Title

Problems and Solutions: Machine Learning Approaches for a Dynamic Ocean

Permalink

https://escholarship.org/uc/item/70j2r271

Author

Walker, Joseph Leslie

Publication Date

2023

UNIVERSITY OF CALIFORNIA SAN DIEGO

Problems and Solutions: Machine Learning Approaches for a Dynamic Ocean

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Oceanography

by

Joseph Leslie Walker

Committee in charge:

Kaitlin Frasier, Chair
Florian Meyer
Stuart Sandin
Nuno Vasconcelos

2023

The Dissertation of Joseph Leslie Walker is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

TABLE OF CONTENTS

DISSERTATION APPROVAL PAGE iii

TABLE OF CONTENTS iv

LIST OF FIGURES v

LIST OF TABLES vii

ACKNOWLEDGEMENTS viii

VITA x

ABSTRACT OF THE DISSERTATION xi

Chapter 1 INTRODUCTION 1

Chapter 2 6

Acknowledgements 16

Chapter 3 17

Acknowledgements 44

Chapter 4 45

Acknowledgements 72

LIST OF FIGURES

Figure 1: Using target and background training datasets, the parameters of a classifier and background image weights are jointly learned using an alternative optimization approach. Testing is then performed using target, novel, and dataset shift datasets………………………………………………………………………………… 6

Figure 2: Three examples, selected by a human annotator, from classes in D^{targ} (left column) and D^{out} (right column) showing the morphological similarity between specimen of these classes... 9

Figure 3: Class distribution of the downsampled background sets generated from each of the downsampling strategies. a) class distribution for the 15 least abundant classes that have at least 100 examples in the original D^{out} set. b) class distribution for the five most abundant classes in D^{out}.. 10

Figure 4: Samples from each target class and their respective hard negatives. Ten background classes are represented by the hard negative examples... 11

Figure 5: Daily average background image weight value associated with each of the hard background classes presented in the order: a) *Dictyocha*; b) *detritus*; c) *Skeletonema*; d) *pennate*. The bars represent the standard deviation of the weight values.. 11

Figure 6: OOD testing results. Left-to-right: Accuracy and F1 score obtained from an average of model runs using ratios of {1: 400, 1: 50, 1: 10, 1: 1}. Error bars reflect standard deviation... 12

Figure 7: Sampling locations, dates, methodology, and a SUIT..................................... 24

Figure 8: Example images collected from the six study sites. The bounding boxes containing the SUITs are shown in orange……………………………………………... 27

Figure 9: Data augmentations. a) An example image from the TAK study site. Images b)-f) show the output of the translation, rotation, perspective transformation, cropping, and distance image augmentation functions respectively using the image in a) as input. 29

Figure 10: a) Overview of our augmented version of the progressive domain adaptation method, using the HM-MVGD-HM color-matching algorithm and data augmentation (PDA+CM+DA). b) Illustration of the adversarial feature alignment process……………………………………………………………………………... 31

Figure 11: Region proposals (red) from four different models for an example image from the target study site of two adaptation experiments: a) Using TAK as source and PSW as target, b) Using PNW as source and PS as target. Ground truth bounding boxes are shown in orange…………………………………………………………… 33

Figure 12: Qualitative comparison of image-to-image translation methods using images from TAK and PS as source and target respectively…………………………… 35

Figure 13: Ablation on the five augmentation functions: translation (trans.), rotation (rot.), perspective (per.), cropping (crop.), and distance (dist.)………………………… 36

Figure 14: Map of the Santa Barbara Channel. Traffic Separation Scheme is shown as black lines and HARP location is shown with red pentagram. The white circle around the HARP denotes the 6 km boundary at which ship transits were considered………………………………………………………………………………. 50

Figure 15: Single-transit model architecture……………………………………………. 55

Figure 16: Our fusion method. Features are passed through a linear layer and a dot product is performed against a learnable token. Orange boxes indicate learnable values. Green box indicates features from encoder. Blue boxes indicate fixed mathematical operation………………………………………………………………. 56

Figure 17: Illustration of the 4-fold cross-validation approach for the baseline and neural network models. For each fold, models are tested on data from a single season in 2017 shown in orange. Note that time regions with mixed colors indicate that data was used to train both models………………………………………………………... 59

Figure 18: Comparison of single-transit and baseline model predictions………………. 63

Figure 19: Performance on testing data of late fusion (concat) model with varying transit sampling number *k*…………………………………………………………… 64

LIST OF TABLES

Table 1: Dataset shift testing results (in % including ± Std Dev.)………………………. 12

Table 2: Dataset statistics for each study site. Average and standard deviation of bounding box sizes are reported and expressed in terms of fraction of total image area, where all images are 500x751 pixels………………………………………………… 27

Table 3: Image feature similarity between study sites. Similarity scores have been rescaled linearly between [0,1] where 1 indicates mean size is identical, and 0 represents the largest observed dissimilarity…………………………………………… 27

Table 4: SUIT size similarity between study sites. similarity scores have been rescaled linearly between [0,1] where 1 indicates mean size is identical, and 0 represents the largest observed dissimilarity…………………………………………………………… 28

Table 5: Leave-one-in cross validation results. Each of the five models is trained using a single source study site (X) where all other study sites are individually treated as a target domain (Y)……………………………………………………………………….. 33

Table 6: Description of predictor variables used in statistical models…………………. 51

Table 7: Model performance of sound speed profile estimation across the four test seasons. Performance is reported in terms of root-mean-square error (m/s). The best performing model for each season is shown in boldface……………………………….. 61

ACKNOWLEDGEMENTS

I am deeply grateful for the immense privilege of having an exceptional support network of friends, family, and colleagues throughout my journey as a PhD student. I would like to acknowledge Dr. Kaitlin Frasier for her support throughout my graduate career. From the day following my departmental exam, Kait demonstrated her belief in my capabilities as a student. Halfway through my graduate career, when I found myself on the brink of quitting the program, I was finally able to join her lab and officially call her my advisor.

Entering an entirely new field of research was incredibly daunting, however, this transition was made so much easier with the mentorship and moral support of Vanessa ZoBell and Michaela Alksne.

During this time, I was introduced to an incredibly motivated undergraduate student Zheng Zeng, who was interested in helping me run experiments for two of my research projects. Working with Zheng not only meant that I was not alone during times of struggle, it also meant that I had someone to celebrate with when (very occasionally) experiments went well.

My research experience began when I was a sophomore undergraduate student at the University of Portland supervised by Dr. Ted Eckmann. Ted pushed me to do things I was not confident I could handle. He also nudged me to enroll in the Introduction to Artificial Intelligence course that the university offered. This course is what sparked my interest in artificial intelligence and its application in the Earth Sciences.

I owe a special thanks to my close group of friends since middle school, Greg Berg, Cannon Maruska, and Jake Cline, who have always been there to provide relief from the drudgery and grind of graduate school in the form of margaritas, surfing, and the occasional Mexico trip.

And, of course, I would like to thank my family for being my unwavering supporters since the very beginning. Finals week and academic deadlines were made so much less stressful knowing that I always had a place to stay and a home-cooked meal to enjoy in the comfort of my family. I cannot imagine having gone through graduate school without this incredible support.

Chapter 2, in full, is a reprint of the material as it appears in International Conference on Computer Vision (ICCV) Workshops. Walker, Joseph; Orenstein, Eric 2021. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part, has been submitted for publication of the material as it may appear in IEEE Oceanic Engineering 2023, Walker, Joseph; Zeng, Zheng; Chengchen, Wu; Jaffe, Jules; Frasier, Kaitlin; Sandin, Stuart. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in part, has been submitted for publication of the material as it may appear in Journal of the Acoustical Society of America 2023, Walker, Joseph; Zeng, Zheng; ZoBell, Vanessa; Frasier, Kaitlin. The dissertation author was the primary investigator and author of this paper.

VITA

2018 B.S., Environmental Science
University of Portland
Portland, OR

2023 Ph.D., Oceanography
Scripps Institution of Oceanography
University of California San Diego
Dissertation title: “Problems and Solutions: Machine Learning Approaches for a Dynamic Ocean”

ABSTRACT OF THE DISSERTATION

Problems and Solutions: Machine Learning Approaches for a Dynamic Ocean

by

Joseph Leslie Walker

Doctor of Philosophy in Oceanography

University of California San Diego, 2023

Kaitlin Frasier, Chair

Advancements in observational methods and data collection techniques have empowered oceanographers to gather extensive data on a wide range of oceanic phenomena. Optical imaging systems have provided unprecedented insight into the microscopic world of marine plankton as well as the structure, health, biodiversity, and ecological dynamics of coral reefs. Advances in low-power autonomous acoustic recording devices have enabled continuous long-term monitoring of marine mammals and ocean noise.

These data-driven methods involve the collection, analysis, and interpretation of large datasets to gain insights. Although machine learning offers the potential for automating the analysis of large oceanographic datasets, its utilization in this context is accompanied by challenges and problems due to the high spatiotemporal variability and noise inherent in these datasets.

This thesis delves into an extensive exploration of state-of-the-art machine learning techniques, specifically tailored to optimize the extraction of valuable information from dynamic oceanographic datasets. To obtain a comprehensive understanding of the problem, instances of dataset shift and noise are examined in three distinct case studies spanning the vision and acoustic domains.

The first case study focuses on the problem of novelty detection and class imbalance in the context of plankton image recognition using Images from the WHOI-Plankton dataset. The second case study explores the problem of object detection when samples are collected from different environments or under varying conditions. Lastly, the third case study aims to develop multi-observational techniques to reduce dataset noise using a dataset of acoustic recordings collected in the Santa Barbara Channel.

In each case, the core technical goal is the same: to train a convolutional neural network-based system to learn a robust feature representation that generalizes to unforeseen environmental conditions. To achieve this goal, techniques from the field of hard negative mining, unsupervised domain adaptation, and multi-view learning are integrated into the workflows. Ultimately, my overarching objective is to drive advancements in the development of robust oceanographic data automation tools.

Chapter 1 INTRODUCTION

The vastness of the world's oceans presents an intricate web of interconnected processes, making it a complex environment to understand. New technology platforms have resulted in an exponential increase in the amount of collected oceanic data, resulting in more data collection on the world's oceans in the year 2018 than the cumulative data gathered throughout the entire twentieth century (Tanhua et al., 2019). In fields such as underwater imaging, ocean acoustics, and physical oceanography, oceanographic research has long revolved around the development of physical models and their application in deducing properties of both the ocean environment and objects within it. However, as we move into the era of "Big Data", this paradigm is beginning to change.

In ocean acoustics, sound event classification and localization methods have traditionally relied upon signal processing techniques and signal/channel models. However, these techniques often perform poorly in common scenarios where noise, reverberation, and multiple simultaneously emitting sound sources are present (Blandin et al., 2012; Evers et al., 2020). In underwater imaging, image processing tools such as Sobel convolution kernels, morphological operations, and thresholding were used together with traditional statistical and rule-based methods to perform classification and detection tasks. However, in situations where the data demonstrates a considerable degree of intra-class variability and/or notable inter-class similarity, these techniques frequently demonstrate subpar performance (Bishop, 2006).

While traditional methods in oceanography have provided valuable insights, the surge in available data has opened exciting new avenues for exploration. For example, developments in underwater imaging tools such as the Imaging FlowCytobot (IFCB) and the Scripps Plankton Camera (SPC) have each collected billions of images of microscopic marine plankton (Olson and

Sosik, 2007; Orenstein et al., 2020, 2015). Similarly, advances in autonomous acoustic recording devices such as the High-frequency Acoustic Recording Package (HARP) have collected petabytes of passive acoustic data (Wiggins and Hildebrand, 2007). By leveraging recent advances in data collection, machine learning, and parallelizable computing technology, we can uncover hidden patterns, extract knowledge, and make accurate predictions. These technologies have already revolutionized our understanding of the ocean and its dynamic ecosystems.

Machine learning refers to the use of algorithms and computational techniques to extract meaningful information from data, without relying on predetermined equations or explicit instructions (Bishop, 2006). It's important to note that while machine learning is a part of artificial intelligence (AI), the latter encompasses a broader range of capabilities, including the integration of machine learning with sensors, autonomous vehicles, and computer-based reasoning. The most common application of machine learning in oceanography is the automation of repetitive sorting of data, usually in the form of classification or detection (Bishop, 2006). In underwater imaging, machine learning has emerged as the predominant technique for object classification and detection. This development can be primarily attributed to the development of automated image recognition architectures such as convolutional neural networks (CNNs). These algorithms can detect and classify objects and organisms in underwater images with remarkable accuracy, saving researchers significant time and effort.

CNNs have been used in applications ranging from the estimation of plankton and fish population densities (Li et al., 2015), biodiversity monitoring of coral reefs (Jaisakthi et al., 2019), unexploded ordnance detection(Czub et al., 2018), and detection of other man-made objects (Olmos et al., 2002; Rizzini et al., 2015). In acoustics, CNNs have emerged as the dominant approach for sound event detection and source localization. In the 2017 Detection and

Classification of Acoustic Scenes and Events (DCASE) challenge, a CNN achieved state-of-the-art results in the sound event detection task (Mesaros et al., 2017). CNNs have also been used for broadband direction of arrival estimation, obtaining competitive results with steered response power phase transform (SRP-PHAT) beamforming (Brandstein and Ward, 2001).

Despite the numerous benefits offered by machine learning, challenges persist when it comes to its application in oceanography due to the presence of noise. There are various sources of noise that can affect oceanographic data, including instrumental errors, measurement uncertainties, environmental disturbances, and data collection and processing artifacts. These sources can introduce random or systematic errors, outliers, missing values, or inconsistencies into the data.

Deploying machine learning-based models effectively is further complicated by the presence of dataset shift, which refers to differences between the statistical properties of the training and deployment data. This shift is problematic because machine learning models learn a joint distribution between the input features and the target variable based on the training data. One of the most common forms of dataset shift is covariate shift, which occurs when the distribution of the input variables (covariates) in the training data is different from the distribution of the covariates encountered during deployment. Other forms of dataset shift, such as prior probability shift, are also prevalent.

Addressing dataset noise and shift has become a crucial challenge. To combat these issues, some researchers have turned to the utilization of multi-view learning and unsupervised domain adaptation techniques. Multi-view learning leverages multiple perspectives or representations of data to enhance prediction accuracy and robustness. By incorporating diverse representations of the same data, it becomes possible to capture a more comprehensive understanding of the data, thereby mitigating the impact of noise and reducing biases.

Unsupervised domain adaptation, on the other hand, focuses on overcoming the covariate shift problem. In underwater image classification, this can arise due to variations in lighting conditions, water quality, or camera settings between different underwater environments. Unsupervised domain adaptation aims to bridge this gap by learning domain-invariant representations that can generalize well across different domains or underwater scenarios. By leveraging unlabeled data from the target domain and aligning it with the labeled source domain, the algorithm can effectively adapt and transfer knowledge, mitigating the effects of covariate shift.

Enhancing the robustness of machine learning-based classifiers in the field of oceanography holds great promise for advancing our understanding of the oceans. By improving the robustness and accuracy of these classifiers, we identify complex patterns within the oceans, ultimately deepening our knowledge of this crucial component of our planet. This knowledge will enable us to make informed decisions, address environmental challenges, and strive towards the sustainable management and conservation of our oceans.

Bishop, C.M., 2006. Pattern recognition and machine learning, Information science and statistics. Springer, New York.

Blandin, C., Ozerov, A., Vincent, E., 2012. Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. Signal Processing 92, 1950–1960. https://doi.org/10.1016/j.sigpro.2011.09.032

Brandstein, M., Ward, D., 2001. Microphone Arrays: Signal Processing Techniques and Applications. Springer Science & Business Media.

Czub, M., Kotwicki, L., Lang, T., Sanderson, H., Klusek, Z., Grabowski, M., Szubska, M., Jakacki, J., Andrzejewski, J., Rak, D., Bełdowski, J., 2018. Deep sea habitats in the chemical warfare dumping areas of the Baltic Sea. Science of The Total Environment 616–617, 1485–1497. https://doi.org/10.1016/j.scitotenv.2017.10.165

Evers, C., Löllmann, H.W., Mellmann, H., Schmidt, A., Barfuss, H., Naylor, P.A., Kellermann, W., 2020. The LOCATA Challenge: Acoustic Source Localization and Tracking. IEEE/ACM Transactions on Audio, Speech, and Language Processing 28, 1620–1643. https://doi.org/10.1109/TASLP.2020.2990485

Jaisakthi, S.M., Mirunalini, P., Aravindan, C., 2019. Coral Reef Annotation and Localization using Faster R-CNN, in: CLEF.

Li, X., Shang, M., Qin, H., Chen, L., 2015. Fast accurate fish detection and recognition of underwater images with Fast R-CNN, in: OCEANS 2015 - MTS/IEEE Washington. Presented at the OCEANS 2015 - MTS/IEEE Washington, pp. 1–5. https://doi.org/10.23919/OCEANS.2015.7404464

Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., Raj, B., Virtanen, T., 2017. DCASE 2017 Challenge setup: Tasks, datasets and baseline system. Presented at the DCASE 2017 - Workshop on Detection and Classification of Acoustic Scenes and Events.

Olmos, A., Trucco, E., Lane, D., 2002. Automatic man-made object detection with intensity cameras, in: OCEANS '02 MTS/IEEE. Presented at the OCEANS '02 MTS/IEEE, pp. 1555–1561 vol.3. https://doi.org/10.1109/OCEANS.2002.1191867

Olson, R.J., Sosik, H.M., 2007. A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging FlowCytobot. Limnology and Oceanography: Methods 5, 195–203. https://doi.org/10.4319/lom.2007.5.195

Orenstein, E.C., Beijbom, O., Peacock, E.E., Sosik, H.M., 2015. WHOI-Plankton- A Large Scale Fine Grained Visual Recognition Benchmark Dataset for Plankton Classification. arXiv:1510.00745 [cs].

Orenstein, E.C., Ratelle, D., Briseño-Avena, C., Carter, M.L., Franks, P.J.S., Jaffe, J.S., Roberts, P.L.D., 2020. The Scripps Plankton Camera system: A framework and platform for in situ microscopy. Limnology and Oceanography: Methods 18, 681–695. https://doi.org/10.1002/lom3.10394

Rizzini, D.L., Kallasi, F., Oleari, F., Caselli, S., 2015. Investigation of Vision-Based Underwater Object Detection with Multiple Datasets. International Journal of Advanced Robotic Systems 12, 77. https://doi.org/10.5772/60526

Tanhua, T., Pouliquen, S., Hausman, J., O'Brien, K., Bricher, P., De Bruin, T., Buck, J.J.H., Burger, E.F., Carval, T., Casey, K.S., Diggs, S., Giorgetti, A., Glaves, H., Harscoat, V., Kinkade, D., Muelbert, J.H., Novellino, A., Pfeil, B., Pulsifer, P.L., Van De Putte, A., Robinson, E., Schaap, D., Smirnov, A., Smith, N., Snowden, D., Spears, T., Stall, S., Tacoma, M., Thijsse, P., Tronstad, S., Vandenberghe, T., Wengren, M., Wyborn, L., Zhao, Z., 2019. Ocean FAIR Data Services. Front. Mar. Sci. 6, 440. https://doi.org/10.3389/fmars.2019.00440

Wiggins, S.M., Hildebrand, J.A., 2007. High-frequency Acoustic Recording Package (HARP) for broad-band, long-term marine mammal monitoring, in: 2007 Symposium on Underwater Technology and Workshop on Scientific Use of Submarine Cables and Related Technologies. Presented at the 2007 Symposium on Underwater Technology and Workshop on Scientific Use of Submarine Cables and Related Technologies, pp. 551–557. https://doi.org/10.1109/UT.2007.370760

Chapter 2

International Conference on Computer Vision (ICCV) Workshops. 2021.

