Finding NEMO: Navigating Environment-Specific Models for Object Detection in Marine Imagery

James Liu, The University of Texas at Austin

Mentors: Laura Chrobak, Kevin Barnard, Giovanna Sainz, Joost Daniels, Kakani Katija

Summer 2026

Keywords: Computer Vision, Object Detection, Supercategories, Data Exploration, Marine Imagery, FathomNet

ABSTRACT

The analysis of marine imagery remains a critical bottleneck for ecological research; complex environments and large data backlogs push the limits of human annotation efforts, highlighting the need for more efficient methods. Our work seeks to train object detection machine learning models to automate the process of identifying and localizing relevant marine animals. For our model, we utilize the FathomNet Database[2] - a public repository of expertly annotated marine imagery - for our training data. To reduce class imbalance and reflect ecological priorities, we collapsed the 2,055 FathomNet labels into 29 benthic and 21 midwater supercategories, balancing morphological similarity, taxonomic structure, data availability, and ecological relevance. Additionally, we sort images into either benthic or midwater categories due to their inherent environmental differences, training models that are specific to each environment. In this paper, we were only able to test our benthic models in our benchmarking framework. For benchmarking, we curated a full coverage benthic dataset of 389 images, ensuring a high-quality ground truth for quantitative metrics. Our best model, a fine-tuned YOLO11x[1], had an 0.539 mAP@0.5 and an F1 score of 0.589 (precision

0.813, recall 0.461). This model was trained with a carefully selected subset of our total data that emphasized data quality over data quantity, highlighting the idea that an increasing dataset size without controlling annotation quality can degrade performance. Qualitative inspection additionally showed that the model occasionally outperformed their own training labels by correctly detecting organisms missing from annotations. However, recall was still limited and misclassifications frequently occurred between morphologically similar supercategories (e.g., sharks vs. fish, eels vs. fish, gastropods vs. sediment). Overall, the model was successful in detecting a significant portion of relevant animals in marine imagery, demonstrating that object detection models, coupled with human-in-the-loop oversight, may offer a potential solution for accelerating marine ecological research. Future work involves expanding our benchmarking framework for greater benthic representation, creating a midwater full-coverage test set, and experimenting with new methods to extract the most performance from a noisy dataset.

INTRODUCTION

The analysis of marine imagery remains a critical bottleneck for ecological research; complex environments produce scenes with dense fauna, poor lighting conditions, and noisy backgrounds that challenge human capabilities. Additionally, advances in underwater imaging technologies have produced an unprecedented volume of data, outpacing the capacity of manual annotation workflows and highlighting their current unscalable nature. Finally, different marine environments contain unique biodiversity, noise, and environmental contexts, requiring unique domain expertise to properly capture meaningful analysis. Through this research, we aim to train and benchmark environment-specific object detection models to evaluate their potential in automating post-processing of underwater imagery, specifically for animal localization and labeling.

PREVIOUS WORK

ORIGINAL SUPERCATEGORY DETECTORS

Prior to my internship, the MBARI and CVisionAI had trained older versions of benthic and midwater supercategory detectors in 2023. The old benthic model had been based on a Detectron2 architecture with a ResNet backbone, and the midwater model had been based on a YOLOv5. Both of these models had been trained using early FathomNet data and internal MBARI data; however, there was no separation of training data between benthic and midwater which is a key development in this project. Additionally, the supercategory mappings used for these older models did not broadly capture the biodiversity in our expanded FathomNet data, thus requiring us to update them accordingly. Finally, there were also taxonomic reclassifications (mostly regarding soft corals, sea fans, and sea pens) that required an update to the supercategory mappings as well. Thus, key differences between the old models and the new models include: the separation of training data into benthic and midwater, updated supercategories for greater biodiversity coverage, and consistent quantitative benchmarking against a full-coverage dataset to best capture performance without the influence of noise in our testset.

MATERIALS AND METHODS

FATHOMNET DATABASE

Computer vision and machine learning models are highly dependent on training using large volumes of data in order for algorithms to converge. For this project, we sourced our data from the FathomNet Database - a public repository of expertly annotated underwater imagery. Some examples of contributors to the FathomNet Database include MBARI, NatGeo, NOAA, Schmidt Ocean Institute, WHOI, and individual contributors. More specifically, we used a snapshot of the FathomNet Database from November 2024 with 107,661 images and 287,877 localizations (bounding boxes). The labels from the database were primarily taxonomically based, ranging from extremely specific labels on the species level, to extremely ambiguous phylum level labels (eg: porifera). Additionally, there were also a few niche label types

such as equipment (eg: paint bucket, drum, light laser), fauna components (eg: detritus, sinkers, inner/outer houses), and undefined species (eg: Aeolidiidae sp. 1). Before using the data for training, we first sanitized the labels by trimming labels down - by removing auxiliary titles, punctuation, unclear species, body part notations, etc - until they matched a noted taxonomic title found in WoRMS (World Register of Marine Species). After cleaning the labels, we were left with 2,055 recovered labels. However, within these labels, we had labels that were children of other labels - for example, a species that belonged to a phylum - which added further complexity. Additionally, the distribution of these labels were also greatly unequal. As seen in figure 1, some labels had thousands of instances while others had single examples; thus, balancing the labels became extremely important which led to our development of supercategories.

