MBARI Monterey Bay Aquarium Research Institute

# Making use of unlabeled data: Comparing strategies for marine animal detection in long-tailed datasets using self-supervised and semi-supervised pre-training

**Tarun Sharma, California Institute of Technology**

***Mentors: Duane Edgington and Danelle E. Cline***
***Summer 2023***

**Keywords: Self-supervised learning, semi-supervised learning, long-tail distribution, unbalanced datasets, rare class detection, contrastive learning, SimCLR, PAWS, underwater animal detection, object detection.**

## ABSTRACT

This paper discusses strategies for object detection in marine images from a practitioner's perspective working with real-world long-tail distributed datasets with a large amount of additional unlabeled data on hand. The paper discusses the benefits of separating the localization and classification stages, making the case for robustness in localization through the amalgamation of additional datasets inspired by a widely used approach by practitioners in the camera-trap literature. For the classification stage, the paper compares strategies to use additional unlabeled data, comparing supervised, supervised iteratively, self-supervised, and semi-supervised pre-training approaches. Our findings reveal that semi-supervised pre-training, followed by supervised fine-tuning, yields a significantly improved balanced performance across the long-tail distribution, albeit occasionally with a trade-off in overall accuracy. These insights are validated through experiments on two real-world long-tailed underwater datasets collected by the Monterey Bay Aquarium Research Institute.

## INTRODUCTION

With the rise of the blue economy, studying ocean community composition and their ecosystems is essential to understanding the ecological impact activities like offshore energy and deep sea mining will have on them. Oceanographic institutes have been surveying parts of the deep oceans using underwater vehicles fitted with video monitoring capabilities for many years now, resulting in a data deluge. Automating the analysis of this video data for biodiversity monitoring using supervised computer vision techniques like object detection has been successful in both ocean and land realms (Tuia et al. 2022; Ditria et al. 2020). This approach however requires expensive manual data annotations by taxonomists, localizing and categorizing every animal in a frame. While these annotations are crucial for our ability to automate video analysis to any extent, this results in the bulk of collected data, the unlabeled data, being completely unused for model training. Moreover, supervised computer vision models exhibit suboptimal performance on imbalanced datasets, particularly struggling with the accurate classification of rare entities. Given that datasets procured from natural environments invariably exhibit a long-tailed distribution, the shortcomings of these models become more pronounced. The erroneous identification of rare species during investigations assessing the ecological impact on oceanic communities holds the potential for significant repercussions. Consequently, an optimal objective entails achieving a balanced performance across all classes within the model's purview.

Self-supervised learning, a machine learning paradigm where a model is trained using implicit labels arising from inherent structures or relationships within the input data alone as a supervisory signal as opposed to relying on human annotations,  has emerged as a viable strategy for leveraging unlabeled data, demonstrating superior performance in certain instances compared to their supervised pre-training counterparts across various downstream tasks (He et al. 2020; Goyal et al. 2019). The scalability of self-supervised learning methods with respect to both data and model size has been substantiated (Goyal et al., 2019), and their proficiency as few-shot learners has been established, particularly when trained with extensive corpora of unlabeled data (Goyal et al. 2021). Notably, these methods are more robust toward unbalanced datasets (Liu et al. 2022), hypothesized to result from a more uniform representation space (Kang et al. 2020). The efficacy of self-supervised learning is most pronounced when unlabeled data originates from the same domain, the availability of supervised data is limited, and the task granularity is relatively coarse (Cole et al. 2022). In the context of automated analysis of deep-sea video footage, this methodology aligns with at least two of these criteria.

Self-supervised learning can be broadly categorized into two main types, task-based methods and contrastive methods (Balestriero et al. 2023). Task-based methodologies involve models predicting various aspects, such as masked-out regions of an image, the color composition of an image (Zhang, Isola, and Efros 2016), or the sequential order of image components (Noroozi and Favaro 2016). Contrastive learning methods minimize the distance in the representation space between two semantically similar images, forming a positive pair, while maximizing the distance between two semantically dissimilar images, forming a negative pair. In the absence of explicit labels, positive pairs often consist of two augmented versions of the same image. One prominent instance of contrastive learning is SimCLR (Chen et al. 2020), where the authors show that the type of data augmentations used is a crucial factor affecting performance. Semi-supervised learning encompasses approaches to train models using a combination of labeled and unlabeled data. A typical semi-supervised strategy is to perform contrastive learning on unlabeled data, also called pre-training, followed by end-to-end supervised fine-tuning using labeled data or exclusively fine-tuning the last few layers. Recent advances in semi-supervised learning have demonstrated enhanced performance and computational efficiency by using a small subset of labeled data during contrastive pre-training, by the addition of an extra term in the loss function such as the SuNCEt loss (Assran et al. 2020). This can also be achieved by minimizing the cross-entropy loss between pseudo labels, assigned based on a small set of supervised support samples, for the positive pair, as demonstrated by the PAWS method (Assran et al. 2021).

Strategies for leveraging unlabeled data to enhance biodiversity monitoring include task-based approaches, like ranking pairs (original image and a crop from the image) of unlabeled noisy sonar images based on the number of fish in them while simultaneously predicting the density maps of fish in a supervised manner using a subset of labeled images (Tarling et al. 2022), pseudo labeling of unlabeled data using a supervised model (Noman et al. 2021), and contrastive learning approaches, like selecting positive pairs of images based on temporal or contextual relatedness from camera trap data as opposed to the standard approach of using two augmented versions of the same image (Pantazis et al. 2021). Most contrastive or task-based pre-training approaches are focused on improving classification performance. Self-supervised object detection pre-training methods, wherein both the region proposal and classification heads are pre-trained, result in only limited enhancements compared to traditional object detection (Huang et al. 2022), often demand substantial computational resources, and do not effectively address the open-world problem. Conversely, a more straightforward localization approach, exemplified by the MegaDetector (Beery, Morris, and Yang 2019), a standard object detector model trained for animal localization in camera trap images, proves robust to unseen data and finds widespread use among non-profits and ecologists globally (“Who Is Using MegaDetector?,” n.d.). The MegaDetector's robustness and practical utility stem from its training on multiple datasets, reducing all classes into three overarching categories: 'animal,' 'vehicle,' and 'human.' This model, characterized by its simplicity, low computational cost, ease of fine-tuning, and adept handling of the open-world problem, successfully localizes previously unseen animals in novel backgrounds. Inspired by the effectiveness of this approach, we further fine-tuned a similar single-class detection model for fish, the MegaFishDetector (Yang et al. 2023), which has been trained using a combination of six underwater datasets.