Improving Rare-Class Recognition of Marine Plankton with Hard Negative Mining

Joseph L. Walker
Scripps Institution of Oceanography
jlwalker@ucsd.edu

Eric C. Orenstein
Monterey Bay Aquarium Research Institute
eorenstein@mbari.org

Abstract

Biological oceanographers are increasingly adopting machine learning techniques to conduct quantitative assessments of marine plankton. Most supervised plankton classifiers are trained on labeled image datasets annotated by domain experts under the closed world assumption: all object classes and their priors are the same during both training and deployment. This assumption, however, is hard to satisfy in the actual ocean where data is subject to dataset shift due to shifting populations and from the introduction of object categories not seen during training. Here we present an alternative approach for training and evaluating plankton classifiers under the more realistic open world scenario. We specifically address the problems of out-of-distribution detection and dataset shift under the class imbalance setting where downsampling is needed to reliably detect and classify relatively rare target classes. We apply a hard negative mining approach called Background Resampling to perform downsampling and compare it to other strategies. We show that Background Resampling improves detection of novel particle classes while simultaneously providing competitive classification performance under dataset shift.

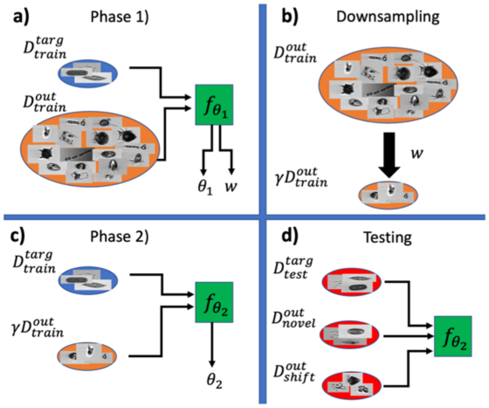


Figure 1. **a)** Using target and background training datasets, denoted as D^{targ}_{train} and D^{out}_{train} respectively, the parameters of a classifier, f_{θ_1}, and background image weights, w, are jointly learned using an alternative optimization approach. **b)** the background dataset is downsampled by interpreting w as resampling probabilities. **c)** the original target dataset and downsampled background dataset is used to train a new classifier, f_{θ_2}. **d)** testing is then performed using target, novel, and dataset shift datasets with f_{θ_2}.

1. Introduction

Marine plankton are a critical component of the biogeochemical processes that are responsible for regulating the climate, supporting the aquatic food web, and producing oxygen [21, 1]. The innumerable ecological roles of plankton make it imperative to monitor their populations as a function of natural and anthropogenic environmental change. Quantifying the fluctuations of individual taxa and the diversity of planktonic communities in response to perturbations is fundamental to understanding planktonic ecosystem dynamics. However, technological limitations constrain our ability to obtain highly temporally resolved time series of individual taxa.

Plankton ecologists are increasingly using *in situ* imaging and deep learning to make population estimates of plankton. Numerous imaging systems have been developed and deployed to study plankton in their natural environment [4, 11, 9, 41]. One of the most widely used plankton imagers is the Imaging FlowCytobot (IFCB), which was developed at Woods Hole Oceanographic Institution (WHOI) to study microorganisms within the 10-100 µm range [36]. The WHOI-Plankton annotated dataset is one of the largest, best maintained labeled plankton image sets available [39].

Together with *in situ* imaging, advances in deep learning have enabled oceanographers to sample ocean populations with higher spatiotemporal resolution and provide the opportunity to produce long, highly resolved time series of individual taxa. Convolutional Neural Networks (CNNs), a family of deep neural network architectures, have been shown to improve classification accuracy on marine plankton imagery versus ensemble or margin-based methods [38, 16, 39]. CNNs obviate the need for defining handcrafted features by learning the feature extraction and classification process end-to-end. This training paradigm enables the learning of feature representations with more discriminative power [26, 25]. CNNs could therefore alleviate the human cost of manually examining the collected data in order to extract ecologically relevant information.

In many cases, biological oceanographers are only interested in identifying organisms that belong to a small set of classes, referred to here as *target classes*. Some projects are specifically formulated to reduce the number of target classes: Harmful Algal Bloom (HAB) monitoring and parasite tracking to name a few [37, 14, 24, 35, 5, 10, 6]. The annotation process requires a trained taxonomist to search through a large set of images obtained from an experiment or deployment and sort them into ecologically meaningful classes. Either by design or due to circumstance, the available data for classifier training will consist of labeled images associated with the target classes and a large pool of unlabeled data often simply called "other". Classifier training is thus often formulated as an *N+1* classification problem, where there are N target classes and all other object types are mapped to an additional *background* class. The term *background* therefore refers to data that the classifier has been trained to distinguish from target examples, whereas *out-of-distribution* (OOD) refers to data from novel classes that the classifier has not been trained on and is only exposed to during the testing stage or deployment.

The combined abundance of objects from the target classes is often much smaller than the prevalence of all other objects that form the background, the so-called class *imbalance* problem [27, 7, 6, 17]. The issue is exacerbated by the size structure of particles in the ocean and design constraints of imaging systems. There are orders of magnitude more small objects near the lower resolution limit of any optical imaging system. As a result, *in situ* optical imaging systems will image more small indistinguishable objects than large easy-to-identify particles [41, 36, 17]. Therefore, the background class will often be populated by many examples of these small undifferentiated particles.

Training on imbalanced data will encourage a machine learning based classifier to minimize its loss by accurately and reliably classifying majority class examples at the price of diminishing recall of minority class examples [33]. A widely adopted strategy for addressing the imbalance problem is to upsample the minority classes via data augmentation and downsample the larger classes via random downsampling [40, 7]. However, random downsampling is likely to lose crucial information regarding the distribution of possible features that are associated with objects belonging to the background class.

Developing effective machine classifiers for plankton imagery is further complicated by the diversity and constant flux of novel taxa present in the sampling environment [20, 44]. *N+1* classifier training implicitly assumes the classifier's learned representations are robust enough that any and all future objects that do not belong to the set of target classes will be mapped to the background class. But this kind of generalization is not explicitly enforced or encouraged when the classifier is trained and evaluated on datasets that share the same set of labels; a common practice in plankton ecology studies [35, 38, 16, 17, 9, 12].

When a classifier is tasked with labeling unlabeled data, another assumption is made: that the class priors and distribution of features characterizing the classes are unchanging. Changes in these distributions are broadly referred to as dataset shift, and have been shown to impact classifier performance. This problem has received a significant amount of attention in both the plankton ecology [40, 19, 2] and machine learning [34, 54, 18] communities.

Plankton recognition in the open ocean is a particularly challenging endeavor because incoming data is almost guaranteed to be imbalanced, composed of novel classes, and subject to dataset shift. In this work, we present an effective solution to this integrated recognition task for the case where the goal is to identify images belonging to relatively uncommon plankton groups. We examine how the construction of the background class training set via downsampling can impact out-of-distribution detection and dataset shift classification performance. We use a hard negative mining approach called Background Resampling to optimize the downsampling procedure to preserve information regarding the set of features associated with the background class. Our study makes the following three particular contributions:

1. We present a new framework for training and evaluating plankton classifiers that addresses the challenges that are encountered in an open ocean deployment, primarily OOD detection and dataset shift.
2. We show that downsampling via hard negative mining can endow models with greater generalization abilities across a range of challenging test scenarios where other approaches are inconsistent.
3. We benchmark a contemporary OOD detection technique on a fine-grained OOD detection problem.

2. Related Work

2.1. Out-of-distribution detection

Out-of-distribution (OOD) detection methods seek to train a classifier to successfully recognize data that does not belong to the set of target classes. In the case of marine plankton classification, OOD data would present as novel object classes. Outlier Exposure (OE), a popular new approach for OOD detection, leverages the fact that deep networks produce an estimate of the posterior class distribution. OE measures the entropy of the posterior class distribution to estimate the likelihood that a given data point is OOD [23]. This is implemented with a softmax network layer, which models the probability of an input x being recognized as class i as

$$P(i|x) = \frac{exp\left(w_i^T g\left(x;\theta\right) + b_i\right)}{\sum_{j=1}^{N} exp\left(w_j^T g\left(x;\theta\right) + b_j\right)} \tag{1}$$

where $i \in \{1, 2, ..., N\}$ indexes one of the N target classes. $g(x;\theta)$ denotes the embedding of example x in feature space as a function of network parameters θ. w_j and b_j denote the weight vector and bias terms for class j respectively. The classifier is trained to output a high entropy (i.e., uniform) distribution $P(i|x)$ for background examples, and a confident low entropy distribution for examples from the target classes. For OE, classification is performed by thresholding the softmax scores, where the threshold T is determined empirically from a validation set and calibrated to provide a desired recall on the target classes. If $T < max_i\, P(i|x)$ then the classification is upheld, otherwise, the example is classified as non-target.

OE has been shown to generalize well to OOD examples that come from an entirely different domain [23, 28, 15]. This inspired a wave of OOD detection models which build from the OE concept. [15] introduced Objectosphere loss which aims to minimize the magnitude of $g(x;\theta)$ for background data, which naturally results in low confidence softmax outputs. [30] incorporated scaling and input preprocessing to further increase the softmax output disparity between target and background data. However, these methods are typically tested using OOD and target data from completely separate domains. This is unlike many real-world applications, where OOD data is from the same domain as, and looks very similar to, target class examples. In the case of marine particle classification, particle classes can be visually very similar, which makes OOD detection a challenging problem [12, 48, 32, 55].

2.2. Hard negative mining

Hard negative mining (HNM) approaches seek to identify a set of negative (or background) examples that are likely to generate a false positive [45, 13, 53]. Focusing classifier training on these hard examples has been shown to improve classification performance relative to other downsampling methods [13, 29]. While similar techniques have been applied to datasets consisting of hand-crafted features to predict phytoplankton blooms, to our knowledge, they have not been applied to plankton image classification [49].

3. Dataset

We use the WHOI-Plankton dataset[1] for all experiments [52]. This fully annotated dataset is comprised of 103 classes totaling over 3.5 million grayscale IFCB images, ranging from millions to as few as four examples per class. The bulk of the images belong to the *mix* category which corresponds to small unidentifiable particles. This dataset was amassed over 9 years (2006-2014) from nearly continuous sampling at the Martha's Vineyard Coastal Observatory. An expert taxonomist labeled all images collected in two randomly selected, non-consecutive, single hour time points from each two-week period. Each hour thus represents a complete, independent sample of the plankton population at that point in time. The image data is sorted into subfolders reflecting the image acquisition year.

4. Methods

4.1. Background resampling

Background Resampling (BR) is a HNM approach which aims to ameliorate the class imbalance problem while improving OOD detection [29]. BR assigns each background training image a weight that is proportional to the confidence with which the image is classified as one of the target classes. Then a subset of the background images is sampled according to the image weights which are interpreted as resampling probabilities. This downsampled set is then used to train a new classifier. BR is from the family of OE methods for OOD detection and therefore requires both background and target training datasets, denoted as D_{train}^{out} and D_{train}^{targ} respectively. D_{train}^{out} and D_{train}^{targ} are used to train the parameters θ_1 of a classifier, denoted as f_{θ_1}, to output high and low entropy distributions over the softmax outputs (eq. 1) respectively. The BR procedure can be broken into two distinct phases:

Phase (1): Using D_{train}^{out} and D_{train}^{targ}, learn the background image weights w and θ_1 with the alternative optimization

$$\theta_1^{(t)} = \underset{\theta_1}{argmin} \left[L_{targ}\left(\theta_1\right) + \alpha L_{out}(\theta_1; w^{(t-1)})\right] \tag{2}$$

$$w^{(t)} = \underset{w}{argmax} \left[L_{targ}(\theta_1^{(t)}) + \alpha L_{out}(\theta_1^{(t)}; w)\right] \tag{3}$$

where L_{targ} is the cross-entropy classification loss term used to penalize incorrect classifications for target class data. L_{out} is the loss term that penalizes overly confident predictions on background examples and is defined as the Kullback-Leibler divergence between the uniform distribution and the softmax outputs. The solution to this system of equations is approximated using a differential relaxation (stochastic gradient descent) and batches of images from

[1] doi:10.1575/1912/7341

both D^{out}_{train} and D^{targ}_{train}. $w^{(t)}$ is defined as the set of image weights that *maximize* the associated loss at time step t, where t denotes the batch number. This ensures that the reweighting algorithm will assign high weight values to images from D^{out}_{train} that are difficult to classify, guaranteeing that the resampling process selects challenging background images that are visually very similar to the target class examples. The adversarial nature of the iterative process – classification vs selection of difficult examples for the classifier – is critical to accurately learning the boundary between target and background classes. The hyperparameter α controls the trade-off between learning to output confident and low-confident predictions for target class and background examples respectively. For all experiments, we set $\alpha = 0.5$ in accordance with the standard OE default [23, 29].

Phase (2): A resampling percentage γ is empirically set and will typically reflect the degree of imbalance between the background and target classes. Using the learned background image weights, the background class is downsampled to γ percent of its original size. This is done by selecting each background image x_i, associated with weight w_i, independently with probability $p_i = min\left(1, \frac{\gamma D^{out}_{train}}{\sum_{j=1} w_j} w_i\right)$. Once the background image weights are obtained, f_{θ_1} is discarded. The downsampled background dataset and full target training dataset are then used to train another classifier, denoted as f_{θ_2}. Testing is then performed with f_{θ_2}. For all experiments, we use γ = 0.05. A schematic diagram of the entire process is shown in Fig. 1.

4.2. Experimental setup

Our experiments were designed to simulate a scenario where a biological oceanographer is interested in tracking the prevalence of a few relatively rare plankton groups. The abundance of these groups can fluctuate over a very large background class whose images are not of interest. We construct subsets of the WHOI-Plankton dataset to perform our experiments:

Target Data. The classes to be detected, or target classes, are *Ceratium*, *Dinobryon*, *Pleurosigma* and *Ephemera*. All the available data for these four classes is denoted as D^{targ}. Both the *Dinobryon* and *Ceratium* genus are associated with algal blooms. The *Pleurosigma* genus is of interest in biomedical applications because they are believed to produce rare but important organic compounds [3, 56]. The *Ephemera* class is taxonomically ambiguous, but previous studies have identified it as difficult to classify because of its visual similarity to other organisms in the WHOI-Plankton dataset [55]. Images from these four classes were drawn from each year of the WHOI-Plankton dataset but capped at 900 examples per class. This was to prevent significant class imbalance within the set of target classes and to provide a realistic amount of data that could be obtained relatively easily from a low-budget data annotation campaign.

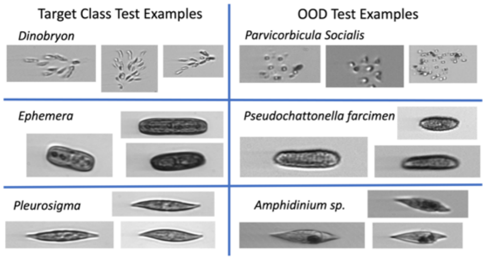


Figure 2. Three examples, selected by a human annotator, from classes in D^{targ}_{test} (left column) and D^{out}_{novel} (right column) showing the morphological similarity between specimen of these classes

Training Data. 55% of data from D^{targ} (1783 images) was randomly selected (stratified by class) to serve as a training set for the target classes and is referred to as D^{targ}_{train}. The background training dataset, denoted as D^{out}_{train}, consists of all images from the year 2006 (totaling 134,293 images) that do not belong to the target classes.

Validation data. 22% of data from $D^{targ} \backslash D^{targ}_{train}$ (311 images), referred to as D^{targ}_{val}, was randomly selected to be used to learn the OOD detection decision threshold T.

Testing Data. We construct three different testing datasets to assess classification performance on target examples, novel object classes, as well as a dataset shift scenario where the prior probabilities of the classes in D^{out}_{train} are subject to change. The remaining 78% of data from $D^{targ} \backslash D^{targ}_{train}$ (1125 images) was selected for testing target class classification and is referred to as D^{targ}_{test}.

There are 13 classes in the WHOI-Plankton dataset that are not present in $D^{targ}_{train} \cup D^{out}_{train}$. These classes were used to form a hold-out set of novel classes to test OOD detection performance. This hold-out set is referred to as D^{out}_{novel} (totaling 1112 images). Many of the classes in D^{out}_{novel} look remarkably similar to the classes in D^{targ}_{test}, underscoring the difficulty of OOD detection in plankton imagery (Fig. 2). For OOD detection testing, we utilize datasets D^{targ}_{test} and D^{out}_{novel}. Since they are approximately the same size, testing on the combination of these sets implicitly assumes that the number of target class and OOD examples is similar. This may be realistic if the novel classes are relatively rare, but in practice the target examples are often rare compared to non-target examples. Therefore, we test classifier performance using several ratios of target to OOD examples using subsamples from D^{targ}_{test} and the full D^{out}_{novel} set.

While we wish to develop classifiers that reliably detect novel examples, it is important that improved OOD detection does not diminish classifier performance on other important aspects of plankton recognition, such as recognition under dataset shift. To simulate a real world deployment,

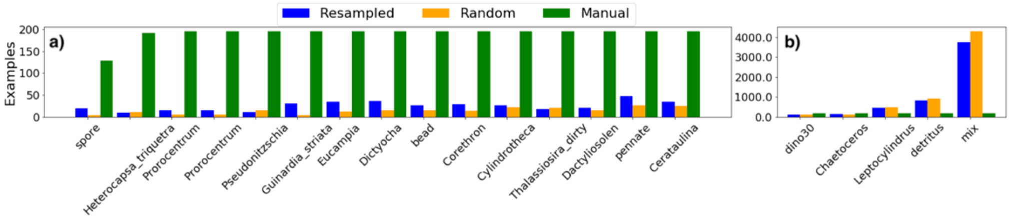


Figure 3. Class distribution of the downsampled background sets generated from each of the downsampling strategies. **a)** class distribution for the 15 least abundant classes that have at least 100 examples in the original D_{train}^{out} set. **b)** class distribution for the five most abundant classes in D_{train}^{out}. Note the change in vertical axis scale between **a)** and **b)**.

each day's worth of data from the WHOI-Plankton 2014 image directory is used to test each classifier under a dataset shift scenario, where the classes are the same as D_{train}^{out}, but the prior probabilities and appearance of the background classes are subject to change. This testing dataset is denoted as D_{shift}^{out} (totaling 329,835 images). For dataset shift testing, each day's worth of data from D_{shift}^{out} is combined with a random 75% sample of D_{test}^{targ}, ensuring that the background and target classes are subject to dataset shift.

4.3. Models and training

The dataset size ratio D_{train}^{targ}:D_{train}^{out} is approximately 1:75. For the class with the fewest examples in D_{train}^{targ}, denoted as $D_{train}^{targ\_min}$, the ratio $D_{train}^{targ\_min}$:D_{train}^{out} is approximately 1:400. Training on data with this degree of imbalance typically results in poor detection for minority class examples [7, 49, 50]. Instead, it is common to downsample the majority classes and upsample the minority classes to improve results [51, 31, 47]. Using our downsampling percentage γ = 0.05, the ratio $D_{train}^{targ\_min}$:γD_{train}^{out} is approximately 1:20. To fully balance the classes, we upsample the four target classes using random image rotations.

It is possible that all images of a rare background object class are lost when using random downsampling. This increases the risk that the classifier trained on the randomly downsampled data will assign examples of that class to one of the target classes. In the case where the images in D_{train}^{out} are assigned their true class label, D_{train}^{out} can be downsampled by taking an even number of examples from each class within D_{train}^{out}, referred to as class-balanced downsampling. This guarantees that every class is represented in the downsampled dataset, therefore maximizing the feature diversity in this new downsampled set. For this reason, we consider the scenario where a fine-grained labeled image set is available and class-balanced downsampling is possible. In this setting, images associated with the background meta-class are assigned their true class label, but still trained as one class.

