3	58.8	57.9	66.8	56.1	60.2	62.5	56.5
Swin	70.5	58.1	55.8	76.8	56.5	48.5	93.9	67.2	67.2	76.7	58.1	67.9	54.8	44.1
Faster R-CNN	88.6	53.4	30.9	83.4	10.2	52.9	86.4	84.8/57.3	88.7	95.3	75.2	86.1	75.1	87.8
Cascade R-CNN	90.0	67.1	63.1	82.7	53.2	48.4	81.3	82.7/85.7	82.3	94.9	80.8	82.8	89.2	82.0
YOLOv11	83.0	81.1	86.2	86.7	82.7	81.4	96.7	82.1/80.7	80.1	92.0	77.8	86.7	72.7	72.8
DETR	81.7	51.1	31.9	59.1	14.8	25.2	75.8	70.6/99.3	57.2	70.1	71.4	70.5	79.8	68.3
SAM	60.5	24.0	00.1	73.8	42.9	09.7	00.5	50.0/95.7	50.1	56.1	49.1	53.1	77.2	50.4
CLIP	77.5	24.5	02.1	02.0	34.7	13.8	71.1	62.9	57.4	58.6	60.9	63.1	74.5	63.7
LLama	52.5	00.0	37.5	05.8	00.4	01.5	01.2	40.4	10.9	11.3	45.0	39.8	49.5	39.5
LLava	65.4	03.4	00.6	02.7	13.8	13.7	00.0	50.9	23.8	26.6	55.6	49.5	63.3	50.5
Janus	00.0	00.0	96.6	07.7	06.0	41.8	50.0	06.2	04.1	04.8	05.8	06.6	01.2	07.3

which achieve average accuracies of 53.0% and 49.7%, respectively. In addition, both high visibility and complete object integrity contribute positively to detection performance. On average, high-visibility scenes improve detection accuracy by 11.9%, while complete object views lead to an average gain of 13.5% across all models. A key observation across detection models is the trade-off between detection rate and classification accuracy. Although all models are trained on the same dataset, some tend to adopt more conservative detection strategies, focusing only on the most confident instances. For example, Faster R-CNN achieves a relatively high classification accuracy of 84.8%, but its detection rate is limited to 57.3%. In contrast, DETR detects nearly all possible objects in each image, achieving the highest coverage of 99.3%, but its classification accuracy drops to 70.6%. YOLO strikes the best balance between the two, achieving both high detection rate with 80.7% and strong classification accuracy with 82.1%. Notably, SAM is a semantic segmentation model and demonstrates a different behavior. While it can segment almost every object instance in the image, its masks often cover only a portion of the actual target object. As a result, the model struggles to provide semantically complete inputs for classification, leading to near-zero accuracy on several categories, such as whelk, roundfish, and eel.

LMMs exhibit a distinct behavior compared to classification and detection models. While these models appear to perform classification by producing plausible predictions and achieving reasonable accuracy on certain categories, they reveal a collapse of predictions into a single dominant class, with limited attention to fine-grained distinctions. For instance, CLIP, LLaMA, and LLaVA actually assign most images to the Scallop category, whereas Janus predicts most images as Whelk. This prediction bias limits category diversity and undermines the reliability of their results.

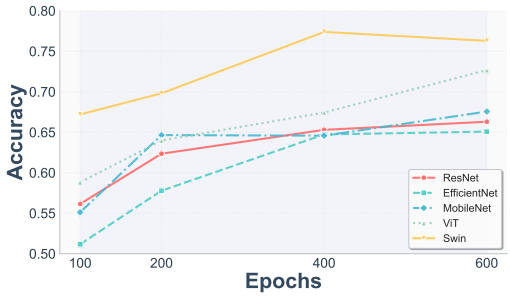
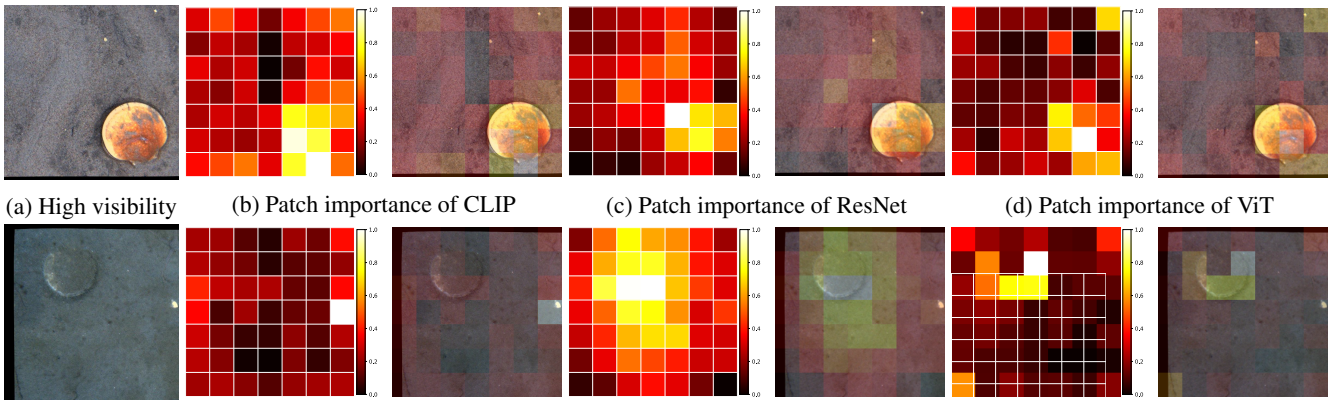


