FathomNet: Building the Dataset

Kevin Barnard, Colorado School of Mines

Mentors: Kakani Katija, Joost Daniels

Summer 2018

Keywords: oceanography, big data, machine learning

ABSTRACT

Annotation and organization of collected oceanographic visual data proves to be a significant challenge for the oceanographic community. As the quantity and quality of collected data has increased, manual annotation approaches have become costlier and more intensive. Machine learning techniques, particularly in the area of deep convolutional neural networks (CNNs), have produced significant advances in the task of image classification. The *Big Ocean, Big Data* project is a joint effort of the oceanographic research community to create a public platform for semi-autonomous video annotation. This paper discusses the construction of *FathomNet*, an image dataset created with imagery from the Monterey Bay Aquarium Research Institute (MBARI) to train machine learning models for image classification of oceanographic video. The *FathomNet* dataset currently contains 77100 images, spread over 38 concepts and 3 high-level categories: Benthic, Midwater, and Geology.

1 INTRODUCTION

The task of oceanographic visual data collection and annotation has been growing at a rate that manual approaches struggle to match. As an increasing number of underwater vehicles are gathering higher-quality, multi-feed imagery, the structured intake of this data has become a significant problem. Oceanographic institutions address this issue through varied techniques (Gomes-Pereira 2016), leading to incongruence across the community and making coordinated efforts more challenging. The *Big Ocean, Big Data* project, as originally proposed in March 2018, looks to solve this big data problem by creating an open ocean community platform for automated image classification in the deep sea.

The current limiting factors of such a process include (1) a lack of dataset standardization, (2) sparse annotation tools, and (3) insufficient formatting of existing imagery (Katija et al. 2018). *FathomNet*, previously referred to as *TBDNet*, is the oceanographic image dataset that has been developed to be the basis for a public platform aimed to obviate these issues. This paper will discuss the materials and methods used in developing *FathomNet* as well as the future of the platform.

2 MATERIALS AND METHODS

2.1 GENERAL WORKFLOW

The process for generating the *FathomNet* dataset is illustrated in Fig. 1. Selected lists of concepts are passed to *vars-download*, a Java application that generates queries to the MBARI Video Annotation and Reference System (VARS) database. The VARS database is comprised of manually derived annotations corresponding to individual frames in the MBARI video archives, which span over 30 years of remotely operated vehicle (ROV) observations. The results of the VARS queries are parsed by the *vars-download* application, which then downloads the corresponding images from the MBARI file server. The saved images are multi-resolution, and some of them contain black letterbox borders. To resolve this, the images are passed to another application (*bobd-borders*) to remove the black borders. Finally, the cropped images are watermarked and tagged by a third Java application (*bobd-watermark*), then uploaded to the *FathomNet* FTP server.