All classifiers are fine-tuned ResNet-18 models [22], pre-trained on ImageNet [46]. Three downsampling methods are used to train the classifiers and compared. Each classifier uses the same training procedure, using D_{train}^{targ} but a different subset of D_{train}^{out}:

1. **Resampled**: Trained on a subset of images from D_{train}^{out} of approximate size γD_{train}^{out} that was selected according to the resampling probabilities described in Sec. 4.1.
2. **Random**: Trained on a subset of D_{train}^{out} of approximate size γD_{train}^{out} drawn randomly. This represents the standard downsampling approach and therefore serves as a baseline for comparison.
3. **Manual (class-balanced)**: Each subclass within the background meta-class is downsampled by capping the number of examples at 196. This upper limit was determined empirically to yield a downsampled background meta-class of approximate size γD_{train}^{out}. Note that this mode of downsampling is only possible if labeled data is available for background examples. Using this classifier as a baseline, we seek to determine whether BR is beneficial when labeled data is available for background examples.

For each phase (defined in Sec. 4.1), we used image batch sizes of 64 from both D_{train}^{out} and D_{train}^{targ}. The weight learning optimization is performed until the loss associated with the background image weights (eq. 3) fails to decrease for 10 epochs. For phase 2 training, each classifier was trained on its respective subset of D_{train}^{out} for 50 epochs, using an initial learning rate of 0.0003 which was reduced by a factor of 0.5 after every 10 epochs.

Predetermining the number of epochs is common for studies involving OE [23, 28, 29, 42] since the validation set is used to learn the decision threshold T rather than to perform early stopping. The values for all other hyperparameters used during training are those of [29]. After training, a decision threshold is calculated for each classifier as the largest threshold that allows for 95% recall of examples from D_{val}^{targ}.

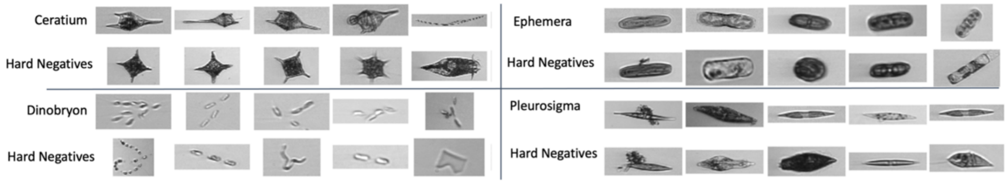


Figure 4. Samples from each target class and their respective hard negatives. Ten background classes are represented by the hard negative examples.

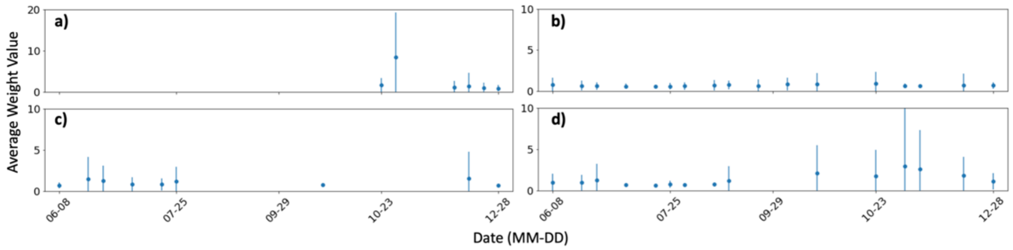


Figure 5. Daily average background image weight value associated with each of the hard background classes presented in the order: **a)** *Dictyocha*; **b)** *detritus*; **c)** *Skeletonema*; **d)** *pennate*. The bars represent the standard deviation of the weight values. Note that the vertical axis scale in **a)** is greater than that of **b)-d)**.

4.4. Performance metrics

We use two metrics to assess model performance:

1. **F1 score** is a metric for binary classification, defined as the harmonic mean of the precision and recall for a given class. The F1 score is calculated for each class, by treating all other classes as a single class. The F1 scores are then uniformly averaged over each class.
2. **Accuracy (overall precision)** is the fraction of correctly classified images from the four target classes and background class.

Using these two metrics, we benchmark each classifier on an OOD detection task (Sec. 5.2) and dataset shift scenario (Sec. 5.3).

4.5. Alternative target classes

To test the generalization of BR, we repeated all experiments for five different sets of four target classes. These classes were randomly selected but were restricted to classes with 600-10,000 examples. This restriction was added to preserve the D^{targ}_{train}:D^{out}_{train} ratio across all sets of target classes. For all target classes considered, the number of examples per class was capped at 900.

5. Results

5.1. Downsampling analysis

Based on the class frequency distribution, BR draws disproportionately more examples from the minority classes compared to random downsampling (Fig. 3). While manual downsampling also samples disproportionately from the minority classes, it creates a class distribution that is radically different than the natural population distribution.

To visualize difficult OOD samples, we drew examples from the background class that were classified into one of the target labels with high confidence (Fig. 4). These "hard negatives" reveal that the classifier confused background examples from more than just a select few classes. Four background classes are represented in these hard negatives more than others: *Dictyocha*, *detritus*, *Skeletonema*, and *pennate*. We refer to these as "hard background classes".

The daily average image weight values associated with the hard background classes vary substantially between classes and over time (Fig. 5). This information allows us to determine whether examples within the hard classes were consistently ascribed higher weight values or if hard negatives are outlier examples for those classes.

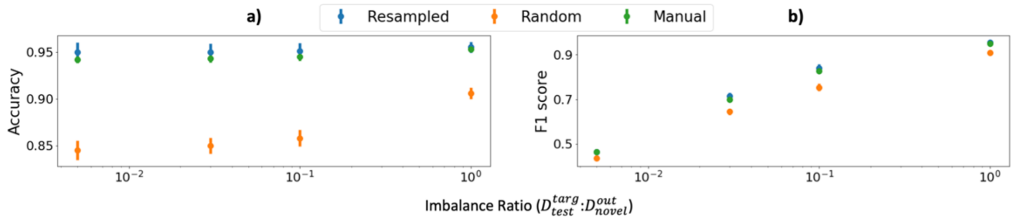


Figure 6. OOD testing results. Left-to-right: Accuracy and F1 score obtained from an average of model runs using D_{test}^{targ}:D_{novel}^{out} ratios of $\{1\!:400, 1\!:50, 1\!:10, 1\!:1\}$. Error bars reflect standard deviation.

5.2. OOD testing

Testing is done using k-fold cross-validation where the number of folds reflects the desired imbalance ratio. We consider D_{test}^{targ}:D_{novel}^{out} ratios of $\{1\!:400, 1\!:50, 1\!:10, 1\!:1\}$ where classifier performance is averaged over each fold (Fig. 6). The ratios considered here are designed to reflect sampling environments that a plankton ecologist may encounter during deployment. The 1:400 ratio represents the most extreme case of population flux, where the deployment environment consists overwhelmingly of novel classes. This is akin to taking a classifier trained on data from the North Atlantic and using it to detect the same four classes in the Tasman Sea. The 1:10 and 1:50 ratios represent a more amenable scenario, where the model is deployed in a similar environment where target class organisms are not as rare. The 1:1 ratio produces a balanced testing set, and assumes an even number of novel and target class examples in the deployment environment.

5.3. Dataset shift testing

When background class population statistics remain roughly constant throughout deployment, the training dataset generated by BR may produce a classifier that is biased against identifying the more common classes. This is because training is disproportionately focused on rare/abnormal examples under the BR procedure (Fig. 3). To assess model performance under a variety of deployment scenarios, we test and average the performance of each classifier over each day's worth of data in D_{shift}^{out} combined with randomly drawn target class examples (Table 1). This assessment measures the classifiers ability to classify under changes in prior distributions and dataset shift. No novel classes were used in this testing scenario.

5.4. Alternative target class testing

For the different sets of target classes, the relative performance among the classifiers was similar to the results shown for the target classes considered in Sec. 4.2.

OOD testing. BR brought the largest performance gains when the target and OOD classes were visually very similar. All OOD detection results are reported for the 1:1 testing ratio. For accuracy, the *Resampled* classifier on average outperformed the *Random* and *Manual* classifiers by 5.1% and 0.2% respectively. For F1 score, the *Resampled* classifier on average outperformed *Random* and *Manual* classifiers by 4.7% and 0.0% respectively.

Method	Accuracy	F1 Score
Resampled	**88 ± 1.4**	**94.1 ± .23**
Random	85 ± 1.3	93.9 ± .12
Manual	79.7 ± 2.8	79.2 ± 1.5

Table 1. Dataset shift testing results (in % including ± Std Dev.).

Dataset shift testing. For accuracy, the *Resampled* classifier on average outperformed the *Random* and *Manual* classifiers by 0.9% and 9.8% respectively. For F1 score, the *Resampled* classifier on average outperformed the *Random* and *Manual* classifiers by 0.1% and 7.1%.

6. Discussion

We have shown that when downsampling is required, OOD detection performance can be improved by selecting an optimal subset of background training images. In each testing scenario, BR slightly outperformed its nearest competitor. However, BR was the only downsampling method to perform well in both testing scenarios, whereas the performance of the other two downsampling methods varied significantly in each regime. This was observed for the alternative target classes as well. This finding underscores the efficacy of BR since an automated plankton classifier deployed on real-time data is almost guaranteed to experience both novel classes and dataset shift.

For some hard background classes, the distribution of image weights appears to have a seasonal dependence (Fig. 5). The image weights associated with the *detritus* class (Fig. 5b) are comparatively low, suggesting that only occasional instances of *detritus* will possess physical attributes that pose a challenge to the classifier. This makes sense considering the large range of shapes and textures that objects described as "detritus" can have. This differs from classes like *Dictyocha*, *Skeletonema*, and *pennate* whose images are assigned comparatively larger weight values. However, these three classes show seasonal, and even daily, with-in class variability as indicated by the standard deviation of the image weights (Fig. 5d). This variability suggests that an optimal background training set for OOD detection has to be curated at the example level - randomly sampling from harder background classes may not produce adequate discrimination between the background and target classes. BR optimizes the downsampling strategy by deliberately selecting hard negatives that are very close to the target classes (Fig. 4).

The F1 score from each test set suggests that the 1:10 and 1:50 imbalance ratio produces greater performance differences (Fig. 6a). Despite the relatively higher accuracy obtained by the *Resampled* and *Manual* classifiers, the target classes become so polluted by false positives that absolute and relative performance – as measured by F1 score – degrades significantly with increasing imbalance ratio.

Overall, BR provides competitive or even slightly better OOD detection performance than the class-balanced downsampling used to train the *Manual* classifiers (Fig. 6). The improved performance can likely be attributed to the fact that class-balanced downsampling, while drawing from each class disproportionality, still performs random downsampling within each class. BR, in contrast, selects disproportionately from each class, while simultaneously select challenging examples from within each class. This may be particularly valuable for taxonomic groups where organisms can express different phenotypes as they go through different life stages.

While the *Manual* classifier yields satisfactory detection on novel classes (Fig. 6), this classifier significantly underperforms on natural population changes compared to the other classifiers (Table 1). This is likely because class-balanced downsampling produces a background class distribution that is significantly different from the background class distribution encountered during deployment. The classifiers trained on randomly drawn subsets perform comparatively well in the dataset shift scenario, likely because the background class distribution generally resembles the class priors encountered during testing. BR will typically draw relatively more examples from the minority classes and fewer examples from the most abundant classes as compared to the *Random* training set (Fig. 3). But BR's disproportionate subsampling is not as extreme as the class-balanced downsampling. The fact that BR preserves a significant amount of information regarding the class priors is perhaps why it performs better than class-balanced downsampling for natural population distributions

7. Comments and recommendations

In order to adequately simulate the deployment of a classifier, our testing procedure involved the use of data from training and novel classes as well as shifting prior distributions. We believe this to be the most rigorous form of testing and hope that this study can serve as a framework for future plankton classifier benchmarking. The high visual similarity between the target and OOD examples makes this a challenging detection problem. Most of the methods introduced in the OOD literature, including BR, test using OOD examples that come from an entirely separate domain. This study is one of the first studies to benchmark the performance of a contemporary OOD detection method on in-domain OOD data. It is our hope that these results will facilitate the development of new tools for HAB species monitoring and early detection systems. The reduced false positive rates demonstrated in our experiments make the output of the algorithm more amenable to quality control for verification.

The BR procedure can be used to improve classification systems that incorporate an ensemble of "one-versus-all", which are popular within the plankton ecology community and have been used for HAB species monitoring [8, 43]. This could be done by training each one-versus-all classifier using a subset of background images that is optimized to produce the best discrimination for the class that each classifier is trying to detect.

We note that the utility of the background weight learning mechanism is not limited to HNM approaches. It can in and of itself be used to communicate potential failure modes to a human supervisor. For example, by examining the background images that were assigned large weight values, a human user could learn prior to deployment which non-target classes are likely to produce false positives. With this knowledge, they may decide to train the model to detect these tough classes as well.

We have shown that obtaining class labels for background objects for the purpose of class-balanced downsampling does not improve OOD detection performance. Therefore, we conclude that for any future plankton classification campaigns similar to this experimental setup, all human annotation efforts should be focused on the target classes. Instead of random or class-balanced downsampling, automatic procedures such as BR should be used to optimally resample the 'other' category.

Acknowledgements This work was supported by the U.S. Office of Naval Research under Award No. N00014-19-1-2851. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research. We would also like to thank Jules S. Jaffe and Nuno Vasconcelos for critical readings of the manuscript and their advisory roles. We thank Yi Li for advising the use of the Background Resampling method.

References

[1] Kevin R. Arrigo. Marine microorganisms and global nutrient cycles. *Nature*, 437(7057):349–355, Sept. 2005. Number: 7057 Publisher: Nature Publishing Group.

[2] Oscar Beijbom, Judy Hoffman, Evan Yao, Trevor Darrell, Alberto Rodriguez-Ramirez, Manuel Gonzalez-Rivero, and Ove Hoegh Guldberg. Quantification in-the-wild: data-sets and baselines. *arXiv:1510.04811 [cs]*, Nov. 2015. arXiv: 1510.04811.

[3] Simon T. Belt, Guy Allard, Guillaume Massé, Steve Rowland, and Jean-Michel Robert. Important sedimentary sesterterpenoids from the diatom Pleurosigma intermedium. *Chemical Communications*, (6):501–502, 2000.

[4] Mark Benfield, Philippe Grosjean, Phil Culverhouse, Xabier Irigolen, Michael Sieracki, Angel Lopez-Urrutia, Hans Dam, Qiao Hu, Cabell Davis, Allen Hanson, Cynthia Pilskaln, Edward Riseman, Howard Schulz, Paul Utgoff, and Gabriel Gorsky. RAPID: Research on Automated Plankton Identification. *Oceanography*, 20(2):172–187, June 2007.

[5] Tristan Biard, Lars Stemmann, Marc Picheral, Nicolas Mayot, Pieter Vandromme, Helena Hauss, Gabriel Gorsky, Lionel Guidi, Rainer Kiko, and Fabrice Not. In situ imaging reveals the biomass of giant protists in the global ocean. *Nature*, 532(7600):504–507, Apr. 2016. Number: 7600 Publisher: Nature Publishing Group.

[6] Erik Bochinski, Ghassen Bacha, Volker Eiselein, Tim J. W. Walles, Jens C. Nejstgaard, and Thomas Sikora. Deep Active Learning for In Situ Plankton Classification. In Zhaoxiang Zhang, David Suter, Yingli Tian, Alexandra Branzan Albu, Nicolas Sidère, and Hugo Jair Escalante, editors, *Pattern Recognition and Information Forensics*, Lecture Notes in Computer Science, pages 5–15, Cham, 2019. Springer International Publishing.

[7] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, Oct. 2018.

[8] Lisa Campbell, Darren W. Henrichs, Robert J. Olson, and Heidi M. Sosik. Continuous automated imaging-in-flow cytometry for detection and early warning of Karenia brevis blooms in the Gulf of Mexico. *Environmental Science and Pollution Research*, 20(10):6896–6902, Oct. 2013.

[9] R W Campbell, P L Roberts, and J Jaffe. The Prince William Sound Plankton Camera: a profiling in situ observatory of plankton and particulates. *ICES Journal of Marine Science*, 77(4):1440–1455, July 2020.

[10] Svenja Christiansen, Henk-Jan Hoving, Florian Schütte, Helena Hauss, Johannes Karstensen, Arne Körtzinger, Simon-Martin Schröder, Lars Stemmann, Bernd Christiansen, Marc Picheral, Peter Brandt, Bruce Robison, Reinhard Koch, and Rainer Kiko. Particulate matter flux interception in oceanic mesoscale eddies by the polychaete Poeobius sp. *Limnology and Oceanography*, 63(5):2093–2109, 2018.

[11] Robert K. Cowen and Cedric M. Guigand. In situ ichthyoplankton imaging system (ISIIS): system design and preliminary results. *Limnology and Oceanography: Methods*, 6(2):126–132, 2008.

[12] Jialun Dai, Zhibin Yu, Haiyong Zheng, Bing Zheng, and Nan Wang. A Hybrid Convolutional Neural Network for Plankton Classification. In Chu-Song Chen, Jiwen Lu, and Kai-Kuang Ma, editors, *Computer Vision – ACCV 2016 Workshops*, Lecture Notes in Computer Science, pages 102–114, Cham, 2017. Springer International Publishing.

[13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, June 2005. ISSN: 1063-6919.

[14] Cabell S. Davis, Qiao Hu, Scott M. Gallager, Xiaoou Tang, and Carin J. Ashjian. Real-time observation of taxa-specific plankton distributions: an optical sampling method. *Marine Ecology Progress Series*, 284:77–96, Dec. 2004.

[15] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing Network Agnostophobia. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9157–9168. Curran Associates, Inc., 2018.

[16] Jeffrey S. Ellen, Casey A. Graff, and Mark D. Ohman. Improving plankton image classification using context metadata. *Limnology and Oceanography: Methods*, page lom3.10324, July 2019.

[17] Robin Faillettaz, Marc Picheral, Jessica Y. Luo, Cédric Guigand, Robert K. Cowen, and Jean-Olivier Irisson. Imperfect automatic image classification successfully describes plankton distribution patterns. *Methods in Oceanography*, 15-16:60–77, Apr. 2016.

[18] George Forman. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206, Oct. 2008.

[19] Pablo González, Eva Álvarez, Jorge Díez, Ángel López-Urrutia, and Juan José del Coz. Validation methods for plankton image classification systems. *Limnology and Oceanography: Methods*, 15(3):221–237, 2017.

[20] L. R. Haury, J. A. McGowan, and P. H. Wiebe. Patterns and Processes in the Time-Space Scales of Plankton Distributions. In John H. Steele, editor, *Spatial Pattern in Plankton Communities*, NATO Conference Series, pages 277–327. Springer US, Boston, MA, 1978.

[21] Graeme C. Hays, Anthony J. Richardson, and Carol Robinson. Climate change and marine plankton. *Trends in Ecology & Evolution*, 20(6):337–344, June 2005.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE.

[23] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep Anomaly Detection with Outlier Exposure. *arXiv:1812.04606 [cs, stat]*, Jan. 2019. arXiv: 1812.04606.

[24] Qiao Hu and Cabell S. Davis. Accurate automatic quantification of taxa-specific plankton abundance using dual classification with correction. 2006.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2017.

[26] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov. 1998. Conference Name: Proceedings of the IEEE.

[27] Hansang Lee, Minseok Park, and Junmo Kim. Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3713–3717, Sept. 2016. ISSN: 2381-8549.

[28] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples. *arXiv:1711.09325 [cs, stat]*, Feb. 2018. arXiv: 1711.09325.

[29] Yi Li and Nuno Vasconcelos. Background Data Resampling for Outlier-Aware Classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13215–13224, Seattle, WA, USA, June 2020. IEEE.

[30] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. *arXiv:1706.02690 [cs, stat]*, Feb. 2018. arXiv: 1706.02690.