Figure 5: Overall accuracy of different classification models across multiple training epochs.

5 In-Depth Analysis and Discussion

5.1 Relationship between training Epochs and accuracy performance

When the feature extractor is frozen and only the classification head is updated, additional training time still leads to noticeable gains in accuracy. Figure 5 presents the overall classification accuracy of five models, MobileNet, EfficientNet, ResNet, ViT, and Swin, trained with 100, 200, 400, and 600 epochs. Across all models, a general upward trend in accuracy is observed as training epochs increase. Swin achieves the highest accuracy at 400 epochs with 77.4%, although a slight decline is observed at 600 epochs. ViT shows the second-best performance and the most consistent improvement, increasing steadily from 58.8% at 100 epochs to 72.7% at 600 epochs. Models like MobileNet and EfficientNet also benefit from extended training, with MobileNet improving from 55.1% to 67.6%, and EfficientNet from 51.1% to 65.0%. ResNet follows a similar trend to Swin, improving up to 65.3% at 400 epochs before slightly declining. These results validate that longer training can frequently improve classification performance, even under constrained fine-tuning settings where only the classification



(a) High visibility (b) Patch importance of CLIP (c) Patch importance of ResNet (d) Patch importance of ViT

(e) Low visibility (f) Patch importance of CLIP (g) Patch importance of ResNet (h) Patch importance of ViT

Figure 6: Patch-level attribution analysis of CLIP, ResNet, and ViT on two scallop images under high and low visibility. Each subfigure shows (from left to right): the patch importance heatmap and attribution overlay

head is trained while the feature extractor remains frozen. Overall, the optimal number of training epochs varies by model architecture, as some models may experience performance saturation or slight degradation when over-trained without further feature-level training. Of particular note, transformer-based architectures such as Swin and ViT outperform convolutional-based models in general, especially under extended training.

5.2 Why LMMs generally fail in benthic imaginary tasks

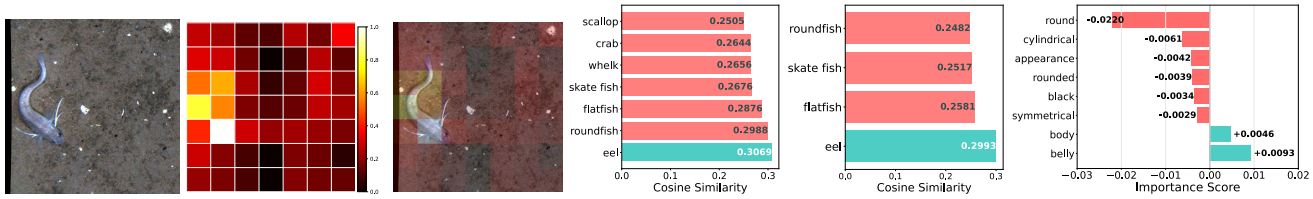
LMMs appear to underperform in benthic species classification relative to classification and object detection models, despite their success on general vision-language tasks and access to extensive pretraining data. Fine-tuning LMMs requires substantial computational resources and large-scale domain-specific data, which are beyond the scope of this study. Instead, we focus on analyzing the underlying factors that may contribute to their suboptimal performance in this domain-specific context.

In this section, we select CLIP as a representative multimodal model to investigate potential reasons for underperformance. CLIP's open-source availability and relatively simple architecture make it well-suited for detailed analysis and interpretability. Its modular design allows us to systematically examine the individual contributions of the image and text encoders to the model's overall performance.

