Open Ocean • Here be Dragons 2018

Big Ocean, Big Data Report

**Kakani Katija¹ Eric Orenstein¹ Brian Schlining¹ Lonny Lundsten¹
Kevin Barnard¹ Giovanna Sainz¹ Océane Boulais² Alexandra Lapidés¹
Benjamin Woodward³ Katy Croff Bell⁴**

¹Monterey Bay Aquarium Research Institute, ²MIT Media Lab, ³CVision AI,

⁴MIT Media Lab & National Geographic Society

Published on: Aug 02, 2019

DOI: <https://doi.org/10.21428/a680be9a.207c346e/853ce745>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Introduction

The Big Ocean, Big Data project emerged from the Here Be Dragons event at the MIT Media Lab in February of 2018. The stated goal was to:

“... establish FathomNet, a new baseline image training set, optimized to accelerate development of modern, intelligent, automated analysis of underwater imagery.”

In addition to the data set, we also aimed to create baseline, modern, deep learning algorithms to complete the following tasks:

1. Weakly supervised localization: It was desired to have an algorithm that could propose bounding boxes to be manually corrected and verified by human experts. This is a workflow that is known to accelerate annotation time, and reduce the strain of producing so many annotations for a user.
2. Image level labels: The structure of MBARI's data set was such that most of the frame grabs had image level labels associated with them. It was desired to create a baseline algorithm for that task.
3. Bounding Box Algorithm: The ultimate goal of the FathomNet data set was to enable modern Convolutional Neural Network (CNN)-based object detection and classification algorithms to be developed for species that existed in MBARI's data sets. This would be the final algorithm that showed the viability and potential of the data set.

Executive Summary

The FathomNet data set consists of frame grabs from MBARI's benthic and midwater remotely operated vehicle (ROV) dives in Monterey Bay. In total there are:

- ~ 60,000 images with single concept labels
- ~3000 images with bounding boxes and labels, spanning 198 classes, and totaling >23,000 total bounding boxes

There have been three high-level algorithms trained using this data, corresponding to the three algorithm goals from above. The weakly supervised localization algorithm did not prove as useful in creating bounding box proposals as anticipated, but it did prove to be a useful algorithm for unsupervised tracking and detecting inputs in certain situations, even on data completely separate from the training set (e.g. National Geographic Society's Deep Sea Camera System and NOAA's ROV *Deep Discoverer* video). Additionally, in the course of processing the image label data set, experiments were run on multi-label image algorithms, pointing the way to interesting future algorithm development. Finally, the object detection algorithm trained on the subset of bounding box images has opened up avenues for hierarchical labeling and crowdsourcing.

Data Investigations

MBARI uses high-resolution video equipment to record hundreds of remotely and autonomously operated vehicle dives each year. This video library contains detailed footage of the biological, chemical, geological, and physical aspects of each deployment. Since 1988, more than 23,000 hours of videotape have been archived, annotated, and maintained as a centralized MBARI resource. This resource is enabled by the Video Annotation and Reference System (VARS), which is a software interface and database system that provides tools for describing, cataloguing, retrieving, and viewing the visual, descriptive, and quantitative data associated with MBARI's deep-sea video archives. All of MBARI's video resources are expertly annotated by members of the Video Lab (VL), and there are currently more than 6 million annotations and 4,000 terms in the VARS knowledgebase, with over 2,000 of those terms belonging to either genera or species.

Using the VARS Query tool, we determined a list of annotations that described different genera and geologic features. Of those annotations, each genus and geologic feature were ranked by the number of associated frame grabs. We then selected the top 18 midwater genera, the top 17 benthic genera, and the top 3 geologic features to incorporate into FathomNet ([Figure 1](#)). As the initial phases of this effort were on automated classification of “things” (or animals) instead of “stuff” (or geological features), we divided the image set into Midwater and Benthic classes.

Figure 1
Benthic (left) and Midwater (right) classes found in FathomNet.

Midwater

The frame grabs and labels for the top midwater concepts were of 18 midwater genres, with a mix of iconic and non-iconic views. We trained a classifier for 15 species, filtered based on the number of images for each species, with a cutoff of approximately 1,000 images per species. This resulted in a data set of 33,064 images.