[31] Charles X Ling and Chenghui Li. Data Mining for Direct Marketing: Problems and Solutions. *KDD'98: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, page 7, 1998.

[32] George B. McManus and Laura A. Katz. Plankton Identification: Morphology or Molecules or Both? *Limnology and Oceanography Bulletin*, 18(4):86–90, 2009.

[33] Giovanna Menardi and Nicola Torelli. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1):92–122, Jan. 2014.

[34] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, Jan. 2012.

[35] Aleksander Borge Nesse. Classifying Dinoflagellates in Palynological Slides Using Convolutional Neural Networks. 2020. Accepted: 2020-09-28T18:52:08Z Publisher: University of Stavanger, Norway.

[36] Robert J. Olson and Heidi M. Sosik. A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging FlowCytobot. *Limnology and Oceanography: Methods*, 5(6):195–203, 2007.

[37] Eric Coughlin Orenstein. *Automated analysis of oceanographic image data*. PhD thesis, UC San Diego, 2018.

[38] Eric C. Orenstein and Oscar Beijbom. Transfer Learning and Deep Feature Extraction for Planktonic Image Data Sets. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1082–1088, Mar. 2017. ISSN: null.

[39] Eric C. Orenstein, Oscar Beijbom, Emily E. Peacock, and Heidi M. Sosik. WHOI-Plankton- A Large Scale Fine Grained Visual Recognition Benchmark Dataset for Plankton Classification. *arXiv:1510.00745 [cs]*, Oct. 2015. arXiv: 1510.00745.

[40] Eric C. Orenstein, Kasia M. Kenitz, Paul L. D. Roberts, Peter J. S. Franks, Jules S. Jaffe, and Andrew D. Barton. Semi- and fully supervised quantification techniques to improve population estimates from machine classifiers. *Limnology and Oceanography: Methods*, 18(12):739–753, 2020.

[41] Eric C. Orenstein, Devin Ratelle, Christian Briseño-Avena, Melissa L. Carter, Peter J. S. Franks, Jules S. Jaffe, and Paul L. D. Roberts. The Scripps Plankton Camera system: A framework and platform for in situ microscopy. *Limnology and Oceanography: Methods*, 18(11):681–695, 2020.

[42] Aristotelis-Angelos Papadopoulos, Mohammad Reza Rajati, Nazim Shaikh, and Jiamian Wang. Outlier Exposure with Confidence Control for Out-of-Distribution Detection. *arXiv:1906.03509 [cs, stat]*, June 2020. arXiv: 1906.03509.

[43] Vito P. Pastore, Thomas G. Zimmerman, Sujoy K. Biswas, and Simone Bianco. Annotation-free learning of plankton for classification and anomaly detection. *Scientific Reports*, 10(1):12142, Dec. 2020.

[44] Emily E. Peacock, Robert J. Olson, and Heidi M. Sosik. Parasitic infection of the diatom Guinardia delicatula, a recurrent and ecologically important phenomenon on the New England Shelf. *Marine Ecology Progress Series*, 503:1–10, Apr. 2014.

[45] Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Human Face Detection in Visual Scenes. Technical report, Carnegie Mellon University, 1995.

[46] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec. 2015.

[47] A.H. Schistad Solberg and R. Solberg. A large-scale evaluation of features for automatic detection of oil spills in ERS SAR images. In *IGARSS '96. 1996 International Geoscience and Remote Sensing Symposium*, volume 3, pages 1484–1486 vol.3, May 1996.

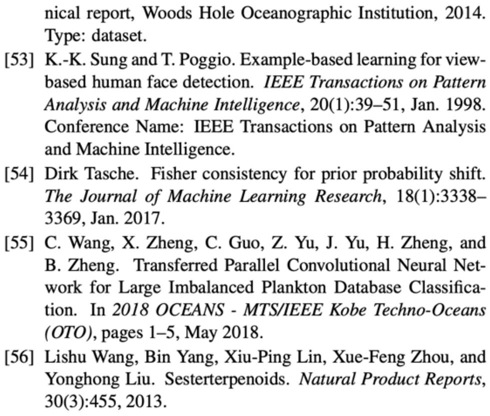
[48] Jan Schulz, Kristina Barz, Patricia Ayon, Andree Ludtke, Oliver Zielinski, Dirk Mengedoht, and Hans-Jurgen Hirche. Imaging of plankton specimens with the lightframe on-sight keyspecies investigation (LOKI) system. *Journal of the European Optical Society: Rapid Publications*, 5:10017s, Apr. 2010.

[49] Jihoon Shin, Seonghyeon Yoon, YoungWoo Kim, Taeho Kim, ByeongGeon Go, and YoonKyung Cha. Effects of class imbalance on resampling and ensemble learning for improved prediction of cyanobacteria blooms. *Ecological Informatics*, 61:101202, Mar. 2021.

[50] Akila Somasundaram and U Srinivasulu Reddy. Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data. (978):8, 2016.

[51] Akila Somasundaram and U. Srinivasulu Reddy. *Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data*. Jan. 2016.

[52] Heidi M. Sosik, Emily E. Peacock, and Emily F. Brownlee. WHOI Plankton: Annotated Plankton Images - Data Set for Developing and Evaluating Classification Methods. Tech-



nical report, Woods Hole Oceanographic Institution, 2014. Type: dataset.

[53] K.-K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, Jan. 1998. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[54] Dirk Tasche. Fisher consistency for prior probability shift. *The Journal of Machine Learning Research*, 18(1):3338–3369, Jan. 2017.

[55] C. Wang, X. Zheng, C. Guo, Z. Yu, J. Yu, H. Zheng, and B. Zheng. Transferred Parallel Convolutional Neural Network for Large Imbalanced Plankton Database Classification. In *2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO)*, pages 1–5, May 2018.

[56] Lishu Wang, Bin Yang, Xiu-Ping Lin, Xue-Feng Zhou, and Yonghong Liu. Sesterterpenoids. *Natural Product Reports*, 30(3):455, 2013.

Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in International Conference on Computer Vision (ICCV) Workshops. Walker, Joseph; Orenstein, Eric 2021. The dissertation author was the primary investigator and author of this paper.

Chapter 3

IEEE Oceanic Engineering, (2023)

Underwater Object Detection Under Domain Shift

Abstract

There is increasing interest in using deep learning-based object recognition algorithms to perform automatic labeling of underwater imagery from marine surveys. However, underwater object detection is a particularly challenging problem due to changes in scattering and absorption of light, and spotty data collection efforts, which rarely capture the broad variability of the marine environment. Using deep learning-based object detection systems for long-term or multi-site marine surveying is further complicated by shifting data distributions between training and testing stages. Using data from the 100 Island Challenge, we investigate how object detection performance is impacted by changes in site characteristics and imaging conditions. We demonstrate that the combined use of data augmentation and unsupervised domain adaptation techniques can mitigate performance drops in the presence of domain shift. The proposed method is broadly applicable to observational datasets in marine and terrestrial environments where a single algorithm needs to adapt to and perform comparably across changing conditions.

1. Introduction

Optical imaging has remained an indispensable tool in oceanographic studies, as it offers detailed descriptions that are easily interpreted by humans. As a result, a myriad of systems have been developed for acquiring optical images in almost every oceanographic context. Autonomous underwater vehicles (AUVs) and unmanned underwater vehicles (UUVs) equipped with optical cameras have been used for the exploration and mapping of the seafloor [1], [2] monitoring invasive species [3], and fisheries management [4]. Imaging systems for in-situ studies of plankton and other marine particles have also been developed [5]–[8]. The rising popularity of these tools have led to an explosion in underwater optical data collection [9]. This increase in data has driven the need to develop object detection systems that can automate the analysis of underwater digital imagery.

Object detection is a computer vision task concerned with locating and classifying objects in images or videos. The most significant advancements in object detection can be attributed to the use of deep convolutional neural networks. Currently, one of the most popular architectures for object detection is the Faster R-CNN [10]. The Faster R-CNN consists of three modules: (1) a feature extractor convolutional neural network to extract features from the entire image, (2) a region proposal network which is trained end-to-end with the rest of the detection network to propose regions of interest in the feature map produced by (1), and (3) two fully-connected networks for classification and bounding box regression.

Underwater object detection is a particularly challenging problem as images are typically of lower quality compared to out-of-water images due to light scattering and absorption. The lack of precise control over the relative imaging depth and orientation to objects in underwater environments can produce high variability in their features. Despite these challenges, numerous applications of underwater object detection exist, ranging from the estimation of plankton and fish population densities [11], biodiversity monitoring of coral reefs [12], unexploded ordnance detection [13], and detection of other man-made objects [14], [15].

A fundamental challenge in incorporating deep learning technology in oceanography (and most other real-world applications) arises from the fact that models tend to overfit to the training data distribution. Differences in the training and testing dataset distributions, referred to as *dataset shift*, have been shown to contribute to diminished model performance [16]–[18]. Changes in the sampling location or methodology can produce dataset shift by altering image appearance (shifts in illumination, color, noise, etc.), background features, or statistical differences in object class features (e.g. new phenotypes or morphologies of species of interest). The specific term used to describe the statistical changes in input features is known as *domain shift* [18]. Other forms of dataset shift include *prior probability shift*, which is where the predicted variables prior probabilities differs between training and testing [19]–[22].

For applications of deep learning in oceanography, model deployment is almost always limited to the same study site and data collection protocol as the training data. However, oceanographic data are most often collected in multiple locations with varying environmental conditions, making application of a model built in a single context insufficient to achieve high performance across use cases. One solution for producing a more generalizable model is to annotate data from all environments in which the model is deployed. However, data annotation is extremely costly and serves as the primary bottleneck in incorporating deep learning in long-term or multi-site studies. It is therefore desirable to develop adaptive deep learning models that can scale to many study sites even when annotated data is limited to a single site.

Data augmentation is often used to artificially increase the volume of available training data [23]. This process involves defining a set of augmentation functions that alter the appearance of the training data while preserving the class label, effectively synthesizing new data examples from existing data. Commonly used augmentation functions include image flipping, cropping, translation, and noise addition. These augmentations are often treated as universal, as they are used across a range of image recognition applications. In some cases, it may be possible to design specialized augmentation functions that address known sources of variability to improve model generalizability. For example, if future data is expected to be collected using an imaging system with higher illumination intensity, then illumination synthesis could be used to simulate the difference between collected and future data. However, this approach may require *a priori* knowledge of the variability and well-defined augmentation functions that accurately model the variability.

The problem of domain shift has received a significant amount of attention in recent years primarily for the development of autonomous driving. In this context, collecting and annotating data from enough environments, weather conditions, and sensor configurations to ensure that future data is not outside the training distribution may be prohibitively costly or impossible. In practice, it may only be possible to collect labeled data from a single source domain, however, acquiring *unlabeled* data from the testing, i.e., target, domain may be more attainable. In such circumstances, it has been shown that by leveraging both labeled data from a source domain and unlabeled data from a target domain can help achieve better performance on data from the target domain [24]–[27]. This technique is referred to as unsupervised domain adaptation (UDA).

Prior works in underwater object recognition under domain shift have largely focused on using domain generalization techniques [28]–[30]. Unlike UDA, domain generalization methods aim to build models that can generalize well across multiple target domains without accessing unlabeled target domain data during training [31]. Domain generalization techniques are particularly beneficial in object detection applications in video streams or other scenarios where the data arrives continuously and needs to be processed in real-time. However, in cases where real-time annotation is not required and unlabeled target data is available during training, UDA has been shown to outperform domain generalization [32], [33]. We describe two commonly used classes of UDA methods in sections 1.1 and 1.2.

1.1. Adversarial feature learning

First introduced in the context of image classification, adversarial feature learning involves the use of a domain classifier to adversarially train the model to learn domain invariant features [25]. Source and target images are given a domain label of 0 and 1 respectively and the domain classifier is tasked with categorizing images from their respective domains. During backpropagation, the weights of the domain classifier are updated, then the gradients are pushed through a gradient reversal layer before being applied to the weights of the convolutional layers. This results in the learning of features that fool the domain classifier, i.e., are domain-independent.

1.2. Image-to-image translation

Another more intuitive approach to adaptation is to match the appearance of the source domain images to that of the target domain (or vice versa). Image-to-image translation reduces the domain discrepancy in the pixel domain, which has the advantage of utilizing human visual inspection for quality assessment. Many contemporary image-to-image translation techniques borrow directly from, or use ideas similar to, the CycleGAN models [34].

1.3. Contributions

The goal of this study is to train a Faster R-CNN model using labeled data from one source environment that can scale to many target environments. To do this, we use a combination of data augmentation and UDA techniques to minimize domain shift between environments. We summarize our primary contributions as follows:

- We present a new underwater object detection dataset for domain adaptation experimentation.
- We present a framework for developing robust underwater object detectors that are more resilient to dataset shift.
- For our task, we show that existing state-of-the-art UDA techniques can be improved by incorporating data augmentation.
- In the case of limited training data (as in our case), we show that the HM-MVGD-HM [35] color-matching algorithm can produce better image-to-image translation results than more sophisticated methods such as CycleGAN.
- Our dataset and code are publicly available and will be used to augment data collection for the 100 Island Challenge project (described in section 2.1) and other projects requiring robust underwater object detection.

2. Material and methods

2.1. Dataset

The 100 Island Challenge (100IC) is an ongoing collaborative effort based at Scripps Institution of Oceanography, UC San Diego, to digitally archive and monitor coral reefs across the globe. Using tools of large-area imaging, high resolution images have been collected and collated to form comprehensive digital mosaics and three-dimensional reconstructions from multiple coral reef sites at each of over 100 islands across the globe. These detailed maps enable the study of benthic dynamics at an unprecedented spatial scale. The 100IC has incorporated the use of Smart Underwater Imaging Telemeters (SUITs) to facilitate in-situ environmental data collection to complement visual surveys of coral reefs [36]. The 100IC has produced a unique dataset of this standardized object (the SUIT) that has been imaged across multiple study sites and imaging conditions. This dataset, consisting of a single annotated class, is therefore particularly well suited for the study of binary underwater object detection under domain shift.

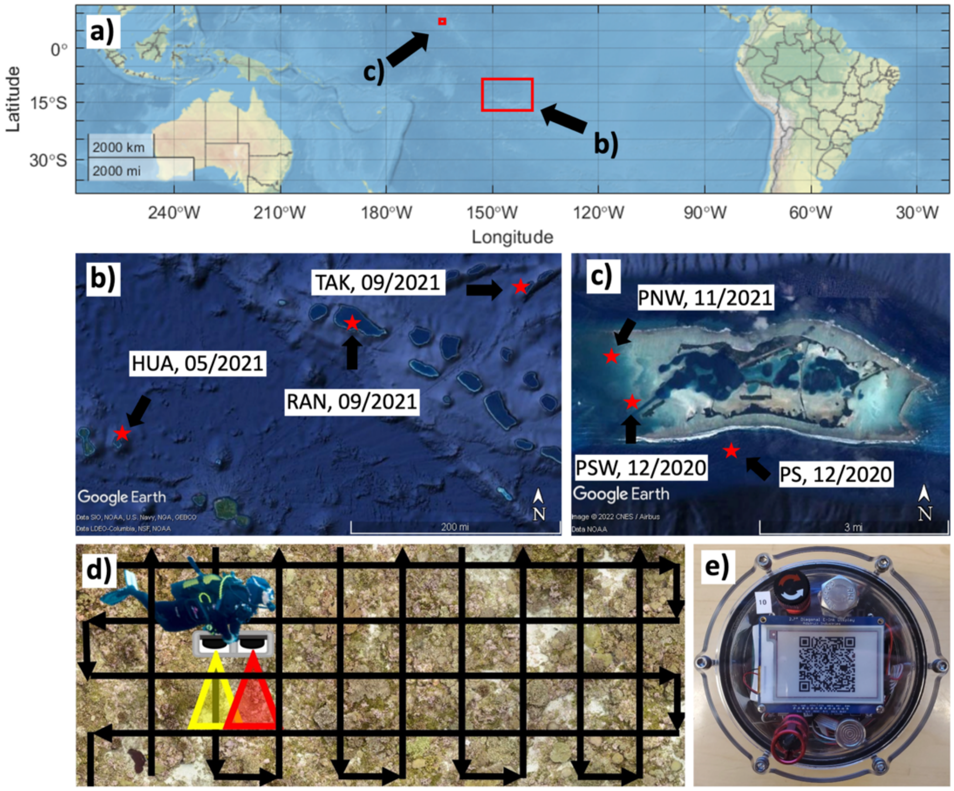


Fig. 1. Sampling locations, dates, methodology, and a SUIT. a) Data was collected from two regions (red boxes) in the tropical Pacific Ocean. b) Study sites from the Tuamotu archipelago region, which include the islands Takapoto (TAK), Rangiroa (RAN), and Huahine (HUA). Sampling date for each study site is reported as MM/YYYY. c) Study sites from the Palmyra Atoll, which include sites in the southern (PS), southwest (PSW), and northwest (PNW) parts of the island. d) For each study site, a survey plot ($100m^2$ or $200m^2$) is defined and is imaged by divers in a grid pattern. e) Photo of the SUIT.

In this study, we consider the subset of 100IC image data that was collected from two regions in the tropical Pacific Ocean. The data come from the islands Huahine (HUA), Takapoto (TAK) and Rangiroa (RAN) in the Tuamotu archipelago region as well as three sites around the Palmyra Atoll, which include sites in the southern (PS), southwest (PSW), and northwest (PNW) parts of the island. The location and sampling date of these sites as well as an illustration of the data collection procedure and SUIT are shown in Fig. 1. Example images from each of the six sites are shown in Fig. 2. The number of bounding box annotations per study site and bounding box size distribution statistics are listed in Table 1. We identify two types of variability across the images which make detection of the SUITs challenging:

2.2. Variability in low-level features.

In the context of image processing, low-level visual information may include brightness, texture, color distribution and noise. In the 100IC imagery, these features can change according to various physical phenomena that influence the image formation process. Due to the large spatiotemporal range of the sampling, it is likely that the inherent optical properties of the seawater are inconsistent across sampling periods. This can lead to different degrees of color distortion and contrast loss. The ambient light field is also subject to change according to weather conditions leading to inconsistent scene illumination. Caustic patterns on the seafloor, especially visible in Fig. 2d, can create bright white regions that are similar to the white pixels of the SUIT display. Lastly, some images have been color corrected while others have not. In cases where color correction is applied, broad assumptions (such as constant scene depth) are made. Therefore, there is high variability among the color corrected images including noticeable depth-related artifacts.

We provide a quantitative measure for low-level feature similarity between the environments by comparing the image features extracted from a VGG16 [37] encoder. For each of the eight study sites, a centroid is calculated by averaging the extracted image features. The pairwise distances between the centroids are used to calculate similarity between the study sites using a cosine similarity measure. The pairwise similarity values are reported in Table 2.

2.3. Variability in SUIT scale.

Another source of variability is created by changes in structural features associated with the SUITs themselves, caused by changes in scale and orientation. The data used in this study was collected using a Nikon D780 or Nikon D7000 camera used in combination with a 24mm and 18mm wide-angle lens respectively. This affects the apparent size of the SUITs in the images. Variability in the distance between the camera and the seafloor can make the SUITs appear to be differently sized. Topography is also highly variable across the environments. In environments with highly textured benthic surfaces, the SUITs are more likely to be imaged at an angle. For each of the eight sampling locations, we calculate the average bounding box size and compute the magnitude of the pairwise differences between each of the averages. We divide these differences by the largest difference to scale the values to be between zero and one, and then subtract each of the values from one to calculate the similarity. All pairwise similarity scores are reported in Table 3.