The Monterey Bay Aquarium Research Institute (MBARI) has been collecting video data from the deep oceans for over 30 years using a suite of autonomous and robotically controlled underwater vehicles. Since 1988, the institute has archived more than 23,000 hours of video footage derived from numerous research expeditions in the Monterey Bay Canyon and other areas including the Pacific Northwest, Northern California, Hawaii, and the Gulf of California, all meticulously organized within MBARI’s in-house Video Annotation and Reference System (“MBARI VARS,” n.d.). As there are often multiple animals in a frame, localization plus classification is essential to determine counts and community composition across varying environmental conditions. A subset of the collected data is fully annotated for the task of object detection with bounding box coordinates and assigned classes (at various taxonomic levels).

In this paper, we compare strategies for making use of unlabeled data, in addition to a subset of labeled data, for biodiversity monitoring in two real-world underwater datasets collected by MBARI exhibiting long-tail distributions. We discuss the benefits of separating the localization and classification stages, training a single-class object detector for the localization step taking inspiration from a widely used and successful study on camera trap images. For the subsequent classification of the localized crops, we compare supervised, supervised iteratively, and semi-supervised approaches, showing that self-supervised and semi-supervised pre-training using unlabeled images followed by supervised fine-tuning results in a more balanced performance across all classes. For both datasets, contrastive methods that use a combination of unlabeled and labeled data for pretraining, such as PAWS and SimCLR with SuNCEt loss, resulted in significantly higher balanced accuracy scores in comparison to contrastive pretraining using unlabeled data alone as is the case with standard SimCLR. On both datasets we achieved the highest balanced accuracy scores using semi-supervised pre-training using PAWS followed by supervised fine-tuning, however, the significant increase in balanced accuracy score comes at a cost of decreased overall accuracy. By proposing a pipeline consisting of a localization approach that is being widely used in practice, along with comparing classification approaches to make use of unlabeled data ranging from straightforward iterative supervision to newer methods such as semi-supervised pre-training, we hope that our results on two noisy real-world long-tailed datasets can serve as a practical guide for practitioners working on similar problems.

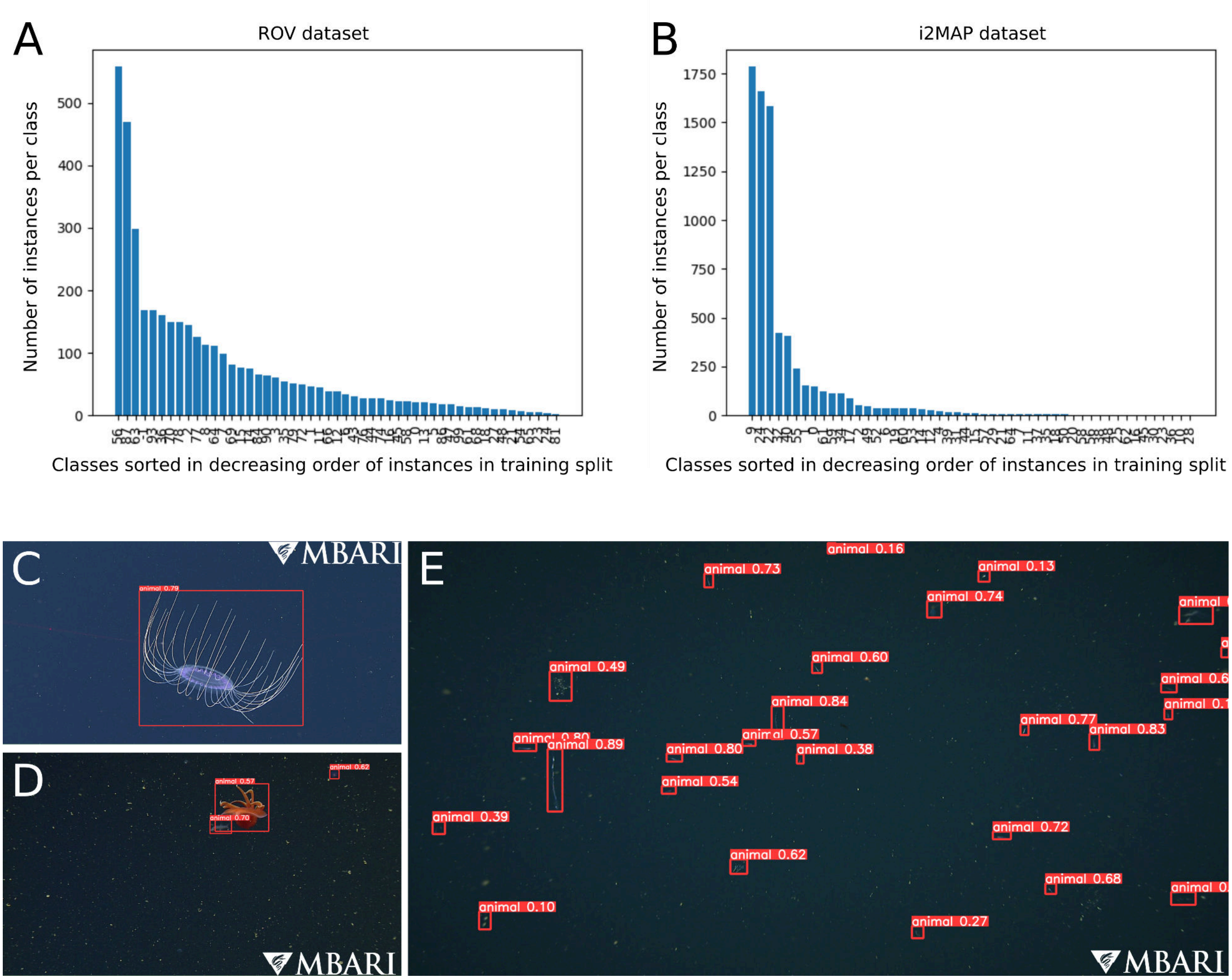

Figure 1. Training data distribution and example images from two different underwater datasets collected by MBARI. (1A and 1B) Number of instances per class (class indices shown in place of taxonomic assignments for data embargo reasons) sorted from highest to lowest number of instances in the training splits of the ROV and i2MAP datasets respectively. These are the extracted crops from both datasets for classification. Data from both datasets follows a long-tailed distribution. (1C and 1D) Example images from the ROV dataset. Overlaid red boxes are localization predictions with confidence scores from our trained single-class animal detection model. (1E) Example image from the i2MAP dataset. Overlaid red boxes are localization predictions with confidence scores from our trained single-class animal detection model. As the  i2MAP dataset was collected by an autonomous vehicle, images generally contain multiple animals at greater distances away.