The nature of the frame grabs and the distribution of species meant that images often only contained a single species of interest, which corresponded to the label that came along with the image, though not always. In fact, it worked well enough that an algorithm was able to be trained for the concepts using only the single image labels, and was able to identify multiple concepts within an image using a Top-N scoring methodology. This methodology can be described as:

1. Run the algorithm and obtain a sorted list of scores for each concept.
2. If the correct label for an image is in the Top N (e.g., $N=3$) entries in the list, it is counted as a true positive
3. If the correct label for an image is not in the Top N it is a false negative, and the image is given a label for the top scoring concept, producing a false positive for that concept.

There are variations to this methodology that adopt thresholds for the confidence scores, as well as other heuristics to improve performance, but the main technique is as described above. A rigorous application of this technique would produce false positives for each of the Top N concepts, but the point was to illustrate relative accuracy, as well as to produce an idea of what classes were being misidentified. [Figure 2](#) shows the resulting accuracies and confusion matrix for the FathomNet data set. Using the aforementioned scoring methodology, we found the Top 1 and Top 3 Accuracies to be 85.7% and 92.9%, respectively.

Figure 2
Confusion matrices for Benthic algorithms (top) and Midwater algorithms (bottom).

These results indicate acceptable performance for the algorithm used, and we next tried to evaluate the algorithm on a Midwater Transect data set. Due to differences in imagery between the Midwater Transect data (e.g., differences in scale and resolution of objects) and the FathomNet training data, we found this performed poorly, thereby necessitating that the intended test set matches your training set to the largest extent possible.

Figure 3
Figure 3 - Example images from the training set (top) and transect set (bottom) for Poebius.

We found that for many midwater species of interest, the FathomNet training set consisted of a fair amount of zoomed in, iconic views of animals. These views contrasted sharply with the limited spatial resolution of targets in the Midwater Transect footage. The difference in spatial scales was akin to showing someone iconic pictures of sports cars, and asking them to identify these vehicles from aerial photos.

Benthic

The Benthic imagery consisted of frame grabs from two different data types. The first data type included geological features (e.g. pillow lava), while the second type were animals to the genus level (e.g. *Sebastolobus*). Drawing on the nomenclature used in the Panoptic Segmentation ([Kirilov et al., 2019](#)) task, these correspond to “stuff” and “things”. For algorithm development in the initial phase, we decided to forego the geological frame grabs (“stuff”), and focus only on animals (“things”).

The frame grabs were selected using the same methods described above, and consisted of 17 classes, where each described a benthic animal genus. We initially trained an algorithm on 12 of the classes, filtering by number of images. We initially set the minimum number of images for the training set at 700, but later removed this limit to try and include classes that were abundant in the test data set, as there was a mismatch between abundances in these two data sets. This resulted in a data set of 15 classes, with a total of 33,064 images. We then tested the training set against a Benthic Transect data set, and using the Top N methodology as described above, we obtained Top 1 and Top 3 Accuracies of 72.4% and 92.8%, respectively. An immediately noticeable difference between the Benthic and Midwater results is the Top 1 vs Top 3 accuracy metrics. There is a marked improvement in the Benthic data set moving to Top 3 accuracy. The reason for this is rooted in the fact that the images in FathomNet for the benthic classes are strongly multi-concept, while the Midwater images tended to be mostly single-concept. In many cases the label assigned to a benthic image was not the dominant concept from the data set within the image. This created a number of challenges, but also opened up interesting research areas in multi-label image algorithms and noisy data sets. An example of a multi-concept image is shown in [Figure 4](#). Therefore, the remainder of our efforts and discussion on algorithm development focuses on the benthic imagery because of the more appropriate nature of the test transect video, as well as the similar nature of video and images to other collaborators’ data.

Figure 4
Multi-concept example, showing Funiculina and Chionoecetes

Algorithms

We experimented with three types of algorithms that share some common features but accomplish slightly different goals. Each algorithm is based on a Convolutional Neural Network (CNN) “backbone”, which can be viewed as a feature extractor. We then built an image classification algorithm (using ResNet50), which assigns a single label from N possible labels to an image. After that, we used weakly supervised localization (using GradCam++) to find object instances with the same training data as used in image classification. Finally, we trained an object detection algorithm (RetinaNet) using extensive, multi-instance localization data generated by the MBARI Video Lab. The details of each algorithm and our experiences with them are described in the following sections.