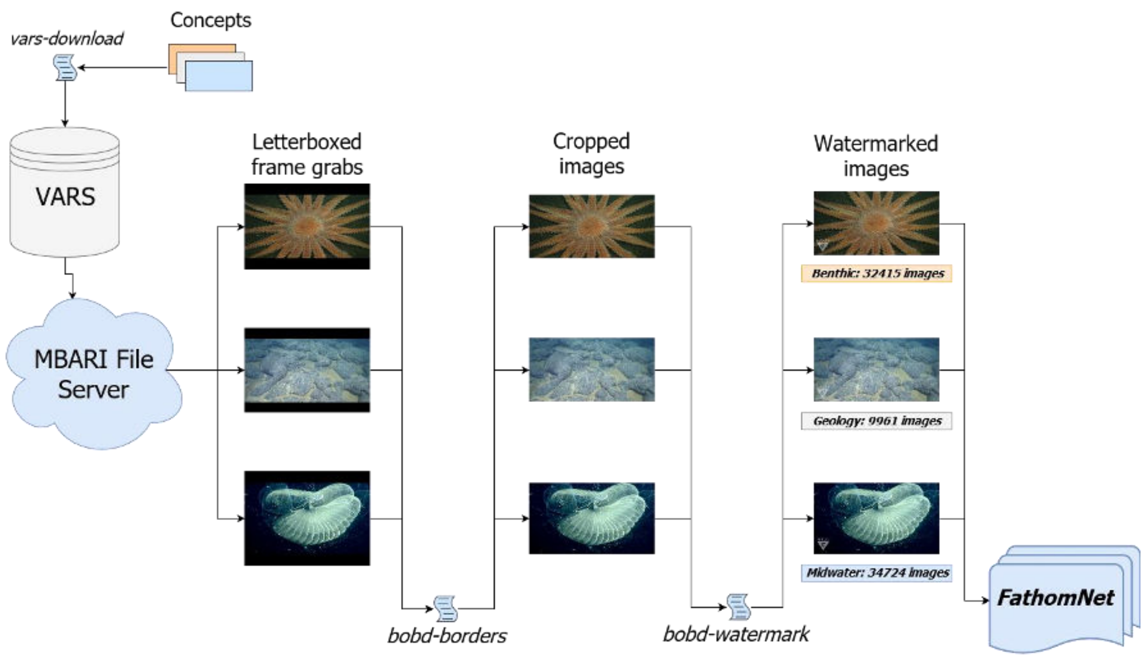

Figure 1. Process diagram for generating the base FathomNet image dataset

2.2 DATABASE CONCEPT QUERYING

An annotation within VARS serves as a link between the sighting of a particular object or animal and its timestamp on a particular video tape, as well as all associated metadata (e.g., depth, temperature, dissolved oxygen, latitude, longitude, etc.) collected by the ROV at that instant. As footage is annotated, certain moments with a view of the concept are frame grabbed, and compressed/uncompressed digital image files are uploaded to the MBARI file server. To determine the concepts to be entered into the first version of the *FathomNet* dataset, VARS was queried to determine the number of annotations on the database with linked images, or frame grabs. On average, the proportion of annotations with associated frame grabs at the genus level was 26.4%. As VARS does not have an internal taxonomic hierarchy, the image count per genus was considered to be the number of frame grabs recorded at the genus level added to the number of frame grabs recorded for each genus' species, where each image was counted distinctly to avoid duplicates. The Java *vars-summary* application was written to perform this process across all concepts listed on the VARS database. The source code is available at https://bitbucket.org/mbari/vars-summary.

The returned concepts were then organized by the number of images associated with the genus. The top 18 midwater genera, top 17 benthic genera, and three geologic concepts were hand-selected to be included in the first version of *FathomNet*. The concepts were then organized into three lists, as shown in Appendix A.

2.3 IMAGE GATHERING AND ORGANIZATION

In order to gather the images, queries to the VARS database had to be generated for each concept. To handle these queries, the Java *vars-download* application (source code for *vars-download* is available at https://bitbucket.org/mbari/vars-download) was written. The application takes an input of a list of concepts to query, then generates a proper query to get a list of all of the image links for each concept. Each image link is translated to its uncompressed PNG file equivalent on the MBARI file server, and invalid links (where files are missing/damaged) are discarded. Each image is then downloaded and saved to a folder corresponding to its genus, and then to a subfolder corresponding to its species.

Additionally, a small web page (currently not publicly available) was written to visualize the dataset. This page includes a display of the file structure of the dataset as well as an application to display basic taxonomic information about each concept. An example usage of the web page is shown in Fig. 2.