# MATERIALS AND METHODS

## 2.1 DATASETS USED

Two separate datasets were used for these experiments henceforth referred to as ROV (remotely operated vehicle) and i2MAP datasets. The ROV dataset consists of ~27,000 images collected from multiple ROV transects by MBARI. The dataset contains a mix of images taken in the benthic and midwater zones and was partially annotated with bounding box coordinates and label assignments (varying degrees of taxonomic assignment level). The ROV dataset contained 100 different animal labels with a subset of animals in every image annotated. This resulted in a total of ~41,000 localizations. Annotations were done by the video lab at MBARI. The images were of varying resolution and consisted of animals of different sizes, ranging from small views of animals in the distance, to zoomed-in close-up shots of animals.

The i2MAP is an autonomous underwater vehicle used by MBARI for running midwater transects. It consists of a high-resolution (2k) camera and moves at a speed of about 10 m/s underwater. The i2MAP dataset consisted of ~11,000 fully annotated images and ~75000 localizations. The i2MAP dataset contained 70 different label assignments at different taxonomic levels. Annotations were done by Danelle E. Cline using a combination of manual labeling, heuristic methods such as blob detection and unsupervised methods such as clustering and manual verification of low-confidence assignments. The higher speed of the i2MAP vehicle and its low-power electric motor versus a slew of loud hydraulic systems on the ROV, resulted in many more animal captures per frame, including those that are fast enough to escape the ROV. The large number of animals per frame along with the typically small size of localizations in this dataset made the annotation task very laborious.

For both datasets only the first 50 animal labels, sorted from highest to lowest number of instances in the training split (discussed below), were retained and the remaining labels were grouped together as the ‘unknown’ class, resulting in 51 classes. This was done to study the effect of including examples from novel classes in our unlabeled split.

## 2.2 DATASET SPLITS AND UNLABELED DATA

In order to simulate the availability of additional unlabeled data, annotations (bounding box coordinates plus assigned labels) from 75% of the data from each dataset were removed and these images were treated as the unlabeled split. The remaining 25% of each dataset was split into train-val-test splits as 10-5-10%. All training was done on the train splits of each dataset and metrics were reported on their respective test split. For a fair comparison, the supervised approach was also trained and evaluated on the same train-val-test splits.

## 2.3 EVALUATION METRICS

For evaluating object detection, the standard metric of mAP50 was used. For evaluating classification performance, overall accuracy (OA) and balanced accuracy score (BA) (from sklearn) were used. As both datasets were unbalanced and followed a long tail distribution, overall accuracy did not give us a good estimate of performance across all classes. Balanced accuracy score,  the macro average of recall per class ranging from 0 to 1, is a better metric giving equal importance to performance on all classes irrespective of the number of instances per class. This is particularly advantageous in the context of unbalanced datasets where the overall accuracy may be skewed by the dominance of the majority class.

## 2.4 OBJECT DETECTION

We trained a YOLOv5 (Jocher et al. 2022) object detector model to predict bounding boxes around objects of interest in a class-agnostic manner by collapsing all classes into a general ‘animal’ class. We initialized this model with YOLOv5 parameters from a previously generalized fish detector called MegaFishDetector (Yang et al. 2023), fine-tuned on the training splits, and evaluated on respective test splits of both datasets. Once training and evaluation were complete, we were able to use the final model to extract crops of animals for the downstream classification task. Crops were extracted from images in the unlabeled splits only of both datasets as we already had annotated bounding box coordinates for train, val, and test splits. We evaluated the generalized object detector on its ability to correctly localize animals as measured by the mAP50 metric. A YOLOv5 medium model was used with the long edge of the image being 1280 pixels. A confidence threshold of 0.2 and an IoU threshold of 0.1 was used for predictions.

## 2.5 CLASSIFICATION

The localizations from our class agnostic animal detector on the unlabeled splits of both datasets were cropped out, resulting in ~31000 and ~56000 extracted crops for the unlabeled splits in the case of the ROV and i2MAP datasets respectively. Annotated bounding box coordinates were cropped out from the train, val, and test splits of both datasets resulting in ~4000, ~2000, ~4000 and ~7500, ~3750, ~7500 extracted crops for the train, val, and test sets of ROV and i2MAP datasets respectively. Crops were resized to a size of 224 x 224 and fed into different classification models. For a fair comparison of the advantages of incorporating unlabeled data, baring small modifications (see section 4.3), the same model architecture, Resnet50 (He et al. 2016) was used to compare approaches and was trained for 200 epochs and a batch size of 128 with weighted cross-entropy loss for all supervised components. Models were initialized either from Imagenet weights or weights after contrastive pre-training on unlabeled splits.

### 2.5.1 SUPERVISED (LABELLED DATA ONLY)

The baseline for comparison is training a Resnet50 following the standard supervised learning approach on the training split for each dataset. Models were initialized using Imagenet weights and were trained for 200 epochs using weighted cross-entropy loss, with weighting based on the number of instances of each class in the training split. The loss was monitored on the training and validation splits. Final evaluation metrics were calculated on the test split for each dataset.

### 2.5.2 SUPERVISED ITERATIVELY

An approach that is not commonly compared to in the semi-supervised learning papers, is the simple approach of using a trained model (supervised model from 2.5.1), to generate predictions on the unlabeled data, thresholding these predictions based on confidence treating them as pseudo labels, and subsequently training a new model with a combination of the known labels plus pseudo labels on training split + unlabeled split that has been assigned a pseudo label (depends on threshold chosen). This iterative supervised approach, although simple, is a fair comparison for approaches making use of additional unlabeled data. Whereas ideally a confidence threshold for assigning pseudo labels would be picked based on a precision recall curve, in this limited study we compared thresholds at two ends, 0.1 and 0.7.

### 2.5.3 SELF-SUPERVISED PRE-TRAINING: SIMCLR

The first approach explored for incorporating additional unlabeled data was a contrastive learning approach, SimCLR. SimCLR has been shown to benefit from large models and large batch sizes. We compared batch sizes 256 and 1024, using a Resnet50 backbone with a prediction head (original model using in the SimCLR paper). Contrastive pre-training using unlabeled data was followed by supervised fine-tuning using the labeled data. The pre-training using contrastive NT-XENT loss was done for 100 epochs on unlabeled crops only using 4 GPUs (a ml.p3.8xlarge AWS instance) in the case of batch size 256 and 8 GPUs (a ml.p3.16xlarge AWS instance) in the case of batch size 1028. We only ran this model for 100 epochs as compared to 200 epochs due to the high financial cost associated with high GPU memory demands originating from the requirement of this method to have a large batch size. Augmentations used for pre-training were the same as in the original paper, random crop and color distortion. Pre-trained weights were then used as an initialization for supervised fine-tuning on the training split, for the same number of epochs (200) as in the supervised case.