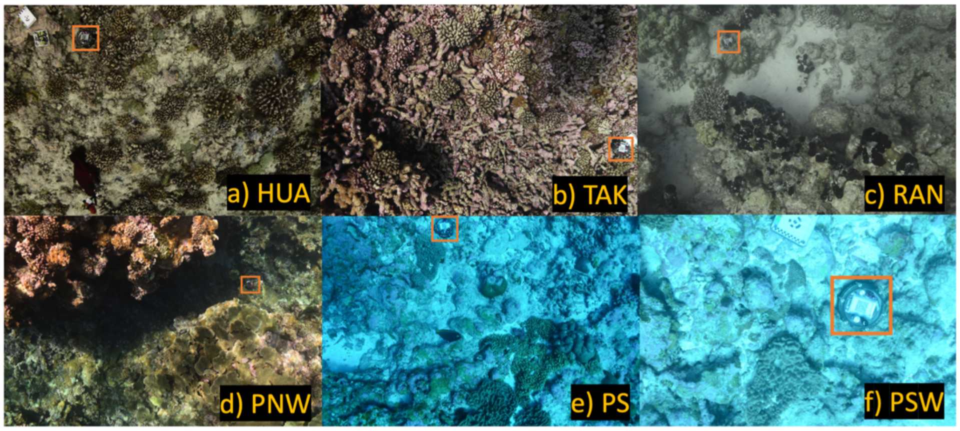


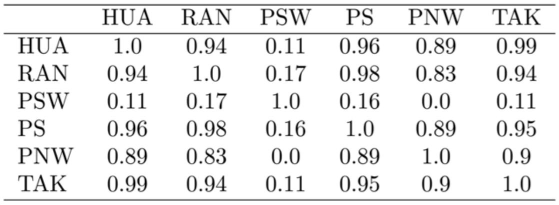
Fig. 2. Example images collected from the six study sites. The bounding boxes containing the SUITs are shown in orange.

Site name	Abbreviation	Num. boxes	Avg. box area	Std. Dev. box area
Huahine island, French Polynesia	HUA	351	0.8%	0.7%
Rangiroa atoll, French Polynesia	RAN	153	1.0%	0.5%
Southwest Palmyra atoll, USA	PSW	239	1.0%	1.0%
South Palmyra atoll, USA	PS	81	4.8%	3.0%
Northwest Palmyra atoll, USA	PNW	450	0.2%	0.1%
Takapoto atoll, French Polynesia	TAK	835	0.7%	0.4%

Table 1. Dataset statistics for each study site. Average and standard deviation of bounding box sizes are reported and expressed in terms of fraction of total image area, where all images are 500x751 pixels.

	HUA	RAN	PSW	PS	PNW	TAK
HUA	1.0	0.94	0.88	0.88	0.95	0.96
RAN	0.94	1.0	0.9	0.84	0.9	0.91
PSW	0.88	0.9	1.0	0.94	0.88	0.85
PS	0.88	0.84	0.94	1.0	0.86	0.85
PNW	0.95	0.9	0.88	0.86	1.0	0.88
TAK	0.96	0.91	0.85	0.85	0.88	1.0

Table 2. Image feature similarity between study sites. Similarity scores have been rescaled linearly between [0,1] where 1 indicates mean size is identical, and 0 represents the largest observed dissimilarity.



	HUA	RAN	PSW	PS	PNW	TAK
HUA	1.0	0.94	0.11	0.96	0.89	0.99
RAN	0.94	1.0	0.17	0.98	0.83	0.94
PSW	0.11	0.17	1.0	0.16	0.0	0.11
PS	0.96	0.98	0.16	1.0	0.89	0.95
PNW	0.89	0.83	0.0	0.89	1.0	0.9
TAK	0.99	0.94	0.11	0.95	0.9	1.0

Table 3. SUIT size similarity between study sites. similarity scores have been rescaled linearly between [0,1] where 1 indicates mean size is identical, and 0 represents the largest observed dissimilarity.

2.4. Progressive domain adaptation

We use UDA techniques to mitigate performance drops caused by image feature differences. Specifically, we use the progressive domain adaptation (PDA) method proposed by Hsu *et al.* [27]. PDA involves a two-stage procedure for aligning the features from the source and target domains. First, a synthetic image dataset is generated by mapping the source images to the target domain using a CycleGAN. In the first stage, the features of the source and synthetic domains are aligned using adversarial feature learning. In the second stage, the features of the synthetic and target domains are aligned using adversarial feature learning.

The original PDA method uses a CycleGAN to produce the synthetic dataset [38]. However, CycleGANs typically require large amounts of training data to produce quality image mappings. For the experiments in Hsu *et al.* [27], the authors used between 3,475 to 41,986 training examples to train the CycleGAN models. This amount of data is often not available for many oceanographic applications where data collection and annotation is difficult. For this study, only 81 to 835 examples were collected from each study site. In our initial experiments using CycleGAN, we found that almost all the translated images contained significant distortions and failed to preserve features of the SUITs. For this reason, we replaced the CycleGAN with the HM-MVGD-HM color-matching algorithm [35]. The HM-MVGD-HM algorithm uses an analytical solution to a Multivariate Gaussian Distribution (MVGD) color transfer equation in addition to classical histogram matching. Example synthetic images using CycleGAN and HM-MVGD-HM are shown in Fig. 6. We also compare object detection performance using CycleGAN and HM-MVGD-HM in Table 4.

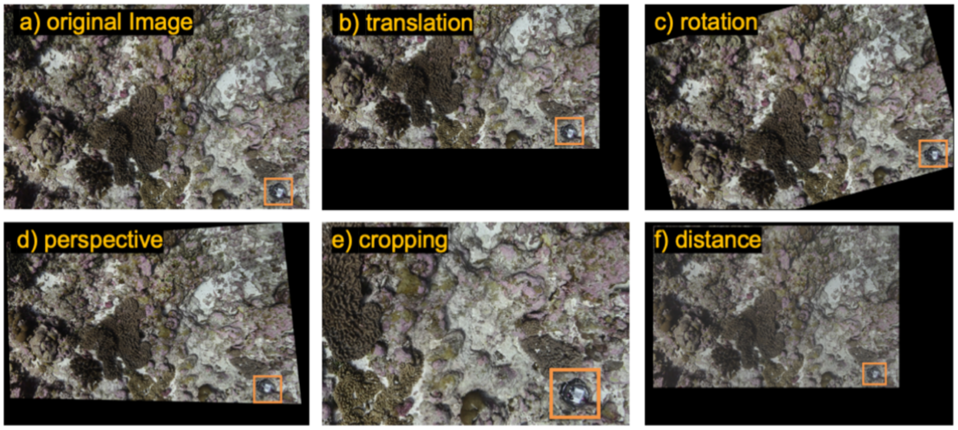


Fig. 3. Data augmentations. a) An example image from the TAK study site. Images b)-f) show the output of the translation, rotation, perspective transformation, cropping, and distance image augmentation functions respectively using the image in a) as input.

2.5. Augmentation functions

To address variability in SUIT object size, we implement five data augmentation techniques, each designed to simulate a potential source of variation (Fig. 3). The (x,y) pixel coordinate of the SUIT center in an image is arbitrary, and is determined only by the SUIT's placement relative to the transect during image collection. To prevent the models from learning irrelevant patterns related to the position of the SUITs, we simulate different SUIT placements by applying random image translations and rotations. We define this set of placement transformations as $T_P = \{\text{translation, rotation}\}$. The distance between the camera and the bottom will affect the apparent size of the SUIT. Simulating imaging at closer range can be approximated by using random cropping. However, imaging at greater distances involves simulating the effects of resolution and contrast loss. To simulate these effects, we created an augmentation function which performs downsampling followed by contrast reduction. The subsequent image is then zero-padded to the original image size. Because symmetrically padding the image would bias SUIT placement towards the image center, padding is followed by a random translation. We refer to this augmentation as *distance*. To simulate different imaging angles, we adopt an approach similar to Huang *et al.* [39] by applying perspective transformations to the images. Because perspective transformation, cropping, and distance augmentations can distort the apparent size of the SUIT, we refer to this set of transformations as $T_S = \{\text{perspective, cropping, distance}\}$. Fig. 4 shows examples of all five augmentations.

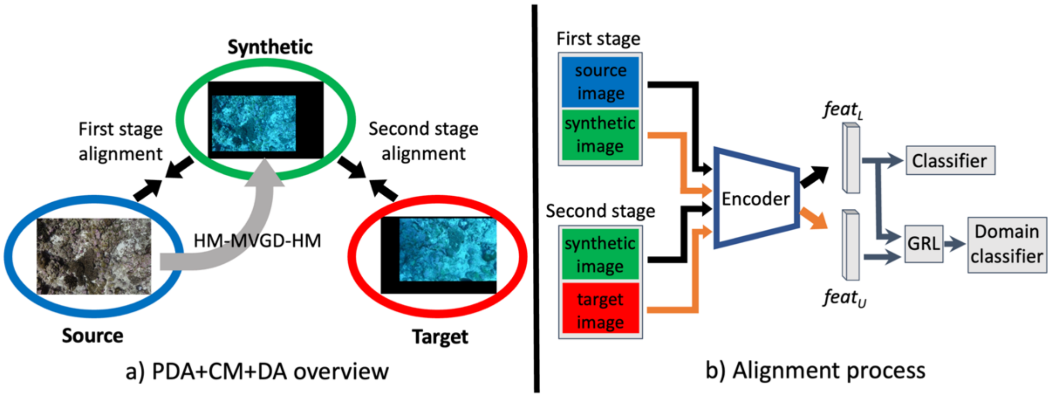


Fig. 4. a) Overview of our augmented version of the progressive domain adaptation method, using the HM-MVGD-HM color-matching algorithm and data augmentation (*PDA+CM+DA*). A source image (blue oval) and target image is drawn from the source and target study sites respectively. A synthetic image is generated by color-matching the source image to the target image using the HM-MVGD-HM algorithm. The synthetic and target images are then augmented, producing the images seen in the green and red ovals respectively. Black arrows represent the feature alignment steps. b) Illustration of the adversarial feature alignment process. In the first stage of training, features are extracted from the labeled source image and unlabeled synthetic image, denoted as *feat*$_{\mathrm{L}}$ and *feat*$_{\mathrm{U}}$ respectively. Supervised object detection is performed using only *feat*$_{\mathrm{L}}$. Adversarial feature learning is performed by passing both *feat*$_{\mathrm{L}}$ and *feat*$_{\mathrm{U}}$ to the domain classifier, whose gradients are reversed during backpropagation when passed through the gradient reversal layer (GRL). In second stage training, features are extracted from the labeled synthetic image and unlabeled target image, where labels for the synthetic image are inherited from the source image.

2.6. Models

All models use a Faster R-CNN architecture with a VGG16 [37] backbone. We consider five different models:

1) *baseline*: Faster R-CNN trained without PDA or data augmentation. Models are trained on data from a single study site and applied directly to a target site.
2) *DA*: Same as *baseline*, but trained using data augmentation.
3) *PDA:* Faster R-CNN trained according to the PDA method using the HM-MVGD-HM color-matching algorithm.
4) *PDA+DA:* Faster R-CNN trained according to the PDA method using the HM-MVGD-HM color-matching algorithm and data augmentation.
5) *PDA+CGAN+DA:* Faster R-CNN trained according to the PDA method using a CycleGAN (CGAN) and data augmentation.

Note that all models trained using the PDA method (*PDA* and *PDA+DA*) use the HM-MVGD-HM color-matching algorithm to perform the image-to-image translation step. Only the *PDA+CGAN+DA* model uses the CycleGAN for image-to-image translation.

2.7. Experimental Setup

All models are trained and tested according to a *leave-one-in cross validation* approach - models are trained on labeled data from one source study site and all other study sites are individually treated as the target domain. All study sites with at least 400 images are used as source and target environments. Study sites with fewer examples are used as target domains only. For models using data augmentation, one augmentation is selected randomly from both T_P and T_S (Defined in Section 2.5). The transformations are applied only during training to both source and target images. During testing, no augmentations are applied. A batch size of one is used during training and the images are resized to 500x751. All other hyperparameter values for models using PDA are the same as Hsu *et al.* [27]. All experiments were run using a Tesla P100 GPU and Intel Xeon 6126 CPU.

3. Results

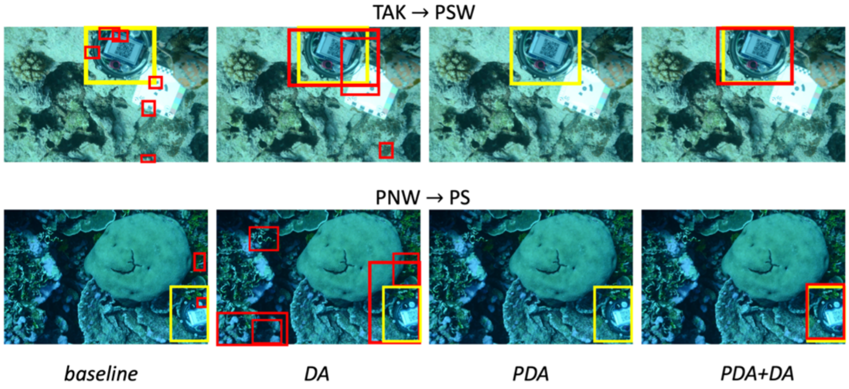


Fig. 5. Region proposals (red) from four different models for an example image from the target study site of two adaptation experiments: a) Using TAK as source and PSW as target, b) Using PNW as source and PS as target. Ground truth bounding boxes are shown in yellow.

Model	Adaptation (source → target)					
	TAK → HUA	TAK → RAN	TAK → PSW	TAK → PS	TAK → PNW	TAK → TAK
baseline	90.0	90.5	53.7	73.5	70.9	-
DA	86.2	88.5	90.8	62.4	76.6	-
PDA	**90.9**	90.7	73.5	90.6	80.9	-
PDA+CGAN+DA	89.7	89.2	**99.7**	89.2	80.7	-
PDA+DA	90.7	**91.0**	98.9	**90.9**	**86.5**	-
	PNW → HUA	PNW → RAN	PNW → PSW	PNW → PS	PNW → PNW	PNW → TAK
baseline	72.6	24.6	1.1	1.2	-	47.0
DA	88.0	90.1	66.7	40.0	-	90.4
PDA	79.9	66.3	0.3	69.5	-	69.6
PDA+CGAN+DA	**89.9**	88.7	74.3	**90.4**	-	86.7
PDA+DA	**89.9**	**90.9**	**80.2**	89.6	-	**90.5**

Table 4. Leave-one-in cross validation results. Each of the five models is trained using a single source study site (X) where all other study sites are individually treated as a target domain (Y). The adaptation of a source study site to a target study site is expressed as X→Y. Two source study sites, TAK and PNW, are considered individually. Results are reported in terms of classification accuracy (higher is better). Best performing model for each adaptation is shown in bold.

Table 4 shows that in all adaptation scenarios, *PDA+DA* trained using either a CycleGAN or the HM-MVGD-HM algorithm outperformed or performed very similarly to the best performing model. Performance of *baseline* varied significantly across the adaptation experiments. Table 2 and Table 3 indicate that the TAK→HUA and TAK→RAN adaptations have relatively high image feature and SUIT size similarity. For both adaptations, *baseline* performed comparatively well with the other three models providing little improvement. For adaptation instances with relatively low image feature similarity but similar SUIT size similarity, which includes TAK→PS, TAK→PNW and PNW→PS, *PDA* outperformed *DA* and *baseline*. This supports the hypothesis that UDA techniques are most effective for bridging differences in low-level features.

For adaptation instances with low SUIT size similarity, including every scenario in which study site PNW is used as source, *DA* outperformed *PDA* except in the case of PNW→PS. We note that this case also exhibits low image feature similarity and that *DA* still brought significant improvements compared to *baseline*. As is shown in Fig. 5, *baseline* is restricted to predicting regions that are of a similar size to the bounding box annotations of the source dataset. The added augmentation functions allow the model to consider a greater range of bounding box predictions. This supports the hypothesis that data augmentation techniques may be more effective for bridging apparent structural differences in the objects of interest.

The PNW→PS and PNW→PSW adaptations are assumed to be the most difficult as they exhibit low image feature and SUIT size similarity. This difficulty is evident by the very low performance of *baseline* in both cases. Despite the large shift in data distributions, the best performing model was able to improve performance on the target domains dramatically.

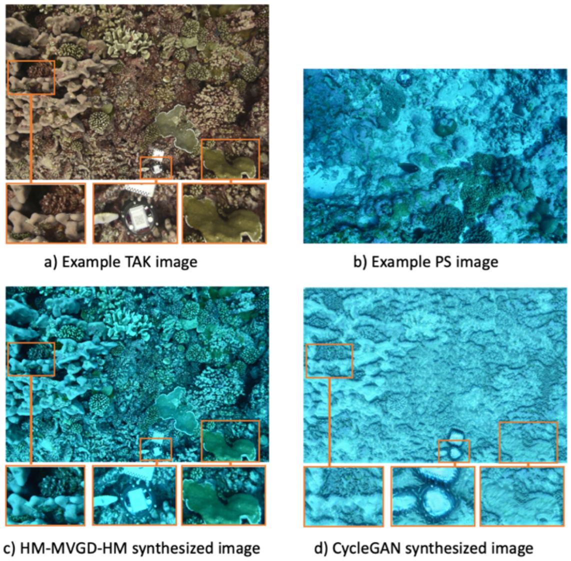

a) Example TAK image
b) Example PS image
c) HM-MVGD-HM synthesized image
d) CycleGAN synthesized image

Fig. 6. Qualitative comparison of image-to-image translation methods using images from TAK and PS as source and target respectively. a) A random image drawn from TAK to be translated to PS. b) An example image from PS. c) The TAK image is color matched to the target image using the HM-MVGD-HM algorithm. d) The TAK image is translated to the PS environment using a CycleGAN model.

Transforming images from the source domain to the target domain using the HM-MVGD-HM algorithm requires no additional training. Generating all 10 synthetic datasets for the 10 adaptation experiments took approximately 40 minutes or approximately 0.5 seconds per image. Generating the synthetic images using a CycleGAN required significantly more memory allocation and increased training time. CycleGAN training took about 5-6 minutes per epoch or 16.6-20 hours in total for each source/target pair. As seen in Table 4, the trainable CycleGAN generally did not provide improved performance compared to the HM-MVGD-HM algorithm. We believe that the relatively small dataset used in this study was insufficient for training a CycleGAN but similar works to this study with larger available datasets may be able to benefit from the flexibility of a trainable image-to-image translation model.

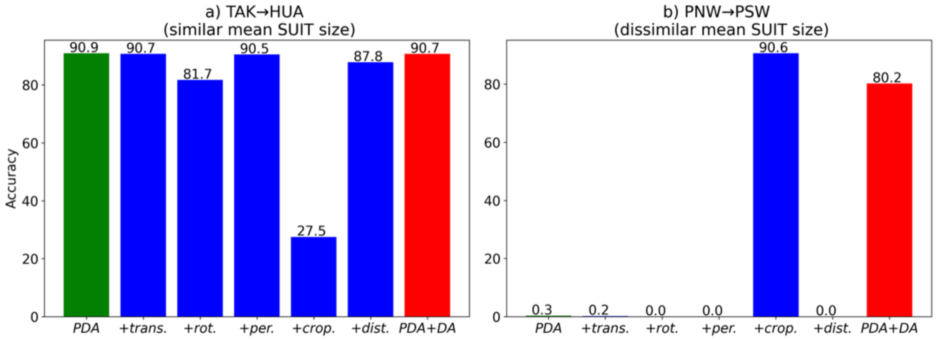


Fig. 7. Ablation on the five augmentation functions: translation (trans.), rotation (rot.), perspective (per.), cropping (crop.), and distance (dist.). a) Ablation results using TAK as source and HUA as target. b) Ablation results using PNW as source and PSW as target. Performance of *PDA*, which uses no data augmentation, is shown in green. Performance of *PDA+DA*, which uses all five augmentations, is shown in red. The ablation applies each one of the five augmentations individually together with *PDA*.