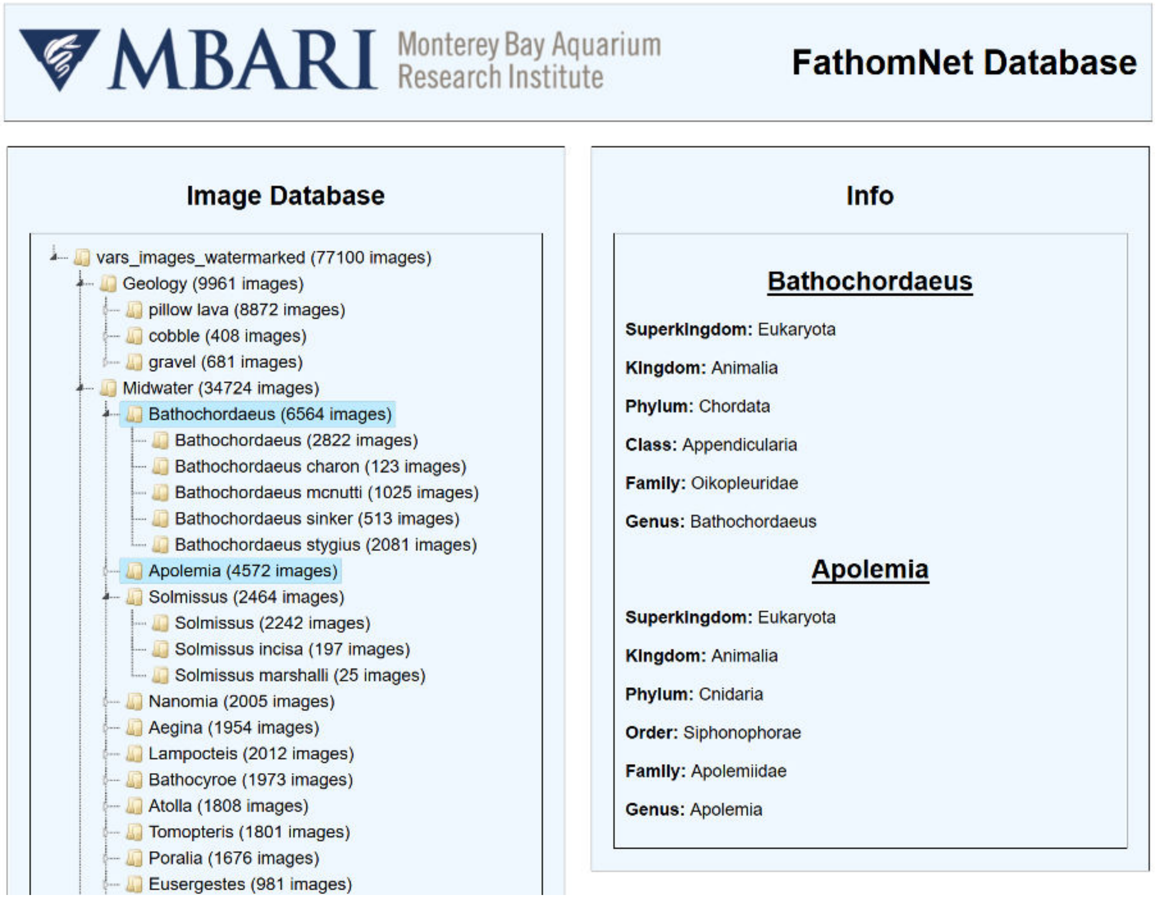

Figure 2. FathomNet dataset visualization web page demonstration

2.4 IMAGE DATA CLEANING

The images downloaded from the MBARI file server span a variety of resolutions (ranging from 640x486 to 1920x1080), and many contain black letterbox borders. In order to make a more consistent dataset, the borders were removed from the images. The borders, however, are not fully black; they often contain bits of color as artefacts of the video tape digitization. In the case of an image with an especially dark background, a strong gradient to define where the letterbox border ends and the image begins is lacking. Additionally, the valid region for each image is not perfectly centered in its frame grab.

Adaptive thresholding proved to be a computationally expensive operation for detecting the black borders, so an alternative method was derived. The central mechanism of the created *bobd-borders* Java application is an iterative approach. Initially, the uppermost row of pixels in an image is iterated. If any pixel in the row has a color component value greater than a specified threshold value, the row is considered a part of the image. Otherwise, it is considered part of the letterbox, and the next row of pixels is considered. A similar process is then performed from the left, bottom, and right. The final product is simply the portion of the original image when each side's first row or column of valid pixels is detected. The process of *bobd-borders* can be seen in Fig. 3. The leftover image is considered to be the inner region of the red, green, blue, and yellow lines.