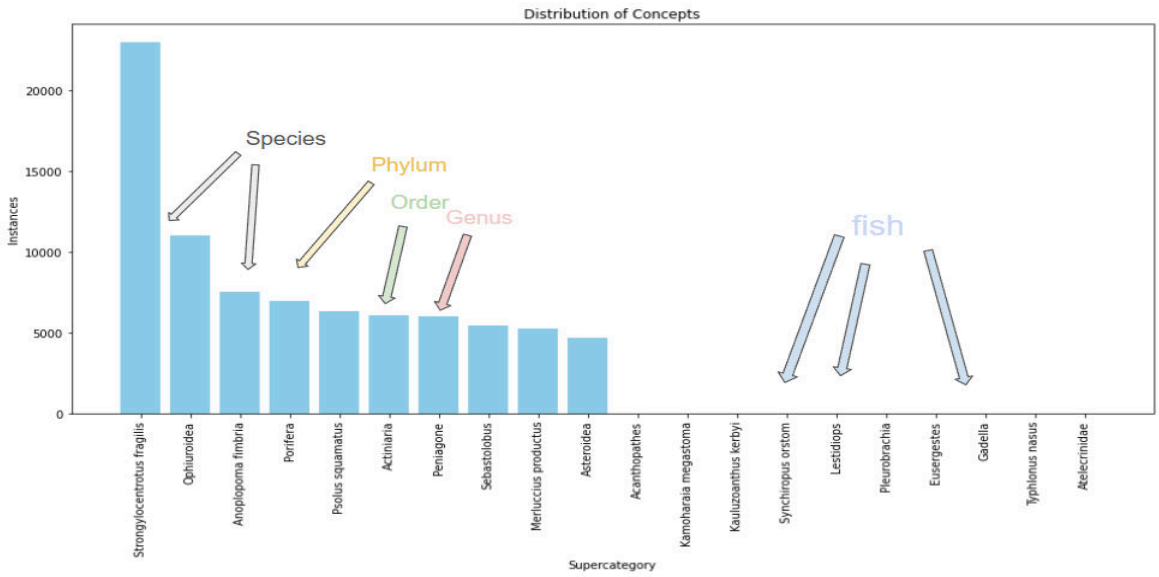

Figure 1. The distribution of a select few labels from the FathomNet Database. Notice the disparity between the most and least represented labels. Also notice the different taxonomic levels of the labels.

SUPERCATEGORIES

Ultimately, we decided to collapse our labels into 29 benthic supercategories and 21 midwater supercategories. These supercategories represent umbrella groupings that contain subsets of the base labels that we got from FathomNet. More specifically, since our labels from FathomNet were taxonomic names, we represented our supercategories as

groupings of highlevel taxonomic classes with all of their children included. Additionally, for each supercategory, we also included taxonomic classes to exclude since some supercategories may be supersets of others (eg: bony fish and flat fish) that we chose to separate for ecological interest reasons. This way, there was no overlap between any of the supercategories in each environment.

```
"Bony fishes": {
      "include": [ "Actinopterygii" ],
      "exclude": [ "Anguilliformes", "Notacanthiformes", "Myxiniformes", "Zoarcidae",
"Pleuronectiformes" ]
}
```

Figure 2. An example supercategory and its included/excluded taxonomic classes. The mappings were stored in the JSON format for ease of use in Python.

In order to determine what supercategories we wanted, we considered a variety of factors including balancing morphological and taxonomic relations, data availability, and ecological interest. The ecological interests were determined by consulting benthic and midwater animal experts (big thanks to the VARS lab) to see which groups of animals were considered relevant for the analysis of marine ecosystems. We also iteratively improved on the list by identifying which FathomNet labels and taxonomic groups were not captured by our initial draft list, and creating new supercategories to represent them. Our final supercategories are shown below:

0: Sea Anemones	0: Bony fishes
1: Bony fishes	1: Eel-like
2: Flatfish	2: Sharks
3: Eels	3: Rays and Skates
4: Gastropods	4: Cephalopods
5: Sharks	5: Pteropods
6: Rays and Skates	6: Shrimps
7: Chimaeras	7: Amphipod-like
8: Sea stars	8: Isopods
9: Feather stars and sea lilies	9: Hydroidolina
10: Sea cucumbers	10: Trachylinae
11: Urchins	11: Scyphozoa
12: Glass sponges	12: Calycophorae
13: Sea fans	13: Physonectae

14: Soft corals 15: Sea pens 16: Stony corals 17: Black corals 18: Crabs 19: Shrimps 20: Squat lobsters 21: Barnacles 22: Sea spiders 23: Worms 24: Brittle Stars 25: Tube-Dwelling Anemones 26: Demosponges 27: Zoanthids 28: Clams	14: Lobata-like 15: Beroidae 16: Appendicularia 17: Pyrosome 18: Salps 19: Worm-like 20: Arrow Worms

Figure 3. All of the supercategories and their mappings to either benthic or midwater environments

Additionally, we had these supercategories mapped to either benthic or midwater environments based on the usual habitat of the specific category. This mapping serves as a resource for future marine ecosystem analysis, and also guides our methods for separating images into benthic vs midwater environments. Below we have the distributions of supercategories for the benthic data and the midwater data respectively:

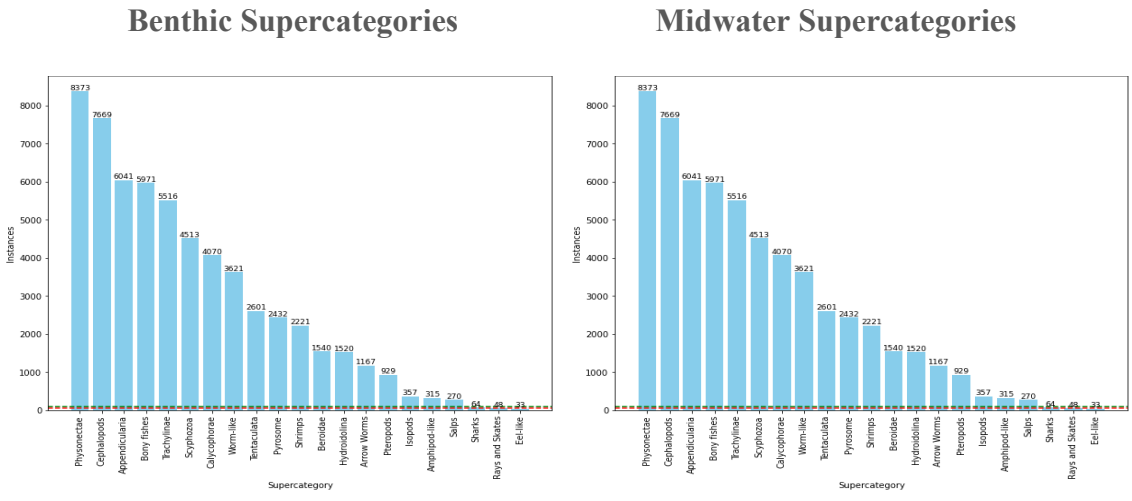

Figure 4. The distributions of respective supercategories for our sorted benthic and midwater images

ENVIRONMENTAL SEPARATION

Beyond just having ecological differences, benthic and midwater are fundamentally different environments with unique animals, lighting, and noise. These differences inherently affect model performance due to background context. Thus, our next step was to separate our images into either benthic or midwater environments. However, with 100,000 images, doing this by hand was not feasible; thus, we had to find ways to automate this process. The first thing we did was hand label ~8,000 images, in order to create a source of truth for either testing methods or machine learning model training. This was done using a custom data annotation platform I programmed, using a NextJS frontend with a GoLang, PostgreSQL, and Redis backend.