We performed ablation on the five data augmentation functions for adaptations TAK→HUA and PNW→PSW and show the results in Fig. 7. These two adaptations were selected for ablation due to their representation of extreme cases, where SUIT size similarity is either very small or very large (Table 3). The results of the ablation reveal that the performance contribution of each of the five augmentation functions is strongly dependent on the variability in apparent size. In cases where there is little to no difference in apparent object size between the source and target study sites, the incorporation of any amount of data augmentation can negatively impact performance (see Fig. 7a). However, if the difference in apparent object size is large, the choice of augmentation functions can have a substantial impact on performance (see Fig. 7b).

The results suggest that in the case where it is known *a priori* that the target domain objects will appear much larger/smaller, then the best results may be achieved by limiting the set of augmentations to a set of function(s) that exclusively model this difference. Both Fig. 7a and Fig. 7b suggest that incorporating augmentation functions that do not directly relate to the sources of variability may negatively impact performance. However, we note that the cases studied in the ablation represent the extreme cases, and that in the absence of *a priori* knowledge of the object variability, *PDA+DA* (trained using all augmentations) still performs the best on average and therefore we conclude that using the entire set of augmentations is a strong default choice.

4. Discussion

Developing generalizable object detection models is complicated by shifts in data distribution. We have shown that domain shift can greatly impact detection performance. However, we demonstrate that by combining data augmentation with existing UDA techniques, performance drops can be significantly reduced. This is a significant finding, as the results provide the possibility for broad spatiotemporal surveying even when annotated data is limited to one study site. We further show that models can be trained to be robust against other sources of variability including color correction and object scale.

Overall, the results of the cross-validation experiments are intuitive - source and target study sites with low visual differences produced higher baseline performance and, in these cases, more sophisticated models produced marginal improvements. However, in many cases, study sites with significant visual differences benefited tremendously from the combined use of UDA and data augmentation. An alternative approach to bridging differences in data distributions could involve the use of light attenuation models that are specific to each environment. However, this would require accurate measurement and prediction of the ambient light field and inherent optical properties of the water column. This approach is likely most appropriate when target domain image data is unavailable during training, which may include real-time detection tasks. In these cases, *a priori* knowledge of future imaging conditions should be leveraged to synthetically generate the training dataset. If target data is available during model training, we propose that the main advantage of using the techniques developed in this study is that they require no prior knowledge of light field or water column properties. Instead, environment agnostic features can be learned through the combined use of image translation and adversarial feature learning. This data-first approach also has the advantage of scaling to many sources of visual variability beyond water column properties which can be difficult to model, including different cameras or lenses, or illumination patterns (e.g. shadows or caustics).

There is growing interest in using video to conduct oceanographic surveys [41], [42], however, directly applying still image-based object detection models to video presents unique challenges. These challenges include increased computational costs as well as motion blur and video defocus. In addition, the methods outlined in this study assume that target domain data is available during training, however, applications of real-time object detection and live target searching may be incompatible with this assumption. In these cases, the models must be able to scale to multiple target domains using source data alone. We conclude that domain randomization techniques remain as the best possible solution when target domain data is unavailable [43]–[45].

In many real-world applications of machine learning, including in oceanography, the available annotated data is insufficient for training models with large parameter spaces, and could result in overfitting [40]. Fig. 6 shows that CycleGAN produced synthetic images of lower visual quality compared to the HM-MVGD-HM color-matching algorithm. This was likely due to a relatively low training dataset size. We conclude that in data limited cases, PDA may be improved by replacing the CycleGAN with HM-MVGD-HM or a similar algorithm with few trainable parameters.

5. Conclusion

Two major present-day obstacles hindering advances in the analysis of oceanographic data include (1) challenges in developing analysis tools that are robust across different conditions, equipment, and locations; and (2) costs associated with trying to collect and annotate variable datasets from which effective models can be trained. Our results indicate that the procedures developed in this study may be a viable solution for improving model robustness while reducing the human data annotation effort. We view this as a critical step for maximizing utility and cost effectiveness of oceanographic field campaigns.

Data availability

All data used in this study are available on the project's GitHub repository (https://github.com/JosephLWalker96/underwater-object-detection).

Acknowledgements

This work was supported by the Office of Naval Research and Task Force Ocean (Robert Headrick) [grant number N00014-19-1-2851]. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

[1] O. Pizarro and H. Singh, "Toward large-area mosaicing for underwater scientific applications," *IEEE Journal of Oceanic Engineering*, vol. 28, no. 4, pp. 651–672, Oct. 2003, doi: 10.1109/JOE.2003.819154.

[2] N. R. Gracias, S. van der Zwaan, A. Bernardino, and J. Santos-Victor, "Mosaic-based navigation for autonomous underwater vehicles," *IEEE Journal of Oceanic Engineering*, vol. 28, no. 4, pp. 609–624, Oct. 2003, doi: 10.1109/JOE.2003.819156.

[3] N. Barrett, J. Seiler, T. Anderson, S. Williams, S. Nichol, and S. Nicole Hill, "Autonomous Underwater Vehicle (AUV) for mapping marine biodiversity in coastal and shelf waters: Implications for marine management," in *OCEANS'10 IEEE SYDNEY*, May 2010, pp. 1–6. doi: 10.1109/OCEANSSYD.2010.5603860.

[4] D. A. Smale *et al.*, "Regional-scale benthic monitoring for ecosystem-based fisheries management (EBFM) using an autonomous underwater vehicle (AUV)," *ICES Journal of Marine Science*, vol. 69, no. 6, pp. 1108–1118, Jul. 2012, doi: 10.1093/icesjms/fss082.

[5] E. C. Orenstein *et al.*, "The Scripps Plankton Camera system: A framework and platform for in situ microscopy," *Limnology and Oceanography: Methods*, vol. 18, no. 11, pp. 681–695, 2020, doi: 10.1002/lom3.10394.

[6] R. J. Olson and H. M. Sosik, "A submersible imaging-in-flow instrument to analyze nano- and microplankton: Imaging FlowCytobot," *Limnology and Oceanography: Methods*, vol. 5, no. 6, pp. 195–203, 2007, doi: https://doi.org/10.4319/lom.2007.5.195.

[7] R. K. Cowen and C. M. Guigand, "In situ ichthyoplankton imaging system (ISIIS): system design and preliminary results," *Limnology and Oceanography: Methods*, vol. 6, no. 2, pp. 126–132, 2008, doi: https://doi.org/10.4319/lom.2008.6.126.

[8] R. W. Campbell, P. L. Roberts, and J. Jaffe, "The Prince William Sound Plankton Camera: a profiling in situ observatory of plankton and particulates," *ICES Journal of Marine Science*, vol. 77, no. 4, pp. 1440–1455, Jul. 2020, doi: 10.1093/icesjms/fsaa029.

[9] H. Lu, Y. Li, Y. Zhang, M. Chen, S. Serikawa, and H. Kim, "Underwater Optical Image Processing: a Comprehensive Review," *Mobile Netw Appl*, vol. 22, no. 6, pp. 1204–1211, Dec. 2017, doi: 10.1007/s11036-017-0863-4.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *arXiv:1506.01497 [cs]*, Jan. 2016, Accessed: Jul. 04, 2021. [Online]. Available: http://arxiv.org/abs/1506.01497

[11] X. Li, M. Shang, H. Qin, and L. Chen, "Fast accurate fish detection and recognition of underwater images with Fast R-CNN," in *OCEANS 2015 - MTS/IEEE Washington*, Oct. 2015, pp. 1–5. doi: 10.23919/OCEANS.2015.7404464.

[12] S. M. Jaisakthi, P. Mirunalini, and C. Aravindan, "Coral Reef Annotation and Localization using Faster R-CNN," in *CLEF*, 2019.

[13] M. Czub *et al.*, "Deep sea habitats in the chemical warfare dumping areas of the Baltic Sea," *Science of The Total Environment*, vol. 616–617, pp. 1485–1497, Mar. 2018, doi: 10.1016/j.scitotenv.2017.10.165.

[14] D. L. Rizzini, F. Kallasi, F. Oleari, and S. Caselli, "Investigation of Vision-Based Underwater Object Detection with Multiple Datasets," *International Journal of Advanced Robotic Systems*, vol. 12, no. 6, p. 77, Jun. 2015, doi: 10.5772/60526.

[15] A. Olmos, E. Trucco, and D. Lane, "Automatic man-made object detection with intensity cameras," in *OCEANS '02 MTS/IEEE*, Oct. 2002, pp. 1555–1561 vol.3. doi: 10.1109/OCEANS.2002.1191867.

[16] C. M. Bishop, *Pattern recognition and machine learning*. in Information science and statistics. New York: Springer, 2006.

[17] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains. In: ECCV." 2010.

[18] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*. MIT Press, 2008.

[19] G. Forman, “Quantifying counts and costs via classification,” *Data Min Knowl Disc*, vol. 17, no. 2, pp. 164–206, Oct. 2008, doi: 10.1007/s10618-008-0097-y.

[20] E. C. Orenstein, K. M. Kenitz, P. L. D. Roberts, P. J. S. Franks, J. S. Jaffe, and A. D. Barton, “Semi- and fully supervised quantification techniques to improve population estimates from machine classifiers,” *Limnology and Oceanography: Methods*, vol. 18, no. 12, pp. 739–753, 2020, doi: https://doi.org/10.1002/lom3.10399.

[21] O. Beijbom *et al.*, “Quantification in-the-wild: data-sets and baselines,” *arXiv:1510.04811 [cs]*, Nov. 2015, Accessed: Mar. 10, 2020. [Online]. Available: http://arxiv.org/abs/1510.04811

[22] P. González, A. Castaño, E. E. Peacock, J. Díez, J. J. Del Coz, and H. M. Sosik, “Automatic plankton quantification using deep features,” *Journal of Plankton Research*, vol. 41, no. 4, pp. 449–463, Jul. 2019, doi: 10.1093/plankt/fbz023.

[23] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *Journal of Big Data*, vol. 6, no. 1, p. 60, Jul. 2019, doi: 10.1186/s40537-019-0197-0.

[24] G. Wilson and D. J. Cook, “A Survey of Unsupervised Deep Domain Adaptation,” *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 5, p. 51:1-51:46, Jul. 2020, doi: 10.1145/3400066.

[25] Y. Ganin and V. Lempitsky, “Unsupervised Domain Adaptation by Backpropagation,” in *Proceedings of the 32nd International Conference on Machine Learning*, PMLR, Jun. 2015, pp. 1180–1189. Accessed: Aug. 05, 2022. [Online]. Available: https://proceedings.mlr.press/v37/ganin15.html

[26] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, “Maximum Classifier Discrepancy for Unsupervised Domain Adaptation,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 3723–3732. doi: 10.1109/CVPR.2018.00392.

[27] H.-K. Hsu *et al.*, “Progressive Domain Adaptation for Object Detection,” presented at the The IEEE Winter Conference on Applications of Computer Vision, Mar. 2020, pp. 738–746. doi: 10.1109/WACV45572.2020.9093358.

[28] H. Liu, P. Song, and R. Ding, “WQT and DG-YOLO: towards domain generalization in underwater object detection.” arXiv, Apr. 14, 2020. Accessed: Jun. 14, 2023. [Online]. Available: http://arxiv.org/abs/2004.06333

[29] H. Liu, P. Song, and R. Ding, “Towards Domain Generalization In Underwater Object Detection,” in *2020 IEEE International Conference on Image Processing (ICIP)*, Oct. 2020, pp. 1971–1975. doi: 10.1109/ICIP40778.2020.9191364.

[30] Y. Chen *et al.*, “Achieving Domain Generalization in Underwater Object Detection by Domain Mixup and Contrastive Learning,” Apr. 2021. doi: 10.48550/arXiv.2104.02230.

[31] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, “Domain Generalization: A Survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–20, 2022, doi: 10.1109/TPAMI.2022.3195549.

[32] F. J. Piva, D. De Geus, and G. Dubbelman, “Empirical Generalization Study: Unsupervised Domain Adaptation vs. Domain Generalization Methods for Semantic Segmentation in the Wild,” in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA: IEEE, Jan. 2023, pp. 499–508. doi: 10.1109/WACV56688.2023.00057.

[33] M. Wang and W. Deng, “Deep visual domain adaptation: A survey,” *Neurocomputing*, vol. 312, pp. 135–153, Oct. 2018, doi: 10.1016/j.neucom.2018.05.083.

[34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” *arXiv:1703.10593 [cs]*, Aug. 2020, Accessed: Mar. 23, 2021. [Online]. Available: http://arxiv.org/abs/1703.10593

[35] C. Hahne and A. Aggoun, “PlenoptiCam v1.0: A light-field imaging framework,” *IEEE Trans. on Image Process.*, vol. 30, pp. 6757–6771, 2021, doi: 10.1109/TIP.2021.3095671.

[36] D. Ratelle, S. Sandin, B. Zgliczynski, C. Edwards, and J. Jaffe, *7630 -Design and Prototyping of the Smart Underwater Imaging Telemeter (SUIT) for Embedding Environmental Data in Imaging Surveys of Benthic Communities*. 2022.

[37] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition.” arXiv, Apr. 10, 2015. Accessed: Aug. 02, 2022. [Online]. Available: http://arxiv.org/abs/1409.1556

[39] H. Huang, H. Zhou, X. Yang, L. Zhang, L. Qi, and A.-Y. Zang, “Faster R-CNN for marine organisms detection and recognition using data augmentation,” *Neurocomputing*, vol. 337, pp. 372–384, Apr. 2019, doi: 10.1016/j.neucom.2019.01.084.

[40] Z. Li, K. Kamnitsas, and B. Glocker, “Overfitting of Neural Nets Under Class Imbalance: Analysis and Improvements for Segmentation,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 402–410. doi: 10.1007/978-3-030-32248-9_45.

[41] J. Giddens, A. Turchik, W. Goodell, M. Rodriguez, and D. Delaney, “The National Geographic Society Deep-Sea Camera System: A Low-Cost Remote Video Survey Instrument to Advance Biodiversity Observation in the Deep Ocean,” *Front. Mar. Sci.*, vol. 7, p. 601411, Jan. 2021, doi: 10.3389/fmars.2020.601411.

[42] A. M. Friedlander *et al.*, “Marine biodiversity from zero to a thousand meters at Clipperton Atoll (Île de La Passion), Tropical Eastern Pacific,” *PeerJ*, vol. 7, p. e7279, Jul. 2019, doi: 10.7717/peerj.7279.

[43] J. Huang, D. Guan, A. Xiao, and S. Lu, “FSDR: Frequency Space Domain Randomization for Domain Generalization,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, Jun. 2021, pp. 6887–6898. doi: 10.1109/CVPR46437.2021.00682.

[44] S. Zakharov, W. Kehl, and S. Ilic, “DeceptionNet: Network-Driven Domain Randomization,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South): IEEE, Oct. 2019, pp. 532–541. doi: 10.1109/ICCV.2019.00062.

[45] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong, “Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization Without Accessing Target Domain Data,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South): IEEE, Oct. 2019, pp. 2100–2110. doi: 10.1109/ICCV.2019.00219.

Acknowledgements

Chapter 3, in part, has been submitted for publication of the material as it may appear in IEEE Oceanic Engineering 2023, Walker, Joseph; Zeng, Zheng; Chengchen, Wu; Jaffe, Jules; Frasier, Kaitlin; Sandin, Stuart. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Journal of the Acoustical Society of America. 2023.

Underwater sound speed profile estimation from vessel traffic recordings and multi-view neural networks

Joseph Walker[1], Zheng Zeng[2], Vanessa ZoBell[1], Kaitlin Frasier[1]

[1]*Marine Physical Laboratory, Scripps Institution of Oceanography, La Jolla, California, 92093-0238, United States*

[2]*Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA, 92093-0238, United States*

jlwalker@ucsd.edu, zhz396@ucsd.edu, vmzobell@ucsd.edu, kfrasier@ucsd.edu

Keywords: Ocean acoustics; Neural networks; Acoustic inversion; Multi-view

ABSTRACT

The potential for exploiting oceanic vessel noise as a sound source of opportunity to estimate ocean sound speed profile is investigated. A deep learning-based inversion scheme, relying upon the underwater radiated noise of moving vessels measured by a single hydrophone, is proposed. The dataset used for this study consists of acoustic recordings of commercial vessels transiting through the Santa Barbara Channel between January of 2015 through December of 2017. To obtain descriptors of the vessels, the recordings have been paired with Automatic Identification System data. The acoustic recordings and vessel descriptors are used as predictors for regressing sound speed for each meter in the top 200 meters of the water column, where sound speeds are most variable. Daily sound speed profiles were obtained from the California State Estimation Short-Term State Estimation model and were used to train and test the models. Multiple (typically ranging between 4-10) transits were recorded each day, therefore, this dataset provides an excellent opportunity for investigating whether multiple acoustic observations can be leveraged together to improve inversion estimates. We compare existing multi-view late fusion

methods against our approach, which generates a salience ranking of the transits to perform a weighted sum over the learned transit features.

I. INTRODUCTION

Acoustic inversion is frequently employed in oceanography for the purpose of inferring ocean sound speed profile (SSP): a parameter that characterizes the dependence of the speed of sound on the temperature, salinity, and pressure of water (Chen e al., 2018; Lovett, 1978). Reliably estimating sound speeds is critical due to the profound effect of these profiles on acoustic propagation. Knowledge of local SSPs is important for improving the performance of underwater acoustic systems such as sonar and for various oceanographic studies involving ocean currents, internal waves, and underwater topography. Oceanic SSPs can be directly measured using autonomous underwater vehicles or surface vessel-based instruments, such as a conductivity-temperature-depth (CTD) sensor. The CTD sensor is lowered into the water column recording the temperature, salinity, and pressure at regular intervals on both the descent and ascent path. The recorded variables are then related to sound speed using a polynomial expression such as the Chen and Millero equation (Chen and Millero, 1977). These direct measurements are typically conducted during periodic field efforts. Hindcast models are used to spatially and temporally interpolate between these local measurements, ingesting observations to estimate oceanographic conditions across a region or period of interest. These models use observational data (opportunistic in situ measurements, satellite observations, and buoy data) and detailed physical oceanographic models, often developed for a specific region of interest (Stammer et al., 2002; Zaba et al., 2018).

In conventional acoustic inversion studies, an active source is employed in conjunction with vertical hydrophone arrays to conduct the inversion process. However, this type of recording setup is costly and requires dedicated facilities to house the systems. In addition, repetitive and high intensity sound waves produced in active acoustics can disturb and potentially harm marine organisms that rely on sound for communication, navigation, and foraging (Richardson et al., 2013; Southall et al., 2019). To address these limitations, there is a need for inversion strategies that can make use of more readily available single sensor passive acoustic recordings.

The introduction of the Automatic Identification System (AIS), which provides precise locations of large commercial vessels, has made it possible to use vessel traffic noise as a source of opportunity. Using vessel traffic noise as a source of opportunity has three main advantages: 1) large vessels produce low frequency noise that can be detected at long distances, 2) commercial vessels are found in almost all areas of the ocean, making them an easily accessible source of data, 3) the regular and frequent movement of commercial vessels makes them a consistent and reliable source of data for long-term studies. Numerous studies have demonstrated the use of propeller noise from passing vessels received by seafloor hydrophones as acoustic sources of opportunity for estimating characteristics of the ocean environment and seafloor through which the signals have traveled (Gemba et al., 2018; Gervaise et al., 2012; Koch and Knobles, 2005; Tollefsen et al., 2020). This strategy has been used to estimate the waveguide invariant property, which represents the dispersive characteristics of the waveguide under variable oceanographic conditions as well as for geoacoustic parameter inversions (Park et al., 2005; Stotts et al., 2010; Verlinden et al., 2017). To the best of the authors' knowledge, the utilization of vessel traffic noise as a source of opportunity to directly estimate ocean SSPs has not been investigated.