Classification Algorithm - ResNet 50

The classification algorithm that we used is based on a CNN architecture known as ResNet. ResNets have architectural features that allow them to have many more layers of filters than other network architectures such as VGG, while still maintaining computational tractability, as detailed in ([He et al., 2016](#)).

We took advantage of a technique known as Network Transfer Learning ([Oquab et al., 2014](#)) to leverage the idea that “...the internal layers of the CNN can act as a generic extractor of mid-level image representation, which can be pre-trained on one dataset (the source task, here ImageNet) and then re-used on other target tasks...”. We downloaded a pre-trained ResNet50 architecture (<https://keras.io/applications/>), and fine-tuned it, retraining the fully connected layers to discriminate between the classes in our training data sets.

As mentioned earlier (see Data Investigations), one of the limitations that we encountered in our training and testing of Benthic imagery was that they were strongly multi-label images, whereas the standard image classification task for ResNet50 architecture is normally applied is strongly single-label. There are ways to train networks, such as ResNet, for multi-label problems ([Gardner & Nichols 2017](#); [Wang et al., 2018](#); [Li & Yeh, 2018](#); [Wang et al., 2017](#)), but our data set only provided single-label imagery ([Figure 4](#)). Future efforts to generate multi-label imagery to address this discrepancy are currently underway.

Weakly Supervised Localization Algorithm - GradCAM++

In an attempt to automate generation of bounding box information from labeled data, we turned to a technique known as weakly supervised localization. This is still a very active area of research, with many potential pathways to pursue ([Gao et al., 2018](#); [Najibi et al., 2018](#); [Papadopoulos et al., 2016](#)). The instantiation that we chose to use is known as GradCam++ ([Chattopadhyay et al., 2018](#)). The basic idea of this technique is to use the backpropagation of the loss function on an image to identify the parts of that image that most strongly contributed to the label determination. This results in a pseudo saliency map that can be combined with morphological operations to automatically propose bounding boxes for objects of certain classes in an image. There are two ways to generate these types of saliency maps and bounding box proposals: class specific search, and dominant class search. Once the class to search for has been selected, the algorithm operates exactly the same for each search technique. [Figure 5](#) shows how this technique works. The operating method for the example in the figure was dominant class.

Figure 5

Example of bounding box proposal generation using GradCAM++ on Sebastolobus with (left) the original image, (center) the GradCam++ saliency heatmap, and (right) a bounding box proposal using segmentation of the GradCam++ output.

When an image was strongly single-concept with limited numbers of object instances, the technique worked remarkably well (e.g., midwater footage). However, since many of the images were multi-concept and contained large amounts of object instances, we found GradCam++ performed inconsistently with these data. Despite these data limitations, these results proved promising, and has emerged as another area for future work.

Object Detection Algorithm - RetinaNet

The final class of algorithms that we employed are known as object detection algorithms. These techniques again use CNNs as feature extractors, with additional network components used to create object detection proposals (usually bounding boxes). Research on object detection algorithms using CNNs started to gain prominence in 2013 with the introduction of Overfeat ([Sermanet et al., 2013](#)). The introduction of R-CNN in 2014 ([Girchick et al., 2014](#)) and YOLO ([Redmon et al., 2016](#)) started a fast pace of object detection research that continues today. An overview of object detection frameworks based on CNNs can be found in [Huang et al., 2017](#), as well as the performance trade-offs inherent in various architectures.

We chose to use RetinaNet ([Lin et al., 2017](#)), which is a single-stage object detection algorithm that uses bounding box anchor proposals to try and simultaneously localize and identify objects of interest. In order to use this algorithm, the MBARI Video Lab exhaustively annotated over 3000 images with bounding box information for over 200 species. This resulted in approximately 23,000 bounding box annotations. As with most data sets of this type, we suffered from the long tail problem, with the large majority of the annotations belonging to only a handful of classes. The results of training on the full 200 species were that the algorithm did fairly well at drawing bounding boxes around many objects of interest within a scene, but labeling was not effective.