First, we try to explore whether CLIP can extract discriminative feature information related to benthic species from the input image. Specifically, we divide each image into a 7×7 patch grid. We occlude each patch by replacing it with a black square and recompute the image embedding. The importance of each patch is quantified by the ℓ_2 distance between the occluded and original embeddings, where a larger distance indicates a greater contribution to the original representation. Figure 6 illustrates two scallop examples to highlight the performance contrast. In the successful case, the model correctly identifies the scallop as the most salient region; its orange shell contrasts with the dark background, making it visually distinctive. In the failed case, however, the scallop is partially obscured by sediment, with only a faint white edge visible. Under these conditions, the image encoder mistakenly assigns high salience to unrelated white regions near the image border—background noise rather than the organism itself. Despite CLIP's exposure to large-scale training data, it struggles to extract meaningful features from low-visibility benthic imagery. Consequently, the resulting representation lacks the visual cues necessary to match the image to the appropriate text prompt, making the misclassification a plausible outcome.

We further repeated the same occlusion-based experiment using standalone ViT and ResNet models. As shown in Figure 6c and Figure 6d, the ResNet model, which is based on a convolutional architecture, demonstrates more robust feature extraction under both high and low-visibility conditions. In particular, it performs noticeably better than CLIP when processing low-visibility images, capturing the relevant object regions more clearly, as shown in Figure 6g. Although ViT also adopts a transformer-based architecture similar to CLIP, it still shows improved recognition in the low visibility case. It is able to partially localize the true scallop in the upper-right region while avoiding distraction from background noise, as shown in Figure 6h. These results suggest that, despite being trained on large-scale datasets, CLIP struggles to extract benthic-relevant features under challenging visibility conditions. The resulting image representation lacks the necessary visual cues to match the image to the correct text prompt, making the misclassification an expected outcome. Furthermore, convolutional architectures may outperform transformer-based models in this domain, possibly because they are better at capturing fine-grained local features that are critical for identifying benthic organisms.

We next assess CLIP's ability to align image and text embeddings within a shared semantic space. As shown in Figure 7b, the image encoder correctly localizes the roundfish in the input image and extracts relevant visual features. To evaluate alignment, we compute the cosine similarity between the image embedding and each text embedding. Despite ac-



(a) Roundfish (b) Patch importance of CLIP (c) Prompt-level and word-level effects on CLIP's image-text matching

Figure 7: **An example of roundfish image.** From left to right: (a) a high-visibility input image, (b) patch-level attribution maps and overlays, and (c) similarity results with simple prompts, more detailed prompts, and word-level importance visualizations for the descriptive prompt.

curately identifying the roundfish visually, CLIP assigns the highest similarity to the 'Eel' prompt, misclassifying the image. This highlights a misalignment between the visual and textual representations, even when the image features are extracted correctly. To probe this further, we extract the four most relevant fish-related categories and replace simple prompts with more descriptive ones:

- *a round fish which has a cylindrical black body with a rounded belly and a symmetrical appearance.*
- *a flatfish which has a flattened body with both eyes on one side, blending well with the ocean floor.*
- *a skate fish which has a flattened, diamond-shaped body with wing-like pectoral fins and a long tail.*
- *an eel which is a long, slender fish with a smooth, snake-like body and small fins.*

However, while the detailed prompts add richer semantic cues, the prediction remains incorrect: the roundfish image is still labeled as being most similar to the eel description, as shown in Figure 7c. This suggests that the open vocabulary capability of CLIP may not necessarily improve classification accuracy, as the model fails to leverage the added semantic detail to distinguish between visually similar benthic species. To further support this observation, we apply a similar occlusion strategy to the text modality. We assess the importance of each word by masking it and measuring the decrease in cosine similarity between the image embedding and the modified text embedding, relative to the similarity from the unmasked sentence. As shown in Figure 7c, descriptive terms such as **round** and **cylindrical**, which are expected to be informative, instead reduce the similarity score, suggesting a misalignment between textual semantics and visual representation. While this provides a single illustrative example, our global accuracy results in Table 2 suggest CLIP struggles to capture the fine-grained semantic distinctions necessary for accurate benthic species recognition. Our analysis reveals this is likely driven by three key limitations. First, the image encoder frequently attends to coarse or background features rather than the organism itself, resulting in the extraction of misleading visual representations. Second, the text encoder exhibits weak grounding, where critical descriptive terms often fail to meaningfully improve image-text alignment. Third, some visual differences between benthic species are inherently difficult to express in natural language, limiting the effectiveness of CLIP's text-based reasoning. Together, these challenges underscore the difficulty of applying large, pre-trained multimodal models directly to domain-specific tasks like benthic imagery classification without targeted adaptation or fine-tuning.