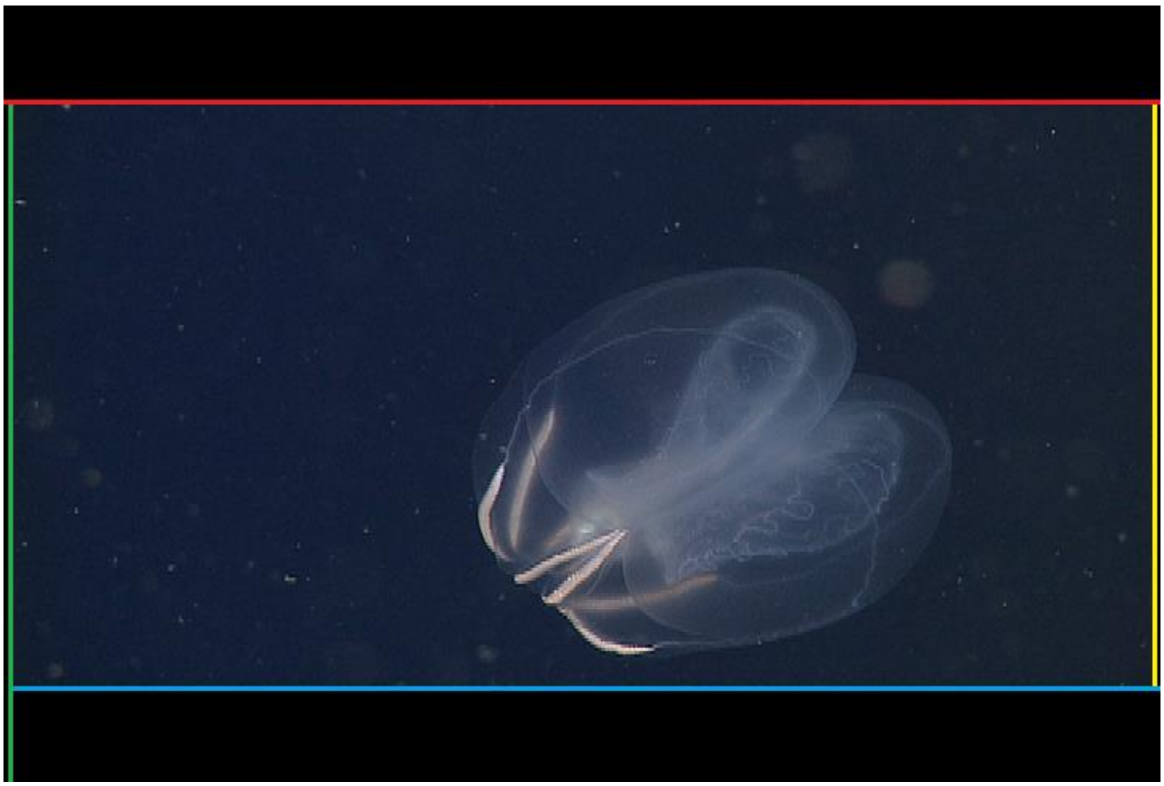

Figure 3. Visualization of the *bobd-borders* algorithm on an image of a ctenophore

Given how the *bobd-borders* algorithm needs to perform calculations over much less of the image than a traditional threshold mask approach, it was more computationally efficient in handling the domain of images within the dataset after a small amount of tuning. The source code for *bobd-borders* is available at https://bitbucket.org/mbari/bobd-borders.

The final step in the data cleaning process is adding the MBARI watermark, which was performed by the Java application *bobd-watermark* written by B. Schlining. Its source code is available at https://bitbucket.org/mbari/bobd-watermark.

3 RESULTS

The baseline version of the *FathomNet* dataset contains 77100 images distributed among 38 concepts, the total counts of which are listed in Appendix A. Before the cropping and watermarking process, 77171 images were downloaded, as 71 images on the file server were invalid. The distribution of the concepts can be seen in Figs. 3, 4, and 5.