### 2.5.4 SEMI-SUPERVISED PRE-TRAINING: PAWS AND SIMCLR WITH SUNCET LOSS

The inclusion of a subset of labeled data during the contrastive pre-training step has been shown to result in faster convergence without the need to have a very large batch size, hence allowing cheaper GPU instances to be used. We explored two approaches that fall in this category, PAWS and SimCLR with SuNCEt loss. For both approaches, we used a relatively small unsupervised batch size of 64 and pre-trained for 200 epochs using 3 GPUs (a ml.g4dn.12xlarge AWS instance; there was a weird bug when trying to use 4 GPUs).

Pre-trained weights were then used as an initialization for supervised fine-tuning on the training split, for the same number of epochs (200) as in the supervised case.

PAWS is based on assigning soft pseudo-labels to unlabeled images based on their distances in feature space from a support set of labeled examples per class. The approach minimizes the cross entropy loss between pseudo-labels assigned to two transformed versions of the same image. The support set of labeled examples is only used for pseudo-label assignment in the pre-training step. We tested the PAWS approach on both the ROV and i2MAP datasets. We tested SuNCEt loss, a semi-supervised loss that combines the SimCLR contrastive loss with an additional term aiming to distinguish labeled examples of different classes, only on the i2MAP dataset.

### 2.5.5 CODE AND PARAMETER FILES USED

The code repository for supervised fine-tuning, including supervised fine-tuning after contrastive pre-training can be found here. The parameter yaml file used for the ROV dataset is here and the yaml for the i2MAP dataset can be found here.

The code repository for multi-GPU implementation of standard SimCLR can be found here and the config file used for the ROV dataset can be found here.

The code repository for PAWS and SimCLR with SuNCEt loss can be found here.

For PAWS, the config file used for either dataset can be found here and the config file used for SimCLR with SuNCEt loss can be found here.

## RESULTS

### 3.0 DATASET DISTRIBUTION

Fig 1A and 1B plot the number of images per class in sorted order from highest to lowest for the ROV and i2MAP train splits respectively. The actual class names (taxonomic assignments at various levels) are omitted for data embargo reasons. The top 50 classes in either dataset were retained and the rest were clubbed into a collective “unknown” class with an assigned index of -1 resulting in a total of 51 classes per dataset. Both datasets, like most datasets collected in the wild, exhibit a long tail distribution with many instances of common classes and some rare classes consisting of 2 or 3 images only. Although not shown, the validation and test sets also exhibit long-tail distributions.

### 3.1 DETECTION RESULTS

Table 1 shows the mAP50 scores of single-class (animal) detection starting from MegaFishDetector weights found online. Training on even a subset of data within the distribution of either dataset greatly increases performance. This is not surprising as deep networks struggle with out-of-distribution data. We use our final model to extract crops from unlabeled images in either dataset. Fig 1C and 1D show examples of predicted bounding boxes on images in the ROV dataset test split and Fig 1E shows the same on an image from the i2MAP dataset test split.

| Model | Tested on | Precision | Recall | mAP50 |
|---|---|---|---|---|
| YOLOv5m Megafishdetector weights | ROV dataset test split | 0.541 | 0.445 | 0.39 |
| YOLOv5m model finetuned on ROV train split initialized from Megafishdetector | ROV dataset test split | 0.74 | 0.74 | 0.783 |
| YOLOv5m model finetuned on ROV train split initialized from Megafishdetector | i2MAP dataset test split | 0.719 | 0.512 | 0.647 |
| YOLOv5m model finetuned on i2MAP train split initialized from model above | i2MAP dataset test split | 0.689 | 0.687 | 0.739 |

Table 1. Comparison of localization performance of single-class animal detection models on ROV and i2MAP test sets showing that initialization from Megafishdetector weights followed by fine-tuning on the respective training sets yields the best results.