Using uncontrolled, opportunistic vessel traffic noise as an acoustic source oceanographic applications poses several challenges. Two of the main challenges are: 1) signal variability: The acoustic signal produced by vessel traffic is highly variable and dependent on factors such as the size of the vessel, speed, load, and environmental conditions. Some of these factors are knowable from AIS, but transit-dependent factors such as load and actual draft are not. Incomplete information can limit our ability to explain observed acoustic variability, 2) background noise: The underwater environment is inherently noisy, and vessel traffic noise can be masked by other sources of noise such as natural sounds from marine life, wind, and waves. Recording systems can also differ in their self-noise characteristics. These challenges are exacerbated when a single hydrophone is employed to sample the acoustic signal, as is typically done in long-term (month to year-long) observational passive acoustic monitoring.

When the underwater radiated noise (URN) of a moving ship is recorded in a shallow water environment, the signal contains characteristic interference patterns when viewed in the time-frequency domain (D’Spain and Kuperman, 1999). Prior works have linked these striation patterns with interference between propagative modes and exploited them to perform geoacoustic inversion (Gervaise et al., 2012). These works relied upon conventional signal processing tools to extract the dispersion patterns. However, these algorithms require a high signal-to-noise ratio in order to be reliably extracted.

In recent years, machine learning approaches have been shown to outperform conventional signal and image processing techniques in a wide range of spectrogram processing applications (Ferguson et al., 2018; Kirsebom et al., 2020; Liu et al., 2021; Tréboutte et al., 2023). One of the advantages of using deep learning for acoustic inversion is that it can learn relationships between the input data and the output properties, even when those relationships are highly nonlinear and difficult to model using traditional methods. Multi-view learning, a machine learning approach that leverages multiple sources of information (i.e., views), can be integrated to learn more robust and accurate models. In the context of acoustic inversion, multi-view learning could theoretically be used to combine multiple recorded transits from the same day to improve the estimation of daily SSPs.

In this study, we investigate whether passive recordings of transiting commercial vessels from a single hydrophone can be used together with deep learning to estimate local SSPs. We summarize our findings and contributions as follows:

1) We show that passive recordings of transiting vessels can be used as sound sources of opportunity to perform ocean acoustic inversion.
2) We show that SSP estimates can be improved by leveraging multiple transits.
3) We present a novel approach for leveraging multiple transits for acoustic inversion and compare performance against existing multi-view learning techniques.
4) We present a newly curated dataset comprising 5,865 real underwater recordings of transiting vessels. These recordings were collected over 899 sampling days, encompassing diverse sea states and noise levels.

II. MATERIALS AND METHODS

A. Study site

The dataset used in this study consists of acoustic recordings of commercial vessels transiting through the Santa Barbara Channel (SBC) between January of 2015 to December of 2017. The shipping lane outside the SBC is approximately 20 nautical miles wide, extending from Point Conception in the north to the Long Beach Harbor. The SBC experiences a high volume of vessel traffic throughout the year, with container ships making up approximately 60% of all transits. Vehicle carriers, bulk carriers and tankers each constitute about 10% of the transits and cruise ships, tugs, research vessels, law enforcement and military vessels combined make up less than 10%.

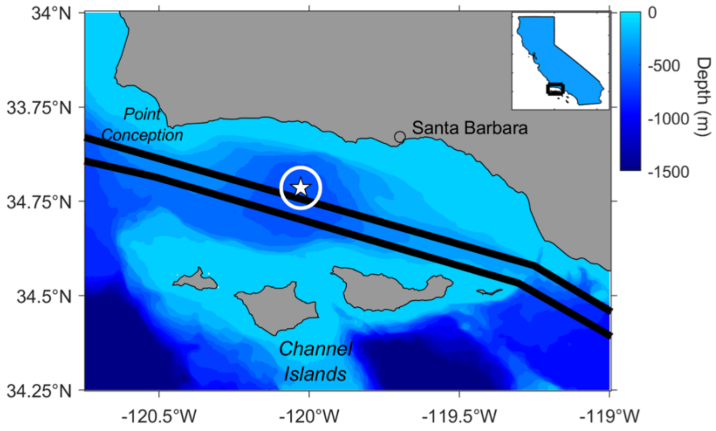


Figure 1: Map of the Santa Barbara Channel. Traffic separation scheme is shown as black lines and HARP location is shown with white pentagram. The white circle around the HARP denotes the 6 km boundary at which ship transits were considered.

B. Automatic identification system dataset

Vessels have been identified through Automatic Identification System (AIS) records, collected at onshore stations located at Coal Oil Point (34.411N 119.877 W) from April 2014 to present and Santa Ynez Peak (34.029 N 119.784 W) from August 2016 to present. The received AIS messages were time-stamped and continuously logged with an on-site computer. All AIS-derived information relevant to understanding vessel signature variability was used in this study. These variables are listed in Table 1.

Predictor Variable	**Abbreviation**	**Description**
Ship Design		
Length	LOA	total length of ship in meters
Type	TYP	numerical value that represents the general category of the vessel's type or purpose
Operational		
Draught	DRT	depth of a vessel below the waterline
Heading	HDG	direction that a vessel's bow is pointing
Course over ground	COG	actual direction of progress of a vessel relative to the Earth's surface
Speed over ground	SOG	speed of a vessel relative to the Earth's surface
Closest point of approach	CPA	point at which the distance between the ship
Oceanographic		
Month	MTH	month of the year

Table 1. Description of predictor variables used in statistical models.

C. Vessel noise dataset

An existing database of 5,865 recordings of identified ships transiting through the Santa Barbara Channel (SBC) between January of 2015 to December of 2017 was used for this study. Acoustic recordings were collected in the SBC using a High-frequency Acoustic Recording Package (34°16.53 N 120°1.11 W; Figure 1) (HARPs; Wiggins and Hildebrand, 2007), which is a bottom-mounted recorder with a hydrophone buoyed 10 m above the seafloor (580 m bottom depth).

The audio was sampled at 200 kHz, which was then decimated by a factor of 20 resulting in a 10 kHz sampling rate, and a Nyquist frequency of 5 kHz. The data were low-pass filtered with an 8th order Chebyshev Type I IIR filter to prevent aliasing during decimation. Each transit recording was clipped to only consider the time period in which the ship was within 6 km of the recording station. These audio clips were converted into spectrograms using 10,000-point short time Fourier transform with no overlap, resulting in a frequency resolution of 1 Hz and time resolution of 1.0 s. Spectrograms were cropped to limit the frequency range under consideration from to 10 to 300 Hz, the range over which local vessel URN is typically the dominant signal in this dataset, and interference patterns are most apparent.

D. Hindcast dataset

Estimating the near surface region of the sound speed profile is challenging because it experiences the highest level of variability. For this reason, daily sound speed profiles were obtained for the top 200 meters of the study region using the California State Estimation Short-Term State Estimation (CASE-STSE) model output (Zaba et al., 2018). This model utilizes hindcast data and integrates the Massachusetts Institute of Technology general circulation model (MITgcm) through a least-square fitting solution. The data used in the integration includes profiles from Spray gliders, High-Resolution expendable bathythermographs, Argo, and satellite measurements of sea surface height and temperature.

E. Models

1. Baseline model

Oceanic sound speed profiles typically manifest seasonal patterns, primarily due to their significant dependence on temperature. Therefore, we first propose a model for sound speed profile estimation that computes seasonal averages from previous years to estimate the sound speed profile for all days within that specific season. This approach does not utilize any of the AIS or acoustic data. This model is from here onwards referred to as the **baseline** model.

This baseline model is used to provide context for the neural network-based model performance. All neural network-based models used in this study use the same information as the baseline model (i.e. season) in addition to the transit data. Therefore, we can evaluate the informativeness of the transit recordings by comparing the performance of the neural network-based models with the baseline. If the transit data contains additional information regarding the local sound speed profile, we would expect incorporation of the transit data to improve the sound speed profile prediction estimate. Conversely, if the transit data is uninformative, we expect the estimation performance to remain unchanged.

2. Single-transit model

We designed a neural network-based model to produce an estimate for sound speed profile from each of the recorded transits using the spectrograms and vessel descriptors. This model is from here onwards referred to as the **single-transit** model.

The data used to train the single-transit model is denoted as $\mathcal{X}_s$, and can be formulated as follows. The data corpus $\mathcal{X}_s = \{(\mathbf{x}^{(1)}, \mathbf{v}^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{x}^{(n)}, \mathbf{v}^{(n)}, \mathbf{y}^{(n)})\}$, where $\mathbf{x}^{(i)}$ and $\mathbf{v}^{(i)}$ are the spectrogram representation of the audio recording and the vessel descriptors respectively for the ith recorded transit.

We seek to lean a model $f : (\mathbf{x}, \mathbf{v}) \rightarrow \mathbf{y}$ that maps input variables $\mathbf{x}$ and $\mathbf{v}$ to $\mathbf{y} \in \mathcal{R}^{200\times1}$ where $\mathbf{y}$ is a vector containing the sound speed estimations for each meter in the upper 200 meters of the water column (Fig. 2A). We conceptualized f as comprising two functions that are applied in sequence: first an encoder function E and then a regression function R. The encoder function $E : (\mathbf{x}, \mathbf{v}) \rightarrow \mathbf{z}$ maps input variables $\mathbf{x}$ and $\mathbf{v}$ to a hidden variable $\mathbf{z} \in \mathcal{R}^{128\times1}$. The regression function $R : \mathbf{z} \rightarrow \mathbf{y}$ produces the sound speed profile estimate from $\mathbf{z}$ (Fig. 2C). Because the input variables are of different modalities (spectrogram image and AIS data), we divide E into two sub-encoders. Spectrograms are encoded using a convolutional neural network, denoted as E_S, while vessel descriptors are encoded using a fully connected network, denoted as E_V. Each encoder returns a vector that is then concatenated together (Fig. 2B). The encoder function E refers to this integrated process of joint encoding and concatenation.

Our proposed method hinges on leveraging recorded vessel noise in conjunction with AIS data to estimate sound speed profile. To validate that our model genuinely learns pertinent features related to sound speed from the combined modalities and avoids relying on any spurious correlations that might exist between AIS data and ocean sound speed, we introduce a modified version of the single-transit model, referred to as **single-transit (noAudio)**. In this variant, we set all spectrogram values to zero, effectively removing the vessel noise data while retaining only the AIS data.

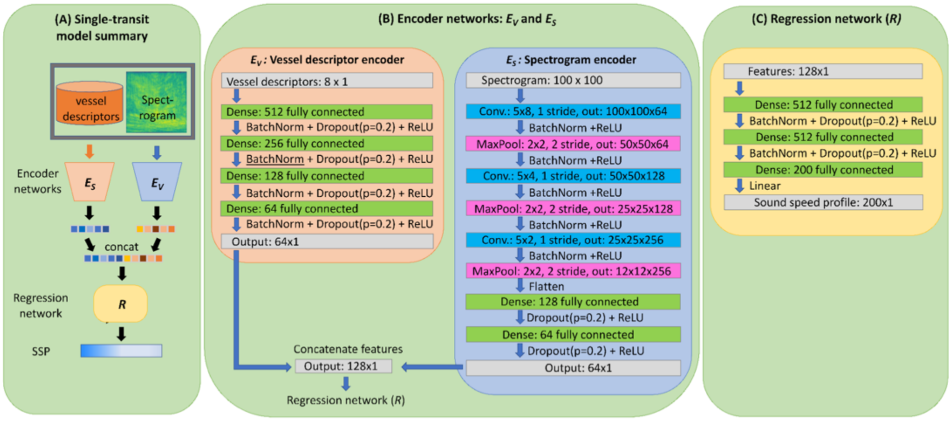

(A) Single-transit model summary
vessel descriptors
Spect-rogram
Encoder networks
E_S
E_V
concat
Regression network
R
SSP
(B) Encoder networks: E_V and E_S
E_V: Vessel descriptor encoder
Vessel descriptors: 8 x 1
Dense: 512 fully connected
BatchNorm + Dropout(p=0.2) + ReLU
Dense: 256 fully connected
BatchNorm + Dropout(p=0.2) + ReLU
Dense: 128 fully connected
BatchNorm + Dropout(p=0.2) + ReLU
Dense: 64 fully connected
BatchNorm + Dropout(p=0.2) + ReLU
Output: 64x1
E_S: Spectrogram encoder
Spectrogram: 100 x 100
Conv.: 5x8, 1 stride, out: 100x100x64
BatchNorm +ReLU
MaxPool: 2x2, 2 stride, out: 50x50x64
BatchNorm +ReLU
Conv.: 5x4, 1 stride, out: 50x50x128
BatchNorm +ReLU
MaxPool: 2x2, 2 stride, out: 25x25x128
BatchNorm +ReLU
Conv.: 5x2, 1 stride, out: 25x25x256
BatchNorm +ReLU
MaxPool: 2x2, 2 stride, out: 12x12x256
Flatten
Dense: 128 fully connected
Dropout(p=0.2) + ReLU
Dense: 64 fully connected
Dropout(p=0.2) + ReLU
Output: 64x1
Concatenate features
Output: 128x1
Regression network (R)
(C) Regression network (R)
Features: 128x1
Dense: 512 fully connected
BatchNorm + Dropout(p=0.2) + ReLU
Dense: 512 fully connected
BatchNorm + Dropout(p=0.2) + ReLU
Dense: 200 fully connected
Linear
Sound speed profile: 200x1

Figure 2: Single-transit model architecture. (A) Summary of the single-transit model. (B) The spectrogram and vessel descriptors are encoded separately using a convolutional neural network E_S and fully connected network E_V respectively. Both encoder networks output a vector which are concatenated. (C) The concatenated vector is then forward propagated through a fully connected regression network R which produces an estimate for sound speed for each meter in the upper 200 meters of the water column.

3. Multi-transit models

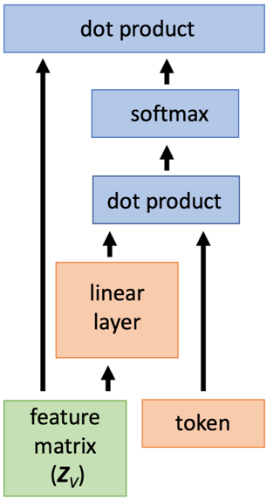


Figure 3: Our fusion method. Features $\mathbf{Z}_V$ are passed through a linear layer and a dot product is performed against a learnable token. Orange boxes indicate learnable values. Green box indicates features from encoder E. Blue boxes indicate fixed mathematical operation.

The last set of models we consider are designed to combine and/or contrast information from multiple transits to improve SSP estimation. The methodologies we examine are inspired by, or are direct implementations of, existing multi-view machine learning techniques. For our application, all transits recorded on the same day are considered as distinct "views" which contain information about the same sound speed profile.

To train the multi-transit models, we organize the acoustic dataset into daily collections denoted as $\mathbf{C}_V$, where each collection consists of multiple transits. Notably, all transits within a collection share the same sound speed profiles, as they were recorded on the same day. This corpus is denoted as $\mathcal{X}_m = \{(\mathbf{C}_V^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{C}_V^{(N)}, \mathbf{y}^{(N)})\}$ where each collection $\mathbf{C}_V^i$ consists of all spectrogram and vessel descriptor pairings that were recorded on the ith day and $N = 899$ is the number of sampling days. Hence, we reformulate the modeling task as

$f : \mathbf{C}_V \rightarrow \mathbf{y} \mid \mathbf{C}_V = \{(\mathbf{x}, \mathbf{v})^{(1)}, (\mathbf{x}, \mathbf{v})^{(2)}, \ldots, (\mathbf{x}, \mathbf{v})^{(T_V)}\}$,

where $(\mathbf{x}, \mathbf{y})^{(v)}$ represents one audio recording and vessel descriptor pair $v \in \{1, \ldots, T_V\}$ and T_V denotes the number of transits in collection $\mathbf{C}_V$ which is variable across the collections. All variables in $\mathcal{X}_m$ are the same as the single-transit data collection $\mathcal{X}_s$.

The simplest way to leverage multiple transits is to average the estimations for each of the transits in a collection using the single-transit model. We refer to this approach as **single-transit (avg)**. However, this approach is not able to leverage complementary information or weigh saliency differences from multiple transits to improve prediction accuracy. To address these concerns, we evaluate three “late fusion” techniques for combining information across the transits within each collection. Late fusion is a technique in multi-view learning that allows the combination of learned features from multiple views at a later stage in the learning process.

Two existing late fusion approaches we consider are: (1) **late fusion (max)**: max value is calculated for each of the features across the transits, and (2) **late fusion (concat)**: a fixed number of feature vectors are concatenated.(Seeland and Mäder, 2021) Lastly, we implement a novel late fusion technique referred to as **late fusion (token)** that is described below.

The forward propagation of a transit collection $\mathbf{C}_V$ into encoder E produces a matrix $\mathbf{Z}_V$ whose columns are the feature vectors of length $D = 128$ from each transit in the collection.

$\mathbf{Z}_V = [\mathbf{z}^{(1)}; \ldots ; \mathbf{z}^{(T_V)}] = E(\mathbf{C}_V) \in \mathcal{R}^{D \times T_V}$

For **late fusion (max)**, an element-wise maximum operation is applied for each of the D features which produces the vector

$\hat{\mathbf{z}}_V = \max_v \mathbf{Z}_V \in \mathcal{R}^{D \times 1}$

For **late fusion (concat)**, we concatenate k columns in $\mathbf{Z}_V$ to form a vector

$\hat{\mathbf{z}}_V = [\mathbf{z}^{(1)}, \ldots , \mathbf{z}^{(k)}] \in \mathcal{R}^{Dk \times 1}$

If $k < T_V$ we subsample the transits by randomly selecting k columns without replacement. If $k > T_V$ we up-sample the transits by randomly selecting $k - T_V$ columns with replacement to duplicate and concatenate all features vectors. We set the value of $k = 8$ for all experiments, as it was determined to yield the highest performance on a validation set.

Late fusion (token) combines ideas from Scaled Dot-Product Attention and prompt tuning with the goal of automatically weighting more informative transits (Jia et al., 2022; Vaswani et al., 2017). A weight matrix $\mathbf{W} \in \mathcal{R}^{h \times D}$ is used to project the features in $\mathbf{Z}_V$ into a lower dimension of size $h = 64$. The projected features are then compared against a learnable token $\mathbf{q} \in \mathcal{R}^{h \times 1}$. The similarity values are then normalized using the softmax function. The normalized values are then used to compute a weighted sum of the original features

$\hat{\mathbf{z}}_V = \text{softmax}(\frac{\mathbf{q}^T \mathbf{W}^T \mathbf{Z}_V}{\sqrt{D}}) \cdot \mathbf{Z}_V^T \in \mathcal{R}^{D \times 1}$

A model trained with this multi-view approach will reduce its loss by learning to assign larger weights (i.e. large similarity with $\mathbf{q}$) to transits that produce more reliable sound speed estimates. An illustration of the late fusion (token) method is shown in Fig. 3.

For all the aforementioned late-fusion methods, the fused feature vector $\hat{\mathbf{z}}_V$ is forward propagated through R to regress sound speed profile.