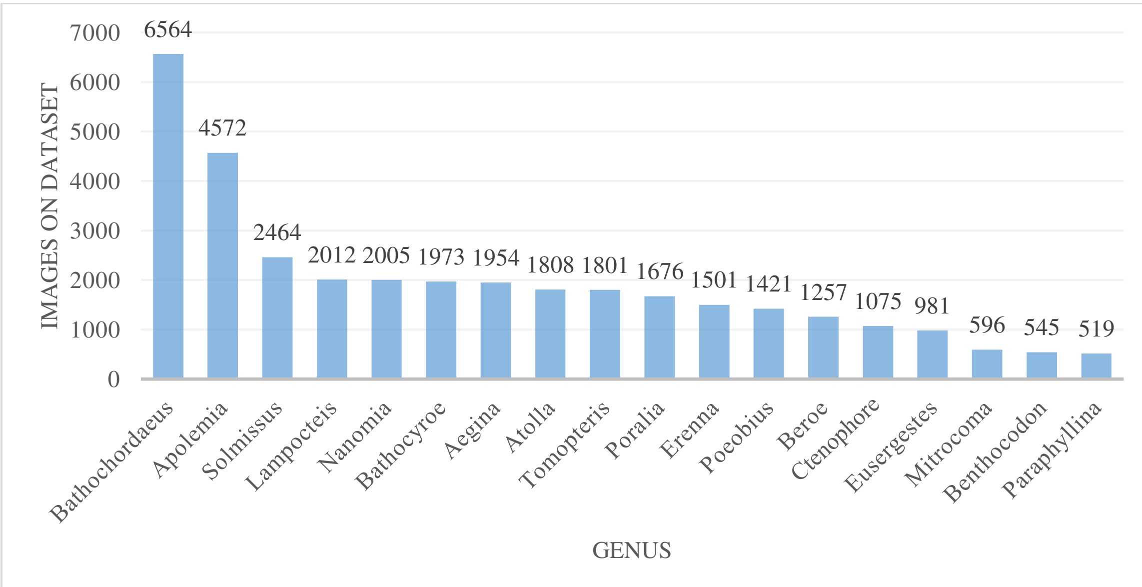

Figure 3. Midwater genera image distribution

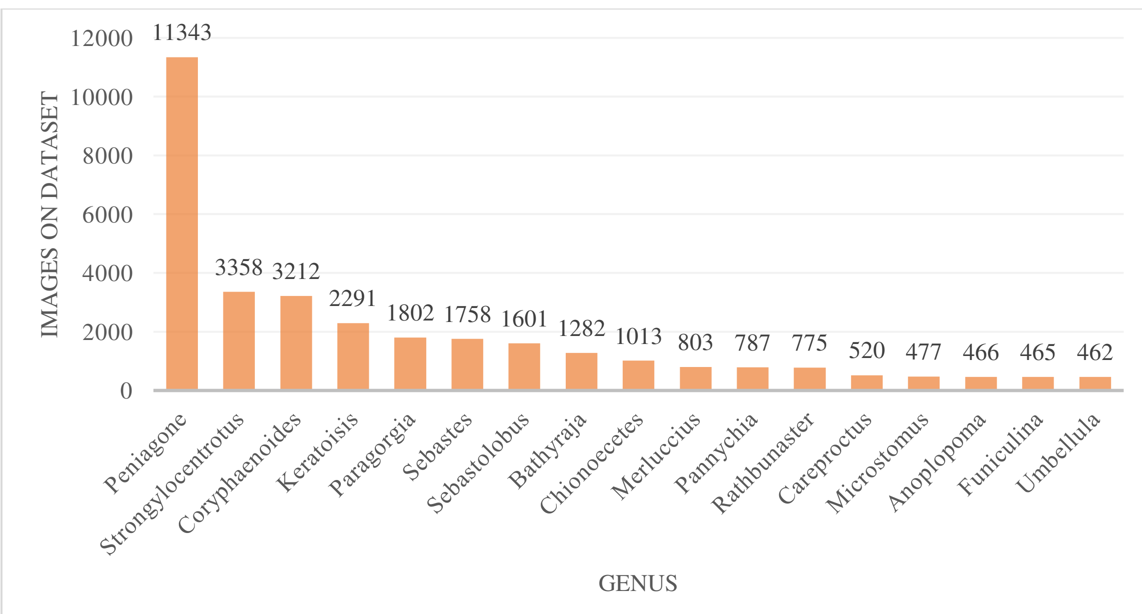

Figure 4. Benthic genera image distribution

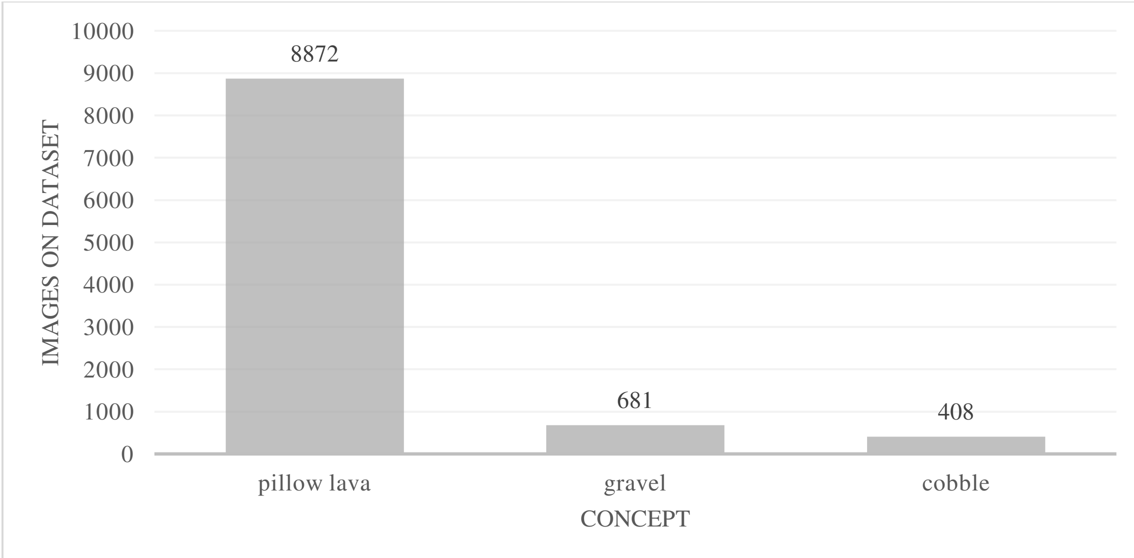

Figure 5. Geology concepts image distribution

The image quality across classes is non-uniform. As the VARS database is over thirty years old, the frame grabs added over time have been sourced from video tapes of varying quality, and thus the frame grabs stored on the MBARI file server are multi-resolution. On top of this, the *bobd-borders* application outputs a wide range of image resolutions when cropping the images. As a result, the images on the *FathomNet* dataset are considerably inconsistent in resolution per class, ranging from 640x364 pixels to 1920x1080 pixels.

4 DISCUSSION

4.1 QUANTITY AND QUALITY

There is significant non-uniformity in the image distribution across concepts in every category. This is a result of an imbalance in the concepts observed during underwater missions which have been annotated on the MBARI VARS database. Because of this skew, machine learning models trained on *FathomNet* could perform varying capacities in identifying different concepts. Though there is no global minimum number of images recommended for a training set, the average class within *FathomNet* contains 2029 images.

In its current state, *FathomNet* is best suited for transfer learning techniques. As transfer learning models train only a shallow portion of the network, they can apply a given pre-trained feature space to a different domain of interest. As a result, transfer learning models require far fewer images to learn a new domain of interest than needed to train a full network from scratch (Pan & Yang 2010).