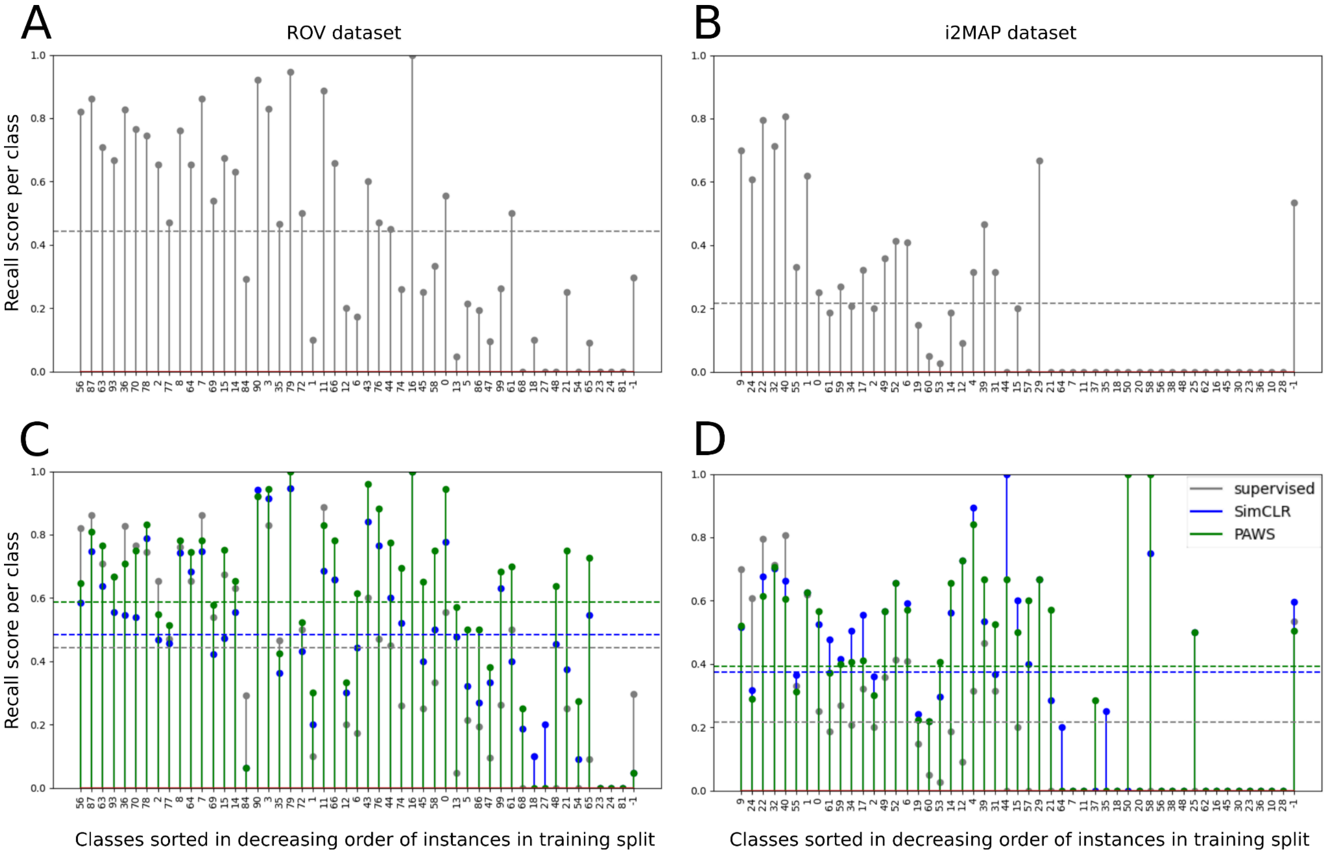

Figure 2. Comparison of supervised and semi-supervised approaches for classification on ROV and i2MAP datasets. (2A and 2B) Per-class recall scores on the test sets of ROV and i2MAP datasets respectively using standard supervised learning (fine-tuning on Imagenet weights). No unlabeled data was used. Classes are sorted in the same order (highest to lowest instances in the training set) as Fig. 1A and Fig. 1B except the unknown (-1) class which is at the end. The dotted line reflects the balanced accuracy score for each dataset. Supervised performance follows a long-tailed distribution. (2C) Per-class recall scores on the test set of the ROV dataset comparing supervised and two semi-supervised approaches, standard SimCLR pre-training on unlabeled split followed by supervised fine-tuning on the training set, and PAWS pseudo label assignment of unlabeled data using labeled examples followed by student-teacher training. (2D) Per-class recall scores on the test set of the i2MAP dataset comparing supervised and two semi-supervised approaches, SimCLR pre-training with SuNCEt loss on a combination of labeled and unlabeled data followed by supervised fine-tuning on the training set, and PAWS pseudo label assignment of unlabeled data using labeled examples followed by student-teacher training.

## 3.2 CLASSIFICATION RESULTS

Once crops of animals are extracted either from our generalized animal detector in the case of unlabeled splits, or annotated coordinates in cases of the train, val, and test splits, they are resized to 224 x 224 and fed into a classification model to assign to one of the 51 classes (50 animal taxonomic assignments plus 1 catch-all unknown class). We compared supervised, supervised iteratively, and self and semi-supervised classification approaches incorporating additional unlabeled data (unlabeled split) on two real-world datasets exhibiting a long tail distribution (Fig 1A, B).

### 3.2.1 SUPERVISED ONLY

To ascertain the upper limit of classification performance for the ROV dataset, we conducted an experiment retaining labels for the unlabeled split. We trained a supervised model on a combination of the training split (10% of the dataset) plus the unlabeled split (75% of the dataset). Tables 2 and 3 present a comprehensive comparison of various classification approaches, encompassing both supervised and semi-supervised learning methods applied to the ROV and i2MAP datasets respectively. The evaluation metrics used were the overall accuracy and the balanced accuracy score. Fig 2A, 2B show the per-class recall scores on the ROV and i2MAP test splits respectively obtained from a supervised Resnet50 model fine-tuned on the training split only, initialized from Imagenet weights. The class indices are sorted in the same order as Fig 1A, 1B, i.e from highest to lowest number of instances in the training split. It is not surprising to see that the per-class performance also follows a long tail, as we know that deep networks perform poorly given a lower number of training examples. The dotted line shows the balanced accuracy score in either case.

### 3.2.2 SUPERVISED ITERATIVELY

To ensure a fair comparison with semi-supervised methodologies utilizing additional unlabeled data, we leveraged the trained supervised model from section 3.2.1 to generate predictions on images from the unlabeled split. This assessment was exclusively conducted for the ROV dataset. Predictions were subjected to a thresholding process based on confidence, with predictions surpassing the threshold considered as pseudo labels. While the optimal threshold selection typically involves a meticulous precision-recall curve analysis on the val set, we pragmatically assessed only two thresholds—0.7 and 0.1—for the sake of expediency. As shown in Table 2, this iterative supervised approach exhibits a modest enhancement in performance when incorporating additional data from the unlabeled split along with their corresponding pseudo labels. However, it is imperative to acknowledge the inherent risk of perpetuating biases learned during the initial supervised stage. Furthermore, any bias associated with the long-tailed nature of the dataset will be further emphasized, as only predictions with confidence exceeding the chosen threshold, usually the head classes, will contribute to additional pseudo labels.

### 3.2.3 SELF-SUPERVISED PRE-TRAINING: STANDARD SIMCLR

For self-supervised contrastive pre-training approaches, we tested SimCLR using the original NT-Xent loss function. This assessment was exclusively conducted for the ROV dataset. Contrastive pretraining was performed using the unlabeled split of the ROV dataset followed by supervised fine-tuning on the ROV training split and evaluation using the ROV test split. As we can see from Table 2 and Fig 1C, this approach yielded only a modest improvement on the balanced accuracy score in comparison to the supervised only approach, improving balanced accuracy score from 0.443 to 0.485 while leading to a decrease in overall accuracy from 64.97 to 56.67 when using a batch size of 256. Increasing the batch size from 256 to 1024 did not result in significant gains. Pre-training was done for 100 epochs as opposed to 200 because of the limited improvements going from batch size 256 to 1024, along with the high financial cost associated with GPU memory requirements for this method that requires large batch sizes.