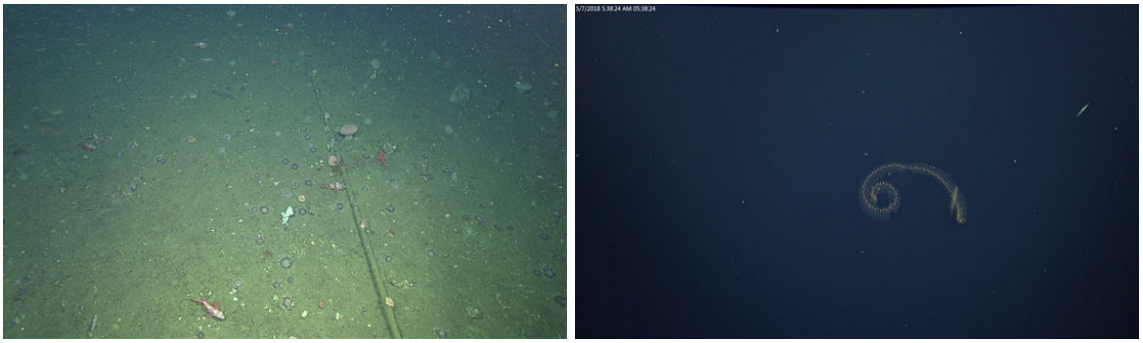

Figure 5. An example of the differences between benthic (left) and midwater (right). Different environmental contexts include lighting differences, presence of sediment, fauna differences, and more.

Our first approach was to extrapolate from the image annotations and our benthic/midwater mappings. We took the annotations in the image, and mapped them to either belonging in either benthic or midwater classes. Then, we simply took the majority count (if there were more benthic instances or midwater instances) to determine the environment. This worked well for around 80% of the images; however, this meant that 20,000 images were still unclassified and therefore excluded from training. Reasons for this included having equal proportions of benthic and midwater entities in one image, or having only supercategories that were present in both benthic and midwater environments (eg: bony fish). Thus, we needed some way to sort these remaining images using the only data we had - annotations and the image itself.

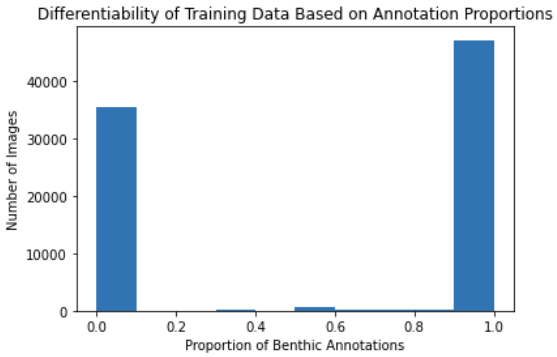

Figure 6. The distribution of benthic vs midwater localization ratios in all of the FathomNet data. As demonstrated, most images contained only midwater examples, or benthic examples. Images with only examples that were mapped to both midwater and benthic were then discarded.

Our second approach was to use image features extracted from a ResNet50 convolutional neural network for classification. These image features are learned by training the ResNet50 on our handlabeled training data; the resulting layers of the network then represent specific features in the image that the model has decided to be relevant when deciding how to classify the image. For example, some of the features that were extracted were shadow and highly contrasted areas. Intuitively this makes sense, as shadows can only exist where there is a surface for light to project onto, which exists only in the benthic zones and not the midwater. This method required only the image - no other metadata was needed - which made it the most convenient to use out of the box without needing to run data cleaning on the annotation metadata. This ended up working for the majority of images, reaching a 99% accuracy on our test data set. However, this method did not take into account the annotation data, which contained helpful discerning information, leading to our third approach to incorporate both data streams into one decision model.

Our final approach was to combine image features (eg: shadows) and annotation features (eg: presence of coral) to classify environments. In this case, we trained a multimodal model that used the ResNet50 CNN backbone to extract image features, which we then combined with the image annotations as additional features for the model to use as input in classification. This way we can combine the benefits of the mostly accurate annotation based heuristic with the nuanced approach of the image feature

method. This model ultimately was the most accurate as tested on our test set; however, it did require more overhead to get the annotations into the correct format to be used as features. In the end, we were left with 43,568 midwater images and 64,093 benthic images.

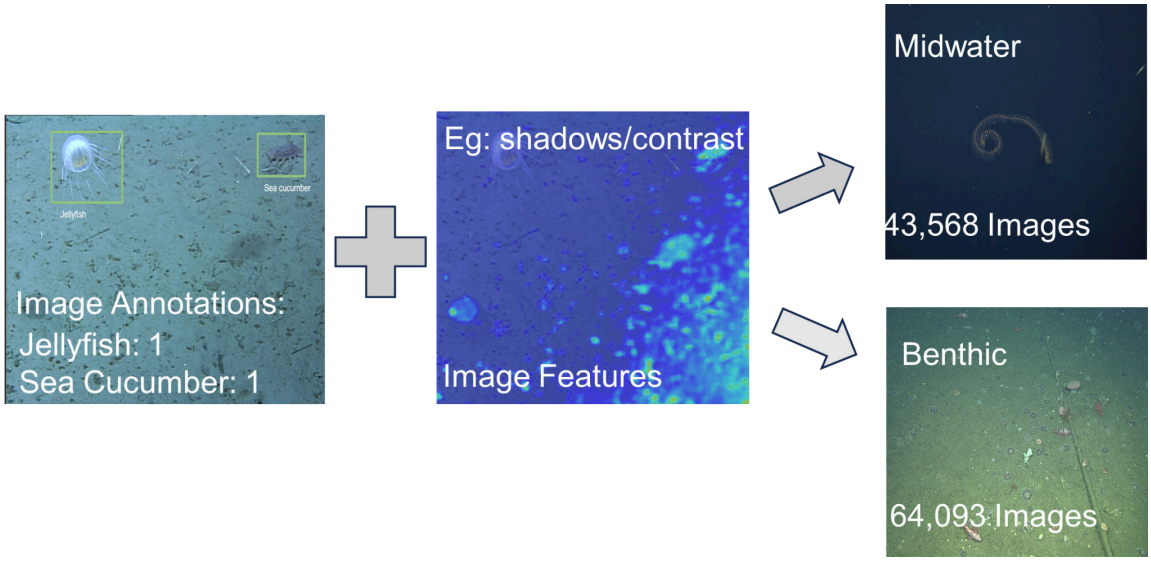

Figure 7. An illustrated example of our multimodal pipeline. This demonstrates how the image annotations were combined with image features (eg: shadows) to sort our images into benthic and midwater.

FULL COVERAGE BENCHMARKING

Finally, in order to test our models in a consistent and reliable manner, we needed a good test dataset. One issue we faced when testing on a split of our full FathomNet data is the fact that the images are mostly noisy. Ideally, for every image we use in training and testing, every entity of interest would be localized and labelled; however, due to human annotation limitations, this is not necessarily possible. Thus, the majority of images in our dataset had many unlabeled positive examples, which would greatly affect our testing results (essentially introducing false negatives). For example, if the model detects a sea urchin that exists in the image but wasn't labeled in the data, then that will be counted as an incorrect prediction even though it was objectively correct.