In order to overcome this limitation, two approaches were approached. The first was to train the object detector to truly only be an object detector, collapsing all labels to a single “object” label; a second algorithm could then be trained to identify the content of each of the bounding boxes. The second approach involved grouping the labels into different hierarchies. In this way, we theorized that we might be able to leverage more annotators with lesser expertise to generate vastly more bounding box labels at a higher taxonomic level, with expert annotators being able to quickly move between bounding boxes of appropriately labeled data, drill down the hierarchy to a more specific label, and eventually be able to train multiple algorithms at different levels of the hierarchy on less data. [Figure 6](#) shows an example of how this workflow would work.

We chose the latter approach because of the attractiveness of a semi-automated workflow that could emerge. Two promising hierarchies were identified, one based on a high-level concept (e.g. fish, crustacean, etc.), and one based on morphological appearance (fish-like, laterally flattened, fan-like, plant-like, etc.). [Figure 7](#) shows an example of results using this the morphological grouping.

Figure 6

Example hierarchy showing how a multi-tiered algorithm could be used. Current algorithm efforts (green) will work to help humans efficiently label down to the more detailed level (blue), and eventually a multi-stage algorithm could be developed to conduct hierarchical classification.

Figure 7

Example image with object detection and classification results using the morphotype grouping. The colors represent the following classes: Orange - Fan-like, Lime Green - Plant-like, Neon Green - Crustacean-like, Pink - Flower-like, Yellow - Star-like.

Results on External Partner Footage

In addition to running these experiments on MBARI footage, we were also able to obtain video from NOAA's ROV Okeanos, as well as National Geographic Society's DropCam. We applied GradCAM++ on videos from each of these sources, and obtained very promising results. [Figure 8](#) shows example screenshots from each of these videos with the GradCAM++ saliency maps overlaid; the QR codes can be used to view the footage. As we continue to iterate on these object detection algorithms, we will also incorporate results on these additional imagery, allowing us to greatly expand our data set for object annotations.

Figure 8

GradCAM++ results from (left) NOAA's ROV Okeanos and (right) National Geographic Society's DropCam respectively. For both images, the model used was a 12 class ResNet50 trained on MBARI's benthic data. For the Okeanos clip (left), the model generated a saliency map for the class *Sebastolobus*, correctly detecting multiple *Sebastolobus* in the footage. For the DropCam clip (right), the model generated a saliency map for the class *Merluccius*, detecting a similar-looking class of animals in the footage.

Future Steps

Here we briefly summarize future potential areas of investigation.

Hierarchical labeling process

We have learned that automated image and video annotation should start at a higher taxonomic depth. We will be investigating how to use algorithmically generated labels at a higher level along with annotators of lesser expertise, to accelerate the annotation of large amounts of data for further refinement and algorithm development. One promising path forward is the use of few-shot learning ([Chen et al., 2019](#)) to help refine taxonomic labels. In this paradigm, an object detection algorithm trained at one level of hierarchy would create

object detections, filtered to the label that we wish to refine. Then using an iterative training and evaluation cycle, we would rapidly train models for sub labels of interest. In this way we would be able to rapidly split a hierarchy, even with relatively few labels. These noisy, single label detectors can then be used in conjunction with the high performing object detection algorithm to build up a new level of the taxonomy. Once sufficient labels have been produced using this technique, a retraining of the object detection algorithm can be performed, providing a new baseline for object detection.

Different types of annotations

The main focus of our algorithm efforts on FathomNet pertained to single-label image annotations or object-level bounding boxes. There are a variety of different types of annotations that can inform other workflows. One such avenue is multi-label annotation, which would reduce a large amount of the noise associated with training single-label algorithms on multi-label imagery. Very large taxonomic multi-label algorithms are an open area of research.

Another annotation type is using segmentation masks, both for instance segmentation of objects, or “things”, as well as semantic segmentation of scenes, or “stuff”. These masks can help characterize benthic scenes for example, or provide more information for recognition algorithms, as seen in Mask R-CNN ([He et al., 2017](#)). One of the difficulties in segmentation approaches is generating training data. This involves drawing appropriate boundaries around every of object of interest, a task even more tedious than drawing bounding boxes. Recent efforts in this area have provided significant speedups for this task ([Ling et al., 2019](#); [Acuna et al., 2019](#)).