### 3.2.4 SEMI-SUPERVISED PRE-TRAINING: SIMCLR WITH SUNCET LOSS AND PAWS

To test semi-supervised contrastive learning approaches, we compared two approaches, SimCLR using SuNCEt loss, and PAWS on both the ROV and i2MAP datasets. Both these approaches use a combination of unlabeled data (from unlabeled split) and a subset of labeled data (from training split) for the pre-training step. Pre-training was followed by supervised fine-tuning as in 3.2.3. As we can see from Tables 2 and 3, performing semi-supervised contrastive pre-training on the unlabeled split, followed by supervised fine-tuning on the training split results in a significantly higher balanced accuracy score sometimes at a cost of lower overall accuracy for either dataset. This is also evident from the per-class performance of these models in Fig 2C and 2D. We see a much more balanced performance, higher performance on rare classes plus slightly lower or the same performance on majority classes. In the case of the ROV dataset, PAWS resulted in a significantly higher balanced accuracy score of 0.587 in comparison to standard SimCLR with NT-XENT loss and supervised only methods yielding balanced accuracy scores of 0.485 and 0.443 respectively. In the case of the i2MAP dataset, we observe a similar significant gain in balanced accuracy score, nearly doubling the balanced accuracy score of supervised only approaches from 0.217 to 0.375 and 0.393 for SimCLR with SuNCEt loss and PAWS respectively, emphasizing the efficacy of these methods in handling dataset imbalances. SimCLR with SuNCEt loss resulted in much closer performance gains to those obtained using PAWS. In summary, for both the ROV and i2MAP long-tailed datasets, PAWS pre-training followed by supervised fine-tuning resulted in the highest balanced accuracy scores sometimes at the cost of a decrease in overall accuracy (i2MAP dataset only and not ROV) compared to supervised fine-tuning only.

| Classification approach ROV dataset | Overall accuracy (%) | Balanced accuracy score (0-1) |
|---|---|---|
| Supervised only - fine-tuning using training split + retaining labels of unlabeled split (max upper limit possible). | 79.8 | 0.684 |
| Supervised only - fine-tuning using training split starting from Imagenet. | 64.97 | 0.443 |
| Supervised iteratively - fine-tuning using training split + pseudo labels on unlabeled split thresholded at 0.1 confidence. | 64.40 | 0.483 |
| Supervised iteratively - fine-tuning using training split + pseudo labels on unlabeled split thresholded at 0.7 confidence. | **68.78** | 0.472 |
| SimCLR - Contrastive pre-training with NTXent loss on unlabeled split with batch size 256 for 100 epochs, followed by supervised fine-tuning using training split. | 56.67 | 0.485 |
| SimCLR - Contrastive pre-training with NTXent loss on unlabeled split with batch size 1028 for 100 epochs followed by supervised fine-tuning using training split. | 53.03 | 0.473 |
| PAWS - Contrastive pre-training on unlabeled split with unsupervised batch size 64 for 200 epochs followed by supervised fine-tuning using training split. | 66.04 | **0.587** |

Table 2. Comparison of supervised and semi-supervised classification performance on the ROV test set. As the data, including the test set, follows a long-tailed distribution, a balanced accuracy score, the macro average of recall scores per class, is a better metric than overall accuracy. As additional unlabeled data was used for the semi-supervised approaches, for fair comparison, we use pseudo labels on unlabeled data obtained by thresholding predictions from a supervised model trained on the training split and retrain a model in a supervised manner using both labeled and pseudo-labeled data. PAWS on unlabeled data followed by supervised

fine-tuning gives us the best performance. Although the overall accuracy of PAWS is similar to the supervised-only approach, we see a significantly higher balanced accuracy score.

| Classification approach i2MAP dataset | Overall accuracy (%) | Balanced accuracy score (0-1) |
|---|---|---|
| Supervised only - Fine-tuning using training split starting from Imagenet | **62.74** | 0.217 |
| SimCLR - Contrastive pre-training using SuNCEt loss on unlabeled split followed by supervised fine-tuning using training split, unsupervised batch size: 64 | 52.56 | 0.375 |
| PAWS - Contrastive pre-training on unlabeled split followed by supervised fine-tuning using training split, unsupervised batch size: 64 | 49.58 | **0.393** |

Table 3. Comparison of supervised and semi-supervised classification performance on the i2MAP test set. As the data, including the test set, follows a long-tailed distribution, a balanced accuracy score, the macro average of recall scores per class, is a better metric than overall accuracy. PAWS on unlabeled data followed by supervised fine-tuning gives us the best performance. Although the overall accuracy of PAWS is lower than the supervised-only approach, we see a significantly higher balanced accuracy score, almost double of the supervised-only model. In comparison to the results on the ROV dataset, we see that SimCLR with a SuNCEt loss performs much better than the standard SimCLR with a NT-XENT loss.

## DISCUSSION

We demonstrate that in the case of classification in large underwater datasets consisting of a subset of labeled data and a large amount of unlabeled data, semi-supervised pre-training methods such as SimCLR with SuNCEt loss and PAWS, followed by supervised fine-tuning using the labeled data, results in a significantly more balanced performance across classes (Fig. 2C, 2D and Table 2 and 3) when compared to supervised only baselines. This is especially apparent in cases of real-world datasets exhibiting a long-tailed distribution (Fig. 1A, 1B) as is most often the case with datasets collected in the wild. We demonstrate that splitting the localization and classification steps allows for training a robust generalized single-class detector (Fig. 1C-E and Table 1) which helps address the open-world problem for localization. This subsequently allows training a suite of different classifiers depending on the task at hand, either supervised only for the best results on common classes, semi-supervised for the most balanced performance, classifiers focusing on few-shot learning for rare classes or classifiers addressing the open-world problem for classification.

## 4.1 ADVANTAGES OF SEPARATING DETECTION AND CLASSIFICATION STAGES

The deliberate separation of the localization step from the classification step presents several advantages compared to the conventional integration of these stages in standard object detectors, whether single-stage or two-stage. Training a single-class detector enables the amalgamation of data from diverse datasets by consolidating labels into a singular 'animal' class. This approach substantially enhances the model's generalizability and robustness. The widespread adoption of Megadetector, a generalized animal detector for camera-trap data on land, underscores the efficacy of this methodology. Beyond facilitating the integration of multiple datasets, this approach proves advantageous in the context of open-world detection. In scenarios involving previously unidentified species, a plausible occurrence in deep ocean exploration, our generalized detection model exhibits a higher likelihood of localizing the animal, having encountered diverse animal types from different backgrounds. Subsequently, the classification model can address the open-world scenario for classification, employing anomaly detection methods. In contrast, standard multi-class object detectors may entirely miss the animal due to a lack of resemblance to a limited training set of animal classes. Unlike self-supervised object detection approaches, which can be computationally intensive and offer marginal improvements over standard object detectors (Huang et al., 2022), the segregation of localization and classification steps not only capitalizes on the robustness of a single-class detector but also enables the exploration of self-supervised and semi-supervised learning strategies for utilizing unlabeled data in classification. These approaches are typically less computationally demanding and have demonstrated promising results. An additional benefit arising from the use of a single-class detector is the potential improvement in downstream tasks such as tracking, attributed to the absence of label switches from a multi-class object detector.