6 Open Problem

For **standard CV models**, we freeze the backbone and region proposal components and only fine-tune the classification heads. This setup is intended to isolate classification capability, but it leaves open the question of whether pre-trained models, which are trained on non-benthic datasets, can effectively extract low-level and mid-level features relevant to benthic imagery. As a result, the full potential of different architectures has not been fully explored. Future work could investigate end-to-end training on benthic datasets to better assess model capacity, although this would involve higher computational costs. For **LMMs**, our evaluation centers on species recognition through image-to-species name matching. Although we explored the influence of descriptive prompts in a limited set of cases, we did not perform a systematic, dataset-level analysis. Future work should evaluate the prompt sensitivity of LMMs more comprehensively, examining whether specific prompt designs consistently improve performance and are more effective.

7 Conclusion

In this work, we introduce **Benthic Benchmark**, a large-scale, high-resolution underwater image dataset designed to evaluate the performance of classification, object detection, and multi-modal models on benthic species recognition. The dataset contains three inherent recognition challenges: limited visibility, incomplete object integrity, and inconsistent object scale. We construct two standard tasks, image classification and object detection, and benchmark a diverse set of representative architectures including CNNs, Transformers, detection models, and LMMs. We further analyze the influence of each challenge through controlled subset evaluation and observe that these factors reduce model performance generally, with varying degrees across different architectures. In particular, LMMs such as CLIP struggle to generalize in benthic settings, showing issues related to semantic mismatch and prompt sensitivity. These findings demonstrate the limitations of current models and establish Benthic Benchmark as a foundation for domain-specific evaluation and future research in benthic visual understanding.

References

Agrawal, A.; Kedia, N.; Mohan, J.; Panwar, A.; Kwatra, N.; Gulavani, B. S.; Ramjee, R.; and Tumanov, A. 2024. Vidur: A large-scale simulation framework for llm inference. *Proceedings of Machine Learning and Systems*, 6: 351–366.

Anwar, S.; and Li, C. 2020. Diving deeper into underwater image enhancement: A survey. *Signal Processing: Image Communication*, 89: 115978.

Cai, Z.; and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 6154–6162.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*, 213–229. Springer.

Chang, J.-H.; Hart, D. R.; Shank, B. V.; Gallager, S. M.; Honig, P.; and York, A. D. 2016. Combining imperfect automated annotations of underwater images with human annotations to obtain precise and unbiased population estimates. *Methods in Oceanography*, 17: 169–186.

Chang, J.-H.; Shank, B. V.; and Hart, D. R. 2017. A comparison of methods to estimate abundance and biomass from belt transect surveys. *Limnology and Oceanography: Methods*, 15(5): 480–494.

Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; and Shan, Y. 2024. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16901–16911.

Davis, C. S.; Gallager, S. M.; and Solow, A. R. 1992. Microaggregations of oceanic plankton observed by towed video microscopy. *Science*, 257(5067): 230–232.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Gallager, S. M.; Singh, H.; Tiwari, S.; Howland, J.; Rago, P.; Overholtz, W.; Taylor, R.; and Vine, N. 2005. High resolution underwater imaging and image processing for identifying essential fish habitat. In *Report of the National Marine Fisheries Service Workshop on Underwater Video Analysis*, 50.

Hart, D. R.; and Rago, P. J. 2006. Long-term dynamics of US Atlantic sea scallop Placopecten magellanicus populations. *North American Journal of Fisheries Management*, 26(2): 490–501.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Howard, A. G. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Jocher, G.; and Qiu, J. 2024. Ultralytics YOLO11.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *ICCV*, 4015–4026.

Kristensen, E.; Penha-Lopes, G.; Delefosse, M.; Valdemarsen, T.; Quintana, C. O.; and Banta, G. T. 2012. What is bioturbation? The need for a precise definition for fauna in aquatic sciences. *Marine ecology progress series*, 446: 285–302.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *NeurIPS*, 36: 34892–34916.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 10012–10022.

NOAA. 2022. Fisheries One Stop Shop (FOSS). https://www.fisheries.noaa.gov/foss.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *CVPR*, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*.

Roman, S.; Lin, C.; and Rudders, D. 2024. A Cooperative High Precision Dredge Survey to Assess the Mid-Atlantic Sea Scallop Resource Area in 2021 and 2022.

Singh, W.; Örnólfsdóttir, E. B.; and Stefansson, G. 2013. A camera-based autonomous underwater vehicle sampling approach to quantify scallop abundance. *Journal of Shellfish Research*, 32(3): 725–732.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 6105–6114.

Taylor, R.; Vine, N.; York, A.; Lerner, S.; Hart, D.; Howland, J.; Prasad, L.; Mayer, L.; and Gallager, S. 2008. Evolution of a benthic imaging system from a towed camera to an automated habitat characterization system. In *OCEANS 2008*, 1–7. IEEE.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wu, C.; Chen, X.; Wu, Z.; Ma, Y.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; Ruan, C.; et al. 2024. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*.