To combat this, we sourced a collection of higher quality images that we call full - coverage. This just means that these images have been extremely carefully annotated, getting as close as possible to localizing all relevant objects in an image. This ensures that our model is not punished for predicting results better than our ground truth dataset. Additionally, having the preset dataset meant that we could compare performance between different models, thus allowing for more quantitative direct comparisons. Ultimately, this full coverage test dataset set was composed of 389 images, spanning only benthic images. As a caveat, while this testset provided a better reflection of model performance by reducing incorrectly punished "false positives", it suffers from its small size, with many supercategories having only a few examples (as little as just one or none).

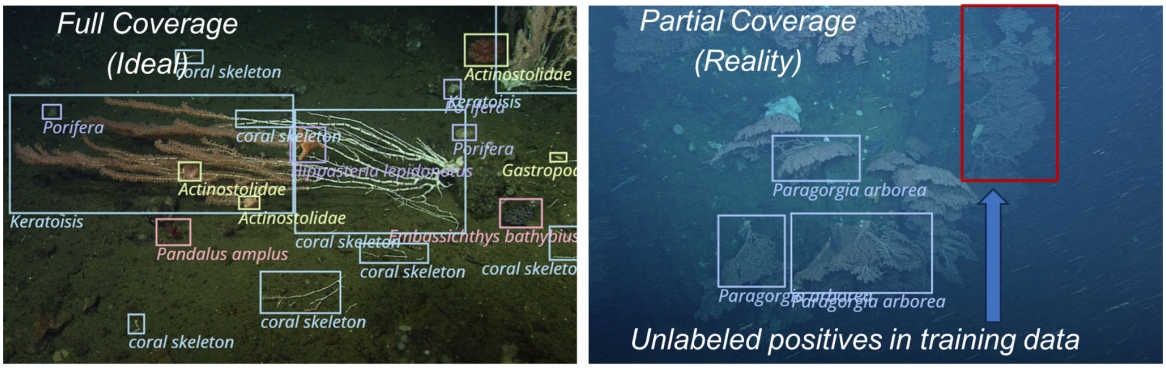

Figure 8. Comparing and contrasting between a full coverage image and a partial coverage image. Full coverage implies that all regions of interest are labeled and localized close to perfectly, while partial

coverage will have missing annotations which introduces noise to our dataset. The red box indicates an example of an unlabeled positive.

Additionally, to properly compare the performances of each model, we had to find the best performing confidence threshold for each model we were testing. This was done using Ultralytics validation framework, focusing on optimizing overall f1 score on the full-coverage test set.

EXPERIMENTS

For this project, we trained both benthic and midwater object detection models, where each model is focused on detecting the respective supercategories defined above. Since we only had a full-coverage dataset for benthic environments, we focused mainly on testing the benthic models. In this case, the models we trained were fine-tuned on the YOLO11x[1] convolutional neural network architecture from Ultralytics (augmented with Albumentations) and logged with CometML. Training was performed on a compute cluster with an AMD Threadripper with two RTX 6000 Ada GPUs. The final benthic model we trained - the ablated dataset model - and a handpicked midwater model were ultimately uploaded to the MBARI HuggingFace repository.

HEURISTIC SPLIT BENTHIC SUPERCATEGORY DETECTOR

For our first model that we trained, we used the data splits that we derived from our annotation heuristic environment sorting method. This means that our dataset was composed of 41,113 images, which was limited by the 20,000 images that our heuristic method could not classify environmentally. Most notably, this dataset underrepresented bony fish because images with only bony fish annotations were unable to be sorted using our heuristic, leading them to be discarded.

FULL DATASET BENTHIC SUPERCATEGORY DETECTOR

The second model that we trained used the environmental splits determined by our multimodal model, which meant that the number of training images increased to 63,000 images with 202,000 localizations. This remediates the underrepresentation of bony fish, and provides more training examples for every supercategory.

ABLATED DATASET BENTHIC SUPERCATEGORY DETECTOR

The final model we trained used a downsampled subset of the full dataset for the above benthic detector with 36,000 images and 164,000 localizations. The reasoning behind this decision relied on the idea that unlabeled positives introduced noise that would be adversarial to training; for example, every unlabeled positive in the image is taught to the model as background, impeding the model's performance on that supercategory. Additionally, our preliminary experiments demonstrated that many of the underrepresented supercategories still achieved high performance, possibly indicating that we could truncate some of our overrepresented categories and still maintain similar levels of performance. Thus, with the knowledge that most images in FathomNet were partial-coverage, and having more images would introduce more unlabeled positives, the goal then became to find a balance between maximizing the number of labeled examples and minimizing the number of unlabeled positives.

The actual process of downsampling is as follows. The initial subset was chosen by first taking images that had examples of underrepresented supercategories. Then for the overrepresented supercategories, we set a threshold of ~15,000 localizations and greedily selected new images to include until we met that threshold. Our greedy criteria was to select images with highest counts of over- represented classes, which maximizes the number of examples while minimizing the number of images, in theory reducing the amount of unlabeled positives.

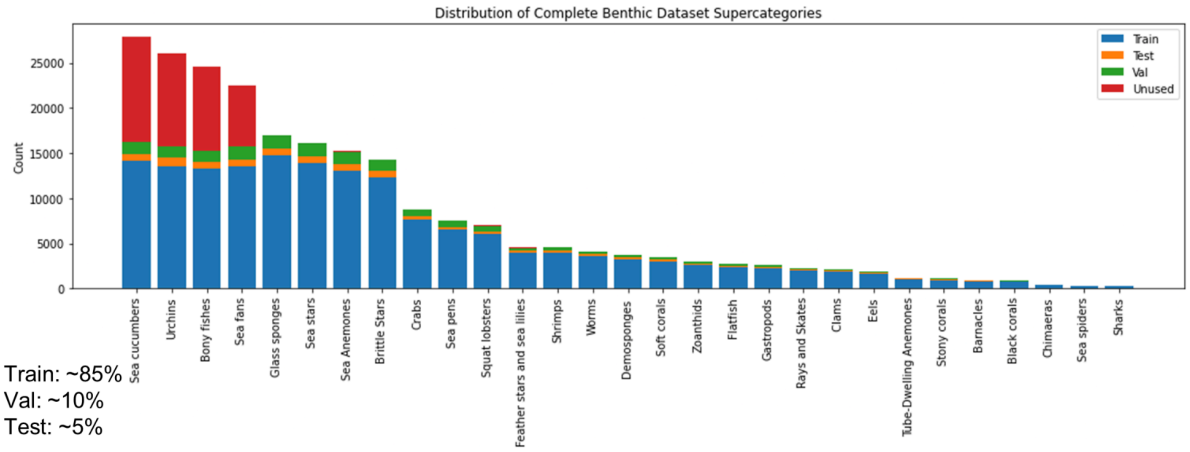

Figure 9. The distribution of the 29 benthic supercategories in the benthic dataset. "Unused" are examples in the full dataset that were downsampled away in the reduced dataset.