## 4.2 A MORE ROBUST AND BALANCED PERFORMANCE FROM PRE-TRAINING

The utilization of unlabeled data for pre-training exposes models to the specific imaging domains they are intended to be trained on, enabling the acquisition of general features unique to marine imaging and marine animals. Incorporating random crop augmentation further facilitates the association of disparate segments of animals, even those that may exhibit gelatinous and structureless characteristics. In contrast to learning exclusively with labeled data, as observed in supervised cases, which compels the model to focus on features for maximal class distinction, incorporating unlabeled data is more likely to foster the learning of more general and robust features. Notably, prior studies have demonstrated that off-the-shelf semi-supervised models exhibit enhanced robustness to class imbalance compared to their fully supervised counterparts (Liu et al., 2022). These models also demonstrate improved performance in out-of-distribution scenarios, cross-task settings, and rare class identification, and exhibit a balanced feature space equidistant from all classes and not dominated by the majority class as in supervised learning (Kang et al., 2020). Our findings, based on two real-world underwater imaging datasets characterized by long-tail distributions, align with these observations. Specifically, our approach involves decoupling localization from classification and subsequently employing semi-supervised learning methods for the classification component. This strategy proves effective in leveraging additional unlabeled data to enhance overall performance. Notably, our results indicate a doubling of balanced accuracy in the case of the i2MAP dataset, which, despite its relatively small size and the small, blob-like appearance of individual animals due to their distance from the vehicle, underscores the efficacy of our approach. Our approach yields balanced results compared to a supervised approach that perpetuates biases, particularly beneficial in low-data scenarios. This is attributed to its capability to extract more instances of rare classes from unlabeled data, thereby addressing the challenges associated with limited data availability.

## 4.3 MODEL ARCHITECTURES DIFFER SLIGHTLY

As detailed in the methods section, it is crucial to note that the original models employed in the SimCLR and PAWS studies, as well as the models utilized for our semi-supervised pre-training, deviate from the standard Resnet50 configuration. Specifically, they feature a Resnet50 architecture augmented with an additional prediction head. While the ideal comparison involves assessing identical architectures across both supervised and semi-supervised approaches, it is improbable that the observed improvement in balanced accuracy scores can be solely attributed to the presence of the supplementary prediction head layer in these models. The exploration of a direct comparison using identical architectures is an ongoing aspect of our research.

## 4.4 SEMI-SUPERVISED PRE-TRAINING WORKS BETTER THAN SELF-SUPERVISED PRE-TRAINING FOR IMBALANCED DATASETS

From section 3.2.4, it is clear that semi-supervised pretraining approaches that use a combination of unlabeled and labeled data for pretraining, such as SimCLR with SuNCEt loss and PAWS, result in significantly higher balanced accuracy scores on the long-tailed distributed test sets for both the ROV and i2MAP datasets. These approaches also require significantly lower compute cost and time in comparison to self-supervised pretraining approaches like SimCLR. One can see how providing some supervisory signal by using a subset of labeled data, can result in faster convergence. We have shown that the weights converged onto by using the additional supervisory signal, result in a greater robustness to dataset imbalance, leading to significantly higher balanced accuracy scores after supervised fine-tuning in comparison to self-supervised pre-training approaches, supervised only and supervised iteratively on two real-world long-tailed underwater datasets.

## 4.5 UNANSWERED QUESTIONS AND FUTURE WORK

An intriguing avenue for exploration is whether semi-supervised pre-training yields a more advantageous starting point for fine-tuning when the unlabeled set encompasses images from classes distinct from those under consideration for classification. To investigate this, we amalgamated classes outside the top 50 classes from the training splits into a unified 'unknown' class. The hypothesis posits that pre-training with other animals from the same domain might yield more favorable initial parameters for fine-tuning compared to entirely dissimilar parameters, such as those derived from Imagenet weights. The forthcoming research will involve experiments with and without the 'unknown' class, probing into the efficacy of more relevant starting points for fine-tuning, such as employing weights from supervised training on the ROV dataset for subsequent fine-tuning on the i2MAP dataset.

Another dimension for exploration involves assessing how the performance disparity between supervised-only models and semi-supervised models evolves with the scaling of labeled data. Previous research indicates that selecting positive pairs for contrastive learning based on temporal and contextual similarities, rather than augmenting the same image twice to form the positive pair, leads to superior performance across various contrastive loss functions

(Pantazis et al., 2021). Given the availability of video data, we contemplate incorporating a similar strategy for positive pair selection to investigate its potential to further enhance balanced performance. These areas of inquiry contribute to our ongoing efforts in refining and advancing the understanding of self-supervised and semi-supervised learning methods applied to the context of underwater image classification.

## CONCLUSIONS/RECOMMENDATIONS

In this paper, we have demonstrated a pipeline for object detection in cases of large underwater datasets exhibiting a long tail distribution. We list the benefits of splitting object detection into its components: single-class localization followed by classification of extracted crops. We showed that we were able to achieve a significantly higher balanced performance across classes when using semi-supervised pre-training on unlabeled data followed by supervised fine-tuning in comparison to supervised fine-tuning only. This method allows the usage of unlabeled data along with a subset of labeled to improve balanced classification performance. This is especially useful for improving rare class classification performance in the cases of underwater datasets where good performance on all classes might be desired, unlabeled data is relatively easy to obtain, and data annotation can be tedious. We demonstrate that our approach works using two real-world underwater datasets. We also show that the semi-supervised pre-training approach PAWS, resulted in doubling the balanced accuracy score of the supervised-only model in the case of the i2MAP dataset, a relatively small dataset wherein individual animals were quite small and blob-like. These results are quite promising and with some additional exploration of scaling up labeled datasets, could be useful to incorporate into the classification models used at MBARI.

## ACKNOWLEDGEMENTS

I would like to thank my mentors Duane Edgington and Danelle E. Cline for supporting me and guiding me through this project, for several engaging brainstorming sessions, being patient with and trusting me, always having an open-door policy to discuss science or anything else, introducing me to colleagues at MBARI, and helping me with whatever I required ranging from data to setting up and training me on using cloud computing resources. I would like to explicitly thank Danelle E. Cline for the work she put into annotating the i2MAP dataset without which this project would not have been possible. I would like to thank George I. Matsumoto for making the internship program possible and for all his help in ensuring the interns were able to have their needs met and access resources needed along with organizing multiple once-in-a lifetime experiences for the interns. I would also like to thank all my fellow interns from the 2023 batch for being so lovely and supportive. I would like to thank the returning intern, Lael Wakamatsu, for organizing many social gatherings for the interns and having multiple check-ins with the interns. Lastly, I would like to thank MBARI for giving me this amazing opportunity to expand my skill set, learn and experience working in this fascinating area of science, and make long-lasting connections.