Open source competitions

One of the reasons that everyday object recognition has improved so rapidly is that there are large scale, public data science competitions held on existing image datasets (e.g. MS COCO). As the scale of our annotated data set increases, we would like to explore hosting a leaderboard-type submission server, as well as potentially organizing similar competitions to invite attention to our growing data set. For instance, a similar effort by iNaturalist (<https://sites.google.com/view/fgvc5/competitions/inaturalist>) has a data set with 8,000 species, and 450,000 images. This competition is limited to single-concept images, and in order to get algorithms that can contribute to existing video analysis workflows, we would want to explore using multi-label annotations.

Adding partner data

Now that we have established a workflow, we plan to start gathering, processing, and annotating other data using the same standards/procedures we developed here. The two immediate sources of data that we would like to start incorporating into FathomNet come from NOAA’s ROV *Deep Discoverer* and National Geographic Society’s Deep Sea Camera System program. Both of these data sources also have parallel analysis efforts associated with them, and we would like to work closely with those teams to try and find a common

representation for different annotations that can support the larger FathomNet data set, as well as develop more robust algorithms for each of these individual projects by leveraging data from FathomNet.

Hosting of data and algorithms with public access

The end goals of this project is to increase access to both video and imagery for underwater research, and to do so by providing easy access to helpful automated processing tools for parsing massive volumes of data. We are going to be investigating the creation of a platform that can host the FathomNet data, as well as a platform for contributing videos, imagery, algorithms, other tools, and accessing already contributed tools to analyze contributed media.

Acknowledgements

Funded by National Geographic Society, NOAA Office of Ocean Exploration & Research, Monterey Bay Aquarium Research Institute, and Open Ocean.

References

- A. Chattopadhyay, A. Sarkar, P. Howlader and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, 2018, pp. 839-847.

[↩](#)

- Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C. F., & Huang, J.-B. (2019). A Closer Look at Few-shot Classification. In International Conference on Learning Representations. Retrieved from <https://openreview.net/forum?id=HkxLXnAcFQ>

[↩](#)

- D. Acuna, A. Kar, & S. Fidler (2019). Devil is in the Edges: Learning Semantic Boundaries from Noisy Annotations. CoRR, abs/1802.07460.

[↩](#)

- D. P. Papadopoulos, J. R. R. Uijlings, F. Keller and V. Ferrari, "We Don't Need No Bounding-Boxes: Training Object Class Detectors Using Only Human Verification," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 854-863.

[↩](#)

- Gao, M., Li, A., Yu, R., Morariu, V.I., & Davis, L.S. (2018). C-WSL: Count-Guided Weakly Supervised Localization. ECCV.

[↩](#)

-

[↑](#)

- H. Ling, J. Gao, A. Kar, W. Chen, & S. Fidler (2019). Fast Interactive Object Annotation with Curve-GCN. CoRR, abs/1903.06874.

[↑](#)

- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778.

[↑](#)

- J. Huang et al., "Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 3296-3297.

[↑](#)

- J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 779-788.

[↑](#)

- K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2980-2988.

[↑](#)

- Kirillov, Alexander, Kaiming He, Ross B. Girshick, Carsten Rother and Piotr Dollár. (2019) Panoptic Segmentation. IEEE/CVF Conference on Computer Vision and Pattern Recognition: 9396-9405.

[↑](#)

- Li, Y., & Yeh, M. (2018). Learning Image Conditioned Label Space for Multilabel Classification. CoRR, abs/1802.07460.

[↑](#)

- M. Najibi, F. Yang, Q. Wang and R. Piramuthu, "Towards the Success Rate of One: Real-Time Unconstrained Salient Object Detection," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, 2018, pp. 1432-1441.

[↑](#)

- Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. pp. 1717–1724. IEEE (2014)

[↑](#)

-

↑

- Sermanet P., Eigen D., Zhang X., Mathieu M., Fergus R., & LeCun Y. (2013, Dec 21). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. Retrieved from <https://arxiv.org/abs/1312.6229>.

↑

- T. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2999-3007.

↑

- Wang, Q., Jia, N., & Breckon, T.P. (2018). A Baseline for Multi-Label Image Classification Using Ensemble Deep CNN. CoRR, abs/1811.08412.

↑

- Z. Wang, T. Chen, G. Li, R. Xu and L. Lin, "Multi-label Image Recognition by Recurrently Discovering Attentional Regions," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 464-472.

↑