RESULTS

QUANTITATIVE RESULTS

We then benchmarked all three benthic models against our full-coverage dataset. Our heuristic split detector achieved a precision of 0.6912 and a recall of 0.3840 for an F1 score of 0.4310. The full benthic dataset detector (63k images, 202k localizations) achieved 0.524 mAP@0.5 and an F1 score of 0.549 (precision 0.846, recall 0.406). The reduced benthic dataset detector (36k images, 164k localizations) surprisingly outperformed the original, with 0.539 mAP@0.5 and an F1 score of 0.589 (precision 0.813, recall 0.461). This reduced dataset was downsampled from the full benthic dataset by greedily selecting images with the highest occurrences of over-represented supercategories until a threshold was met, maximizing the number of examples while minimizing number of images, ideally reducing false negatives. This result highlights that increasing dataset size without controlling annotation quality can degrade performance. Ultimately, the reduced benthic detector proved to be the most performant, with the highest mAP@0.5 and F1 scores.

Additionally, we also examined the confusion matrices of our final model to see individual supercategory performance. These confusion matrices were benchmarked on a validation dataset, not the full-coverage dataset. This was done because the validation

dataset has much greater representation for the supercategories (the full-coverage dataset is much too small for this purpose), so it better represents individual class performance. Because of this, it is expected for the confusion matrix to indicate large amounts of false positives, since these are localizations that exist in reality, but are not reflected in the validation data due to the partial coverage issue.

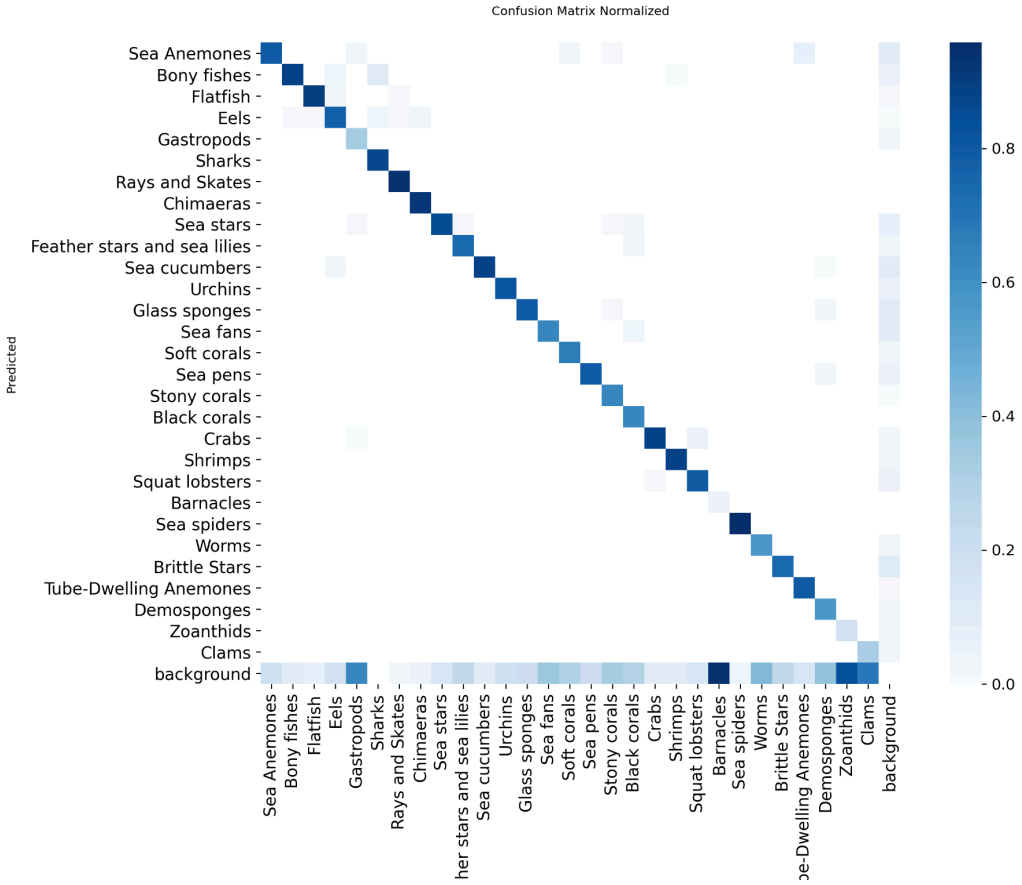

Figure 10. The normalized confusion matrix for the benthic supercategory detector. Normalization was performed column-wise.

From the confusion matrix, we can see that the model performs well for almost every supercategory. A few outliers include barnacles, zoanthinds, and clams which were commonly mistaken for the background. Additional areas of confusion included the fish-like animals (bony fish, flat fish, eels, sharks, etc) and corals (black corals, sea fans, etc). These could be attributed to morphological similarities between the supercategories.

Since these confusion matrices were benchmarked on a validation set, we have a confusion matrix for the midwater detector also, as shown below:

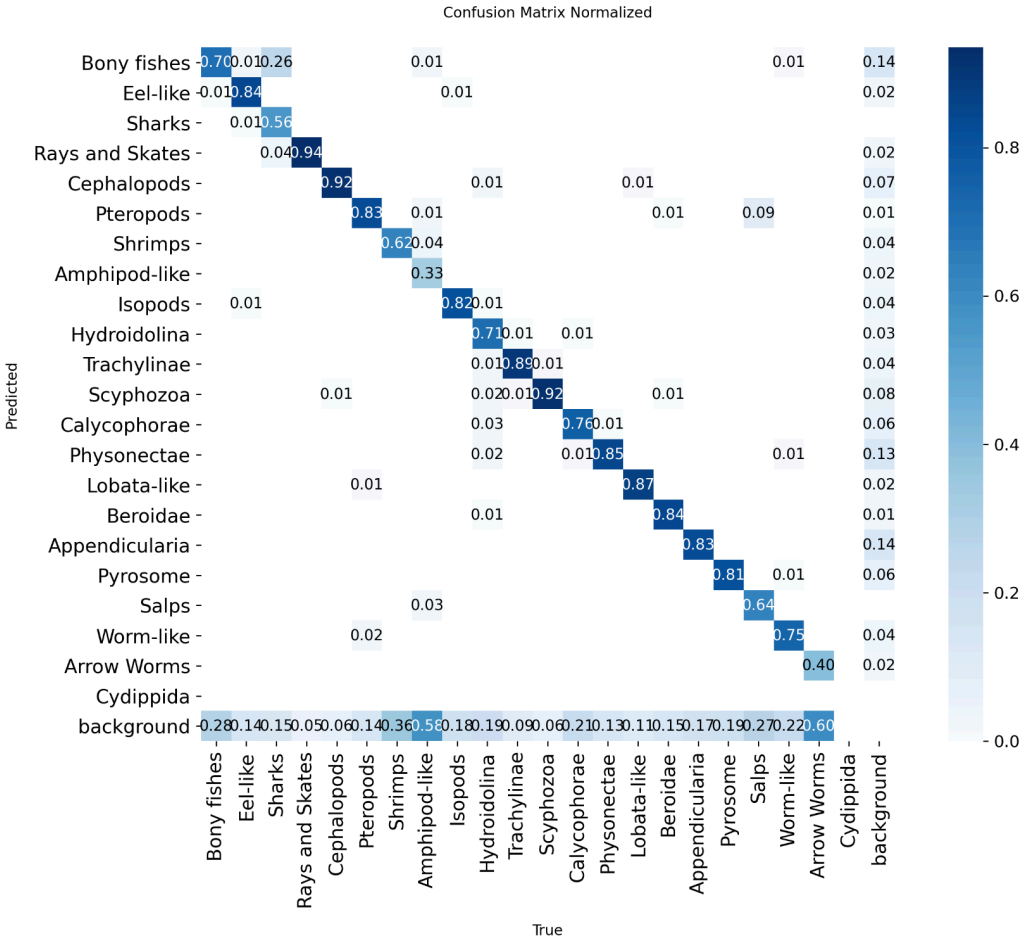

Figure 11. The normalized confusion matrix for the midwater supercategory detector. Normalization was performed column-wise. The Cydippida supercategory was an artifact accidentally leftover from a previous iteration of the supercategories and should not be considered.

QUALITATIVE RESULTS

From a visual inspection of the model prediction mosaics below, it was seen that both the benthic and midwater models perform well in identifying and localizing the majority of relevant animals. Additionally, the differences between midwater and benthic environments are highlighted through these mosaics; images from these different environments are wholly different, with benthic environments being noisier and denser, while the midwater images are much cleaner and usually are focused on a single organism. The results of the models are qualitatively as good, and on occasion better, than the hand-annotated training data that we had used. However, there were still shortcomings that we will discuss in the upcoming sections.

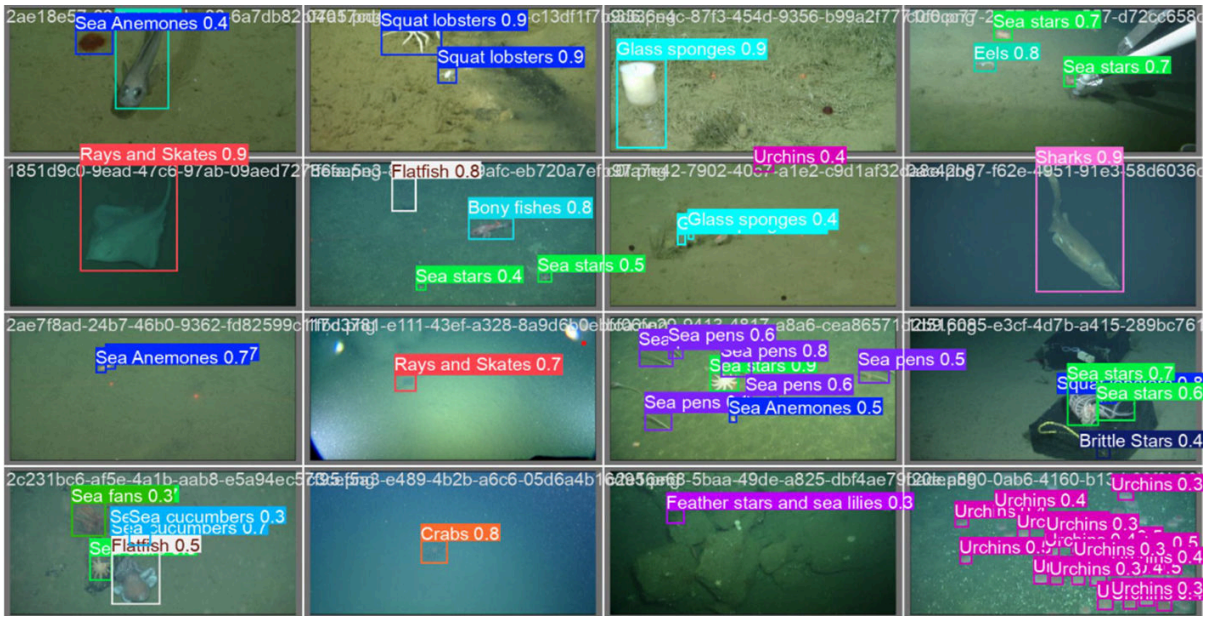

Figure 12. Mosaic of benthic model predictions with an IoU threshold of 0.7, confidence threshold of 0.25, and image size of 640x640.