F. Experimental set-up

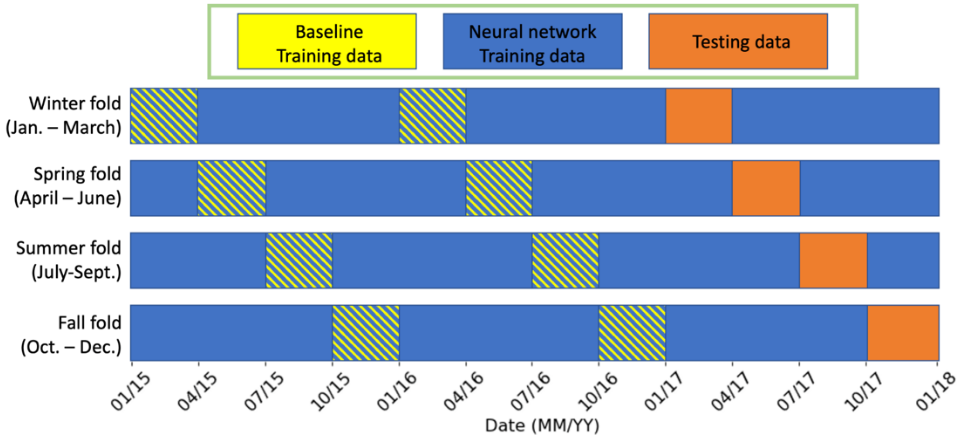


Figure 4. Illustration of the 4-fold cross-validation approach for the baseline and neural network models. For each fold, models are tested on data from a single season in 2017 shown in orange. Note that time regions with mixed colors indicate that data was used to train both models.

Our partitioning of the training and testing data was deliberately crafted to emulate a real-world scenario, where the model is trained on historical data and subsequently deployed on future data. Data from the year 2017 was divided into four distinct testing sets, each corresponding to a specific season. These testing sets were created to ensure non-overlapping periods and were defined as follows: winter (January to March), spring (April to June), summer (July to September), and fall (October to December). This partitioning allowed for the assessment of model performance in relation to the different seasons of the year. We then perform 4-fold cross-validation where for each fold, one season from 2017 is used for testing and the remaining data is used for training.

The multi-transit models were trained using a two-stage approach: 1) for first-stage training, the encoder network and regression network are trained to estimate sound speed profile from each transit, i.e. trained exactly as the single-transit model, 2) for second-stage training, the layers of the encoder network are frozen, and the parameters of the multi-view learning mechanism (if applicable) and the regression network are trained. For all neural network-based models, 25% of each training set was allocated for validation-based early stopping with a patience of 30 epochs. Optimization was performed with the ADAM optimizer using a learning rate of 1e-4 and a scheduler that decays this learning rate by a factor of .75 every 10 epochs. Regression loss is computed as root mean square error (RMSE).

III. RESULTS

The proposed single-transit model provided an average error reduction of about 36% compared to the baseline model across the testing folds. We attribute the observed performance improvement to the inclusion of the acoustic data, as its exclusion (noAudio model) led to a performance level comparable to the baseline model, with predictions akin to historical averaging. We note that the performance improvement of the single-transit model was variable across the folds. Specifically, during the summer and fall testing seasons, the single-transit model achieved substantial reductions in estimation error, with improvements of 44% and 43%

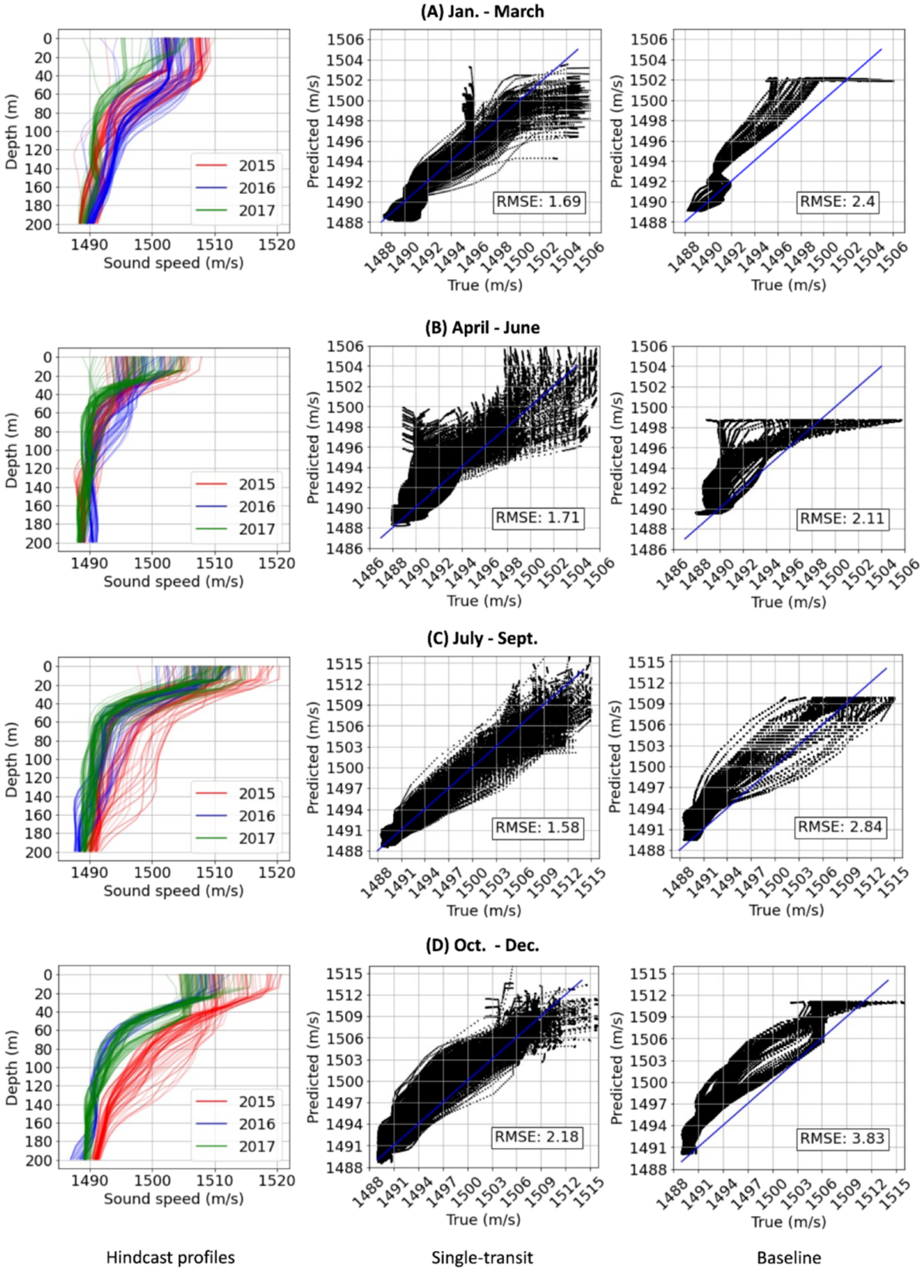
respectively. In contrast, its performance improvement was relatively modest during the spring testing season, with only a 13% reduction in error observed (Table 2).

Test fold	*baseline*	*single-transit*	*single-transit*	Multi-transit			
		(noAudio)		*single-transit*	*late fusion*	*late fusion*	*late fusion*
				(avg)	*(max)*	*(concat)*	*(token)*
Jan. - March	2.4	2.43	1.69	1.59	1.59	1.58	**1.52**
April - June	2.11	2.07	1.71	1.48	1.52	1.52	**1.47**
July - Sept.	2.84	2.1	1.58	**1.45**	1.55	1.55	1.48
Oct. - Dec.	3.83	3.91	2.18	2.04	1.93	**1.69**	1.72
Average	2.8	2.63	1.79	1.64	1.65	1.59	**1.55**

Table 2. Model performance of sound speed profile estimation across the four test seasons. Performance is reported in terms of root-mean-square error (m/s). The best performing model for each season is shown in boldface.

Despite the inherent seasonal regularity, oceanic sound speed profiles can manifest year-to-year variability. The reconstructed hindcast profiles at the study site reflect this variability, and is particularly noticeable comparing profiles from the year 2015 to other years (Fig. 5). We hypothesized that the presence of annual variability in the data would result in estimation bias in both the baseline and single-transit models. Although both models exhibited estimation bias, the scatter plots in Fig. 5 indicate that the deep learning-based approach experiences comparatively less estimation bias than the baseline model. For example, in Fig. 5D, the distribution of points from the two models generally follow the same shape, but the estimations from the single-transit model are centered more along the blue line than the baseline.

Table 2 indicates that model error can be reduced by an additional 8% by averaging estimates obtained from multiple transits. The best performing multi-transit model was late fusion (token), which provided an average error reduction of 13.5% compared to the single-transit model and a 5.5% error reduction compared to the single-transit (avg) model.



(A) Jan. - March

(B) April - June

(C) July - Sept.

(D) Oct. - Dec.

Figure 5. Comparison of single-transit and baseline model predictions. Left: Test set sound speeds to be predicted. Center: Predictions from the single-transit model vs. hindcast values (black dots) plotted against each other, such that perfect predictions would fall along the diagonal (blue line). Right: Predictions from the baseline model plotted against the hindcast values. Prediction bias is visible when points fall primarily to one side of the diagonal. RMSE values in units of m/s summarize the average root mean squared difference between predictions and hindcast values, with higher values indicating poorer predictions.

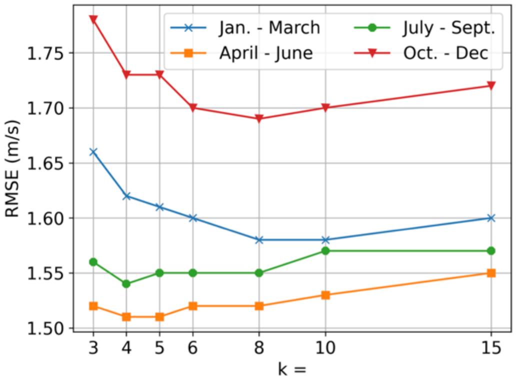


Figure 6. Performance on testing data of late fusion (concat) model with varying transit sampling number k.

IV. DISCUSSION

The relatively lower estimation bias of the neural network models compared to the baseline suggests that the neural network is able to learn relevant patterns and relationships that can generalize across the seasons. The estimation error was found to be highest in the near-surface regions of the sound speed profile. In order to mitigate these errors, we propose that future work should explore the integration of additional observational modalities, such as satellite-derived sea surface temperature estimates.

As indicated by the performance of single-transit (avg), averaging multiple estimates helps to mitigate the effects of random errors or outliers by simply leveraging a larger sample. However, calculating an average considers all estimates to have equal weight in the final average and provides no mechanism for leveraging complimentary information or discard outliers. Multi-view learning techniques provide the opportunity for extracting such information.

Our results using the traditional multi-view techniques (max and concat) were similar to Seeland and Mäder (2021), where the multi-view methods with learnable solutions provided the best results. However, using feature concatenation is complicated in this application because the number of transits is variable. This means that a fixed number, k, of transits need to be sampled, which introduces the trade-off: if k is too small, there is less information to leverage, but if k is too large, the number of trainable parameters grows potentially leading to over-fitting. This produces a U-shape error curve with variable k where the optimal value for k needs to be found through experimentation (Fig. 6).

Our proposed token learning method has the advantage of scaling to arbitrary input size while maintaining a fixed and relatively small number of trainable parameters, which may improve generalizability. Moreover, in this application, we hypothesize that employing a weighted sum, where all features within the transit receive the same weight, rather than using feature-specific fusion, is more ideal. Most multi-view models are developed under the implicit assumption that each observation is *uniquely* informative regarding the object/event of interest. In other words, each observation contains predictive information that the other observations do not contain (e.g. consider two images of the same plant, one image captures the detail of the leaf and the other captures the flower). For our application, it is unlikely that different transits contain this kind of complementary information. Instead, some transits exhibit higher saliency compared to others, and the goal of leveraging multiple observations is to rank the salience, in contrast to pooling information across the transits. This approach may have broad applicability in oceanographic acoustic observing problems involving large amounts of weakly-curated data in which feature salience is variable in time and space, particularly if the salience of relevant features is difficult to estimate a priori. If multiple sensors were available, fusion approaches could be used to incorporate simultaneous views.

An important limitation of this approach is the availability of sound speed profile estimates for model training. Although quarterly in situ measurements were available from a nearby CalCOFI station and periodic local glider transits, these were determined to be too infrequent for training, therefore this study used data assimilative hindcasts for training. Agreement between these regionally-specific hindcasts, and the available in situ measurements, was high for this well-sampled, highly-studied region. Further experimentation is needed to evaluate whether this approach could be used to refine or improve hindcast estimates, particularly in under-sampled regions. Additionally, the proposed method represents a preliminary exploration aimed at evaluating the feasibility of extracting sound-speed relevant features from single sensor acoustic recordings. Further development will be required to adapt this method for use across different recording environments.

V. CONCLUSION

In this paper, a neural network-based model, which uses acoustic recordings of URN of transiting ships and their transit metadata, is proposed to predict SSP. Additionally, we propose a data fusion strategy suitable for large observational acoustic datasets, in which data are weakly-curated and feature salience differs between observations used for prediction. Our results show that the addition of vessel transit recordings markedly improved the estimation of SSP compared to the use of historical averages. We show that multiple transit recordings can be leveraged together to improve SSP estimation and compare multiple techniques for combining available information. We note that this work serves as a first approach in estimating oceanic sound speed profiles from vessel URN, and there still exist sources of error in the estimations of the best performing model. We believe that future work incorporating other data modalities and alternative hydrophone configurations can help further reduce this estimation error.

Acknowledgments

This work was supported by the Office of Naval Research and Task Force Ocean program manager Robert Headrick, Grant No. 200249. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

Bibliography

Chen, C., Ma, Y., and Liu, Y. (**2018**). "Reconstructing Sound speed profiles worldwide with Sea surface data," Applied Ocean Research, **77**, 26–33. doi:10.1016/j.apor.2018.05.002

Chen, C.-T., and Millero, F. J. (**1977**). "Speed of sound in seawater at high pressures," The Journal of the Acoustical Society of America, **62**, 1129–1135. doi:10.1121/1.381646

D'Spain, G. L., and Kuperman, W. A. (**1999**). "Application of waveguide invariants to analysis of spectrograms from shallow water environments that vary in range and azimuth," The Journal of the Acoustical Society of America, **106**, 2454–2468. doi:10.1121/1.428124

Ferguson, E. L., Williams, S. B., and Jin, C. T. (**2018**). "Sound Source Localization in a Multipath Environment Using Convolutional Neural Networks," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2386–2390. Presented at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi:10.1109/ICASSP.2018.8462024

Gemba, K. L., Sarkar, J., Cornuelle, B., Hodgkiss, W. S., and Kuperman, W. A. (**2018**). "Estimating relative channel impulse responses from ships of opportunity in a shallow water environment," The Journal of the Acoustical Society of America, **144**, 1231–1244. doi:10.1121/1.5052259

Gervaise, C., Kinda, B. G., Bonnel, J., Stéphan, Y., and Vallez, S. (**2012**). "Passive geoacoustic inversion with a single hydrophone using broadband ship noise," The Journal of the Acoustical Society of America, **131**, 1999–2010. doi:10.1121/1.3672688

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., and Lim, S.-N. (**2022**). "Visual Prompt Tuning.," Retrieved from http://arxiv.org/abs/2203.12119

Kirsebom, O. S., Frazao, F., Simard, Y., Roy, N., Matwin, S., and Giard, S. (**2020**). "Performance of a deep neural network at detecting North Atlantic right whale upcallsa)," The Journal of the Acoustical Society of America, **147**, 2636–2646. doi:10.1121/10.0001132

Koch, R. A., and Knobles, D. P. (**2005**). “Geoacoustic inversion with ships as sources,” The Journal of the Acoustical Society of America, **117**, 626–637. doi:10.1121/1.1848175

Liu, F., Shen, T., Luo, Z., Zhao, D., and Guo, S. (**2021**). “Underwater target recognition using convolutional recurrent neural networks with 3-D Mel-spectrogram and data augmentation,” Applied Acoustics, **178**, 107989. doi:10.1016/j.apacoust.2021.107989

Lovett, J. R. (**1978**). “Merged seawater sound-speed equations,” The Journal of the Acoustical Society of America, **63**, 1713–1718. doi:10.1121/1.381909

Park, C., Seong, W., and Gerstoft, P. (**2005**). “Geoacoustic inversion in time domain using ship of opportunity noise recorded on a horizontal towed array,” The Journal of the Acoustical Society of America, **117**, 1933–1941. doi:10.1121/1.1862574

Richardson, W. J., Jr, C. R. G., Malme, C. I., and Thomson, D. H. (**2013**). *Marine Mammals and Noise*, Academic Press, 593 pages.

Seeland, M., and Mäder, P. (**2021**). “Multi-view classification with convolutional neural networks,” PLOS ONE, **16**, e0245230. doi:10.1371/journal.pone.0245230

Southall, B. L., Finneran, J. J., Reichmuth, C., Nachtigall, P. E., Ketten, D. R., Bowles, A. E., Ellison, W. T., et al. (**2019**). “Marine Mammal Noise Exposure Criteria: Updated Scientific Recommendations for Residual Hearing Effects,” Aquat Mamm, **45**, 125–232. doi:10.1578/AM.45.2.2019.125

Stammer, D., Wunsch, C., Giering, R., Eckert, C., Heimbach, P., Marotzke, J., Adcroft, A., et al. (**2002**). “Global ocean circulation during 1992–1997, estimated from ocean observations and a general circulation model,” Journal of Geophysical Research: Oceans, **107**, 1-1-1–27. doi:10.1029/2001JC000888

Stotts, S. A., Koch, R. A., Joshi, S. M., Nguyen, V. T., Ferreri, V. W., and Knobles, D. P. (**2010**). “Geoacoustic Inversions of Horizontal and Vertical Line Array Acoustic Data From a

Surface Ship Source of Opportunity," IEEE Journal of Oceanic Engineering, **35**, 79–102. Presented at the IEEE Journal of Oceanic Engineering. doi:10.1109/JOE.2009.2032256

Tollefsen, D., Dosso, S. E., and Knobles, D. P. (**2020**). "Ship-of-Opportunity Noise Inversions for Geoacoustic Profiles of a Layered Mud-Sand Seabed," IEEE Journal of Oceanic Engineering, **45**, 189–200. Presented at the IEEE Journal of Oceanic Engineering. doi:10.1109/JOE.2019.2908026

Tréboutte, A., Carli, E., Ballarotta, M., Carpentier, B., Faugère, Y., and Dibarboure, G. (**2023**). "KaRIn Noise Reduction Using a Convolutional Neural Network for the SWOT Ocean Products," Remote Sensing, **15**, 2183. doi:10.3390/rs15082183

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., et al. (**2017**). "Attention is All you Need," Advances in Neural Information Processing Systems, Curran Associates, Inc., Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html, (date last viewed: 01-Aug-23). Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Verlinden, C. M. A., Sarkar, J., Cornuelle, B. D., and Kuperman, W. A. (**2017**). "Determination of acoustic waveguide invariant using ships as sources of opportunity in a shallow water marine environment," J. Acoust. Soc. Am., **141**, EL102–EL107. doi:10.1121/1.4976112

Wiggins, S. M., and Hildebrand, J. A. (**2007**). "High-frequency Acoustic Recording Package (HARP) for broad-band, long-term marine mammal monitoring," 2007 Symposium on Underwater Technology and Workshop on Scientific Use of Submarine Cables and Related Technologies, 551–557. Presented at the 2007 Symposium on Underwater Technology and Workshop on Scientific Use of Submarine Cables and Related Technologies. doi:10.1109/UT.2007.370760

Zaba, K. D., Rudnick, D. L., Cornuelle, B. D., Gopalakrishnan, G., and Mazloff, M. R. (**2018**).

“Annual and Interannual Variability in the California Current System: Comparison of an Ocean State Estimate with a Network of Underwater Gliders,” Journal of Physical Oceanography, **48**, 2965–2988. doi:10.1175/JPO-D-18-0037.1

Acknowledgements

Chapter 4, in part, has been submitted for publication of the material as it may appear in IEEE Oceanic Engineering 2023, Walker, Joseph; Zeng, Zheng; Chengchen, Wu; Jaffe, Jules; Frasier, Kaitlin; Sandin, Stuart. The dissertation author was the primary investigator and author of this paper.