Most of the popular pre-trained models for transfer learning require a smaller minimum resolution than the smallest-resolution image in the *FathomNet* dataset; for example, the *VGG16* and *ResNet* models require 224x224 pixel images as inputs (Simonyan & Zisserman 2014). Therefore, it can be concluded that all of the images on *FathomNet* are of sufficient spatial resolution for most of the current transfer learning models. It is important to note that the location and size of concepts within the images on *FathomNet* are currently unspecified. In addition, lighting conditions, pose, and orientation of concepts within the dataset are largely varied within each class. An example comparison for these differences can be seen in Fig. 6. Analysis beyond the scope of this paper would be necessary to determine the impact of these factors on different transfer learning models.

Resolution: 720 x 366
Box size: 32 x 38

Resolution: 1920 x 1080
Box size: 315 x 706

Figure 6. Size comparison between concepts in two images of *Bathochordaeus mcnutti*

4.2 DATASET COMPARISON

ImageNet is a large and well-known image dataset that contains over 14 million labeled images (*ImageNet* 2010). It is commonly used as training data in the machine learning community, particularly in conjunction with CNNs. Some of the most popular CNNs for transfer learning, such as *VGG16*, *InceptionV3*, and *ResNet*, have been pre-trained exclusively on the *ImageNet* dataset (Simonyan & Zisserman 2014). The average non-empty large concept, or *synset*, within *ImageNet* contains 650 images (*ImageNet* 2010), which is less than one third of the average image count per concept in *FathomNet*. On the other hand, *ImageNet* contains 21841 concepts (*ImageNet* 2010), compared to the 38 concepts currently distinguished by *FathomNet*.

5 CONCLUSIONS / RECOMMENDATIONS

The *FathomNet* dataset has been given a strong foundation through the process described by this paper. Though it is incomplete, every concept currently included in the dataset has a significant quantity of relatively high-resolution images. For these reasons, the *FathomNet* dataset is sufficient for training transfer learning models for image classification.