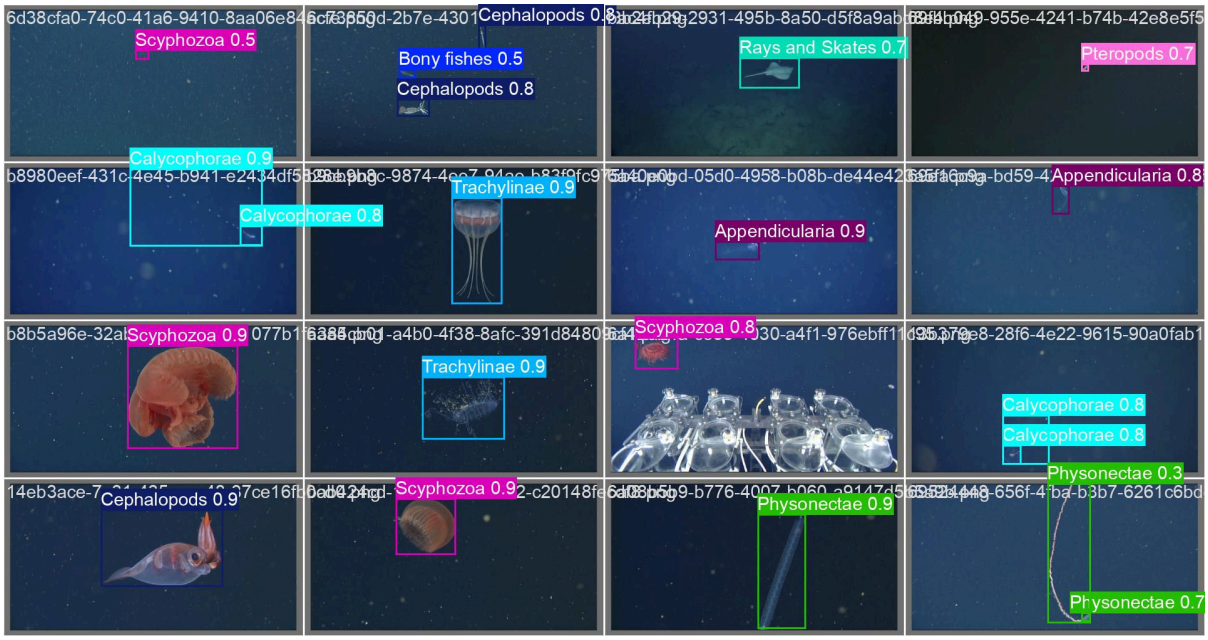

Figure 13. Mosaic of midwater model predictions with an IoU threshold of 0.7, confidence threshold of 0.25, and image size of 640x640.

DISCUSSION

IMPROVEMENTS

With an inference rate of ~5ms per image on a RTX 6000 Ada Generation, these models can provide accurate labels and localizations at a much faster rate than the traditional human-annotation pipeline. From the confusion matrices, we can see that the models do a good job of labelling most supercategories correctly, as well as properly localizing them. Additionally, the model predictions can be seen to be as good or better than the initial FathomNet database training data. For example, figure 14 demonstrates how the benthic model was able to pick out all the sea pens and sea anemones that were originally unannotated in our training data.

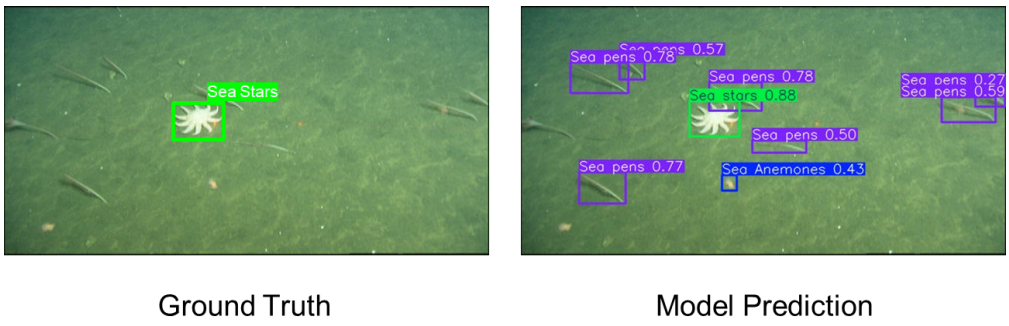

Figure 14. Contrast between the training-data annotations (left) and the model predictions (right). We can see that the model can generate results that are more complete than our training data, and considerably faster too.

SHORTCOMINGS

While the models perform well in labeling and localizing our supercategories, there are still a couple limitations to its abilities. One major limitation is the tendency for the model to mix-up supercategories that are morphologically similar, such as bony fish and sharks. This also extends to smaller organisms that are usually found on the sediment interface; these organisms are commonly mistaken with the background.

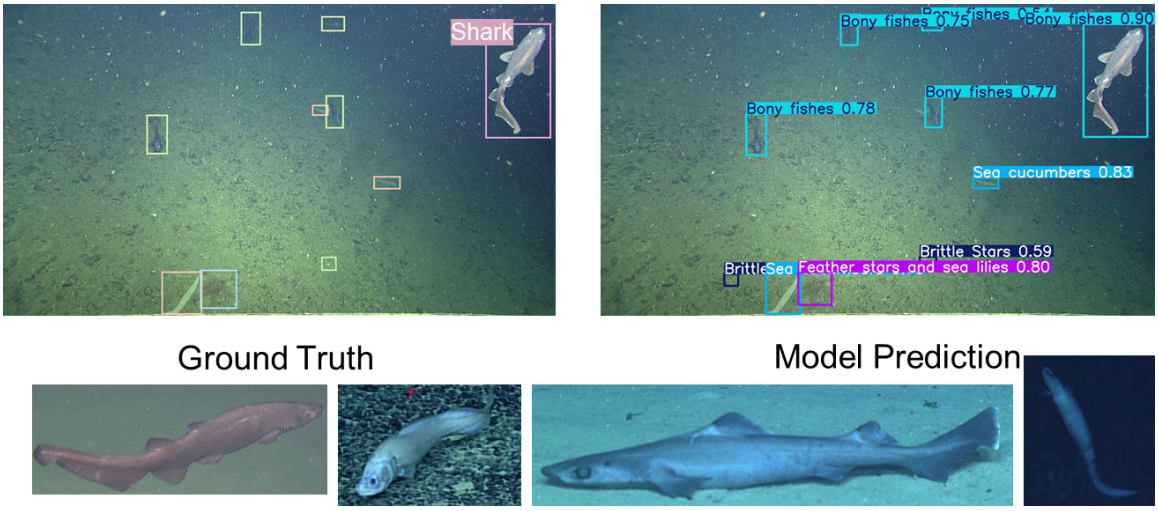

Figure 15. A visual demonstration of the model mixing up bony fish and sharks (top row, right most bounding box). The bottom row of images demonstrates how examples of bony fish and sharks can be morphologically similar (from left to right: shark, fish, shark, fish).

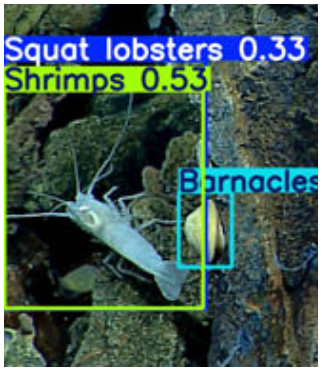

Figure 16. Another image demonstrating the model conflating squat lobsters and shrimps. This is an interesting case where one specimen was actually given two bounding boxes and two labels.