## References

Assran, Mahmoud, Nicolas Ballas, Lluis Castrejon, and Michael Rabbat. 2020. “Supervision Accelerates Pre-Training in Contrastive Semi-Supervised Learning of Visual Representations.” arXiv. http://arxiv.org/abs/2006.10803.

Assran, Mahmoud, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat. 2021. “Semi-Supervised Learning of Visual Features by Non-Parametrically Predicting View Assignments with Support Samples.” In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8443–52. http://openaccess.thecvf.com/content/ICCV2021/html/Assran_Semi-Supervised_Learning_of_Visual_Features_by_Non-Parametrically_Predicting_View_Assignments_ICCV_2021_paper.html.

Balestriero, Randall, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, and Yuandong Tian. 2023. “A Cookbook of Self-Supervised Learning.” *arXiv Preprint arXiv:2304.12210*. https://arxiv.org/abs/2304.12210.

Beery, Sara, Dan Morris, and Siyu Yang. 2019. “Efficient Pipeline for Camera Trap Image Review.” arXiv. http://arxiv.org/abs/1907.06772.

Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. “A Simple Framework for Contrastive Learning of Visual Representations.” In *International Conference on Machine Learning*, 1597–1607. PMLR. http://proceedings.mlr.press/v119/chen20j.html.

Cole, Elijah, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. 2022. “When Does Contrastive Visual Representation Learning Work?” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14755–64. http://openaccess.thecvf.com/content/CVPR2022/html/Cole_When_Does_Contrastive_Visual_Representation_Learning_Work_CVPR_2022_paper.html.

Ditria, Ellen M., Sebastian Lopez-Marcano, Michael Sievers, Eric L. Jinks, Christopher J. Brown, and Rod M. Connolly. 2020. “Automating the Analysis of Fish Abundance Using Object Detection: Optimizing Animal Ecology With Deep Learning.” *Frontiers in Marine Science* 7 (June): 429. https://doi.org/10.3389/fmars.2020.00429.

Goyal, Priya, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, and Armand Joulin. 2021.

"Self-Supervised Pretraining of Visual Features in the Wild." *arXiv Preprint arXiv:2103.01988*. https://arxiv.org/abs/2103.01988.

Goyal, Priya, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. 2019. "Scaling and Benchmarking Self-Supervised Visual Representation Learning." In *Proceedings of the Ieee/Cvf International Conference on Computer Vision*, 6391–6400. http://openaccess.thecvf.com/content_ICCV_2019/html/Goyal_Scaling_and_Benchmarking_Self-Supervised_Visual_Representation_Learning_ICCV_2019_paper.html.

He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. "Momentum Contrast for Unsupervised Visual Representation Learning." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–38. http://openaccess.thecvf.com/content_CVPR_2020/html/He_Momentum_Contrast_for_Unsupervised_Visual_Representation_Learning_CVPR_2020_paper.html.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep Residual Learning for Image Recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–78. http://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Huang, Gabriel, Issam Laradji, David Vazquez, Simon Lacoste-Julien, and Pau Rodriguez. 2022. "A Survey of Self-Supervised and Few-Shot Object Detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (4): 4071–89.

Jocher, Glenn, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, et al. 2022. "Ultralytics/Yolov5: V7.0 - YOLOv5 SOTA Realtime Instance Segmentation." *Zenodo*, November. https://doi.org/10.5281/zenodo.7347926.

Kang, Bingyi, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. 2020. "Exploring Balanced Feature Spaces for Representation Learning." In *International Conference on Learning Representations*. https://openreview.net/forum?id=OqtLIabPTit.

Liu, Hong, Jeff Z. HaoChen, Adrien Gaidon, and Tengyu Ma. 2022. "Self-Supervised Learning Is More Robust to Dataset Imbalance." arXiv. http://arxiv.org/abs/2110.05025.

"MBARI VARS." n.d. https://www.mbari.org/technology/video-annotation-and-reference-system-vars/.

Noman, Md Kislu, Syed Mohammed Shamsul Islam, Jumana Abu-Khalaf, and Paul Lavery. 2021. "Multi-Species Seagrass Detection Using Semi-Supervised Learning." In *2021 36th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 1–6. IEEE. https://ieeexplore.ieee.org/abstract/document/9653222/?casa_token=mpckPuUmZz4AAAAA:xTV_Ionre-wZINaqodAdOcAKUiXoy890LP1qdZVrDGlWL2THIEwuHglK9lfVtlZn_fjcTvGWW9g.

Noroozi, Mehdi, and Paolo Favaro. 2016. "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles." In *Computer Vision – ECCV 2016*, edited by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, 9910:69–84. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-46466-4_5.

Pantazis, Omiros, Gabriel J. Brostow, Kate E. Jones, and Oisin Mac Aodha. 2021. "Focus on the Positives: Self-Supervised Learning for Biodiversity Monitoring." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10583–92. http://openaccess.thecvf.com/content/ICCV2021/html/Pantazis_Focus_on_the_Positives_Self-Supervised_Learning_for_Biodiversity_Monitoring_ICCV_2021_paper.html.

Tarling, Penny, Mauricio Cantor, Albert Clapés, and Sergio Escalera. 2022. "Deep Learning with Self-Supervision and Uncertainty Regularization to Count Fish in Underwater Images." *PloS One* 17 (5): e0267759.

Tuia, Devis, Benjamin Kellenberger, Sara Beery, Blair R. Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, et al. 2022. "Seeing Biodiversity: Perspectives in Machine Learning for Wildlife Conservation." *Nature Communications* 13 (1): 792.

https://doi.org/10.1038/s41467-022-27980-y.

“Who Is Using MegaDetector?” n.d. https://github.com/agentmorris/MegaDetector/#who-is-using-megadetector.

Yang, Daniel, Levi Cai, Stewart Jamieson, and Yogesh Girdhar. 2023. “Biological Hotspot Mapping in Coral Reefs with Robotic Visual Surveys.” *arXiv Preprint arXiv:2305.02330*. https://arxiv.org/abs/2305.02330.

Zhang, Richard, Phillip Isola, and Alexei A. Efros. 2016. “Colorful Image Colorization.” In *Computer Vision – ECCV 2016*, edited by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, 9907:649–66. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-46487-9_40.