To improve the dataset, more images must be added. In its current state, *FathomNet* is built solely from individual frame grabs annotated in the MBARI VARS database. As MBARI's video assets contain over 23,000 hours of archived tapes, MBARI possesses much more data that could potentially be digitized and added to the *FathomNet* dataset (Katija et al. 2018). As mentioned above, only 26.4% of the annotations currently listed on VARS have been digitized and added to the dataset; the remaining 73.6% are lacking frame grabs. Currently, however, gathering these frames individually is a manual, time-consuming process; for each annotation, the tape must be located, fast-forwarded to the annotated timestamp, then frame grabbed and uploaded. A solution to digitize and add more data to *FathomNet* in an automated fashion should be explored further. Finally, additional imagery from other oceanographic institutions (e.g., NOAA, Ocean Exploration Trust, WHOI) could also further improve the size and variety of the dataset.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my mentors, Kakani Katija and Joost Daniels, for guiding me through this process and providing me with invaluable assistance. I would not have been able to make the progress that I have made on this project without their help and support.

I would especially like to thank George Matsumoto and Linda Kuhnz, the MBARI Summer Internship coordinators, for organizing the 2018 program and presenting the interns with countless opportunities throughout.

I would also like to thank Lonny Lundsten and Brian Schlining for mentoring me in the operation of the VARS database. Thanks to Brian Schlining for his development of the *bobd-watermark* application.

Finally, I would like to thank NSF-EAGER #1812535 (GEO-OCE-OTIC to K. Katija), the David and Lucile Packard Foundation, and the Monterey Bay Aquarium Research Institute for affording me the opportunity and resources to work on this project.

References

Gomes-Pereira J.N., V. Auger, K. Beisiegel, R. Benjamin, M. Bergmann, D. Bowden, P. Buhl-Mortensen, F.C. De Leo, G. Dionísio, J.M. Durden, L. Edwards, A. Friedman, J. Greinert, N. Jacobsen-Stout, S. Lerner, M. Leslie, T.W. Nattkemper, J.A. Sameoto, T. Schoening, R. Schouten, J. Seager, H. Singh, O. Soubigou, I. Tojeira, I. van den Beld, F. Dias, F. Tempera, R.S. Santos (2016). Current and future trends in marine image annotation software. *Progress in Oceanography*, 149: 106-120. Available at: https://doi.org/10.1016/j.pocean.2016.07.005

ImageNet. (2010). *Summary and Statistics*. [online] Available at: http://image-net.org/about-stats

Katija, K., G. Montague, G. Young, G. Flaspohler, B. Woodward, J. Gyllinsky, and K.C. Bell (2018). Big Ocean, Big Data. *Journal of Open Exploration*. Retrieved from https://explore.pubpub.org/pub/big-ocean-big-data.

Pan S.J. and Q. Yang (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22: 1345-1359. Available at: https://doi.org/10.1109/TKDE.2009.191

Simonyan K. and A. Zisserman (2014). Very Deep Convolution Networks for Large-Scale Image Recognition. Available at: https://arxiv.org/abs/1409.1556

Appendix A. VARS Concept Lists

Concept	Images	Type
Peniagone	11343	Benthic
Coryphaenoides	3212	Benthic
Keratoisis	2291	Benthic
Paragorgia	1802	Benthic
Sebastes	1758	Benthic
Bathyraja	1282	Benthic
Careproctus	520	Benthic
Anoplopoma	466	Benthic
Chionoecetes	1013	Benthic
Funiculina	465	Benthic
Merluccius	803	Benthic
Microstomus	477	Benthic
Pannychia	787	Benthic
Rathbunaster	775	Benthic
Sebastolobus	1601	Benthic
Strongylocentrotus	3358	Benthic
Umbellula	462	Benthic

Concept	Images	Type
pillow lava	8872	Geology
cobble	408	Geology
gravel	681	Geology

Concept	Images	Type
Bathochordaeus	6564	Midwater
Apolemia	4572	Midwater
Solmissus	2464	Midwater
Nanomia	2005	Midwater
Aegina	1954	Midwater
Lampocteis	2012	Midwater
Bathocyroe	1973	Midwater
Atolla	1808	Midwater
Tomopteris	1801	Midwater
Poralia	1676	Midwater
Eusergestes	981	Midwater
Erenna	1501	Midwater
Poeobius	1421	Midwater
Beroe	1257	Midwater
Ctenophore	1075	Midwater
Mitrocoma	596	Midwater
Benthocodon	545	Midwater
Paraphyllina	519	Midwater