Figure 16 also demonstrates an interesting case where one specimen was given two bounding boxes and labels. The case of one specimen having two bounding boxes generally does not occur due to enacting non-maximum suppression when determining the resulting bounding boxes; however, this was only enabled for overlapping bounding boxes with the same possible label. In this case, where the overlapping boxes had different labels, both boxes were mistakenly kept due to having high enough confidence scores. This highlights both the ability of our model to detect overlapping specimens (eg:

a starfish on coral), while also showcasing its weakness in morphologically similar supercategories.

Another interesting drawback of our models can be seen primarily in the midwater supercategory detector. Due to our original data-cleaning process, where we truncated labels regarding specific body parts of certain specimens, we end up double counting certain supercategories. For example, the FathomNet label “Calycophorae body” and “Calycophorae body” were both truncated down to just “Calycophorae” in our data-cleaning process to maintain our strict taxonomic labels to build supercategories from. As seen in figure 17, this leads to both the specimen's head and body to both be counted as the supercategory, leading to a double count.

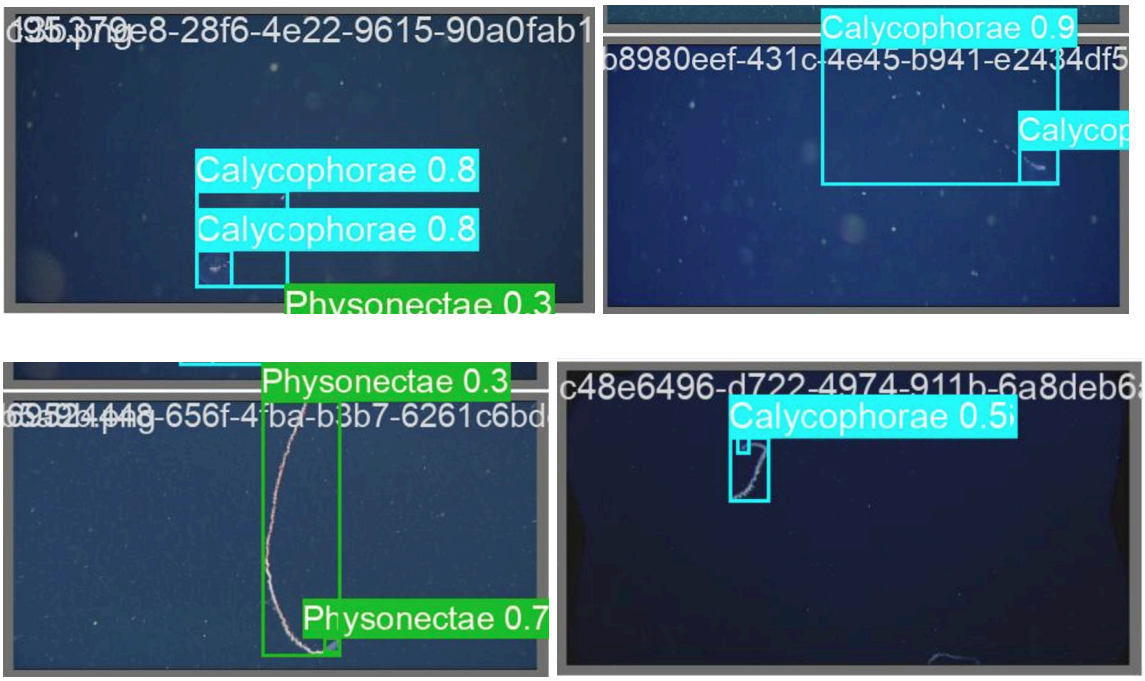

Figure 17. Examples of double counting in the midwater model due to the truncation of FathomNet labels in our data-cleaning process.

Finally, recall is still the model's greatest weakness which can be seen in examples where the model misses very obvious examples. As seen from our quantitative results, the recall of our best benthic model is still around 0.461, indicating that the model is still mispredicting many of the relevant regions as background, likely a result of the partial-coverage training data. However,

Figure 18. An extreme example of the model missing an obvious specimen (the claim in the center of the image)

CONCLUSIONS/RECOMMENDATIONS

For future recommendations, we suggest expanding the full-coverage framework to both include more benthic examples for greater biodiversity and midwater examples. Additionally, we hope to experiment with new methods to extract the most performance from a noisy dataset such as using pseudolabels, artificial generated examples, upsampling, and more. Finally, it would be prudent to work on improving model inference speeds by trying out smaller models (such as a YOLO11n or YOLO11s) or TensorRT optimizations in order to enable real time detections for a video stream.

Overall, we conclude that object detection models can optimize the annotation and analysis workflows for ocean scientists and ecologists. These models can be created by finetuning YOLO models with large, expertly-labeled, public repositories as long as the noise in training images is accounted for. Ultimately, the models were successful in detecting a significant portion of relevant animals in marine imagery, demonstrating that

object detection models, coupled with human-in-the-loop oversight to review proposed annotations, offer a potential solution for accelerating marine ecological research.

ACKNOWLEDGEMENTS

First of all, I'd like to thank the entire Bioinspiration lab for taking me in this summer and creating such a welcoming environment. To Laura Chrobak, Kevin Barnard, Giovanna Sainz, Joost Daniels, Kakani Katija and everyone else in the lab, thank you for being the GOAT mentors this summer; I learned so much and have never felt more inspired in the work I've been doing. Next, I'd like to give a huge shoutout to my roommates Korneel Somers, Tim Sananikone, and the occasional Conor Gagliardi for making Simple Pleasures feel like a home this summer. I'd also like to thank the rest of the interns for making this summer an absolute blast; y'all are incredible and are going to do great things. Finally, I'd like to thank George Matsumoto, Jessica Chapman, and Megan Bassett for having me and running this incredible program - this was one of the most impactful summers I've ever had, and it was truly life-changing!

The MBARI Summer Internship Program is generously supported through a gift from the Dean and Helen Witter Family Fund and the Rentschler Family Fund in memory of former MBARI board member Frank Roberts (1920-2019) and by the David and Lucile Packard Foundation. Additional funding is provided by the Maxwell/Hanrahan Foundation. For my specific project, I was also funded by an NSF grant for which I am extremely thankful.

References:

1. Jocher, G., Qiu, J., & Chaurasia, A. (2023). Ultralytics YOLO (Version 8.0.0) [Computer software]. https://github.com/ultralytics/ultralytics

2. Katija, K., Orenstein, E., Schlining, B., Lundsten, L., Barnard, K., Sainz, G., Boulais, O., Cromwell, M., Butler, E., Woodward, B., & Bell, K. L. C. (2022). FathomNet: A global image database for enabling artificial intelligence in the ocean. Scientific Reports, 12(1), 15914. https://doi.org/10.1038/s41598-022-19939-2